Monographs on Statistics and Applied Probability 135

Hierarchical Modeling and Analysis for Spatial Data

Second Edition

Sudipto Banerjee
Bradley P. Carlin
Alan E. Gelfand

CRC Press
Taylor & Francis Group
A CHAPMAN & HALL BOOK

Hierarchical Modeling and Analysis for Spatial Data

Second Edition

MONOGRAPHS ON STATISTICS AND APPLIED PROBABILITY

General Editors

F. Bunea, V. Isham, N. Keiding, T. Louis, R. L. Smith, and H. Tong

1. Stochastic Population Models in Ecology and Epidemiology *M.S. Barlett* (1960)
2. Queues *D.R. Cox and W.L. Smith* (1961)
3. Monte Carlo Methods *J.M. Hammersley and D.C. Handscomb* (1964)
4. The Statistical Analysis of Series of Events *D.R. Cox and P.A.W. Lewis* (1966)
5. Population Genetics *W.J. Ewens* (1969)
6. Probability, Statistics and Time *M.S. Barlett* (1975)
7. Statistical Inference *S.D. Silvey* (1975)
8. The Analysis of Contingency Tables *B.S. Everitt* (1977)
9. Multivariate Analysis in Behavioural Research *A.E. Maxwell* (1977)
10. Stochastic Abundance Models *S. Engen* (1978)
11. Some Basic Theory for Statistical Inference *E.J.G. Pitman* (1979)
12. Point Processes *D.R. Cox and V. Isham* (1980)
13. Identification of Outliers *D.M. Hawkins* (1980)
14. Optimal Design *S.D. Silvey* (1980)
15. Finite Mixture Distributions *B.S. Everitt and D.J. Hand* (1981)
16. Classification *A.D. Gordon* (1981)
17. Distribution-Free Statistical Methods, 2nd edition *J.S. Maritz* (1995)
18. Residuals and Influence in Regression *R.D. Cook and S. Weisberg* (1982)
19. Applications of Queueing Theory, 2nd edition *G.F. Newell* (1982)
20. Risk Theory, 3rd edition *R.E. Beard, T. Pentikäinen and E. Pesonen* (1984)
21. Analysis of Survival Data *D.R. Cox and D. Oakes* (1984)
22. An Introduction to Latent Variable Models *B.S. Everitt* (1984)
23. Bandit Problems *D.A. Berry and B. Fristedt* (1985)
24. Stochastic Modelling and Control *M.H.A. Davis and R. Vinter* (1985)
25. The Statistical Analysis of Composition Data *J. Aitchison* (1986)
26. Density Estimation for Statistics and Data Analysis *B.W. Silverman* (1986)
27. Regression Analysis with Applications *G.B. Wetherill* (1986)
28. Sequential Methods in Statistics, 3rd edition *G.B. Wetherill and K.D. Glazebrook* (1986)
29. Tensor Methods in Statistics *P. McCullagh* (1987)
30. Transformation and Weighting in Regression *R.J. Carroll and D. Ruppert* (1988)
31. Asymptotic Techniques for Use in Statistics *O.E. Bandorff-Nielsen and D.R. Cox* (1989)
32. Analysis of Binary Data, 2nd edition *D.R. Cox and E.J. Snell* (1989)
33. Analysis of Infectious Disease Data *N.G. Becker* (1989)
34. Design and Analysis of Cross-Over Trials *B. Jones and M.G. Kenward* (1989)
35. Empirical Bayes Methods, 2nd edition *J.S. Maritz and T. Lwin* (1989)
36. Symmetric Multivariate and Related Distributions *K.T. Fang, S. Kotz and K.W. Ng* (1990)
37. Generalized Linear Models, 2nd edition *P. McCullagh and J.A. Nelder* (1989)
38. Cyclic and Computer Generated Designs, 2nd edition *J.A. John and E.R. Williams* (1995)
39. Analog Estimation Methods in Econometrics *C.F. Manski* (1988)
40. Subset Selection in Regression *A.J. Miller* (1990)
41. Analysis of Repeated Measures *M.J. Crowder and D.J. Hand* (1990)
42. Statistical Reasoning with Imprecise Probabilities *P. Walley* (1991)
43. Generalized Additive Models *T.J. Hastie and R.J. Tibshirani* (1990)
44. Inspection Errors for Attributes in Quality Control *N.L. Johnson, S. Kotz and X. Wu* (1991)
45. The Analysis of Contingency Tables, 2nd edition *B.S. Everitt* (1992)

46. The Analysis of Quantal Response Data *B.J.T. Morgan* (1992)
47. Longitudinal Data with Serial Correlation—A State-Space Approach *R.H. Jones* (1993)
48. Differential Geometry and Statistics *M.K. Murray and J.W. Rice* (1993)
49. Markov Models and Optimization *M.H.A. Davis* (1993)
50. Networks and Chaos—Statistical and Probabilistic Aspects *O.E. Barndorff-Nielsen, J.L. Jensen and W.S. Kendall* (1993)
51. Number-Theoretic Methods in Statistics *K.-T. Fang and Y. Wang* (1994)
52. Inference and Asymptotics *O.E. Barndorff-Nielsen and D.R. Cox* (1994)
53. Practical Risk Theory for Actuaries *C.D. Daykin, T. Pentikäinen and M. Pesonen* (1994)
54. Biplots *J.C. Gower and D.J. Hand* (1996)
55. Predictive Inference—An Introduction *S. Geisser* (1993)
56. Model-Free Curve Estimation *M.E. Tarter and M.D. Lock* (1993)
57. An Introduction to the Bootstrap *B. Efron and R.J. Tibshirani* (1993)
58. Nonparametric Regression and Generalized Linear Models *P.J. Green and B.W. Silverman* (1994)
59. Multidimensional Scaling *T.F. Cox and M.A.A. Cox* (1994)
60. Kernel Smoothing *M.P. Wand and M.C. Jones* (1995)
61. Statistics for Long Memory Processes *J. Beran* (1995)
62. Nonlinear Models for Repeated Measurement Data *M. Davidian and D.M. Giltinan* (1995)
63. Measurement Error in Nonlinear Models *R.J. Carroll, D. Rupert and L.A. Stefanski* (1995)
64. Analyzing and Modeling Rank Data *J.J. Marden* (1995)
65. Time Series Models—In Econometrics, Finance and Other Fields *D.R. Cox, D.V. Hinkley and O.E. Barndorff-Nielsen* (1996)
66. Local Polynomial Modeling and its Applications *J. Fan and I. Gijbels* (1996)
67. Multivariate Dependencies—Models, Analysis and Interpretation *D.R. Cox and N. Wermuth* (1996)
68. Statistical Inference—Based on the Likelihood *A. Azzalini* (1996)
69. Bayes and Empirical Bayes Methods for Data Analysis *B.P. Carlin and T.A Louis* (1996)
70. Hidden Markov and Other Models for Discrete-Valued Time Series *I.L. MacDonald and W. Zucchini* (1997)
71. Statistical Evidence—A Likelihood Paradigm *R. Royall* (1997)
72. Analysis of Incomplete Multivariate Data *J.L. Schafer* (1997)
73. Multivariate Models and Dependence Concepts *H. Joe* (1997)
74. Theory of Sample Surveys *M.E. Thompson* (1997)
75. Retrial Queues *G. Falin and J.G.C. Templeton* (1997)
76. Theory of Dispersion Models *B. Jørgensen* (1997)
77. Mixed Poisson Processes *J. Grandell* (1997)
78. Variance Components Estimation—Mixed Models, Methodologies and Applications *P.S.R.S. Rao* (1997)
79. Bayesian Methods for Finite Population Sampling *G. Meeden and M. Ghosh* (1997)
80. Stochastic Geometry—Likelihood and computation *O.E. Barndorff-Nielsen, W.S. Kendall and M.N.M. van Lieshout* (1998)
81. Computer-Assisted Analysis of Mixtures and Applications—Meta-Analysis, Disease Mapping and Others *D. Böhning* (1999)
82. Classification, 2nd edition *A.D. Gordon* (1999)
83. Semimartingales and their Statistical Inference *B.L.S. Prakasa Rao* (1999)
84. Statistical Aspects of BSE and vCJD—Models for Epidemics *C.A. Donnelly and N.M. Ferguson* (1999)
85. Set-Indexed Martingales *G. Ivanoff and E. Merzbach* (2000)
86. The Theory of the Design of Experiments *D.R. Cox and N. Reid* (2000)
87. Complex Stochastic Systems *O.E. Barndorff-Nielsen, D.R. Cox and C. Klüppelberg* (2001)
88. Multidimensional Scaling, 2nd edition *T.F. Cox and M.A.A. Cox* (2001)
89. Algebraic Statistics—Computational Commutative Algebra in Statistics *G. Pistone, E. Riccomagno and H.P. Wynn* (2001)
90. Analysis of Time Series Structure—SSA and Related Techniques *N. Golyandina, V. Nekrutkin and A.A. Zhigljavsky* (2001)
91. Subjective Probability Models for Lifetimes *Fabio Spizzichino* (2001)
92. Empirical Likelihood *Art B. Owen* (2001)
93. Statistics in the 21st Century *Adrian E. Raftery, Martin A. Tanner, and Martin T. Wells* (2001)

94. Accelerated Life Models: Modeling and Statistical Analysis
Vilijandas Bagdonavicius and Mikhail Nikulin (2001)
95. Subset Selection in Regression, Second Edition *Alan Miller* (2002)
96. Topics in Modelling of Clustered Data *Marc Aerts, Helena Geys, Geert Molenberghs, and Louise M. Ryan* (2002)
97. Components of Variance *D.R. Cox and P.J. Solomon* (2002)
98. Design and Analysis of Cross-Over Trials, 2nd Edition *Byron Jones and Michael G. Kenward* (2003)
99. Extreme Values in Finance, Telecommunications, and the Environment
Bärbel Finkenstädt and Holger Rootzén (2003)
100. Statistical Inference and Simulation for Spatial Point Processes
Jesper Møller and Rasmus Plenge Waagepetersen (2004)
101. Hierarchical Modeling and Analysis for Spatial Data
Sudipto Banerjee, Bradley P. Carlin, and Alan E. Gelfand (2004)
102. Diagnostic Checks in Time Series *Wai Keung Li* (2004)
103. Stereology for Statisticians *Adrian Baddeley and Eva B. Vedel Jensen* (2004)
104. Gaussian Markov Random Fields: Theory and Applications *Håvard Rue and Leonhard Held* (2005)
105. Measurement Error in Nonlinear Models: A Modern Perspective, Second Edition
Raymond J. Carroll, David Ruppert, Leonard A. Stefanski, and Ciprian M. Crainiceanu (2006)
106. Generalized Linear Models with Random Effects: Unified Analysis via H-likelihood
Youngjo Lee, John A. Nelder, and Yudi Pawitan (2006)
107. Statistical Methods for Spatio-Temporal Systems
Bärbel Finkenstädt, Leonhard Held, and Valerie Isham (2007)
108. Nonlinear Time Series: Semiparametric and Nonparametric Methods *Jiti Gao* (2007)
109. Missing Data in Longitudinal Studies: Strategies for Bayesian Modeling and Sensitivity Analysis
Michael J. Daniels and Joseph W. Hogan (2008)
110. Hidden Markov Models for Time Series: An Introduction Using R
Walter Zucchini and Iain L. MacDonald (2009)
111. ROC Curves for Continuous Data *Wojtek J. Krzanowski and David J. Hand* (2009)
112. Antedependence Models for Longitudinal Data *Dale L. Zimmerman and Vicente A. Núñez-Antón* (2009)
113. Mixed Effects Models for Complex Data *Lang Wu* (2010)
114. Intoduction to Time Series Modeling *Genshiro Kitagawa* (2010)
115. Expansions and Asymptotics for Statistics *Christopher G. Small* (2010)
116. Statistical Inference: An Integrated Bayesian/Likelihood Approach *Murray Aitkin* (2010)
117. Circular and Linear Regression: Fitting Circles and Lines by Least Squares *Nikolai Chernov* (2010)
118. Simultaneous Inference in Regression *Wei Liu* (2010)
119. Robust Nonparametric Statistical Methods, Second Edition
Thomas P. Hettmansperger and Joseph W. McKean (2011)
120. Statistical Inference: The Minimum Distance Approach
Ayanendranath Basu, Hiroyuki Shioya, and Chanseok Park (2011)
121. Smoothing Splines: Methods and Applications *Yuedong Wang* (2011)
122. Extreme Value Methods with Applications to Finance *Serguei Y. Novak* (2012)
123. Dynamic Prediction in Clinical Survival Analysis *Hans C. van Houwelingen and Hein Putter* (2012)
124. Statistical Methods for Stochastic Differential Equations
Mathieu Kessler, Alexander Lindner, and Michael Sørensen (2012)
125. Maximum Likelihood Estimation for Sample Surveys
R. L. Chambers, D. G. Steel, Suojin Wang, and A. H. Welsh (2012)
126. Mean Field Simulation for Monte Carlo Integration *Pierre Del Moral* (2013)
127. Analysis of Variance for Functional Data *Jin-Ting Zhang* (2013)
128. Statistical Analysis of Spatial and Spatio-Temporal Point Patterns, Third Edition *Peter J. Diggle* (2013)
129. Constrained Principal Component Analysis and Related Techniques *Yoshio Takane* (2014)
130. Randomised Response-Adaptive Designs in Clinical Trials *Anthony C. Atkinson and Atanu Biswas* (2014)
131. Theory of Factorial Design: Single- and Multi-Stratum Experiments *Ching-Shui Cheng* (2014)
132. Quasi-Least Squares Regression *Justine Shults and Joseph M. Hilbe* (2014)
133. Data Analysis and Approximate Models: Model Choice, Location-Scale, Analysis of Variance, Nonparametric Regression and Image Analysis *Laurie Davies* (2014)
134. Dependence Modeling with Copulas *Harry Joe* (2014)
135. Hierarchical Modeling and Analysis for Spatial Data, Second Edition *Sudipto Banerjee, Bradley P. Carlin, and Alan E. Gelfand* (2014)

Monographs on Statistics and Applied Probability 135

Hierarchical Modeling and Analysis for Spatial Data

Second Edition

Sudipto Banerjee
Division of Biostatistics, School of Public Health
University of Minnesota, Minneapolis, USA

Bradley P. Carlin
Division of Biostatistics, School of Public Health
University of Minnesota, Minneapolis, USA

Alan E. Gelfand
Department of Statistical Science
Duke University, Durham, North Carolina, USA

CRC Press
Taylor & Francis Group
Boca Raton London New York

CRC Press is an imprint of the
Taylor & Francis Group, an **informa** business
A CHAPMAN & HALL BOOK

CRC Press
Taylor & Francis Group
6000 Broken Sound Parkway NW, Suite 300
Boca Raton, FL 33487-2742

© 2015 by Taylor & Francis Group, LLC
CRC Press is an imprint of Taylor & Francis Group, an Informa business

No claim to original U.S. Government works
Version Date: 20140527

International Standard Book Number-13: 978-1-4398-1918-0 (eBook - PDF)

This book contains information obtained from authentic and highly regarded sources. Reasonable efforts have been made to publish reliable data and information, but the author and publisher cannot assume responsibility for the validity of all materials or the consequences of their use. The authors and publishers have attempted to trace the copyright holders of all material reproduced in this publication and apologize to copyright holders if permission to publish in this form has not been obtained. If any copyright material has not been acknowledged please write and let us know so we may rectify in any future reprint.

Except as permitted under U.S. Copyright Law, no part of this book may be reprinted, reproduced, transmitted, or utilized in any form by any electronic, mechanical, or other means, now known or hereafter invented, including photocopying, microfilming, and recording, or in any information storage or retrieval system, without written permission from the publishers.

For permission to photocopy or use material electronically from this work, please access www.copyright.com (http://www.copyright.com/) or contact the Copyright Clearance Center, Inc. (CCC), 222 Rosewood Drive, Danvers, MA 01923, 978-750-8400. CCC is a not-for-profit organization that provides licenses and registration for a variety of users. For organizations that have been granted a photocopy license by the CCC, a separate system of payment has been arranged.

Trademark Notice: Product or corporate names may be trademarks or registered trademarks, and are used only for identification and explanation without intent to infringe.

Visit the Taylor & Francis Web site at
http://www.taylorandfrancis.com

and the CRC Press Web site at
http://www.crcpress.com

to Sharbani, Caroline, and Mariasun

Contents

Preface to the Second Edition **xvii**

Preface to the First Edition **xix**

1 Overview of spatial data problems **1**
1.1 Introduction to spatial data and models 1
1.1.1 Point-level models 4
1.1.2 Areal models 5
1.1.3 Point process models 6
1.1.4 Software and datasets 7
1.2 Fundamentals of cartography 8
1.2.1 Map projections 8
1.2.2 Calculating distances on the earth's surface 13
1.3 Maps and geodesics in R 16
1.4 Exercises 19

2 Basics of point-referenced data models **23**
2.1 Elements of point-referenced modeling 23
2.1.1 Stationarity 23
2.1.2 Variograms 24
2.1.3 Isotropy 25
2.1.4 Variogram model fitting 30
2.2 Anisotropy 31
2.2.1 Geometric anisotropy 31
2.2.2 Other notions of anisotropy 32
2.3 Exploratory approaches for point-referenced data 32
2.3.1 Basic techniques 32
2.3.2 Assessing anisotropy 37
2.3.2.1 Directional semivariograms and rose diagrams 37
2.3.2.2 Empirical semivariogram contour (ESC) plots 38
2.4 Classical spatial prediction 40
2.4.0.3 Noiseless kriging 43
2.5 Computer tutorials 44
2.5.1 EDA and spatial data visualization in R 44
2.5.2 Variogram analysis in R 47
2.6 Exercises 50

3 Some theory for point-referenced data models **53**
3.1 Formal modeling theory for spatial processes 53
3.1.1 Some basic stochastic process theory for spatial processes 55
3.1.2 Covariance functions and spectra 57
3.1.2.1 More general isotropic correlation functions 60

3.1.3 Constructing valid covariance functions 60
3.1.4 Smoothness of process realizations 61
3.1.5 Directional derivative processes 63
3.2 Nonstationary spatial process models ⋆ 63
3.2.1 Deformation 64
3.2.2 Nonstationarity through kernel mixing of process variables 65
3.2.3 Mixing of process distributions 69
3.3 Exercises 70

4 Basics of areal data models 73
4.1 Exploratory approaches for areal data 74
4.1.1 Measures of spatial association 75
4.1.2 Spatial smoothers 77
4.2 Brook's Lemma and Markov random fields 78
4.3 Conditionally autoregressive (CAR) models 80
4.3.1 The Gaussian case 81
4.3.2 The non-Gaussian case 84
4.4 Simultaneous autoregressive (SAR) models 85
4.4.1 CAR versus SAR models 87
4.4.2 STAR models 87
4.5 Computer tutorials 88
4.5.1 Adjacency matrices from maps using `spdep` 89
4.5.2 Moran's I and Geary's C in `spdep` 90
4.5.3 SAR and CAR model fitting using `spdep` in `R` 90
4.6 Exercises 95

5 Basics of Bayesian inference 97
5.1 Introduction to hierarchical modeling and Bayes' Theorem 97
5.1.1 Illustrations of Bayes' Theorem 98
5.2 Bayesian inference 100
5.2.1 Point estimation 100
5.2.2 Interval estimation 101
5.2.3 Hypothesis testing and model choice 101
5.2.3.1 Bayes factors 102
5.2.3.2 The DIC criterion 103
5.2.3.3 Posterior predictive loss criterion 105
5.2.3.4 Model assessment using hold out data 106
5.3 Bayesian computation 107
5.3.1 The Gibbs sampler 108
5.3.2 The Metropolis-Hastings algorithm 109
5.3.3 Slice sampling 111
5.3.4 Convergence diagnosis 112
5.3.5 Variance estimation 113
5.4 Computer tutorials 115
5.4.1 Basic Bayesian modeling in `R` 115
5.4.2 Advanced Bayesian modeling in `WinBUGS` 116
5.5 Exercises 118

6 Hierarchical modeling for univariate spatial data 123
6.1 Stationary spatial process models 123
6.1.1 Isotropic models 124
6.1.1.1 Prior specification 124
6.1.2 Bayesian kriging in `WinBUGS` 127
6.1.3 More general isotropic correlation functions, revisited 129
6.1.4 Modeling geometric anisotropy 133
6.2 Generalized linear spatial process modeling 136
6.3 Fitting hierarchical models for point-referenced data in `spBayes` 139
6.3.1 Gaussian spatial regression models 139
6.3.1.1 Prediction 143
6.3.1.2 Model selection 145
6.3.2 Non-Gaussian spatial GLM 146
6.4 Areal data models 150
6.4.1 Disease mapping 150
6.4.2 Traditional models and frequentist methods 151
6.4.3 Hierarchical Bayesian methods 152
6.4.3.1 Poisson-gamma model 152
6.4.3.2 Poisson-lognormal models 153
6.4.3.3 CAR models and their difficulties 155
6.4.4 Extending the CAR model 159
6.5 General linear areal data modeling 160
6.6 Comparison of point-referenced and areal data models 160
6.7 Exercises 161

7 Spatial misalignment 165
7.1 Point-level modeling 166
7.1.1 Gaussian process models 166
7.1.1.1 Motivating data set 166
7.1.1.2 Model assumptions and analytic goals 167
7.1.2 Methodology for the point-level realignment 168
7.2 Nested block-level modeling 173
7.2.1 Methodology for nested block-level realignment 173
7.2.2 Individual block group estimation 177
7.2.3 Aggregate estimation: Block groups near the Ithaca, NY, waste site 179
7.3 Nonnested block-level modeling 179
7.3.1 Motivating data set 180
7.3.2 Methodology for nonnested block-level realignment 182
7.3.2.1 Total population interpolation model 186
7.3.2.2 Age and sex effects 188
7.4 A data assimilation example 189
7.5 Misaligned regression modeling 190
7.6 Exercises 195

8 Modeling and Analysis for Point Patterns 199
8.1 Introduction 199
8.2 Theoretical development 203
8.2.1 Counting measure 204
8.2.2 Moment measures 205
8.3 Diagnostic tools 207
8.3.1 Exploratory data analysis; investigating complete spatial randomness 208

8.3.2 G and F functions 208
8.3.3 The K function 210
8.3.4 Empirical estimates of the intensity 213
8.4 Modeling point patterns; NHPP's and Cox processes 213
8.4.1 Parametric specifications 215
8.4.2 Nonparametric specifications 216
8.4.3 Bayesian modeling details 217
8.4.3.1 The "poor man's" version; revisiting the ecological fallacy 218
8.4.4 Examples 218
8.5 Generating point patterns 220
8.6 More general point pattern models 221
8.6.1 Cluster processes 221
8.6.1.1 Neyman-Scott processes 222
8.6.2 Shot noise processes 223
8.6.3 Gibbs processes 224
8.6.4 Further Bayesian model fitting and inference 225
8.6.5 Implementing fully Bayesian inference 226
8.6.6 An example 226
8.7 Marked point processes 228
8.7.1 Model specifications 228
8.7.2 Bayesian model fitting for marked point processes 229
8.7.3 Modeling clarification 230
8.7.4 Enriching intensities 231
8.7.4.1 Introducing non-spatial covariate information 233
8.7.4.2 Results of the analysis 235
8.8 Space-time point patterns 237
8.8.1 Space-time Poisson process models 238
8.8.2 Dynamic models for discrete time data 238
8.8.3 Space-time Cox process models using stochastic PDE's 239
8.8.3.1 Modeling the house construction data for Irving, TX 240
8.8.3.2 Results of the data analysis 242
8.9 Additional topics 242
8.9.1 Measurement error in point patterns 242
8.9.1.1 Modeling details 244
8.9.2 Presence-only data application 246
8.9.2.1 Probability model for presence locations 247
8.9.3 Scan statistics 251
8.9.4 Preferential sampling 252
8.10 Exercises 253

9 Multivariate spatial modeling for point-referenced data 257
9.1 Joint modeling in classical multivariate geostatistics 257
9.1.1 Co-kriging 259
9.1.2 Intrinsic multivariate correlation and nested models 260
9.2 Some theory for cross-covariance functions 261
9.3 Separable models 263
9.4 Spatial prediction, interpolation, and regression 264
9.4.1 Regression in the Gaussian case 266
9.4.2 Avoiding the symmetry of the cross-covariance matrix 268
9.4.3 Regression in a probit model 268
9.4.4 Examples 269

9.4.5 Conditional modeling 272
9.4.6 Spatial regression with kernel averaged predictors 275
9.5 Coregionalization models ⋆ 278
9.5.1 Coregionalization models and their properties 278
9.5.2 Unconditional and conditional Bayesian specifications 281
9.5.2.1 Equivalence of likelihoods 281
9.5.2.2 Equivalence of prior specifications 282
9.6 Spatially varying coefficient models 283
9.6.1 Approach for a single covariate 285
9.6.2 Multivariate spatially varying coefficient models 286
9.6.3 Spatially varying coregionalization models 288
9.6.4 Model-fitting issues 288
9.6.4.1 Fitting the joint model 288
9.6.4.2 Fitting the conditional model 289
9.7 Other constructive approaches ⋆ 293
9.7.1 Generalized linear model setting 296
9.8 Illustrating multivariate spatial modeling with `spBayes` 297
9.9 Exercises 301

10 Models for multivariate areal data 305
10.1 The multivariate CAR (MCAR) distribution 306
10.2 Modeling with a proper, non-separable MCAR distribution 308
10.3 Conditionally specified Generalized MCAR (GMCAR) distributions 311
10.4 Modeling using the GMCAR distribution 314
10.5 Illustration: Fitting conditional GMCAR to Minnesota cancer data 315
10.6 Coregionalized MCAR distributions 319
10.6.1 Case 1: Independent and identical latent CAR variables 319
10.6.2 Case 2: Independent but not identical latent CAR variables 320
10.6.3 Case 3: Dependent and not identical latent CAR variables 321
10.7 Modeling with coregionalized MCAR's 322
10.8 Illustrating coregionalized MCAR models with three cancers from Minnesota 324
10.9 Exercises 327

11 Spatiotemporal modeling 329
11.1 General modeling formulation 330
11.1.1 Preliminary analysis 330
11.1.2 Model formulation 331
11.1.3 Associated distributional results 333
11.1.4 Prediction and forecasting 335
11.2 Point-level modeling with continuous time 339
11.3 Nonseparable spatiotemporal models ⋆ 343
11.4 Dynamic spatiotemporal models ⋆ 344
11.4.1 Brief review of dynamic linear models 345
11.4.2 Formulation for spatiotemporal models 345
11.4.3 Spatiotemporal data 348
11.5 Fitting dynamic spatiotemporal models using `spBayes` 352
11.6 Geostatistical space-time modeling driven by differential equations 355
11.7 Areal unit space-time modeling 361
11.7.1 Aligned data 361

11.7.2 Misalignment across years 365
11.7.3 Nested misalignment both within and across years 367
11.7.4 Nonnested misalignment and regression 370
11.8 Areal-level continuous time modeling 373
11.8.1 Areally referenced temporal processes 374
11.8.2 Hierarchical modeling 376
11.9 Exercises 378

12 Modeling large spatial and spatiotemporal datasets 381
12.1 Introduction 381
12.2 Approximate likelihood approaches 381
12.2.1 Spectral methods 381
12.2.2 Lattice and conditional independence methods 382
12.2.3 INLA 383
12.2.4 Approximate likelihood 384
12.2.5 Variational Bayes algorithm for spatial models 384
12.2.6 Covariance tapering 386
12.3 Models for large spatial data: low rank models 386
12.3.1 Kernel-based dimension reduction 387
12.3.2 The Karhunen-Loéve representation of Gaussian processes 388
12.4 Predictive process models 390
12.4.1 The predictive process 390
12.4.2 Properties of the predictive process 392
12.4.3 Biases in low-rank models and the bias-adjusted modified predictive process 393
12.4.4 Selection of knots 395
12.4.5 A simulation example using the two step analysis 397
12.4.6 Non-Gaussian first stage models 397
12.4.7 Spatiotemporal versions 398
12.4.8 Multivariate predictive process models 399
12.5 Modeling with the predictive process 400
12.6 Fitting a predictive process model in `spBayes` 404
12.7 Exercises 411

13 Spatial gradients and wombling 413
13.1 Introduction 413
13.2 Process smoothness revisited ⋆ 415
13.3 Directional finite difference and derivative processes 417
13.4 Distribution theory for finite differences and directional gradients 418
13.5 Directional derivative processes in modeling 420
13.6 Illustration: Inference for differences and gradients 422
13.7 Curvilinear gradients and wombling 424
13.7.1 Gradients along curves 424
13.7.2 Wombling boundary 426
13.8 Distribution theory for curvilinear gradients 427
13.9 Illustration: Spatial boundaries for invasive plant species 429
13.10 Areal wombling 432
13.10.1 Review of existing methods 433
13.10.2 Simple MRF-based areal wombling 434
13.10.2.1 Adding covariates 437
13.10.3 Joint site-edge areal wombling 438

13.10.3.1 Edge smoothing and random neighborhood structure 439
13.10.3.2 Two-level CAR model 439
13.10.3.3 Site-edge (SE) models 440
13.10.4 FDR-based areal wombling 444
13.11 Wombling with point process data 445
13.12 Concluding remarks 445

14 Spatial survival models 447
14.1 Parametric models 448
14.1.1 Univariate spatial frailty modeling 448
14.1.1.1 Bayesian implementation 449
14.1.2 Spatial frailty versus logistic regression models 453
14.2 Semiparametric models 454
14.2.1 Beta mixture approach 455
14.2.2 Counting process approach 456
14.3 Spatiotemporal models 457
14.3.1 Results for the full model 459
14.3.2 Bayesian model choice 460
14.4 Multivariate models ⋆ 462
14.4.1 Static spatial survival data with multiple causes of death 462
14.4.2 MCAR specification, simplification, and computing 462
14.4.3 Spatiotemporal survival data 463
14.5 Spatial cure rate models ⋆ 466
14.5.1 Models for right- and interval-censored data 468
14.5.1.1 Right-censored data 468
14.5.1.2 Interval-censored data 471
14.5.2 Spatial frailties in cure rate models 471
14.5.3 Model comparison 472
14.6 Exercises 475

15 Special topics in spatial process modeling 479
15.1 Data assimilation 479
15.1.1 Algorithmic and pseudo-statistical approaches in weather prediction 479
15.1.2 Fusion modeling using stochastic integration 480
15.1.3 The downscaler 482
15.1.4 Spatiotemporal versions 484
15.1.5 An illustration 485
15.2 Space-time modeling for extremes 486
15.2.1 Possibilities for modeling maxima 487
15.2.2 Review of extreme value theory 488
15.2.3 A continuous spatial process model 489
15.2.4 Using copulas 490
15.2.5 Hierarchical modeling for spatial extreme values 491
15.3 Spatial CDF's 492
15.3.1 Basic definitions and motivating data sets 492
15.3.2 Derived-process spatial CDF's 495
15.3.2.1 Point- versus block-level spatial CDF's 495
15.3.2.2 Covariate weighted SCDF's for misaligned data 496
15.3.3 Randomly weighted SCDF's 496

Appendices **501**

A Spatial computing methods **503**
A.1 Fast Fourier transforms 503
A.2 Slice Gibbs sampling for spatial process model fitting 504
A.2.1 Constant mean process with nugget 507
A.2.2 Mean structure process with no pure error component 508
A.2.3 Mean structure process with nugget 509
A.3 Structured MCMC sampling for areal model fitting 509
A.3.1 SMCMC algorithm basics 510
A.3.2 Applying structured MCMC to areal data 510
A.3.3 Algorithmic schemes 512
A.4 `spBayes`: Under the hood 513

B Answers to selected exercises **515**

Bibliography **529**

Preface to the Second Edition

In the ten years that have passed since the first edition of this book, we believe the statistical landscape has changed substantially, even more so for analyzing space and space-time data. Apart from the remarkable growth in data collection, with datasets now of enormous size, the fields of statistics and biostatistics are also witnessing a change toward examination of observational data, rather than being restricted to carefully-collected experimentally designed data. We are witnessing an increased examination of complex systems using such data, requiring synthesis of multiple sources of information (empirical, theoretical, physical, etc.), necessitating the development of multi-level models. We are seeing repeated exemplification of the hierarchical framework $[data|process, parameters][process|parameters][parameters]$. The role of the statistician is evolving in this landscape to that of an integral participant in team-based research: A participant in the framing of the questions to be investigated, the determination of data needs to investigate these questions, the development of models to examine these questions, the development of strategies to fit these models, and the analysis and summarization of the resultant inference under these specifications. It is an exciting new world for modern statistics, and spatial analysis is a particularly important player in this new world due to the increased appreciation of the information carried in spatial locations, perhaps across temporal scales, in learning about these complex processes. Applications abound, particularly in the environmental sciences but also in public health, real estate, and many other fields.

We believe this new edition moves forward in this spirit. The first edition was intended as a research monograph, presenting a state-of-the-art treatment of hierarchical modeling for spatial data. It has been a delightful success, far exceeding our expectations in terms of sales and reception by the community. However, reflecting on the decade that has passed, we have made consequential changes from the first edition. Not surprisingly, the new volume is more than 50% bigger, reflecting the major growth in spatial statistics as a research area and as an area of application.

Rather than describing the contents, chapter by chapter, we note the following major changes. First, we have added a much needed chapter on spatial point patterns. This is a subfield that is finding increased importance but, in terms of application, has lagged behind the use of point-referenced and areal unit data. We offer roughly 80 new pages here, developed primarily from a modeling perspective, introducing as much current hierarchical and Bayesian flavor as we could. Second, reflecting the ubiquitous increases in the sizes of datasets, we have developed a "big data" chapter. Here, we focus on the predictive process in its various forms, as an attractive tool for handling reasonably large datasets. Third, near the end of the book we have added a new chapter on spatial and spatiotemporal gradient modeling, with associated developments by us and others in spatial boundary analysis and wombling. As elsewhere in the book, we divide our descriptions here into those appropriate for point-referenced data (where underlying spatial processes guarantee the existence of spatial derivatives) and areal data (where processes are not possible but boundaries can still be determined based on alternate ways of hierarchically smoothing the areal map). Fourth, since geostatistical (point-referenced) modeling is still the most prevalent setting for spatial analysis, we have chosen to present this material in two separate chapters. The first (Chapter 2) is a basic introduction, presented for the reader who is more focused on the

practical side of things. In addition, we have developed a more theoretical chapter (Chapter 3) which provides much more insight into the scope of issues that arise in the geostatistical setting and how we deal with them formally. The presentation of this material is still gentle compared with that in many stochastic processes texts, and we hope it provides valuable model-building insight. At the same time, we recognize that Chapter 3 may be somewhat advanced for more introductory courses, so we marked it as a starred chapter. In addition to these four new chapters, we have greatly revised and expanded the multivariate and spatio-temporal chapters, again in response to the growth of work in these areas. We have also added two new special topics sections, one on data fusion/assimilation, and one on spatial analysis for data on extremes. We have roughly doubled the number of exercises in the book, and also include many more color figures, now integrated appropriately into the text. Finally, we have updated the computational aspects of the book. Specially, we work with the newest version of `WinBUGS`, the new flexible `spBayes` software, and we introduce other suitable `R` packages as needed, especially for exploratory data analysis.

In addition to those to whom we expressed our gratitude in the preface to the first edition, we now extend this list to record (in alphabetical order) the following colleagues, current and former postdoctoral researchers and students: Dipankar Bandyopadhyay, Veronica Berrocal, Avishek Chakraborty, Jim Clark, Jason (Jun) Duan, David Dunson, Andrew Finley, Souparno Ghosh, Simone Gray, Rajarshi Guhaniyogi, Michele Guindani, Xiaoping Jin, Giovanna Jona Lasinio, Matt Heaton, Dave Holland, Thanasis Kottas, Andrew Latimer, Tommy Leininger, Pei Li, Shengde Liang, Haolan Lu, Kristian Lum, Haijun Ma, Marshall McBean, Marie Lynn Miranda, Joao Vitor Monteiro, XuanLong Nguyen, Lucia Paci, Sonia Petrone, Gavino Puggioni, Harrison Quick, Cavan Reilly, Qian Ren, Abel Rodriguez, Huiyan Sang, Sujit Sahu, Maria Terres, Beth Virnig, Fangpo Wang, Adam Wilson, Gangqiang Xia, and Kai Zhu. In addition, we much appreciate the continuing support of CRC/Chapman and Hall in helping to bring this new edition to fruition, in particular the encouragement of the steadfast and indefatigable Rob Calver.

Sudipto Banerjee — Minneapolis, Minnesota
Bradley P. Carlin — Durham, North Carolina
Alan E. Gelfand — July 2013

Preface to the First Edition

As recently as two decades ago, the impact of hierarchical Bayesian methods outside of a small group of theoretical probabilists and statisticians was minimal at best. Realistic models for challenging data sets were easy enough to write down, but the computations associated with these models required integrations over hundreds or even thousands of unknown parameters, far too complex for existing computing technology. Suddenly, around 1990, the "Markov chain Monte Carlo (MCMC) revolution" in Bayesian computing took place. Methods like the Gibbs sampler and the Metropolis algorithm, when coupled with ever-faster workstations and personal computers, enabled evaluation of the integrals that had long thwarted applied Bayesians. Almost overnight, Bayesian methods became not only feasible, but the method of choice for almost any model involving multiple levels incorporating random effects or complicated dependence structures. The growth in applications has also been phenomenal, with a particularly interesting recent example being a Bayesian program to delete spam from your incoming email (see `popfile.sourceforge.net`).

Our purpose in writing this book is to describe hierarchical Bayesian methods for one class of applications in which they can pay substantial dividends: spatial (and spatiotemporal) statistics. While all three of us have been working in this area for some time, our motivation for writing the book really came from our experiences teaching courses on the subject (two of us at the University of Minnesota, and the other at the University of Connecticut). In teaching we naturally began with the textbook by Cressie (1993), long considered the standard as both text and reference in the field. But we found the book somewhat uneven in its presentation, and written at a mathematical level that is perhaps a bit high, especially for the many epidemiologists, environmental health researchers, foresters, computer scientists, GIS experts, and other users of spatial methods who lacked significant background in mathematical statistics. Now a decade old, the book also lacks a current view of hierarchical modeling approaches for spatial data.

But the problem with the traditional teaching approach went beyond the mere need for a less formal presentation. Time and again, as we presented the traditional material, we found it wanting in terms of its flexibility to deal with realistic assumptions. Traditional Gaussian kriging is obviously the most important method of point-to-point spatial interpolation, but extending the paradigm beyond this was awkward. For areal (block-level) data, the problem seemed even more acute: CAR models should most naturally appear as priors for the parameters in a model, not as a model for the observations themselves.

This book, then, attempts to remedy the situation by providing a fully Bayesian treatment of spatial methods. We begin in Chapter 1 by outlining and providing illustrative examples of the three types of spatial data: point-level (geostatistical), areal (lattice), and spatial point process. We also provide a brief introduction to map projection and the proper calculation of distance on the earth's surface (which, since the earth is round, can differ markedly from answers obtained using the familiar notion of Euclidean distance). Our statistical presentation begins in earnest in Chapter 2, where we describe both exploratory data analysis tools and traditional modeling approaches for point-referenced data. Modeling approaches from traditional geostatistics (variogram fitting, kriging, and so forth) are covered here. Chapter 4 offers a similar presentation for areal data models, again starting

with choropleth maps and other displays and progressing toward more formal statistical models. This chapter also presents Brook's Lemma and Markov random fields, topics that underlie the conditional, intrinsic, and simultaneous autoregressive (CAR, IAR, and SAR) models so often used in areal data settings.

Chapter 5 provides a review of the hierarchical Bayesian approach in a fairly generic setting, for readers previously unfamiliar with these methods and related computing and software. (The penultimate sections of Chapters 2, 4, and 5 offer tutorials in several popular software packages.) This chapter is not intended as a replacement for a full course in Bayesian methods (as covered, for example, by Carlin and Louis, 2000, or Gelman et al., 2004), but should be sufficient for readers having at least some familiarity with the ideas. In Chapter 6 then we are ready to cover hierarchical modeling for univariate spatial response data, including Bayesian kriging and lattice modeling. The issue of nonstationarity (and how to model it) also arises here.

Chapter 7 considers the problem of spatially misaligned data. Here, Bayesian methods are particularly well suited to sorting out complex interrelationships and constraints and providing a coherent answer that properly accounts for all spatial correlation and uncertainty. Methods for handling multivariate spatial responses (for both point- and block-level data) are discussed in Chapter 9. Spatiotemporal models are considered in Chapter 11, while Chapter 14 presents an extended application of areal unit data modeling in the context of survival analysis methods. Chapter 15 considers novel methodology associated with spatial process modeling, including spatial directional derivatives, spatially varying coefficient models, and spatial cumulative distribution functions (SCDF's). Finally, the book also features two useful appendices. Appendix A reviews elements of matrix theory and important related computational techniques, while Appendix B contains solutions to several of the exercises in each of the book's chapters.

Our book is intended as a research monograph, presenting the "state of the art" in hierarchical modeling for spatial data, and as such we hope readers will find it useful as a desk reference. However, we also hope it will be of benefit to instructors (or self-directed students) wishing to use it as a textbook. Here we see several options. Students wanting an introduction to methods for point-referenced data (traditional geostatistics and its extensions) may begin with Chapter 1, Chapter 2, Chapter 5, and Section 6.1 to Section 3.2. If areal data models are of greater interest, we suggest beginning with Chapter 1, Chapter 4, Chapter 5, Section 6.4, and Section 6.5. In addition, for students wishing to minimize the mathematical presentation, we have also marked sections containing more advanced material with a star ($\star$). These sections may be skipped (at least initially) at little cost to the intelligibility of the subsequent narrative. In our course in the Division of Biostatistics at the University of Minnesota, we are able to cover much of the book in a 3-credit-hour, single-semester (15-week) course. We encourage the reader to check `http://www.biostat.umn.edu/~brad/` on the web for many of our data sets and other teaching-related information.

We owe a debt of gratitude to those who helped us make this book a reality. Kirsty Stroud and Bob Stern took us to lunch and said encouraging things (and more importantly, picked up the check) whenever we needed it. Cathy Brown, Alex Zirpoli, and Desdamona Racheli prepared significant portions of the text and figures. Many of our current and former graduate and postdoctoral students, including Yue Cui, Xu Guo, Murali Haran, Xiaoping Jin, Andy Mugglin, Margaret Short, Amy Xia, and Li Zhu at Minnesota, and Deepak Agarwal, Mark Ecker, Sujit Ghosh, Hyon-Jung Kim, Ananda Majumdar, Alexandra Schmidt, and Shanshan Wu at the University of Connecticut, played a big role. We are also grateful to the Spring 2003 *Spatial Biostatistics* class in the School of Public Health at the University of Minnesota for taking our draft for a serious "test drive." Colleagues Jarrett Barber, Nicky Best, Montserrat Fuentes, David Higdon, Jim Hodges, Oli Schabenberger, John Silander, Jon Wakefield, Melanie Wall, Lance Waller, and many others provided valuable input and

assistance. Finally, we thank our families, whose ongoing love and support made all of this possible.

Sudipto Banerjee — Minneapolis, Minnesota
Bradley P. Carlin — Durham, North Carolina
Alan E. Gelfand — October 2003

Chapter 1

Overview of spatial data problems

1.1 Introduction to spatial data and models

Researchers in diverse areas such as climatology, ecology, environmental health, and real estate marketing are increasingly faced with the task of analyzing data that are:

- highly multivariate, with many important predictors and response variables,
- geographically referenced, and often presented as maps, and
- temporally correlated, as in longitudinal or other time series structures.

For example, for an epidemiological investigation, we might wish to analyze lung, breast, colorectal, and cervical cancer rates by county and year in a particular state, with smoking, mammography, and other important screening and staging information also available at some level. Public health professionals who collect such data are charged not only with surveillance, but also statistical *inference* tasks, such as *modeling* of trends and correlation structures, *estimation* of underlying model parameters, *hypothesis testing* (or comparison of competing models), and *prediction* of observations at unobserved times or locations.

In this text we seek to present a practical, self-contained treatment of hierarchical modeling and data analysis for complex spatial (and spatiotemporal) datasets. Spatial statistics methods have been around for some time, with the landmark work by Cressie (1993) providing arguably the only comprehensive book in the area. However, recent developments in Markov chain Monte Carlo (MCMC) computing now allow fully Bayesian analyses of sophisticated multilevel models for complex geographically referenced data. This approach also offers full inference for non-Gaussian spatial data, multivariate spatial data, spatiotemporal data, and, for the first time, solutions to problems such as geographic and temporal misalignment of spatial data layers.

This book does not attempt to be fully comprehensive, but does attempt to present a fairly thorough treatment of hierarchical Bayesian approaches for handling all of these problems. The book's mathematical level is roughly comparable to that of Carlin and Louis (2000). That is, we sometimes state results rather formally, but spend little time on theorems and proofs. For more mathematical treatments of spatial statistics (at least on the geostatistical side), the reader is referred to Cressie (1993), Wackernagel (1998), Chiles and Delfiner (1999), and Stein (1999a). For more descriptive presentations the reader might consult Bailey and Gattrell (1995), Fotheringham and Rogerson (1994), or Haining (1990). Our primary focus is on the issues of *modeling* (where we offer rich, flexible classes of hierarchical structures to accommodate both static and dynamic spatial data), *computing* (both in terms of MCMC algorithms and methods for handling very large matrices), and *data analysis* (to illustrate the first two items in terms of inferential summaries and graphical displays). Reviews of both traditional spatial methods (Chapters 2, 3 and 4) and Bayesian methods (Chapter 5) attempt to ensure that previous exposure to either of these two areas is not required (though it will of course be helpful if available).

Figure 1.1 *Map of PM2.5 sampling sites over three midwestern U.S. states; plotting character indicates range of average monitored PM2.5 level over the year 2001.*

Following convention, we classify spatial data sets into one of three basic types:

- *point-referenced data*, where $Y(\mathbf{s})$ is a random vector at a location $\mathbf{s} \in \Re^r$, where $\mathbf{s}$ varies *continuously* over D, a fixed subset of $\Re^r$ that contains an r-dimensional rectangle of positive volume;
- *areal data*, where D is again a fixed subset (of regular or irregular shape), but now partitioned into a finite number of areal units with well-defined boundaries;
- *point pattern data*, where now D is itself random; its index set gives the locations of random events that are the spatial point pattern. $Y(\mathbf{s})$ itself can simply equal 1 for all $\mathbf{s} \in D$ (indicating occurrence of the event), or possibly give some additional covariate information (producing a *marked point pattern process*).

The first case is often referred to as *geocoded* or *geostatistical* data, names apparently arising from the long history of these types of problems in mining and other geological sciences. Figure 1.1 offers an example of this case, showing the locations of 114 air-pollution monitoring sites in three midwestern U.S. states (Illinois, Indiana, and Ohio). The plotting character indicates the 2001 annual average PM2.5 level (measured in ppb) at each site. PM2.5 stands for particulate matter less than 2.5 microns in diameter, and is a measure of the density of very small particles that can travel through the nose and windpipe and into the lungs, potentially damaging a person's health. Here we might be interested in a model of the geographic distribution of these levels that account for spatial correlation and perhaps underlying covariates (regional industrialization, traffic density, and the like). The use of colors makes it somewhat easier to read, since the color allows the categories to be ordered more naturally, and helps sharpen the contrast between the urban and rural areas. Again, traditional analysis methods for point level data like this are described in Chapter 2, while Chapter 6 introduces the corresponding hierarchical modeling approach.

The second case above (areal data) is often referred to as *lattice* data, a term we find misleading since it connotes observations corresponding to "corners" of a checkerboard-like grid. Of course, there *are* data sets of this type; for example, as arising from agricultural field trials (where the plots cultivated form a regular lattice) or image restoration (where the data correspond to pixels on a screen, again in a regular lattice). However, in practice most areal data are summaries over an *irregular* lattice, like a collection of county or other

Figure 1.2 `ArcView` *map of percent of surveyed population with household income below 200% of the federal poverty limit, regional survey units in Hennepin County, MN.*

regional boundaries, as in Figure 1.2. Here we have information on the percent of a surveyed population with household income falling below 200% of the federal poverty limit for a collection of regions comprising Hennepin County, MN. Note that we have no information on any single household in the study area, only regional summaries for each region. Figure 1.2 is an example of a *choropleth map*, meaning that it uses shades of color (or greyscale) to classify values into a few broad classes (six in this case), like a histogram (bar chart) for nonspatial data. Choropleth maps are visually appealing (and therefore, also common), but of course provide a rather crude summary of the data, and one that can be easily altered simply by manipulating the class cutoffs.

As with any map of the areal units, choropleth maps *do* show reasonably precise *boundaries* between the regions (i.e., a series of exact spatial coordinates that when connected in the proper order will trace out each region), and thus we also know which regions are adjacent to (touch) which other regions. Thus the "sites" $\mathbf{s} \in D$ in this case are actually the regions (or *blocks*) themselves, which in this text we will denote not by $\mathbf{s}_i$ but by B_i, $i = 1, \ldots, n$, to avoid confusion between points $\mathbf{s}_i$ and blocks B_i. It may also be illuminating to think of the county centroids as forming the vertices of an irregular lattice, with two lattice points being connected if and only if the counties are "neighbors" in the spatial map, with physical adjacency being the most obvious (but not the only) way to define a region's neighbors.

Some spatial data sets feature *both* point- and areal-level data, and require their simultaneous display and analysis. Figure 1.3 offers an example of this case. The first component of this data set is a collection of eight-hour maximum ozone levels at 10 monitoring sites in the greater Atlanta, GA, area for a particular day in July 1995. Like the observations in Figure 1.1, these were made at fixed monitoring stations for which exact spatial coordinates (say, latitude and longitude) are known. (That is, we assume the $Y(\mathbf{s}_i)$, $i = 1, \ldots, 10$ are random, but the $\mathbf{s}_i$ are not.) The second component of this data set is the number of children in the area's zip codes (shown using the irregular subboundaries on the map) that reported at local emergency rooms (ERs) with acute asthma symptoms on the following day; confidentiality of health records precludes us from learning the precise address of any of the children. These are areal summaries that could be indicated by shading the zip codes, as in Figure 1.2. An obvious question here is whether we can establish a connection between high ozone and subsequent high pediatric ER asthma visits. Since the data are misaligned (point-level ozone but block-level ER counts), a formal statistical investigation of this question requires a preliminary *realignment* of the data; this is the subject of Chapter 7.

Figure 1.3 *Zip code boundaries in the Atlanta metropolitan area and 8-hour maximum ozone levels (ppm) at 10 monitoring sites for July 15, 1995.*

The third case above (spatial point pattern data) could be exemplified by residences of persons suffering from a particular disease, or by locations of a certain species of tree in a forest. Here the response Y is often fixed (occurrence of the event), and only the locations $\mathbf{s}_i$ are thought of as random. In some cases this information might be supplemented by age or other covariate information, producing a *marked* point pattern). Such data are often of interest in studies of event *clustering*, where the goal is to determine whether an observed spatial point pattern is an example of a clustered process (where points tend to be spatially close to other points), or merely the result of a random event process operating independently and homogeneously over space. Note that in contrast to areal data, where no individual points in the data set could be identified, here (and in point-referenced data as well) precise locations are known, and so must often be protected to protect the privacy of the persons in the set.

In the remainder of this initial section, we give a brief outline of the basic models most often used for each of these three data types. Here we only intend to give a flavor of the models and techniques to be fully described in the remainder of this book.

Even though our preferred inferential outlook is Bayesian, the statistical inference tools discussed in Chapters 2 through 4 are entirely classical. While all subsequent chapters adopt the Bayesian point of view, our objective here is to acquaint the reader with the classical techniques first, since they are more often implemented in standard software packages. Moreover, as in other fields of data analysis, classical methods can be easier to compute, and produce perfectly acceptable results in relatively simple settings. Classical methods often have interpretations as limiting cases of Bayesian methods under increasingly vague prior assumptions. Finally, classical methods can provide insight for formulating and fitting hiearchical models.

1.1.1 Point-level models

In the case of point-level data, the location index $\mathbf{s}$ varies *continuously* over D, a fixed subset of $\Re^d$. Suppose we assume that the covariance between the random variables at two

locations depends on the *distance* between the locations. One frequently used association specification is the exponential model. Here the covariance between measurements at two locations is an exponential function of the interlocation distance, i.e., $Cov(Y(\mathbf{s}_i), Y(\mathbf{s}_{i'})) \equiv C(d_{ii'}) = \sigma^2 e^{-\phi d_{ii'}}$ for $i \neq i'$, where $d_{ii'}$ is the distance between sites s_i and $s_{i'}$, and σ^2 and ϕ are positive parameters called the *partial sill* and the *decay parameter*, respectively ($1/\phi$ is called the *range parameter*). A plot of the covariance versus distance is called the *covariogram*. When $i = i'$, $d_{ii'}$ is of course 0, and $C(d_{ii'}) = Var(Y(\mathbf{s}_i))$ is often expanded to $\tau^2 + \sigma^2$, where $\tau^2 > 0$ is called a *nugget effect*, and $\tau^2 + \sigma^2$ is called the *sill*. Of course, while the exponential model is convenient and has some desirable properties, many other parametric models are commonly used; see Section 2.1 for further discussion of these and their relative merits.

Adding a joint distributional model to these variance and covariance assumptions then enables likelihood inference in the usual way. The most convenient approach would be to assume a multivariate *normal* (or *Gaussian*) distribution for the data. That is, suppose we are given observations $\mathbf{Y} \equiv \{Y(\mathbf{s}_i)\}$ at known locations $\mathbf{s}_i$, $i = 1, \ldots, n$. We then assume that

$$\mathbf{Y} \mid \mu, \boldsymbol{\theta} \sim N_n(\mu \mathbf{1}\ ,\ \Sigma(\boldsymbol{\theta}))\ , \tag{1.1}$$

where N_n denotes the n-dimensional normal distribution, μ is the (constant) mean level, $\mathbf{1}$ is a vector of ones, and $(\Sigma(\boldsymbol{\theta}))_{ii'}$ gives the covariance between $Y(\mathbf{s}_i)$ and $Y(\mathbf{s}_{i'})$. For the variance-covariance specification of the previous paragraph, we have $\boldsymbol{\theta} = (\tau^2, \sigma^2, \phi)^T$, since the covariance matrix depends on the nugget, sill, and range.

In fact, the simplest choices for Σ are those corresponding to *isotropic* covariance functions, where we assume that the spatial correlation is a function solely of the distance $d_{ii'}$ between $\mathbf{s}_i$ and $\mathbf{s}_{i'}$. As mentioned above, exponential forms are particularly intuitive examples. Here,

$$(\Sigma(\boldsymbol{\theta}))_{ii'} = \sigma^2 \exp(-\phi d_{ii'}) + \tau^2 I(i = i'),\ \sigma^2 > 0,\ \phi > 0,\ \tau^2 > 0\ , \tag{1.2}$$

where I denotes the indicator function (i.e., $I(i = i') = 1$ if $i = i'$, and 0 otherwise). Many other choices are possible for $Cov(Y(\mathbf{s}_i), Y(\mathbf{s}_{i'}))$, including for example the powered exponential,

$$(\Sigma(\boldsymbol{\theta}))_{ii'} = \sigma^2 \exp(-\phi d_{ii'}^{\kappa}) + \tau^2 I(i = i'),\ \sigma^2 > 0,\ \phi > 0,\ \tau^2 > 0,\ \kappa \in (0, 2]\,,$$

the spherical, the Gaussian, and the Matérn (see Subsection 2.1.3 for a full discussion). In particular, while the latter requires calculation of a modified Bessel function, Stein (1999a, p. 51) illustrates its ability to capture a broader range of local correlation behavior despite having no more parameters than the powered exponential. We shall say much more about point-level spatial methods and models in Chapters 2, 3 and 6 and also provide illustrations using freely available statistical software.

1.1.2 Areal models

In models for areal data, the geographic regions or *blocks* (zip codes, counties, etc.) are denoted by B_i, and the data are typically sums or averages of variables over these blocks. To introduce spatial association, we define a *neighborhood* structure based on the arrangement of the blocks in the map. Once the neighborhood structure is defined, models resembling autoregressive time series models are considered. Two very popular models that incorporate such neighborhood information are the *simultaneously* and *conditionally autoregressive* models (abbreviated SAR and CAR), originally developed by Whittle (1954) and Besag (1974), respectively. The SAR model is computationally convenient for use with likelihood methods. By contrast, the CAR model is computationally convenient for Gibbs sampling used

in conjunction with Bayesian model fitting, and in this regard is often used to incorporate spatial correlation through a vector of spatially varying random effects $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_n)^T$. For example, writing $Y_i \equiv Y(B_i)$, we might assume $Y_i \overset{ind}{\sim} N(\phi_i, \sigma^2)$, and then impose the CAR model

$$\phi_i | \boldsymbol{\phi}_{(-i)} \sim N\left(\mu + \sum_{j=1}^{n} a_{ij}(\phi_j - \mu)\ ,\ \tau_i^2\right)\ , \tag{1.3}$$

where $\boldsymbol{\phi}_{(-i)} = \{\phi_j : j \neq i\}$, τ_i^2 is the conditional variance, and the a_{ij} are known or unknown constants such that $a_{ii} = 0$ for $i = 1, \ldots, n$. Letting $A = (a_{ij})$ and $M = Diag(\tau_1^2, \ldots, \tau_n^2)$, by Brook's Lemma (c.f. Section 4.2), we can show that

$$p(\boldsymbol{\phi}) \propto \exp\{-(\boldsymbol{\phi} - \mu\mathbf{1})^T M^{-1}(I - A)(\boldsymbol{\phi} - \mu\mathbf{1})/2\}\ , \tag{1.4}$$

where $\mathbf{1}$ is an n-vector of 1's, and I is a $n \times n$ identity matrix.

A common way to construct A and M is to let $A = \rho\, Diag(1/w_{i+})W$ and $M^{-1} = \tau^{-2} Diag(w_{i+})$. Here ρ is referred to as the *spatial correlation* parameter, and $W = (w_{ij})$ is a neighborhood matrix for the areal units, which can be defined as

$$w_{ij} = \begin{cases} 1 & \text{if subregions } i \text{ and } j \text{ share a common boundary, } i \neq j \\ 0 & \text{otherwise} \end{cases}\ . \tag{1.5}$$

Thus $Diag(w_{i+})$ is a diagonal matrix with (i, i) entry equal to $w_{i+} = \sum_j w_{ij}$. Letting $\boldsymbol{\alpha} \equiv (\rho, \tau^2)$, the covariance matrix of $\boldsymbol{\phi}$ then becomes $C(\boldsymbol{\alpha}) = \tau^2[Diag(w_{i+}) - \rho W]^{-1}$, where the inverse exists for an appropriate range of ρ values; see Subsection 4.3.1.

In the context of Bayesian hierarchical areal modeling, when choosing a prior distribution $\pi(\boldsymbol{\phi})$ for a vector of spatial random effects $\boldsymbol{\phi}$, the CAR distribution (1.3) is often used with the 0–1 *weight* (or *adjacency*) *matrix* W in (1.5) and $\rho = 1$. While this results in an *improper* (nonintegrable) prior distribution, this problem is remedied by imposing a sum-to-zero constraint on the ϕ_i (which turns out to be easy to implement numerically using Gibbs sampling). In this case the more general conditional form (1.3) is replaced by

$$\phi_i \mid \boldsymbol{\phi}_{(-i)} \sim N(\bar{\phi}_i\ ,\ \tau^2/m_i)\ , \tag{1.6}$$

where $\bar{\phi}_i$ is the average of the $\phi_{j \neq i}$ that are adjacent to ϕ_i, and m_i is the number of these adjacencies (see, e.g., Besag, York, and Mollié, 1991). We discuss areal models in greater detail in Chapters 4 and 6.

1.1.3 Point process models

In the point process model, the spatial domain D is itself random, so that the elements of the index set D are the locations of random events that constitute the spatial point pattern. $Y(\mathbf{s})$ then normally equals the constant 1 for all $\mathbf{s} \in D$ (indicating occurrence of the event), but it may also provide additional covariate information, in which case the data constitute a marked point process.

Questions of interest with data of this sort typically center on whether the data are *clustered* more or less than would be expected if the locations were determined completely by chance. Stochastically, such uniformity is often described through a *homogeneous Poisson process*, which implies that the expected number of occurrences in region A is $\lambda|A|$, where λ is the *intensity* parameter of the process and $|A|$ is the area of A. To investigate this in practice, plots of the data are typically a good place to start, but the tendency of the human eye to see clustering or other structure in virtually every point pattern renders a strictly graphical approach unreliable. Instead, statistics that measure clustering, and

perhaps even associated significance tests, are often used. The most common of these is *Ripley's K function*, given by

$$K(d) = \frac{1}{\lambda} E[\text{number of points within } d \text{ of an arbitrary point}], \tag{1.7}$$

where again λ is the intensity of the process, i.e., the mean number of points per unit area.

The theoretical value of K is known for certain spatial point process models. For instance, for point processes that have no spatial dependence at all, we would have $K(d) = \pi d^2$, since in this case the number of points within d of an arbitrary point should be proportional to the area of a circle of radius d; the K function then divides out the average intensity λ. However, if the data are clustered we might expect $K(d) > \pi d^2$, while if the points follow some regularly spaced pattern we would expect $K(d) < \pi d^2$. This suggests a potential inferential use for K; namely, comparing an estimate of it from a data set to some theoretical quantities, which in turn suggests whether clustering is present, and if so, which model might be most plausible. The usual estimator for K is given by

$$\widehat{K}(d) = n^{-2}|A| \sum\sum_{i \neq j} p_{ij}^{-1} I_d(d_{ij}) \ , \tag{1.8}$$

where n is the number of points in A, d_{ij} is the distance between points i and j, p_{ij} is the proportion of the circle with center i and passing through j that lies within A, and $I_d(d_{ij})$ equals 1 if $d_{ij} < d$, and 0 otherwise.

We provide an extensive account for point processes in Chapter 8. Other useful texts focusing primarily upon point processes and patterns include Diggle (2003), Lawson and Denison (2002), and Møller and Waagepetersen (2004) for treatments of spatial point processes and related methods in spatial cluster detection and modeling.

1.1.4 Software and datasets

This text extensively uses the `R` (`www.r-project.org`) software programming language and environment for statistical computing and graphics. `R` is released under the GNU open-source license and can be downloaded for free from the Comprehensive R Archive Network (CRAN), which can be accessed from `http://cran.us.r-project.org/`. The capabilities of `R` are easily extended through "libraries" or "packages" that perform more specialized tasks. These packages are also available from CRAN and can be downloaded and installed from within the `R` software environment.

There are a variety of spatial packages in `R` that perform modeling and analysis for the different types of spatial data. For example, the `gstat` and `geoR` packages provide functions to perform traditional (classical) analysis for point-level data; the latter also offers simpler Bayesian models. The packages `spBayes` and `sptimer` have much more elaborate Bayesian functions, the latter focusing primarily upon space-time data. We will provide illustrations using some of these `R` packages in Chapters 2 and 6.

The `spdep` package in `R` provides several functions for analyzing areal-level data, including basic descriptive statistics for areal data as well as fitting areal models using classical likelihood methods. For Bayesian analysis, the `BUGS` language and the `WinBUGS` software is still perhaps the most widely used engine to fit areal models. We will discuss areal models in greater detail in Chapters 4 and 6.

Turning to point-process models, a popular spatial `R` package, `spatstat`, allows computation of K for any data set, as well as the approximate 95% intervals for it so the significance of departure from some theoretical model may be judged. However, full inference likely requires use of the `R` package `Splancs`, or perhaps a fully Bayesian approach with user-specific coding (also see Wakefield and Morris, 2001). We provide some examples of `R` packages for point-process models in Chapter 8.

We will use a number of spatial and spatiotemporal datasets for illustrating the modeling and software implementation. While some of these datasets are included in the `R` packages we will be using, others are available from `www.biostat.umn.edu/~brad/data2.html`. We remark that the number of `R` packages performing spatial analysis is already too large to be discussed in this text. We refer the reader to the CRAN Task View `http://cran.r-project.org/web/views/Spatial.html` for an exhaustive list of such packages and brief descriptions regarding their capabilities.

1.2 Fundamentals of cartography

In this section we provide a brief introduction to how geographers and spatial statisticians understand the geometry of (and determine distances on) the surface of the earth. This requires a bit of thought regarding cartography (mapmaking), especially map projections, and the meaning of latitude and longitude, which are often understood informally (but incorrectly) as being equivalent to Cartesian x and y coordinates.

1.2.1 Map projections

A map projection is a systematic representation of all or part of the surface of the earth on a plane. This typically comprises lines delineating meridians (longitudes) and parallels (latitudes), as required by some definitions of the projection. A well-known fact from topology is that it is impossible to prepare a distortion-free flat map of a surface curving in all directions. Thus, the cartographer must choose the characteristic (or characteristics) that are to be shown accurately in the map. In fact, it cannot be said that there is a "best" projection for mapping. The purpose of the projection and the application at hand lead to projections that are appropriate. Even for a single application, there may be several appropriate projections, and choosing the "best" projection can be subjective. Indeed there are an infinite number of projections that can be devised, and several hundred have been published.

Since the sphere cannot be flattened onto a plane without distortion, the general strategy for map projections is to use an intermediate surface that can be flattened. This intermediate surface is called a *developable surface* and the sphere is first projected onto the this surface, which is then laid out as a plane. The three most commonly used surfaces are the cylinder, the cone and the plane itself. Using different orientations of these surfaces leads to different classes of map projections. Some examples are given in Figure 1.4. The points on the globe are projected onto the wrapping (or tangential) surface, which is then laid out to form the map. These projections may be performed in several ways, giving rise to different projections.

Before the availability of computers, the above orientations were used by cartographers in the physical construction of maps. With computational advances and digitizing of cartography, analytical formulae for projections were desired. Here we briefly outline the underlying theory for equal-area and conformal (locally shape-preserving) maps. A much more detailed and rigorous treatment may be found in Pearson (1990).

The basic idea behind deriving equations for map projections is to consider a sphere with the geographical coordinate system (λ, ϕ) for longitude and latitude and to construct an appropriate (rectangular or polar) coordinate system (x, y) so that

$$x = f(\lambda, \phi), \;\; y = g(\lambda, \phi) \; ,$$

where f and g are appropriate functions to be determined, based upon the properties we want our map to possess. We will study map projections using differential geometry concepts, looking at infinitesimal patches on the sphere (so that curvature may be neglected

Figure 1.4 *The geometric constructions of projections using developable surfaces (figure courtesy of the U.S. Geological Survey).*

and the patches are closely approximated by planes) and deriving a set of (partial) differential equations whose solution will yield f and g. Suitable initial conditions are set to create projections with desired geometric properties.

Thus, consider a small patch on the sphere formed by the infinitesimal quadrilateral, $ABCD$, given by the vertices,

$$A = (\lambda, \phi),\ B = (\lambda, \phi + d\phi),\ C = (\lambda + d\lambda, \phi),\ D = (\lambda + d\lambda, \phi + d\phi).$$

So, with R being the radius of the earth, the horizontal differential component along an arc of latitude is given by $|AC| = (R\cos\phi)d\lambda$ and the vertical component along a great circle of longitude is given by $|AB| = Rd\phi$. Note that since AC and AB are arcs along the latitude and longitude of the globe, they intersect each other at right angles. Therefore, the area of the patch $ABCD$ is given by $|AC||AB|$. Let $A'B'C'D'$ be the (infinitesimal) image of the patch $ABCD$ on the map. Then, we see that

$$\begin{aligned} A' &= (f(\lambda, \phi), g(\lambda, \phi)), \\ C' &= (f(\lambda + d\lambda, \phi), g(\lambda + d\lambda, \phi)), \\ B' &= (f(\lambda, \phi + d\phi), g(\lambda, \phi + d\phi)), \\ \text{and } D' &= (f(\lambda + d\lambda, \phi + d\phi), g(\lambda + d\lambda, \phi + d\phi))\ . \end{aligned}$$

This in turn implies that

$$\overrightarrow{A'C'} = \left(\frac{\partial f}{\partial \lambda}, \frac{\partial g}{\partial \lambda}\right) d\lambda \text{ and } \overrightarrow{A'B'} = \left(\frac{\partial f}{\partial \phi}, \frac{\partial g}{\partial \phi}\right) d\phi\ .$$

If we desire an equal-area projection we need to equate the area of the patches $ABCD$ and $A'B'C'D'$. But note that the area of $A'B'C'D'$ is given by the area of parallelogram formed by vectors $\overrightarrow{A'C'}$ and $\overrightarrow{A'B'}$. Treating them as vectors in the xy plane of an xyz system, we see that the area of $A'B'C'D'$ is the cross-product,

$$(\overrightarrow{A'C'}, 0) \times (\overrightarrow{A'B'}, 0) = \left(\frac{\partial f}{\partial \lambda}\frac{\partial g}{\partial \phi} - \frac{\partial f}{\partial \phi}\frac{\partial g}{\partial \lambda}\right) d\lambda d\phi\ .$$

Figure 1.5 *The sinusoidal projection.*

Therefore, we equate the above to $|AC||AB|$, leading to the following partial differential equation in f and g:

$$\left(\frac{\partial f}{\partial \lambda}\frac{\partial g}{\partial \phi} - \frac{\partial f}{\partial \phi}\frac{\partial g}{\partial \lambda}\right) = R^2 \cos\phi\,.$$

Note that this is the equation that must be satisfied by any equal-area projection. It is an underdetermined system, and further conditions need to be imposed (that ensure other specific properties of the projection) to arrive at f and g.

Example 1.1 Equal-area maps are used for statistical displays of areal-referenced data. An easily derived equal-area projection is the sinusoidal projection, shown in Figure 1.5. This is obtained by specifying $\partial g/\partial \phi = R$, which yields equally spaced straight lines for the parallels, and results in the following analytical expressions for f and g (with the 0 degree meridian as the central meridian):

$$f(\lambda, \phi) = R\lambda \cos\phi; g(\lambda, \phi) = R\phi.$$

Another popular equal-area projection (with equally spaced straight lines for the meridians) is the Lambert cylindrical projection given by

$$f(\lambda, \phi) = R\lambda;\ g(\lambda, \phi) = R\sin\phi\,.$$

For conformal (angle-preserving) projections we set the angle $\angle(AC, AB)$ equal to $\angle(A'C', A'B')$. Since $\angle(AC, AB) = \pi/2$, $\cos(\angle(AC, AB)) = 0$, leading to

$$\frac{\partial f}{\partial \lambda}\frac{\partial f}{\partial \phi} + \frac{\partial g}{\partial \lambda}\frac{\partial g}{\partial \phi} = 0$$

or, equivalently, the Cauchy-Riemann equations of complex analysis,

$$\left(\frac{\partial f}{\partial \lambda} + i\frac{\partial g}{\partial \lambda}\right)\left(\frac{\partial f}{\partial \phi} - i\frac{\partial g}{\partial \phi}\right) = 0\,.$$

A sufficient partial differential equation system for conformal mappings of the Cauchy-Riemman equations that is simpler to use is

$$\frac{\partial f}{\partial \lambda} = \frac{\partial g}{\partial \phi}\cos\phi;\ \frac{\partial g}{\partial \lambda} = \frac{\partial f}{\partial \phi}\cos\phi\,.$$

Example 1.2 The Mercator projection shown in Figure 1.6 is a classical example of a conformal projection. It has the interesting property that rhumb lines (curves that intersect the meridians at a constant angle) are shown as straight lines on the map. This is particularly useful for navigation purposes. The Mercator projection is derived by letting $\partial g/\partial \phi =$

Figure 1.6 *The Mercator projection.*

$R \sec \phi$. After suitable integration, this leads to the analytical equations (with the 0 degree meridian as the central meridian),

$$f(\lambda, \phi) = R\lambda;\ g(\lambda, \phi) = R \ln \tan \left(\frac{\pi}{4} + \frac{\phi}{2} \right) .$$

As is seen above, even the simplest map projections lead to complex transcendental equations relating latitude and longitude to positions of points on a given map. Therefore, rectangular grids have been developed for use by surveyors. In this way, each point may be designated merely by its distance from two perpendicular axes on a flat map. The y-axis usually coincides with a chosen central meridian, y increasing north, and the x-axis is perpendicular to the y-axis at a latitude of origin on the central meridian, with x increasing east. Frequently, the x and y coordinates are called "eastings" and "northings," respectively, and to avoid negative coordinates, may have "false eastings" and "false northings" added to them. The grid lines usually do not coincide with any meridians and parallels except for the central meridian and the equator.

One such popular grid, adopted by The National Imagery and Mapping Agency (NIMA) (formerly known as the Defense Mapping Agency), and used especially for military use throughout the world, is the Universal Transverse Mercator (UTM) grid; see Figure 1.7. The UTM divides the world into 60 north-south zones, each of width six degrees longitude. Starting with Zone 1 (between 180 degrees and 174 degrees west longitude), these are numbered consecutively as they progress eastward to Zone 60, between 174 degrees and 180 degrees east longitude. Within each zone, coordinates are measured north and east in meters, with northing values being measured continuously from zero at the equator, in a northerly direction. Negative numbers for locations south of the equator are avoided by assigning an arbitrary false northing value of 10,000,000 meters (as done by NIMA's cartographers). A central meridian cutting through the center of each 6 degree zone is assigned an easting value of 500,000 meters, so that values to the west of the central meridian are less than 500,000 while those to the east are greater than 500,000. In particular, the conterminous 48 states of the United States are covered by 10 zones, from Zone 10 on the west coast through Zone 19 in New England.

In practice, the UTM is used by overlaying a transparent grid on the map, allowing distances to be measured in meters at the map scale between any map point and the nearest grid lines to the south and west. The northing of the point is calculated as the sum

Figure 1.7 *Example of a UTM grid over the United States (figure courtesy of the U.S. Geological Survey).*

Figure 1.8 *Finding the easting and northing of a point in a UTM projection (figure courtesy of the U.S. Geological Survey).*

of the value of the nearest grid line south of it and its distance north of that line. Similarly, its easting is the value of the nearest grid line west of it added to its distance east of that line. For instance, in Figure 1.8, the grid value of line A-A is 357,000 meters east, while that of line B-B is 4,276,000 meters north. Point P is 800 meters east and 750 meters north of the grid lines resulting in the grid coordinates of point P as north 4,276,750 and east 357,800.

Finally, since spatial modeling of point-level data often requires computing distances between points on the earth's surface, one might wonder about a *planar* map projection, which would preserve distances between points. Unfortunately, the existence of such a map is precluded by Gauss' Theorema Eggregium in differential geometry (see, e.g., Guggenheimer, 1977, pp. 240–242). Thus, while we have seen projections that preserve area and shapes, distances are always distorted. The *gnomonic* projection (Snyder, 1987, pp. 164–168) gives the correct distance from a single reference point, but is less useful for the practicing spatial analyst who needs to obtain complete intersite distance matrices (since this would require not one but many such maps). Banerjee (2005) explores different strategies for computing distances on the earth and their impact on statistical inference. We present a brief summary below.

1.2.2 Calculating distances on the earth's surface

Distance computations are indispensable in spatial analysis. Precise inter-site distance computations are used in variogram analysis to assess the strength of spatial association. They help in setting starting values for the non-linear least squares algorithms in classical analysis (more in Chapter 2) and in specifying priors on the range parameter in Bayesian modeling (more in Chapter 5), making them crucial for correct interpretation of spatial range and the convergence of statistical algorithms. For data sets covering relatively small spatial domains, ordinary Euclidean distance offers an adequate approximation. However, for larger domains (say, the entire continental U.S.), the curvature of the earth causes distortions because of the difference in differentials in longitude and latitude (a unit increment in degree longitude is not the same length as a unit increment in degree latitude except at the equator).

Suppose we have two points on the surface of the earth, $P_1 = (\theta_1, \lambda_1)$ and $P_2 = (\theta_2, \lambda_2)$. We assume both points are represented in terms of latitude and longitude. That is, let θ_1 and λ_1 be the latitude and longitude, respectively, of the point P_1, while θ_2 and λ_2 are those for the point P_2. The main problem is to find the shortest distance (*geodesic*) between the points. The solution is obtained via the following formulae:

$$D = R\phi$$

where R is the radius of the earth and ϕ is an angle (measured in *radians*) satisfying

$$\cos\phi = \sin\theta_1 \sin\theta_2 + \cos\theta_1 \cos\theta_2 \cos(\lambda_1 - \lambda_2) \ . \tag{1.9}$$

These formulae are derived as follows. The geodesic is actually the arc of the great circle joining the two points. Thus, the distance will be the length of the arc of a *great circle* (i.e., a circle with radius equal to the radius of the earth). Recall that the length of the arc of a circle equals the angle subtended by the arc at the center multiplied by the radius of the circle. Therefore it suffices to find the angle subtended by the arc; denote this angle by ϕ.

Let us form a three-dimensional Cartesian coordinate system (x, y, z), with the origin at the center of the earth, the z-axis along the North and South Poles, and the x-axis on the plane of the equator joining the center of the earth and the Greenwich meridian. Using the left panel of Figure 1.9 as a guide, elementary trigonometry provides the following relationships between (x, y, z) and the latitude-longitude (θ, λ):

$$\begin{aligned} x &= R\cos\theta\cos\lambda, \\ y &= R\cos\theta\sin\lambda, \\ \text{and } z &= R\sin\theta \ . \end{aligned}$$

Now form the vectors $\mathbf{u}_1 = (x_1, y_1, z_1)$ and $\mathbf{u}_2 = (x_2, y_2, z_2)$ as the Cartesian coordinates corresponding to points P_1 and P_2. Hence ϕ is the angle between $\mathbf{u}_1$ and $\mathbf{u}_2$. From

Figure 1.9 *Diagrams illustrating the geometry underlying the calculation of great circle (geodesic) distance.*

standard analytic geometry, the easiest way to find this angle is therefore to use the following relationship between the cosine of this angle and the dot product of $\mathbf{u}_1$ and $\mathbf{u}_2$:

$$\cos\phi = \frac{\langle \mathbf{u}_1, \mathbf{u}_2 \rangle}{||\mathbf{u}_1||\,||\mathbf{u}_2||} \ .$$

We then compute $\langle \mathbf{u}_1, \mathbf{u}_2 \rangle$ as

$$\begin{aligned} &R^2\left[\cos\theta_1\cos\lambda_1\cos\theta_2\cos\lambda_2 + \cos\theta_1\sin\lambda_1\cos\theta_2\sin\lambda_2 + \sin\theta_1\sin\theta_2\right] \\ &\quad = R^2\left[\cos\theta_1\cos\theta_2\cos(\lambda_1-\lambda_2) + \sin\theta_1\sin\theta_2\right] \ . \end{aligned}$$

But $||\mathbf{u}_1|| = ||\mathbf{u}_2|| = R$, so the result in (1.9) follows. Looking at the right panel of Figure 1.9, our final answer is thus

$$D = R\phi = R\arccos[\sin\theta_1\sin\theta_2 + \cos\theta_1\cos\theta_2\cos(\lambda_1-\lambda_2)] \ . \tag{1.10}$$

While calculating (1.10) is straightforward, Euclidean metrics are popular due to their simplicity and easier interpretability. More crucially, statistical modeling of spatial correlations proceed from *correlation functions* that are often valid only with Euclidean metrics. For example, using (1.10) to calculate the distances in general covariance functions may not result in a positive definite $\Sigma(\boldsymbol{\theta})$ in (1.1). We consider a few different approaches for computing distances on the earth using Euclidean metrics, classifying them as those arising from the classical spherical coordinates, and those arising from planar projections.

Equation (1.10) clearly reveals that the relationship between the Euclidean distances and the geodetic distances is not just a matter of scaling. We cannot mutiply one by a constant number to obtain the other. A simple scaling of the geographical coordinates results in a "naive Euclidean" metric obtained directly in degree units, and converted to kilometer units as: $\|P_1 - P_2\|\pi R/180$. This metric performs well on small domains but always overestimates the geodetic distance, *flattening out* the meridians and parallels, and stretching the curved domain onto a plane, thereby stretching distances as well. As the domain increases, the estimation deteriorates.

Banerjee (2005) also explores a more natural metric, which is along the "chord" joining the two points. This is simply the Euclidean metric $||\mathbf{u}_2 - \mathbf{u}_1||$, yielding a "burrowed through the earth" distance — the chordal length between P_1 and P_2. The slight underestimation of the geodetic distance is expected, since the chord "penetrates" the domain, producing a straight line approximation to the geodetic arc.

The first three rows of Table 1.1 compare the geodetic distance with the "naive Eucidean" and chordal metrics. The next three rows show distances computed by using three planar projections: the Mercator, the sinusoidal and a centroid-based data projection, which is developed in Exercise 10. The first column corresponds to the distance between the farthest

Methods	Colorado	Chicago-Minneapolis	New York-New Orleans
geodetic	741.7	562.0	1897.2
naive Euclidean	933.8	706.0	2172.4
chord	741.3	561.8	1890.2
Mercator	951.8	773.7	2336.5
sinusoidal	742.7	562.1	1897.7
centroid-based	738.7	562.2	1901.5

Table 1.1 *Comparison of different methods of computing distances (in kms). For Colorado, the distance reported is the maximum inter-site distance for a set of* 50 *locations.*

points in a spatially referenced data set comprising 50 locations in Colorado (we will revisit this dataset later in Chapter 11), while the next two present results for two differently spaced pairs of cities. The overestimation and underestimation of the "naive Euclidean" and "chordal" metrics respectively is clear, although the chordal metric excels even for distances over 2000 kms (New York and New Orleans). We find that the sinusoidal and centroid-based projections seem to be distorting distances much less than the Mercator, which performs even worse than the naive Euclidean.

This approximation of the chordal metric has an important theoretical implication for the spatial modeler. A troublesome aspect of geodetic distances is that they are *not* necessarily valid arguments for correlation functions defined on Euclidean spaces (see Chapter 2 for more general forms of correlation functions). However, the excellent approximation of the chordal metric (which is Euclidean) ensures that in most practical settings valid correlation functions in $\Re^3$ such as the Matérn and exponential yield positive definite correlation matrices with geodetic distances and enable proper convergence of the statistical estimation algorithms.

Schoenberg (1942) develops a necessary and sufficient representation for valid positive-definite functions on spheres in terms of normalized Legendre polynomials P_k of the form:

$$\psi(t) = \sum_{k=0}^{\infty} a_k P_k(\cos t),$$

where a_k's are positive constants such that $\sum_{k=0}^{\infty} a_k$ converges. An example is given by

$$\psi(t) = \frac{1}{\sqrt{1+\alpha^2 - 2\alpha\cos t}}, \quad \alpha \in (0,1),$$

which can be easily shown to have the Legendre polynomial expansion $\sum_{k=0}^{\infty} \alpha^k P_k(\cos t)$. The chordal metric also provides a simpler way to construct valid correlation functions over the sphere using a sinusoidal composition of any valid correlation function on Euclidean space. To see this, consider a unit sphere ($R = 1$) and note that

$$\|\mathbf{u}_1 - \mathbf{u}_2\| = \sqrt{2 - 2\langle \mathbf{u}_1, \mathbf{u}_2\rangle} = 2\sin(\phi/2).$$

Therefore, a correlation function $\rho(d)$ (suppressing the range and smoothness parameters) on the Euclidean space transforms to $\rho(2\sin(\phi/2))$ on the sphere, thereby *inducing* a valid correlation function on the sphere. This has some advantages over the Legendre polynomial approach of Schoenberg: (1) we retain the interpretation of the smoothness and decay parameters, (2) it is simpler to construct and compute, and (3) it builds upon a rich legacy of investigations (both theoretical and practical) of correlation functions on Euclidean spaces (again, see Chapter 2 for different correlation functions).

1.3 Maps and geodesics in R

The R statistical software environment today offers excellent interfaces with Geographical Information Systems (GIS) through a number of libraries (also known as packages). At the core of R's GIS capabilities is the maps library originally described by Becker and Wilks (1993). This maps library contains the geographic boundary files for several maps, including county boundaries for every state in the U.S. For example, creating a map of the state of Minnesota with its county boundaries is as simple as the following line of code:

```
> library(maps)
> mn.map <- map(database="county", region="minnesota")
```

If we do not want the county boundaries, we simply write

```
> mn.map <- map("state", "minnesota"),
```

which produces a map of Minnesota with only the state boundary. The above code uses the boundaries from R's own maps database. However, other important regional boundary types (say, zip codes) and features (rivers, major roads, and railroads) are generally not available, although topographic features and an enhanced GIS interface is available through the library RgoogleMaps . While in some respects R is perhaps not nearly as versatile as ArcView or other purely GIS packages, it does offer a rare combination of GIS and statistical analysis capabilities.

It is possible to import shapefiles from other GIS software (e.g. ArcView) into R using the maptools package. We invoke the readShapePoly function in the maptools package to read the external shapefile and store the output in minnesota.shp. To produce the map, we apply plot to this output.

```
> library(maptools)
> minnesota.shp <- readShapePoly("minnesota.shp",
+                     proj4string=CRS("+proj=longlat"))
> plot(minnesota.shp).
```

For the above to work, you will need three files with extensions ".shp", ".shx" and ".dbf". They must have the same name and differ only in the extension. The "minnesota.shp" file contains the geometry data, the "minnesota.shx" file contains the spatial index, and the "minnesota.dbf" file contains the attribute data. These are read using the readShapePoly() function to produce a spatial polygon object.

The above is an example of how to draw bare maps of a state within the USA using either R's own database or an external shapefile. We can also draw maps of other countries using the mapdata package, which has some world map data, in conjunction with maps. For example, to draw a map of Canada, we write

```
> library(maps)
> library(mapdata)
> map("worldHires","Canada",
+     xlim=c(-141,-53), ylim=c(40,85),
+     col="gray90", fill=TRUE)
```

We leave the reader to experiment further with these examples.

In practice, we are not interested in bare maps but would want to plot spatially referenced data on the map. Let us return to the counties in Minnesota. Consider a new file newdata.csv that includes information on the population of each county of Minnesota along with the number of influenza A (H1N1) cases from each county. We first merge our new dataset with the minnesota.shp object already created using the county names.

```
> newdata <- read.csv("newdata.csv")
> minnesota.shp@data <- merge(minnesota.shp@data, newdata,
+                             by="NAME").
```

Figure 1.10 *Map showing Influenza A (H1N1) rates (cases/population)×100 in different counties of Minnesota for 1999.*

To plot the data in a visually appealing way, we use two additional packages: `RColorBrewer`, which creates nice color schemes, and `classInt`, which facilitates classifying the data into classes which will correspond to the plotting colors. Below, we present the code to plot H1N1 influenza rates per 100 after classifying the data into equal intervals.

```
> library(RColorBrewer)
> library(classInt)
> CASES <- minnesota.shp@data$cases
> POP1999 <- minnesota.shp@data$POP1999
> var <- (CASES/POP1999)*100
> nclr <- 7
> plotclr <- brewer.pal(nclr,"Reds")
> class <- classIntervals(var, nclr, style="equal", dataPrecision=4)
> colcode <- findColours(class, plotclr)
> plot(minnesota.shp)
> plot(minnesota.shp, col=colcode, add=T)
> title(main="Influenza A (H1N1) - Rate per 100")
> legend("bottomright", legend=names(attr(colcode, "table")),
+    fill=attr(colcode, "palette"), cex=0.6, bty="n").
```

The resulting map is shown in Figure 1.10 and is called a *choropleth map*.

Map projections in R can be performed using the `mapproj` package. For example, we can create a variety of cartographic map projections of any part of the world using `xlim` and `ylim` to specify the bounds (using longitudes and latitudes) of the plotting region and use `projection` to specify the type of projection we desire. In fact, the `maps` package uses `mapproj` for its map projections. Therefore, loading `maps` automatically invokes `mapproj`.

```
> library(maps) ## automatically loads mapproj
> sinusoidal.proj = map(database= "world", ylim=c(45,90), xlim=c(-160,-50),
+     col="grey80", fill=TRUE, plot=FALSE, projection="sinusoidal")
```

```
> map(sinusoidal.proj)
```

produces a Sinusoidal map projection (Example 1.1) between latitudes 45 and 90 degrees and longitudes −160 and −150 degrees. (Run the above code and test your geography to see what part of the world this is!) Repeating the above with `projection` set to `"mercator"` will result in a Mercator projection (Example 1.2). For higher resolution maps, one can load the `mapdata` package and use `"worldHires"` instead of `"world"` in the database.

For distance computations using a map projection, as was done for the Mercator and sinusoidal in Table 1.1, it is convenient to use the `mapproject` function in package `mapproj`. This simply converts latitude and longitude to rectangular coordinates. Suppose `LON` and `LAT` are two vectors containing longitudes and latitudes on the earth's surface to be projected. Then, a simple command such as

```
> xy.sinusoidal <- mapproject(LON, LAT, projection="sinusoidal")
```

produces the projected coordinates in Euclidean space. The coordinates are accessed by `xy.sinusoidal$x` and `xy.sinusoidal$y`, respectively. For distance computations that will approximate the great circle distance, these coordinates need to be multiplied by the radius of the earth.

While `mapproject` does not offer the UTM projections, the `rgdal` package can be used to construct UTM projections. This is particularly useful when plotting point-level data through the `RgoogleMaps` package, another exciting GIS interface offered by `R`. It is especially useful for plotting points such as GPS (Global Positioning System) locations on maps. For example, the Colorado data used in Table 1.1 is the file `ColoradoS-T.dat` and can be downloaded from `www.biostat.umn.edu/~brad/data2.html`. We can read the file in `R` as

```
> coloradoST <- read.table("ColoradoS-T.dat", header = TRUE).
```

Next, we use the `GetMap.bbox` function in the `RgoogleMaps` package to set the region within which our coordinates will be plotted. A ``maptype'' argument provides some options for the type of background map we want. The region can also be plotted using the `PlotOnStaticMap` function. This is sometimes useful to obtain an idea whether the underlying map will be useful with reference to our plotting coordinates.

```
> library(RgoogleMaps)
> MyMap <- GetMap.bbox(lonR = range(coloradoST$Longitude),
+                      latR = range(coloradoST$Latitude),
+                      size=c(640,640), maptype = "hybrid")
> PlotOnStaticMap(MyMap).
```

We now convert the longitude and latitude to the same coordinate system as in the `MyMap` object we created above.

```
> convert_points <- LatLon2XY.centered(MyMap,
+                       coloradoST$Latitude,
+                       coloradoST$Longitude)
> points(convert_points$newX, convert_points$newY,
+        col = 'red', pch=19).
```

Finally, we convert our points to UTM coordinates using the `rgdal` package. The code below performs the conversion from latitude-longitude to UTM coordinates and then plots these coordinates on the map created above. This will be achieved in two steps. First, we convert out longitudes and latitudes in the Colorado dataset into a special data type known as `SpatialPoints`. We store this in `SP_longlat`, which is converted by `spTransform` to UTM coordinates. The `zone` parameter specifies the zone of the UTM needs to be supplied explicitly.

```
> library(sp)
> library(rgdal)
```

(a) (b)

Figure 1.11 *(a) A plot of the* 50 *locations in Colorado; (b) a plot of locations in the New York/New Jersey Bight reporting scallop catches.*

```
> SP_longlat <- SpatialPoints(coords =
+                 cbind(coloradoST$Longitude,
+                       coloradoST$Latitude),
+            proj4string = CRS("+proj=longlat +ellps=WGS84"))
> SP_utm <- spTransform(SP_longlat,
+               CRS("+proj=utm +zone=13 +datum=WGS84"))
> plot(SP_utm).
```

We repeat the above exercise with another well known spatial data set involving the catches of scallops in the New York/New Jersey Bight. This data set also has coordinates in terms of latitude and longitude and can be downloaded from `www.biostat.umn.edu/~brad/data/scallops.txt`. The resulting plots are presented in Figure 1.11. The map features, including the background, can be altered by changing the parameter ``maptype" in the function `GetMap.bbox`. The options are ``roadmap'', ``mobile'', ``satellite'', ``terrain" and ``hybrid".

Finally, we mention the `fields` package in `R`, which offers several useful functions for spatial analysis. In particular, it includes two functions `rdist` and `rdist.earth` that conveniently compute inter-site distances. Let `X1` and `X2` be two matrices representing two different sets of locations. Then,

```
> library(fields)
> euclidean.dist = rdist(X1, X2)
> spherical.dist = rdist.earth(X1,X2)
```

computes the inter-site distance matrices between the locations in `X1` and `X2`. The function `rdist` uses the Euclidean distance, while `rdist.earth` uses the spherical or geodetic distance. The latter should be used only when `X1` and `X2` contain latitude-longitude coordinates.

1.4 Exercises

1. What sorts of areal unit variables can you envision that could be viewed as arising from point-referenced variables? What sorts of areal unit variables can you envision whose mean could be viewed as arising from a point-referenced surface? What sorts of areal unit variables fit neither of these scenarios?

2. What sorts of sensible properties should characterize association between point-referenced measurements? What sorts of sensible properties should characterize association between areal unit measurements?
3. Suggest some regional-level covariates that might help explain the spatial pattern evident in Figure 1.2. (*Hint:* The roughly rectangular group of regions located on the map's eastern side is the city of Minneapolis, MN.)
4. (a) Suppose you recorded elevation and average daily temperature on a particular day for a sample of locations in a region. If you were given the elevation at a new location, how would you make a plausible estimate of the average daily temperature for that location?
   (b) Why might you expect spatial association between selling prices of single-family homes in this region to be weaker than that between the observed temperature measurements?
5. For what sorts of point-referenced spatial data would you expect measurements across time to be essentially independent? For what sorts of point-referenced data would you expect measurements across time to be strongly dependent?
6. For point-referenced data, suppose the means of the variables are spatially associated. Would you expect the association between the variables themselves to be weaker than, stronger than, or the same as the association between the means?
7. (a) Write your own `R` function that will compute the distance between two points P_1 and P_2 on the surface of the earth. The function should take the latitude and longitude of the P_i as input, and output the geodesic distance D given in (1.10). Use $R = 6371$ km.
   (b) Use your program to obtain the geodesic distance between Chicago (87.63W, 41.88N) and Minneapolis (93.22W, 44.89N), and between New York (73.97W, 40.78N) and New Orleans (90.25W, 29.98N).
8. A "naive Euclidean" distance may be computed between two points by simply applying the Euclidean distance formula to the longitude-latitude coordinates, and then multiplying by $(R\pi/180)$ to convert to kilometers. Find the naive Euclidean distance between Chicago and Minneapolis, and between New York and New Orleans, comparing your results to the geodesic ones in the previous problem.
9. The *chordal* ("burrowing through the earth") distance separating two points is given by the Euclidean distance applied to the Cartesian spherical coordinate system given in Subsection 1.2.2. Find the chordal distance between Chicago and Minneapolis, and between New York and New Orleans, comparing your results to the geodesic and naive Euclidean ones above.
10. A two-dimensional projection, often used to approximate geodesic distances by applying Euclidean metrics, sets up rectangular axes along the centroid of the observed locations, and scales the points according to these axes. Thus, with N locations having geographical coordinates $(\lambda_i, \theta_i)_{i=1}^N$, we first compute the centroid $\left(\bar{\lambda}, \bar{\theta}\right)$ (the mean longitude and latitude). Next, two distances are computed. The first, d_X, is the geodesic distance (computed using (1.10) between $\left(\bar{\lambda}, \theta_{\min}\right)$ and $\left(\bar{\lambda}, \theta_{\max}\right)$, where $\theta_{\min}$ and $\theta_{\max}$ are the minimum and maximum of the observed latitudes. Analogously, d_Y is the geodesic distance computed between $\left(\lambda_{\min}, \bar{\theta}\right)$ and $\left(\lambda_{\max}, \bar{\theta}\right)$. These actually scale the axes in terms of true geodesic distances. The projection is then given by

$$x = \frac{\lambda - \bar{\lambda}}{\lambda_{\max} - \lambda_{\min}} d_X; \text{ and } y = \frac{\theta - \bar{\theta}}{\theta_{\max} - \theta_{\min}} d_Y \ .$$

Applying the Euclidean metric to the projected coordinates yields a good approximation to the inter-site geodesic distances. This projection is useful for entering coordinates in spatial statistics software packages that require two-dimensional coordinate input and uses Euclidean metrics to compute distances (e.g., the variogram functions in `geoR`, the `spatial.exp` function in `WinBUGS`, etc.).

(a) Compute the above projection for Chicago and Minneapolis ($N = 2$) and find the Euclidean distance between the projected coordinates. Compare with the geodesic distance. Repeat this exercise for New York and New Orleans.

(b) When will the above projection fail to work?

11. Use the `sp`, `rgdal` and `RgoogleMaps` packages to create an UTM projection for the locations in the scallops data and produce the picture in Figure 1.11(b).

12. Use the `fields` package to produce the inter-site distance matrix for the locations in the scallops data. Compute this matrix using the `rdist.earth` function, which yields the geodetic distances. Next project the data to UTM coordinates and use the `rdist` function to compute the inter-site Euclidean distance matrix. Draw histograms of the inter-site distances and comment on any notable discrepancies resulting from the map projection.

Chapter 2

Basics of point-referenced data models

In this chapter we present the essential elements of spatial models and classical analysis for point-referenced data. As mentioned in Chapter 1, the fundamental concept underlying the theory is a stochastic process $\{Y(\mathbf{s}) : \mathbf{s} \in D\}$, where D is a fixed subset of r-dimensional Euclidean space. Note that such stochastic processes have a rich presence in the time series literature, where $r = 1$. In the spatial context, usually we encounter r to be 2 (say, northings and eastings) or 3 (e.g., northings, eastings, and altitude above sea level). For situations where $r > 1$, the process is often referred to as a *spatial process*. For example, $Y(\mathbf{s})$ may represent the level of a pollutant at site $\mathbf{s}$. While it is conceptually sensible to assume the existence of a pollutant level at all possible sites in the domain, in practice the data will be a partial realization of that spatial process. That is, it will consist of measurements at a finite set of locations, say $\{\mathbf{s}_1, \ldots, \mathbf{s}_n\}$, where there are monitoring stations. The problem facing the statistician is inference about the spatial process $Y(\mathbf{s})$ and prediction at new locations, based upon this partial realization. The remarkable feature of the models we employ here is that, despite only seeing the process, equivalently, the spatial surface at a finite set of locations, we can infer about the surface at an uncountable number of locations. The reason is that we specify association through *structured* dependence which enables this broad interpolation.

This chapter is organized as follows. We begin with a survey of the building blocks of point-level data modeling, including stationarity, isotropy, and variograms (and their fitting via traditional moment-matching methods). We defer theoretical discussion of the spatial (typically Gaussian) process modeling that enables likelihood (and Bayesian) inference in these settings to Chapters 3 and 6. Then, we illustrate helpful exploratory data analysis tools, as well as traditional classical methods, especially kriging (point-level spatial prediction). We view all of these activities in an exploratory fashion, i.e., as a prelude to fully model-based inference under a hierarchical model. We close with some short tutorials in `R` using easy to use and widely available point-level spatial statistical analysis packages.

The material we cover in this chapter is traditionally known as *geostatistics*, and could easily fill many more pages than we devote to it here. While we prefer the more descriptive term "point-level spatial modeling," we will at times still use "geostatistics" for brevity and perhaps consistency when referencing the literature.

2.1 Elements of point-referenced modeling

2.1.1 Stationarity

For our discussion we assume that our spatial process has a mean, say $\mu(\mathbf{s}) = E(Y(\mathbf{s}))$, associated with it and that the variance of $Y(\mathbf{s})$ exists for all $\mathbf{s} \in D$. The process $Y(\mathbf{s})$ is said to be *Gaussian* if, for any $n \geq 1$ and any set of sites $\{\mathbf{s}_1, \ldots, \mathbf{s}_n\}$, $\mathbf{Y} = (Y(\mathbf{s}_1), \ldots, Y(\mathbf{s}_n))^T$ has a multivariate normal distribution. The process is said to be *strictly stationary* (sometimes *strong* stationarity) if, for any given $n \geq 1$, any set of n sites $\{\mathbf{s}_1, \ldots, \mathbf{s}_n\}$ and any $\mathbf{h} \in \Re^r$,

the distribution of $(Y(\mathbf{s}_1), \ldots, Y(\mathbf{s}_n))$ is the same as that of $(Y(\mathbf{s}_1 + \mathbf{h}), \ldots, Y(\mathbf{s}_n + \mathbf{h}))$. Here D is envisioned as $\Re^r$ as well.

A less restrictive condition is given by *weak stationarity* (also called second-order stationarity). A spatial process is called weakly stationary if $\mu(\mathbf{s}) \equiv \mu$ (i.e., it has a constant mean) and $\mathrm{Cov}(Y(\mathbf{s}), Y(\mathbf{s}+\mathbf{h})) = C(\mathbf{h})$ for all $\mathbf{h} \in \Re^r$ such that $\mathbf{s}$ and $\mathbf{s}+\mathbf{h}$ both lie within D. In fact, for stationarity as a second-order property we will need only the second property; $E(Y(\mathbf{s}))$ need not equal $E(Y(\mathbf{s}+\mathbf{h}))$. But since we will apply the definition only to a mean 0 spatial residual process, this distinction is not important for us. Weak stationarity implies that the covariance relationship between the values of the process at any two locations can be summarized by a covariance function $C(\mathbf{h})$, and this function depends only on the separation vector $\mathbf{h}$. Note that with all variances assumed to exist, strong stationarity implies weak stationarity. The converse is not true in general, but it *does* hold for Gaussian processes; see Exercise 4.

We offer a simple illustration of a weakly stationary process that is not strictly stationary. It is easy to see in the one-dimensional case. Suppose the process $Y_t, t = 1, 2, ...$ consists of a sequence of independent variables such that for t odd, Y_t is a binary variable taking the values 1 and -1 each with probability .5 while for t even, Y_t is normal with mean 0 and variance 1. We have weak stationarity since $\mathrm{cov}(Y(t), Y(t')) = 0, t \neq t', = 1, t = t'$. That is, we only need to know the value of $t - t'$ to specify the covariance. However, clearly Y_1, Y_3 does not have the same distribution as Y_2, Y_4.

2.1.2 Variograms

There is a third type of stationarity called *intrinsic* stationarity. Here we assume $E[Y(\mathbf{s}+\mathbf{h}) - Y(\mathbf{s})] = 0$ and define

$$E[Y(\mathbf{s}+\mathbf{h}) - Y(\mathbf{s})]^2 = Var(Y(\mathbf{s}+\mathbf{h}) - Y(\mathbf{s})) = 2\gamma(\mathbf{h}) \ . \tag{2.1}$$

Equation (2.1) makes sense only if the left-hand side depends *solely* on $\mathbf{h}$ (so that the right-hand side can be written at all), and not the particular choice of $\mathbf{s}$. If this is the case, we say the process is *intrinsically stationary.* The function $2\gamma(\mathbf{h})$ is then called the *variogram*, and $\gamma(\mathbf{h})$ is called the *semivariogram.* We can offer some intuition behind the variogram but it really arose simply as a result of its appearance in traditional kriging where one seeks the best linear unbiased predictor, as we clarify below. Behaviorally, at short distances (small $||\mathbf{h}||$), we would expect $Y(\mathbf{s}+\mathbf{h})$ and $Y(\mathbf{h})$ to be very similar, that is, $(Y(\mathbf{s}+\mathbf{h}) - Y(\mathbf{s}))^2$ to be small. As $||\mathbf{h}||$ grows larger, we expect less similarity between $Y(\mathbf{s}+\mathbf{h})$ and $Y(\mathbf{h})$, i.e., we expect $(Y(\mathbf{s}+\mathbf{h}) - Y(\mathbf{s}))^2$ to be larger. So, a plot of $\gamma(\mathbf{h})$ would be expected to increase with $||\mathbf{h}||$, providing some insight into spatial behavior. (The covariance function $C(\mathbf{h})$ is sometimes referred to as the *covariogram*, especially when plotted graphically.) Note that intrinsic stationarity defines only the first and second moments of the differences $Y(\mathbf{s}+\mathbf{h}) - Y(\mathbf{s})$. It says nothing about the joint distribution of a collection of variables $Y(\mathbf{s}_1), \ldots, Y(\mathbf{s}_n)$, and thus provides no likelihood.

In fact, it says nothing about the moments of the $Y(\mathbf{s})$'s, much less their distribution. It only describes the behavior of differences rather than the behavior of the data that we observe, clearly unsatisfying from the perspective of data analysis. In fact, the $Y(\mathbf{s})$'s need not have any moments. For example, we might have $Y(\mathbf{s}) = W(\mathbf{s}) + V$ where $W(\mathbf{s})$ is a collection of i.i.d. normal variables and, independently, V is a Cauchy random variable. Then $Y(\mathbf{s})$ is intrinsically stationary but the $Y(\mathbf{s})$'s have no moments. Even more disconcerting, the distribution of $Y(\mathbf{s}) - Y(\mathbf{s}')$ may be proper while the distribution of $Y(\mathbf{s})$ and of $Y(\mathbf{s}')$ may be improper. For instance, suppose the joint distribution, $f(Y(\mathbf{s}), Y(\mathbf{s}')) \propto e^{-(Y(\mathbf{s})-Y(\mathbf{s}'))^2/2}$. Then, $Y(\mathbf{s}) - Y(\mathbf{s}') \sim N(0,1)$ and, in fact, $f(Y(\mathbf{s})|Y(\mathbf{s}'))$ and $f(Y(\mathbf{s}')|Y(\mathbf{s}))$ are proper, $N(0,1)$ but the joint distribution is improper. How could we employ a probability specification that could not possibly have generated the data we observe?

It is easy to see the relationship between the variogram and the covariance function:

$$\begin{aligned} 2\gamma(\mathbf{h}) &= Var\left(Y\left(\mathbf{s}+\mathbf{h}\right)-Y\left(\mathbf{s}\right)\right) \\ &= Var(Y(\mathbf{s}+\mathbf{h}))+Var(Y(\mathbf{s}))-2Cov(Y(\mathbf{s}+\mathbf{h}),Y(\mathbf{s})) \\ &= C(\mathbf{0})+C(\mathbf{0})-2C\left(\mathbf{h}\right) \\ &= 2\left[C\left(\mathbf{0}\right)-C\left(\mathbf{h}\right)\right] . \end{aligned}$$

Thus,

$$\gamma\left(\mathbf{h}\right)=C\left(\mathbf{0}\right)-C\left(\mathbf{h}\right) . \tag{2.2}$$

From (2.2) we see that given C, we are able to recover γ easily. But what about the converse; in general, can we recover C from γ? Here it turns out we need to assume a bit more: if the spatial process is *ergodic*, then $C\left(\mathbf{h}\right)\rightarrow 0$ as $||\mathbf{h}||\rightarrow\infty$, where $||\mathbf{h}||$ denotes the length of the $\mathbf{h}$ vector. This is an intuitively sensible condition, since it means that the covariance between the values at two points vanishes as the points become further separated in space. But taking the limit of both sides of (2.2) as $||\mathbf{h}||\rightarrow\infty$, we then have that $lim_{||\mathbf{h}||\rightarrow\infty}\gamma\left(\mathbf{h}\right)=C\left(\mathbf{0}\right)$. Thus, using the dummy variable $\mathbf{u}$ to avoid confusion, we have

$$C\left(\mathbf{h}\right)=C(\mathbf{0})-\gamma(\mathbf{h})=lim_{||\mathbf{u}||\rightarrow\infty}\gamma\left(\mathbf{u}\right)\ -\ \gamma\left(\mathbf{h}\right) . \tag{2.3}$$

In general, the limit on the right-hand side need not exist, but if it does, then the process is weakly (second-order) stationary with $C\left(\mathbf{h}\right)$ as given in (2.3). We therefore have a way to determine the covariance function C from the semivariogram γ. Thus weak stationarity implies intrinsic stationarity, but the converse is not true; indeed, the next section offers examples of processes that are intrinsically stationary but not weakly stationary.

A valid variogram necessarily satisfies a negative definiteness condition. In fact, for any set of locations $\mathbf{s}_1,\ldots,\mathbf{s}_n$ and any set of constants $a_1,\ldots,a_n$ such that $\sum_i a_i=0$, if $\gamma(\mathbf{h})$ is valid, then

$$\sum_i\sum_j a_ia_j\gamma(\mathbf{s}_i-\mathbf{s}_j)\leq 0 . \tag{2.4}$$

To see this, note that

$$\begin{aligned} \sum_i\sum_j a_ia_j\gamma(\mathbf{s}_i-\mathbf{s}_j) &= \frac{1}{2}E\sum_i\sum_j a_ia_j(Y(\mathbf{s}_i)-Y(\mathbf{s}_j))^2 \\ &= -E\sum_i\sum_j a_ia_jY(\mathbf{s}_i)Y(\mathbf{s}_j) \\ &= -E\left[\sum_i a_iY(\mathbf{s}_i)\right]^2\ \leq\ 0 . \end{aligned}$$

We remark that, despite the suggestion of expression (2.2), there is no relationship between this result and the positive definiteness condition for covariance functions (see Subsection 3.1.2). Cressie (1993) discusses further necessary conditions for a valid variogram. Lastly, the condition (2.4) emerges naturally in ordinary kriging (see Section 2.4).

2.1.3 Isotropy

Another important related concept is that of isotropy. If the semivariogram function $\gamma\left(\mathbf{h}\right)$ depends upon the separation vector only through its length $||\mathbf{h}||$, then we say that the variogram is *isotropic*; that is, if $\gamma\left(\mathbf{h}\right)$ is a real-valued function of a univariate argument, and can be written as $\gamma\left(||\mathbf{h}||\right)$. If not, we say it is *anisotropic*. Because of the foregoing issues with intrinsic stationarity, in the absence of a full probabilistic specification (as detailed in

Figure 2.1 *Theoretical semivariograms for three models: (a) linear, (b) spherical, and (c) exponential.*

Chapter 3), we are reluctant to associate an isotropic variogram with a stochastic process. Nonetheless, in the literature, we find terminology stating that, if a process is intrinsically stationary and isotropic, it is also called *homogeneous*.

Isotropic variograms are popular because of their simplicity, interpretability, and, in particular, because a number of relatively simple parametric forms are available as candidates for the semivariogram. Denoting $||\mathbf{h}||$ by d for notational simplicity, we now consider a few of the more important such forms.

1. *Linear:*

$$\gamma(d) = \begin{cases} \tau^2 + \sigma^2 d & \text{if } d > 0,\ \tau^2 > 0,\ \sigma^2 > 0 \\ 0 & \text{otherwise} \end{cases} .$$

Note that $\gamma(d) \to \infty$ as $d \to \infty$, and so this semivariogram does not correspond to a weakly stationary process (although it is intrinsically stationary). This semivariogram is plotted in Figure 2.1(a) using the parameter values $\tau^2 = 0.2$ and $\sigma^2 = 0.5$.

2. *Spherical:*

$$\gamma(d) = \begin{cases} \tau^2 + \sigma^2 & \text{if } d \geq 1/\phi, \\ \tau^2 + \sigma^2 \left\{ \frac{3\phi d}{2} - \frac{1}{2}(\phi d)^3 \right\} & \text{if } 0 < d \leq 1/\phi, \\ 0 & \text{otherwise} \end{cases} .$$

The spherical semivariogram is valid in $r = 1, 2$, or 3 dimensions, but for $r \geq 4$ it fails to correspond to a spatial variance matrix that is positive definite (as required to specify a valid joint probability distribution). This variogram owes its popularity largely to the fact that it offers clear illustrations of the *nugget*, *sill*, and *range*, three characteristics traditionally associated with variograms. Specifically, consider Figure 2.1(b), which plots the spherical semivariogram using the parameter values $\tau^2 = 0.2$, $\sigma^2 = 1$, and $\phi = 1$. While $\gamma(0) = 0$ by definition, $\gamma(0^+) \equiv lim_{d \to 0^+} \gamma(d) = \tau^2$; this quantity is the *nugget*. Next, $\lim_{d \to \infty} \gamma(d) = \tau^2 + \sigma^2$; this asymptotic value of the semivariogram is called the *sill*. (The sill minus the nugget, which is simply σ^2 in this case, is called the *partial sill*.) Finally, the value $d = 1/\phi$ at which $\gamma(d)$ first reaches its ultimate level (the sill) is called the *range*. It is for this reason that many of the variogram models of this subsection are often parametrized through $R \equiv 1/\phi$. Confusingly, both R and ϕ are sometimes referred to as the *range* parameter, although ϕ is often more accurately referred to as the *decay* parameter.

Note that for the linear semivariogram, the nugget is τ^2 but the sill and range are both infinite. For other variograms (such as the next one we consider), the sill is finite, but only reached asymptotically.

3. *Exponential:*

$$\gamma(d) = \begin{cases} \tau^2 + \sigma^2 \left(1 - \exp\left(-\phi d\right)\right) & \text{if } d > 0, \\ 0 & \text{otherwise} \end{cases} .$$

The exponential has an advantage over the spherical in that it is simpler in functional form while still being a valid variogram in all dimensions (and without the spherical's finite range requirement). However, note from Figure 2.1(c), which plots this semivariogram assuming $\tau^2 = 0.2$, $\sigma^2 = 1$, and $\phi = 2$, that the sill is only reached asymptotically; strictly speaking, the range $R = 1/\phi$ is infinite. In cases like this, the notion of an *effective range* is often used, i.e., the distance at which there is essentially no lingering spatial correlation. To make this notion precise, we must convert from γ scale to C scale (possible here since $\lim_{d\to\infty} \gamma(d)$ exists; the exponential is not only intrinsically but also weakly stationary). From (2.3) we have

$$\begin{aligned} C(d) &= lim_{u\to\infty}\gamma(u) - \gamma(d) \\ &= \tau^2 + \sigma^2 - \left[\tau^2 + \sigma^2(1 - \exp(-\phi d))\right] \\ &= \sigma^2 \exp(-\phi d) \ . \end{aligned}$$

Hence

$$C(t) = \begin{cases} \tau^2 + \sigma^2 & \text{if } d = 0 \\ \sigma^2 \exp(-\phi d) & \text{if } d > 0 \end{cases} . \tag{2.5}$$

If the nugget $\tau^2 = 0$, then this expression reveals that the correlation between two points d units apart is $\exp(-\phi d)$; note that $\exp(-\phi d) = 1^-$ for $d = 0^+$ and $\exp(-\phi d) = 0$ for $d = \infty$, both in concert with this interpretation.

A common definition of the *effective range*, d_0, is the distance at which this correlation is *negligible*, customarily taken as having dropped to only 0.05. Setting $\exp(-\phi d_0)$ equal to this value we obtain $t_0 \approx 3/\phi$, since $\log(0.05) \approx -3$. The range will be discussed in more detail in Subsection 3.1.2.

Finally, the form of (2.5) gives a clear example of why the nugget (τ^2 in this case) is often viewed as a "nonspatial effect variance," and the partial sill (σ^2) is viewed as a "spatial effect variance." That is, we have two variance components. Along with ϕ, a statistician would likely view fitting this model to a spatial data set as an exercise in estimating these three parameters. We shall return to variogram model fitting in Subsection 2.1.4.

4. *Gaussian:*

$$\gamma(d) = \begin{cases} \tau^2 + \sigma^2 \left(1 - \exp\left(-\phi^2 d^2\right)\right) & \text{if } d > 0 \\ 0 & \text{otherwise} \end{cases} . \tag{2.6}$$

The Gaussian variogram is an analytic function and yields very smooth realizations of the spatial process. We shall say much more about process smoothness in Subsection 3.1.4.

5. *Powered exponential:*

$$\gamma(d) = \begin{cases} \tau^2 + \sigma^2 \left(1 - \exp\left(-\left|\phi d\right|^p\right)\right) & \text{if } d > 0 \\ 0 & \text{otherwise} \end{cases} . \tag{2.7}$$

Here $0 < p \leq 2$ yields a family of valid variograms. Note that both the Gaussian and the exponential forms are special cases.

6. *Rational quadratic:*

$$\gamma(d) = \begin{cases} \tau^2 + \frac{\sigma^2 d^2}{(\phi + d^2)} & \text{if } d > 0 \\ 0 & \text{otherwise} \end{cases} .$$

Model	Covariance function, $C(d)$
Linear	$C(d)$ does not exist
Spherical	$C(d) = \begin{cases} 0 & \text{if } d \geq 1/\phi \\ \sigma^2\left[1 - \frac{3}{2}\phi d + \frac{1}{2}(\phi d)^3\right] & \text{if } 0 < d \leq 1/\phi \\ \tau^2 + \sigma^2 & d = 0 \end{cases}$
Exponential	$C(d) = \begin{cases} \sigma^2 \exp(-\phi d) & \text{if } d > 0 \\ \tau^2 + \sigma^2 & d = 0 \end{cases}$
Powered exponential	$C(d) = \begin{cases} \sigma^2 \exp(-\lvert\phi d\rvert^p) & \text{if } d > 0 \\ \tau^2 + \sigma^2 & d = 0 \end{cases}$
Gaussian	$C(d) = \begin{cases} \sigma^2 \exp(-\phi^2 d^2) & \text{if } d > 0 \\ \tau^2 + \sigma^2 & d = 0 \end{cases}$
Rational quadratic	$C(d) = \begin{cases} \sigma^2\left(1 - \frac{d^2}{(\phi + d^2)}\right) & \text{if } d > 0 \\ \tau^2 + \sigma^2 & d = 0 \end{cases}$
Wave	$C(d) = \begin{cases} \sigma^2 \frac{\sin(\phi d)}{\phi d} & \text{if } d > 0 \\ \tau^2 + \sigma^2 & d = 0 \end{cases}$
Power law	$C(d)$ does not exist
Matérn	$C(d) = \begin{cases} \frac{\sigma^2}{2^{\nu-1}\Gamma(\nu)} (\phi d)^\nu K_\nu(\phi d) & \text{if } d > 0 \\ \tau^2 + \sigma^2 & d = 0 \end{cases}$
Matérn at $\nu = 3/2$	$C(d) = \begin{cases} \sigma^2 (1 + \phi d) \exp(-\phi d) & \text{if } d > 0 \\ \tau^2 + \sigma^2 & d = 0 \end{cases}$

Table 2.1 *Summary of common isotropic parametric covariance functions (covariograms).*

7. *Wave:*

$$\gamma(d) = \begin{cases} \tau^2 + \sigma^2\left(1 - \frac{\sin(\phi d)}{\phi d}\right) & \text{if } d > 0 \\ 0 & \text{otherwise} \end{cases}.$$

Note this is an example of a variogram that is not monotonically increasing. The associated covariance function is $C(d) = \sigma^2 \sin(\phi d)/(\phi d)$. Bessel functions of the first kind include the wave covariance function and are discussed in detail in Subsections 3.1.2 and 6.1.3.

8. *Power law:*

$$\gamma(d) = \begin{cases} \tau^2 + \sigma^2 d^\lambda & \text{of } d > 0 \\ 0 & d = 0 \end{cases}.$$

This generalizes the linear case and produces valid intrinsic (albeit not weakly) stationary semivariograms provided $0 \leq \lambda < 2$.

9. *Matérn:* The variogram for the Matérn class is given by

$$\gamma(d) = \begin{cases} \tau^2 + \sigma^2\left[1 - \frac{(2\sqrt{\nu} d\phi)^\nu}{2^{\nu-1}\Gamma(\nu)} K_\nu(2\sqrt{\nu} d\phi)\right] & \text{if } d > 0 \\ 0 & d = 0 \end{cases}. \tag{2.8}$$

This class was originally suggested by Matérn (1960, 1986). Interest in it was revived by Handcock and Stein (1993) and Handcock and Wallis (1994), who demonstrated attractive interpretations for ν as well as ϕ. In particular, $\nu > 0$ is a parameter controlling the smoothness of the realized random field (see Subsection 3.1.4) while ϕ is a spatial decay parameter. The function $\Gamma(\cdot)$ is the usual gamma function while K_ν is the modified Bessel function of order ν (see, e.g., Abramowitz and Stegun, 1965, Chapter 9). Implementations of this function are available in several `C/C++` libraries and also in `R` packages such as `spBayes` and `geoR`. Note that special cases of the above are the exponential ($\nu = 1/2$) and the Gaussian ($\nu \to \infty$). At $\nu = 3/2$ we obtain a closed form as well, namely $\gamma(d) =$

Model	Variogram, $\gamma(d)$
Linear	$\gamma(d) = \begin{cases} \tau^2 + \sigma^2 d & \text{if } d > 0 \\ 0 & d = 0 \end{cases}$
Spherical	$\gamma(d) = \begin{cases} \tau^2 + \sigma^2 & \text{if } d \geq 1/\phi \\ \tau^2 + \sigma^2 \left[\frac{3}{2}\phi d - \frac{1}{2}(\phi d)^3\right] & \text{if } 0 < d \leq 1/\phi \\ 0 & d = 0 \end{cases}$
Exponential	$\gamma(d) = \begin{cases} \tau^2 + \sigma^2(1 - \exp(-\phi d)) & \text{if } d > 0 \\ 0 & d = 0 \end{cases}$
Powered exponential	$\gamma(d) = \begin{cases} \tau^2 + \sigma^2(1 - \exp(-\lvert\phi d\rvert^p)) & \text{if } d > 0 \\ 0 & d = 0 \end{cases}$
Gaussian	$\gamma(d) = \begin{cases} \tau^2 + \sigma^2(1 - \exp(-\phi^2 d^2)) & \text{if } d > 0 \\ 0 & d = 0 \end{cases}$
Rational quadratic	$\gamma(d) = \begin{cases} \tau^2 + \frac{\sigma^2 d^2}{(\phi + d^2)} & \text{if } d > 0 \\ 0 & d = 0 \end{cases}$
Wave	$\gamma(d) = \begin{cases} \tau^2 + \sigma^2(1 - \frac{\sin(\phi d)}{\phi d}) & \text{if } d > 0 \\ 0 & d = 0 \end{cases}$
Power law	$\gamma(d) = \begin{cases} \tau^2 + \sigma^2 d^\lambda & \text{if } d > 0 \\ 0 & d = 0 \end{cases}$
Matérn	$\gamma(d) = \begin{cases} \tau^2 + \sigma^2 \left[1 - \frac{(\phi d)^\nu}{2^{\nu-1}\Gamma(\nu)} K_\nu(\phi d)\right] & \text{if } d > 0 \\ 0 & d = 0 \end{cases}$
Matérn at $\nu = 3/2$	$\gamma(d) = \begin{cases} \tau^2 + \sigma^2 \left[1 - (1 + \phi d)\exp(-\phi d)\right] & \text{if } d > 0 \\ 0 & d = 0 \end{cases}$

Table 2.2 *Summary of common parametric isotropic variograms.*

$\tau^2 + \sigma^2 [1 - (1 + \phi d)\exp(-\phi d)]$ for $t > 0$. In fact, we obtain a polynomial times exponential from for the Matérn for all ν of the form, $\nu = k + \frac{1}{2}$ with k a non-negative integer. The Matérn covariance function is often reparametrized to $\alpha = 2\sqrt{\nu}\phi$ along with $\eta = \sigma^2\phi^{2\nu}$ and ν. This transformation is helpful in providing better behaved model fitting, particularly using Markov chain Monte Carlo (see Chapter 6 for further discussion).

The covariance functions and variograms we have described in this subsection are conveniently summarized in Tables 2.1 and 2.2, respectively. An important point which the reader may wonder about is the fact that every presented covariance function and variogram has a discontinuity at 0. That is, the limit as $d \to 0$ for the covariance function is σ^2 not $\sigma^2 + \tau^2$ and for the variogram is τ^2, not 0. Evidently, this is not a typo! What is the reason? We elaborate this in greater detail in Chapter 3. Here, we offer a simple explanation. Consider the form of the residual for a spatial model. We might write it as $r(\mathbf{s}) = w(\mathbf{s}) + \epsilon(\mathbf{s})$ where $w(\mathbf{s})$ is a spatial process, say a stationary Gaussian process with mean 0 and covariance function $\sigma^2\rho(\mathbf{s} - \mathbf{s}')$ and $\epsilon(\mathbf{s})$ is a pure error process, i.e., the $\epsilon(\mathbf{s})$'s are i.i.d., say $N(0, \tau^2)$ with the ϵ's independent of the w's. Then, it is straightforward to compute $\text{cov}(r(\mathbf{s}), r(\mathbf{s}')) = \sigma^2\rho(\mathbf{s} - \mathbf{s}')$ while $\text{var}Y(\mathbf{s}) = \sigma^2 + \tau^2$, whence the discontinuity at 0 emerges. And, in fact, $\text{corr}(r(\mathbf{s}), r(\mathbf{s}'))$ is bounded by $\sigma^2/(\sigma^2 + \tau^2)$. The rationale for including the $\epsilon(\mathbf{s})$'s into the model is that we don't want to insist that all model error is spatial. Of course, we certainly want to include the w's in order to be able to capture a spatial story. Possible explanations for the pure error contribution include: (i) measurement error associated with the data collection at a given location, (ii) replication error to express the possibility that repeated measurements at the same location might not provide identical observations, or (iii) micro-scale error to acknowledge that, though we never see

observations closer to each other than the minimum pairwise distance in our sample, there might be very fine scale structure represented as noise.

2.1.4 Variogram model fitting

Having seen a fairly large selection of models for the variogram, one might well wonder how we choose one of them for a given data set, or whether the data can really distinguish them (see Subsection 6.1.3 in this latter regard). Historically, a variogram model is chosen by plotting the *empirical semivariogram* (Matheron, 1963), a simple nonparametric estimate of the semivariogram, and then comparing it to the various theoretical shapes available from the choices in the previous subsection. The customary empirical semivariogram is

$$\widehat{\gamma}(d) = \frac{1}{2N(d)} \sum_{(\mathbf{s}_i,\mathbf{s}_j)\in N(d)} \left[Y(\mathbf{s}_i) - Y(\mathbf{s}_j)\right]^2 \ , \tag{2.9}$$

where $N(d)$ is the set of pairs of points such that $||\mathbf{s}_i - \mathbf{s}_j|| = d$, and $|N(d)|$ is the number of pairs in this set. Notice that, unless the observations fall on a regular grid, the distances between the pairs will all be different, so this will not be a useful estimate as it stands. Instead we would "grid up" the d-space into intervals $I_1 = (0, d_1), I_2 = (d_1, d_2)$, and so forth, up to $I_K = (d_{K-1}, d_K)$ for some (typically regular) grid $0 < d_1 < \cdots < d_K$. Representing the d values in each interval by the midpoint of the interval, we then alter our definition of $N(d)$ to

$$N(d_k) = \{(\mathbf{s}_i, \mathbf{s}_j) : ||\mathbf{s}_i - \mathbf{s}_j|| \in I_k\} \ , \ k = 1, \ldots, K \ .$$

Selection of an appropriate number of intervals K and location of the upper endpoint t_K is reminiscent of similar issues in histogram construction. Journel and Huijbregts (1979) recommend bins wide enough to capture at least 30 pairs per bin.

Clearly (2.9) is nothing but a method of moments (MOM) estimate, the semivariogram analogue of the usual sample variance estimate s^2. While very natural, there is reason to doubt that this is the best estimate of the semivariogram. Certainly it will be sensitive to outliers, and the sample average of the squared differences may be rather badly behaved since under a Gaussian distributional assumption for the $Y(\mathbf{s}_i)$, the squared differences will have a distribution that is a scale multiple of the heavily skewed χ^2_1 distribution. In this regard, Cressie and Hawkins (1980) proposed a robustified estimate that uses sample averages of $|Y(\mathbf{s}_i) - Y(\mathbf{s}_j)|^{1/2}$; this estimate is available in several software packages (see Section 2.3 below). Perhaps more uncomfortable is the fact that (2.9) uses data differences, rather than the data itself. Also of concern is the fact that the components of the sum in (2.9) will be dependent within and across bins, and that $N(d_k)$ will vary across bins.

In any case, an empirical semivariogram estimate can be plotted, viewed, and an appropriately shaped theoretical variogram model can be fitted to this "data." Since any empirical estimate naturally carries with it a significant amount of noise in addition to its signal, this fitting of a theoretical model has traditionally been as much art as science: in any given real data setting, any number of different models (exponential, Gaussian, spherical, etc.) may seem equally appropriate. Indeed, fitting has historically been done "by eye," or at best by using trial and error to choose values of nugget, sill, and range parameters that provide a good match to the empirical semivariogram (where the "goodness" can be judged visually or by using some least squares or similar criterion); again see Section 2.3. More formally, we could treat this as a statistical estimation problem, and use nonlinear maximization routines to find nugget, sill, and range parameters that minimize some goodness-of-fit criterion.

If we also have a distributional model for the data, we could use maximum likelihood (or restricted maximum likelihood, REML) to obtain sensible parameter estimates; see, e.g., Smith (2001) for details in the case of Gaussian data modeled with the various parametric

variogram families outlined in Subsection 2.1.3. In Chapter 5 and Chapter 6 we shall see that the hierarchical Bayesian approach is broadly similar to this latter method, although it will often be easier and more intuitive to work directly with the covariance model $C(d)$, rather than changing to a partial likelihood in order to introduce the semivariogram. In addition, we will gain full inference, e.g., posterior distributions for all unknowns of interest as well as more accurate assessment of uncertainty than appealing to arguably inappropriate asymptotics (see Chapter 5).

2.2 Anisotropy

Stationary correlation functions extend the class of correlation functions from isotropy where association only depends upon distance to association that depends upon the separation vector between locations. As a result, association depends upon direction. Here, we explore covariance functions that are stationary but not isotropic. (We defer discussion of nonstationary covariance functions to Chapter 3.) A simple example is the class where we *separate* the components. That is, suppose we write the components of $\mathbf{s}$ as s_{lat}, s_{lon}, similarly for $\mathbf{h}$. Then, we can define $\text{corr}(Y(\mathbf{s}+\mathbf{h}), Y(\mathbf{s})) = \rho_1(h_{lat})\rho_2(h_{lon})$ where ρ_1 and ρ_2 are valid correlation functions on $\Re^1$. Evidently, this correlation function is stationary but depends on direction. In particular, if we switch h_{lat} and h_{lon} we will get a different value for the correlation even though $||\mathbf{h}||$ is unchanged. A common choice is $e^{\phi_1(|h_{lat}|)+\phi_2(|h_{lon}|)}$.

The separable correlation function is usually extended to a covariance function by introducing the multiplier σ^2, as we have done above. This covariance function tends to be used, for instance, with Gaussian processes in computer model settings (Santner, Williams and Notz, 2003; Rasmussen and Williams, 2006; Oakley and O'Hagan, 2004; Kennedy and O'Hagan, 2001). Here we seek a response surface over covariate space and, since the covariates live on their own spaces with their own scales, component-wise dependence seems appropriate. In the spatial setting, since lat and lon are on the same scale, it may be less suitable.

2.2.1 Geometric anisotropy

A commonly used class of stationary covariance functions is the geometric anisotropic covariance functions where we set

$$C(\mathbf{s}-\mathbf{s}') = \sigma^2 \rho((\mathbf{s}-\mathbf{s}')^T B(\mathbf{s}-\mathbf{s}')) \ . \tag{2.10}$$

In (2.10), B is positive definite with ρ a valid correlation function in $\Re^r$ (say, from Table 2.1). We would omit the range/decay parameter since it can be incorporated into B. When $r = 2$ we obtain a specification with three parameters rather than one. Contours of constant association arising from c in (2.10) are elliptical. In particular, the contour corresponding to $\rho = .05$ provides the range in each spatial direction. Ecker and Gelfand (1997) provide the details for Bayesian modeling and inference incorporating (2.10); see also Subsection 6.1.4.

Following the discussion in Subsection 3.1.2, we can extend geometric anisotropy to *product* geometric anisotropy. In the simplest case, we would set

$$C(\mathbf{s}-\mathbf{s}') = \sigma^2 \ \rho_1((\mathbf{s}-\mathbf{s}')^T B_1(\mathbf{s}-\mathbf{s}')) \ \rho_2((\mathbf{s}-\mathbf{s}')^T B_2(\mathbf{s}-\mathbf{s}')) \ , \tag{2.11}$$

noting that c is valid since it arises as a product of valid covariance functions. See Ecker and Gelfand (2003) for further details and examples. Evidently, we can extend to a product of more than two geometric anisotropy forms and we can create rich directional range behavior. However, a challenge with (2.11) is that it introduces 7 parameters into the covariance function and it will be difficult to identify and learn about all of them unless we have many, many locations.

2.2.2 Other notions of anisotropy

In a more general discussion, Zimmerman (1993) suggests three different notions of anisotropy: *sill* anisotropy, *nugget* anisotropy, and *range* anisotropy. More precisely, working with a variogram $\gamma(\mathbf{h})$, let $\mathbf{h}$ be an arbitrary separation vector so that $\mathbf{h}/\|\mathbf{h}\|$ is a unit vector in $\mathbf{h}$'s direction. Consider $\gamma(a\mathbf{h}/\|\mathbf{h}\|)$. Let $a \to \infty$ and suppose $\lim_{a\to\infty} \gamma(a\mathbf{h}/\|\mathbf{h}\|)$ depends upon $\mathbf{h}$. This situation is naturally referred to as sill anisotropy. If we work with the usual relationship $\gamma(a\mathbf{h}/\|\mathbf{h}\|) = \tau^2 + \sigma^2\left(1 - \rho\left(a\frac{\mathbf{h}}{\|\mathbf{h}\|}\right)\right)$, then, in some directions, ρ must not go to 0 as $a \to \infty$. If this can be the case, then ergodicity assumptions (i.e., convergence assumptions associated with averaging) will be violated. If so, then perhaps the constant mean assumption, implicit for the variogram, does not hold. Alternatively, it is also possible that the constant nugget assumption fails.

Instead, let $a \to 0$ and suppose $\lim_{a\to 0} \gamma(a\mathbf{h}/\|\mathbf{h}\|)$ depends upon $\mathbf{h}$. This situation is referred to as nugget anisotropy. Since, by definition, ρ must go to 1 as $a \to 0$, this case says that the assumption of uncorrelated measurement errors with common variance may not be appropriate. In particular, a simple white noise process model with constant nugget for the nonspatial errors is not appropriate.

A third type of anisotropy is range anisotropy where the range depends upon direction. Zimmerman (1993) asserts that "this is the form most often seen in practice." Geometric anisotropy and the more general product geometric anisotropy from the previous subsections are illustrative cases. However, given the various constructive strategies offered in Subsection 3.1.2 to create more general stationary covariance functions, we can envision nongeometric range anisotropy, implying general correlation function or variogram contours in $\Re^2$. However, due to the positive definiteness restriction on the correlation function, the extent of possible contour shapes is still rather limited.

Lastly, motivated by directional variograms (see Subsection 2.3.2), some authors propose the idea of nested models (see Zimmerman, 1993, and the references therein). That is, for each separation vector there is an associated angle with, say, the x-axis, which by symmetry considerations can be restricted to $[0, \pi)$. Partitioning this interval into a set of angle classes, a different variogram model is assumed to operate for each class. In terms of correlations, this would imply a different covariance function is operating for each angle class. But evidently this does not define a valid process model: the resulting covariance matrix for an arbitrary set of locations need not be positive definite.

This can be seen with as few as three points and two angle classes. Let $(\mathbf{s}_1, \mathbf{s}_2)$ belong to one angle class with $(\mathbf{s}_1, \mathbf{s}_3)$ and $(\mathbf{s}_2, \mathbf{s}_3)$ in the other. With exponential isotropic correlation functions in each class by choosing ϕ_1 and ϕ_2 appropriately we can make $\rho(\mathbf{s}_1 - \mathbf{s}_2) \approx 0$ while $\rho(\mathbf{s}_1 - \mathbf{s}_3) = \rho(\mathbf{s}_2 - \mathbf{s}_3) \approx 0.8$. A quick calculation shows that the resulting 3×3 covariance (correlation) matrix is not positive definite. So, in terms of being able to write proper joint distributions for the resulting data, nested models are inappropriate; they do not provide an extension of isotropy that allows for likelihood based inference.

2.3 Exploratory approaches for point-referenced data

2.3.1 Basic techniques

Exploratory data analysis (EDA) tools are routinely implemented in the process of analyzing one- and two-sample data sets, regression studies, generalized linear models, etc. (see, e.g., Chambers et al., 1983; Hoaglin, Mosteller, and Tukey, 1983, 1985; Aiktin et al., 1989). Similarly, such tools are appropriate for analyzing point-referenced spatial data.

For continuous spatial data, the starting point is the so-called "first law of geostatistics." Figure 2.2 illustrates this "law" in a one-dimensional setting. The data is partitioned into a mean term and an error term. The mean corresponds to global (or *first-order*) behavior,

Figure 2.2 *Illustration of the first law of geostatistics.*

Figure 2.3 *Illustrative three-dimensional "drop line" scatterplot, Stockton data.*

while the error captures local (or *second-order*) behavior through a covariance function. EDA tools are available to examine both first- and second-order behavior.

The law also clarifies that spatial association in the data, $Y(\mathbf{s})$, need not resemble spatial association in the residuals, $\epsilon(\mathbf{s})$. That is, spatial association in the $Y(\mathbf{s})$ corresponds to looking at $E(Y(\mathbf{s}) - \mu)(Y(\mathbf{s}') - \mu)$, while spatial structure in the $\epsilon(\mathbf{s})$ corresponds to looking at $E(Y(\mathbf{s}) - \mu(\mathbf{s}))(Y(\mathbf{s}') - \mu(\mathbf{s}'))$. The difference between the former and the latter is $(\mu - \mu(\mathbf{s}))(\mu - \mu(\mathbf{s}'))$, which, if interest is in spatial regression, we would not expect to be negligible.

Certainly an initial exploratory display should be a simple map of the locations themselves. We need to assess how *regular* the arrangement of the points is and also whether there is a much larger maximum distance between points in some directions than in others. Next, some authors would recommend a stem-and-leaf display of the $Y(\mathbf{s})$. This plot is evidently nonspatial and is customarily for observations which are i.i.d. We expect both nonconstant mean and spatial dependence, but such a plot may at least suggest potential outliers. Next

Figure 2.4 *Illustrative three-dimensional surface ("perspective") plot, Stockton real estate data.*

Figure 2.5 *Illustrative contour plot, Stockton real estate data.*

we might develop a three-dimensional "drop line" scatterplot of $Y(\mathbf{s}_i)$ versus $\mathbf{s}_i$, alternatively, a three-dimensional surface plot or perhaps, a contour plot as a *smoothed* summary. Examples of these three plots are shown for a sample of 120 log-transformed home selling prices in Stockton, CA, in Figures 2.3, 2.4, and 2.5, respectively. Of these three, we find the contour plot to be the most effective in revealing the entire spatial surface. However, as the preceding paragraph clarifies, such displays may be deceiving. They may show spatial pattern that will disappear after $\mu(\mathbf{s})$ is fitted, or perhaps vice versa. It seems more sensible to study spatial pattern in the residuals.

Figure 2.6 *Illustrative row box plots, Diggle and Ribeiro (2002) surface elevation data.*

In exploring $\mu(\mathbf{s})$ we may have two types of information at location $\mathbf{s}$. One is the purely geographic information, i.e., the geocoded location expressed in latitude and longitude or as projected coordinates such as eastings and northings (Subsection 1.2.1 above). The other will be features relevant for explaining the $Y(\mathbf{s})$ at $\mathbf{s}$. For instance, if $Y(\mathbf{s})$ is a pollution concentration, then elevation, temperature, and wind information at $\mathbf{s}$ could well be useful and important. If, instead, $Y(\mathbf{s})$ is the selling price of a single-family home at $\mathbf{s}$, then characteristics of the home (square feet, age, number of bathrooms, etc.) would be useful.

When the mean is described purely through geographic information, $\mu(\mathbf{s})$ is referred to as a *trend surface*. When $\mathbf{s} \in \Re^2$, the surface is usually developed as a low-dimensional bivariate polynomial. For data that is roughly gridded (or can be assigned to row and column bins by overlaying a regular lattice on the points), we can make row and column boxplots looking for trend. Displaying these boxplots versus their center could clarify the existence and nature of such trend. In fact, median polishing (see, e.g., Hoaglin, Mosteller, and Tukey, 1985) could be used to extract row and column effects, and also to see if a multiplicative trend surface term is useful; see Cressie (1983, pp. 46–48) in this regard.

Figures 2.6 and 2.7 illustrate the row and column boxplot approach for a data set previously considered by Diggle and Ribeiro (2002). The response variable is the surface elevation ("height") at 52 locations on a regular grid within a 310-foot square (and where the mesh of the grid is 50 feet). The plots reveals some evidence of spatial pattern as we move along the rows, but not down the columns of the regular grid.

To assess small-scale behavior, some authors recommend creating the *semivariogram cloud*, i.e., a plot of $(Y(\mathbf{s}_i) - Y(\mathbf{s}_j))^2$ versus $||\mathbf{s}_i - \mathbf{s}_j||$. Usually this cloud is too "noisy" to reveal very much; see, e.g., Figure 6.1. The empirical semivariogram (2.9) is preferable in terms of reducing some of the noise, and can be a helpful tool in assessing the presence of spatial structure. Again, the caveat above suggests employing it for residuals (not the data itself) unless a constant mean is appropriate.

Figure 2.7 *Illustrative column box plots, Diggle and Ribeiro (2002) surface elevation data.*

An empirical (nonparametric) covariance estimate, analogous to (2.9), is also available. Creating bins as in this earlier approach, define

$$\widehat{C}(t_k) = \frac{1}{N_k} \sum_{(\mathbf{s}_i,\mathbf{s}_j)\in N(t_k)} (Y(\mathbf{s}_i) - \bar{Y})(Y(\mathbf{s}_j) - \bar{Y}) \ , \tag{2.12}$$

where again $N(t_k) = \{(\mathbf{s}_i, \mathbf{s}_j) : ||\mathbf{s}_i - \mathbf{s}_j|| \in I_k\}$ for $k = 1, \ldots, K$, I_k indexes the kth bin, and there are N_k pairs of points falling in this bin. Equation (2.12) is a spatial generalization of a lagged autocorrelation in time series analysis. Since $\widehat{c}$ uses a common $\bar{Y}$ for all $Y(\mathbf{s}_i)$, it may be safer to employ (2.12) on the residuals. We note that the empirical covariance function is not used much in practice; the empirical variogram is much more common. Moreover, two further issues arise: first, how should we define $\widehat{C}(0)$, and second, regardless of this choice, we note that $\widehat{\gamma}(t_k)$ does *not* equal $\widehat{C}(0) - \widehat{C}(t_k)$, $k = 1, \ldots, K$. Details for both of these issues are left to Exercise 5. Again, since we are only viewing $\widehat{(\gamma)}$ and $\widehat{C}$ as exploratory tools, there seems to be no reason to pursue the latter further.

Again, with a regular grid or binning we can create "same-lag" scatterplots. These are plots of $Y(\mathbf{s}_i + h\mathbf{e})$ versus $Y(\mathbf{s}_i)$ for a fixed h and a fixed unit vector $\mathbf{e}$. Comparisons among such plots may reveal the presence of anisotropy and perhaps nonstationarity.

Lastly, suppose we attach a neighborhood to each point. We can then compute the sample mean and variance for the points in the neighborhood, and even a sample correlation coefficient using all pairs of data in the neighborhood. Plots of each of them versus location can be informative. The first may give some idea regarding how the mean structure changes across the study region. Plots of the second and third may provide evidence of nonstationarity. Implicit in extracting useful information from these plots is a roughly constant local mean. If $\mu(\mathbf{s})$ is to be a trend surface, this is plausible. But if $\mu(\mathbf{s})$ is a function of some geographic variables at $\mathbf{s}$ (say, home characteristics), then use of residuals would be preferable.

Figure 2.8 *Sites sampled in the Atlantic Ocean for 1990 and 1993 scallop catch data.*

2.3.2 Assessing anisotropy

We illustrate various EDA techniques to assess anisotropy using sampling of scallop abundance on the continental shelf off the coastline of the northeastern U.S. The data comes from a survey conducted by the Northeast Fisheries Science Center of the National Marine Fisheries Service. Figure 2.8 shows the sampling sites for 1990 and 1993. We see much more sampling in the southwest to northeast direction than in the northwest to southeast direction. Evidently, it is more appropriate to follow the coastline in searching for scallops.

2.3.2.1 Directional semivariograms and rose diagrams

The most common EDA technique for assessing anisotropy involves use of directional semivariograms. Typically, one chooses angle classes $\eta_i \pm \epsilon$, $i = 1, \ldots, L$ where ϵ is the halfwidth of the angle class and L is the number of angle classes. For example, a common choice of angle classes involves the four cardinal directions measured counterclockwise from the x-axis ($0°$, $45°$, $90°$, and $135°$) where ϵ is $22.5°$. Journel and Froidevaux (1982) display directional semivariograms at angles $35°$, $60°$, $125°$, and $150°$ in deducing anistropy for a tungsten deposit. While knowledge of the underlying spatial characteristics of region D is invaluable in choosing directions, often the choice of the number of angle classes and the directions seems to be arbitrary.

For a given angle class, the Matheron empirical semivariogram (2.9) can be used to provide a directional semivariogram for angle η_i. Theoretically, all types of anisotropy can be assessed from these directional semivariograms; however, in practice determining whether the sill, nugget, and/or range varies with direction can be difficult. In particular, it is unclear how much variability will arise in directional variograms generated under

Figure 2.9 *Directional semivariograms (a) and a rose diagram (b) for the 1990 scallop data.*

isotropy. Figure 2.9(a) illustrates directional semivariograms for the 1990 scallop data in the four cardinal directions. Note that the semivariogram points are connected only to aid comparison.

Possible conclusions from the figure are: the variability in the 45° direction (parallel to the coastline) is significantly less than in the other three directions and the variability perpendicular to the coastline (135°) is very erratic, possibly exhibiting sill anisotropy. We caution however that it is dangerous to read too much significance and interpretation into directional variograms. For example, from Figure 2.8, as noted above, we see far more sampling at greater distances in the southwest-northeast direction than in the northwest-southeast direction. Moreover, no sample sizes (and thus no assessments of variability) are attached to these pictures. Directional variograms from data generated under a simple isotropic model will routinely exhibit differences of the magnitudes seen in Figure 2.9(a). Furthermore, it seems difficult to draw any conclusions regarding the presence of geometric anisotropy from this figure.

A rose diagram (Isaaks and Srivastava, 1989, pp. 151–154) can be created from the directional semivariograms to evaluate geometric anisotropy. At an arbitrarily selected γ^*, for a directional semivariogram at angle η, the distance d^* at which the directional semivariogram attains γ^* can be interpolated. Then, the rose diagram is a plot of angle η and corresponding distance d^* in polar coordinates. If an elliptical contour describes the extremities of the rose diagram reasonably well, then the process exhibits geometric anisotropy. For instance, the rose diagram for the 1990 scallop data is presented in Figure 2.9(b) using the γ^* contour of 4.5. It is approximately elliptical, oriented parallel to the coastline ($\approx 45°$) with a ratio of major to minor ellipse axes of about 4.

2.3.2.2 Empirical semivariogram contour (ESC) plots

A more informative method for assessing anisotropy is a contour plot of the empirical semivariogram surface in $\Re^2$. Such plots are mentioned informally in Isaaks and Srivastava (1989, pp. 149–151) and in Haining (1990, pp. 284–286); the former call them contour maps of the grouped variogram values, the latter an isarithmic plot of the semivariogram.

Figure 2.10 *ESC plot for the 1993 scallop data.*

Following Ecker and Gelfand (1999), we formalize such a plot here, calling it an *empirical semivariogram contour* (ESC) plot. For each of the $\frac{N(N-1)}{2}$ pairs of sites in $\Re^2$, calculate h_x and h_y, the separation distances along each axis. Since the sign of h_y depends upon the arbitrary order in which the two sites are compared, we demand that $h_y \geq 0$. (We could alternatively demand that $h_x \geq 0$.) That is, we take $(-h_x, -h_y)$ when $h_y < 0$. These separation distances are then aggregated into rectangular bins B_{ij} where the empirical semivariogram values for the (i,j)th bin are calculated by

$$\gamma^*_{ij} = \frac{1}{2N_{B_{ij}}} \sum_{\{(k,l):(\mathbf{s}_k - \mathbf{s}_l) \in B_{ij}\}} (Y(\mathbf{s}_k) - Y(\mathbf{s}_l))^2 \,, \tag{2.13}$$

where $N_{B_{ij}}$ equals the number of sites in bin B_{ij}. Because we force $h_y \geq 0$ with h_x unrestricted, we make the bin width on the y-axis half of that for the x-axis. We also force the middle class on the x-axis to be centered around zero. Upon labeling the center of the (i,j)th bin by (x_i, y_j), a three dimensional plot of γ^*_{ij} versus (x_i, y_j) yields an empirical semivariogram surface. Smoothing this surface using interpolation algorithms (e.g., Akima, 1978) produces a contour plot that we call the ESC plot. A symmetrized version of the ESC plot can be created by reflecting the upper left quadrant to the lower right and the upper right quadrant to the lower left.

The ESC plot can be used to assess departures from isotropy; isotropy is depicted by circular contours while elliptical contours capture geometric anisotropy. A rose diagram traces only one arbitrarily selected contour of this plot. A possible drawback to the ESC plot is the occurrence of sparse counts in extreme bins. However, these bins may be trimmed before smoothing if desired. Concerned that use of geographic coordinates could introduce artificial anisotropy (since 1° latitude $\neq$ 1° longitude in the northeastern United States), we have employed a Universal Transverse Mercator (UTM) projection to kilometers in the E-W and N-S axes (see Subsection 1.2.1).

Figure 2.10 is the empirical semivariogram contour plot constructed using x-axis width of 30 kilometers for the 1993 scallop data. We have overlaid this contour plot on the bin centers with their respective counts. Note that using empirical semivariogram values in the row of the ESC plot for which $h_y \approx 0$ provides an alternative to the usual 0° directional

Figure 2.11 *Directional semivariograms for the 1993 scallop data.*

semivariogram. The latter directional semivariograms are based on a polar representation of the angle and distance. For a chosen direction η and tolerance ϵ, the area for a class fans out as distance increases (see Figure 7.1 of Isaaks and Srivastava, 1989, p. 142). Attractively, a directional semivariogram based on the rectangular bins associated with the empirical semivariogram in $\Re^2$ has bin area remaining constant as distance increases. In Figure 2.11, we present the four customary directional (polar representation) semivariograms for the 1993 scallop data. Clearly, the ESC plot is more informative, particularly in suggesting evidence of geometric anisotropy.

2.4 Classical spatial prediction

In this section we describe the classical (i.e., minimum mean-squared error) approach to spatial prediction in the point-referenced data setting. The approach is commonly referred to as *kriging*, so named by Matheron (1963) in honor of D.G. Krige, a South African mining engineer whose seminal work on empirical methods for geostatistical data (Krige, 1951) inspired the general approach (and indeed, inspired the convention of using the terms "point-level spatial analysis" and "geostatistical analysis" interchangeably!). The problem is one of optimal spatial prediction: given observations of a random field $\mathbf{Y} = (Y(\mathbf{s}_1), \ldots, Y(\mathbf{s}_n))'$, how do we predict the variable Y at a site $\mathbf{s}_0$ where it has not been observed? In other words, what is the best predictor of the value of $Y(\mathbf{s}_0)$ based upon the data $\mathbf{y}$?

A linear predictor for $Y(\mathbf{s}_0)$ based on $\mathbf{y}$ would take the form $\sum \ell_i Y(\mathbf{s}_i) + \delta_0$. Using squared error loss, the best linear prediction would minimize $E[Y(\mathbf{s}_0) - (\sum \ell_i Y(\mathbf{s}_i) + \delta_0)]^2$ over δ_0 and the ℓ_i. Under the intrinsic stationarity specification (2.1) we would take $\sum \ell_i = 1$ in order that $E(Y(\mathbf{s}_0) - \sum \ell_i Y(\mathbf{s}_i)) = 0$. As a result, we would minimize $E[Y(\mathbf{s}_0) - \sum \ell_i Y(\mathbf{s}_i)]^2 + \delta_0^2$, and clearly δ_0 would be set to 0. Now letting $a_0 = 1$ and $a_i = -\ell_i$ we see that the criterion becomes $E[\sum_{i=0}^n a_i Y(\mathbf{s}_i)]^2$ with $\sum a_i = 0$. But from (2.4) this expectation becomes $-\sum_i \sum_j a_i a_j \gamma(\mathbf{s}_i - \mathbf{s}_j)$, revealing how, historically, the variogram arose in kriging within the geostatistical framework. Indeed, the optimal ℓ's can be obtained by solving this constrained optimization. In fact, it is a routine optimization of a quadratic from under a linear constraint and is customarily handled using Lagrange multipliers. The solution will be a function of $\gamma(\mathbf{h})$, in fact of the set of $\gamma_{ij} = \gamma(\mathbf{s}_i - \mathbf{s}_j)$ and $\gamma_{0j} = \gamma(\mathbf{s}_0 - \mathbf{s}_j)$ With an

estimate of γ, one immediately obtains the so-called *ordinary kriging* estimate. Other than the intrinsic stationarity model (Subsection 2.1.2), no further distributional assumptions are required for the $Y(\mathbf{s})$'s.

To provide a bit more detail, restoring the ℓ_i's, we obtain

$$-\sum_i \sum_j a_i a_j \gamma(\mathbf{s}_i - \mathbf{s}_j) = -\sum_i \sum_j \ell_i \ell_j \gamma_{ij} + 2\sum_i \ell_i \gamma_{0i}$$

Adding the constraint, $\sum \ell_i = 1$, times the Lagrange multiplier, λ, we find that the partial derivative of this expression with regard to ℓ_i becomes $-\sum_j \ell_j \gamma_{ij} + \gamma_{0i} - \lambda = 0$. Letting Γ be the $n \times n$ matrix with entries $\Gamma_{ij} = \gamma_{ij}$ and $\boldsymbol{\gamma}_0$ be the $n \times 1$ vector with $(\boldsymbol{\gamma}_0)_i = \gamma_{0i}$, with $\boldsymbol{\ell}$ the $n \times 1$ vector of coefficients, we obtain the system of *kriging* equations, $\Gamma\boldsymbol{\ell} + \lambda\mathbf{1} = \boldsymbol{\gamma}_0$ and $\mathbf{1}^T\boldsymbol{\ell} = 1$. The solution is $\boldsymbol{\ell} = \Gamma^{-1}\left(\boldsymbol{\gamma}_0 + \frac{(1 - \mathbf{1}^T\Gamma^{-1}\boldsymbol{\gamma}_0)}{\mathbf{1}^T\Gamma^{-1}\mathbf{1}}\mathbf{1}\right)$ and $\boldsymbol{\ell}^T\mathbf{Y}$ becomes the best linear unbiased predictor (BLUP). Again, with $\gamma(\mathbf{h})$ unknown, we have to estimate it in order to calculate this estimator. Then, $\boldsymbol{\ell}$ is a function of the data and the estimator is no longer linear. Continuing in this regard, the usual estimator of the uncertainty in the prediction is the predictive mean square error (PMSE), $E(Y(\mathbf{s}_0) - \boldsymbol{\ell}^T\mathbf{Y})^2$, rather than $\mathrm{var}(\boldsymbol{\ell}^T\mathbf{Y})$. There is a closed form expression for the former, i.e., ${\boldsymbol{\gamma}_0}^T\Gamma^{-1}\boldsymbol{\gamma}_0 - (\boldsymbol{\ell}^T\Gamma^{-1}\boldsymbol{\gamma}_0 - 1)^2/\boldsymbol{\ell}^T\Gamma^{-1}\boldsymbol{\ell}$, which, also requires an estimate of $\gamma(\mathbf{h})$ in order to be calculated.

Let us now take a formal look at kriging in the context of Gaussian processes. Consider first the case where we have no covariates, but only the responses $Y(\mathbf{s}_i)$. This is developed by means of the following model for the observed data:

$$\mathbf{Y} = \mu\mathbf{1} + \boldsymbol{\epsilon}, \text{ where } \boldsymbol{\epsilon} \sim N(\mathbf{0}, \Sigma) \ .$$

For a spatial covariance structure having no nugget effect, we specify Σ as

$$\Sigma = \sigma^2 H(\phi) \text{ where } (H(\phi))_{ij} = \rho(\phi; d_{ij}) \ ,$$

where $d_{ij} = ||\mathbf{s}_i - \mathbf{s}_j||$, the distance between $\mathbf{s}_i$ and $\mathbf{s}_j$ and ρ is a valid correlation function on $\Re^r$ such as those in Table 2.1. For a model having a nugget effect, we instead set

$$\Sigma = \sigma^2 H(\phi) + \tau^2 I \ ,$$

where τ^2 is the nugget effect variance.

When covariate values $\mathbf{x} = (x(\mathbf{s}_1), \ldots, x(\mathbf{s}_n))'$ and $x(\mathbf{s}_0)$ are available for incorporation into the analysis, the procedure is often referred to as *universal kriging*, though we caution that some authors (e.g., Kaluzny et al., 1998) use the term "universal" in reference to the case where only latitude and longitude are available as covariates. The model now takes the more general form

$$\mathbf{Y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon}, \text{ where } \boldsymbol{\epsilon} \sim N(\mathbf{0}, \Sigma) \ ,$$

with Σ being specified as above, either with or without the nugget effect. Note that ordinary kriging may be looked upon as a particular case of universal kriging with X being the $n \times 1$ matrix (i.e., column vector) $\mathbf{1}$, and $\boldsymbol{\beta}$ the scalar μ.

We now pose our prediction problem as follows: we seek the function $h(\mathbf{y})$ that minimizes the mean-squared prediction error,

$$E\left[(Y(\mathbf{s}_0) - h(\mathbf{y}))^2 \,\middle|\, \mathbf{y}\right] \ . \tag{2.14}$$

By adding and subtracting the conditional mean $E[Y(\mathbf{s}_0)|\mathbf{y}]$ inside the square, grouping terms, and squaring we obtain

$$\begin{aligned} &E\left[(Y(\mathbf{s}_0) - h(\mathbf{y}))^2 \,\middle|\, \mathbf{y}\right] \\ &\quad = E\left\{(Y(\mathbf{s}_0) - E[Y(\mathbf{s}_0)|\mathbf{y}])^2 \,\middle|\, \mathbf{y}\right\} + \{E[Y(\mathbf{s}_0)|\mathbf{y}] - h(\mathbf{y})\}^2 \ , \end{aligned}$$

since (as often happens in statistical derivations like this) the expectation of the cross-product term equals zero. But, since the second term on the right-hand side is nonnegative, we have

$$E\left[(Y(\mathbf{s}_0)-h(\mathbf{y}))^2 \,\middle|\, \mathbf{y}\right] \geq E\left\{(Y(\mathbf{s}_0)-E[Y(\mathbf{s}_0)|\mathbf{y}])^2 \,\middle|\, \mathbf{y}\right\}$$

for any function $h(\mathbf{y})$. Equality holds if and only if $h(\mathbf{y}) = E[Y(\mathbf{s}_0)|\mathbf{y}]$, so it must be that the predictor $h(\mathbf{y})$ which minimizes the error is the conditional expectation of $Y(\mathbf{s}_0)$ given the data. This result is familiar since we know that the mean minimizes expected squared error loss. From a Bayesian point of view, this $h(\mathbf{y})$ is just the *posterior mean* of $Y(\mathbf{s}_0)$, and it is well known that the posterior mean is the Bayes rule (i.e., the minimizer of posterior risk) under squared error loss functions of the sort adopted in (2.14) above as our scoring rule. Note that the posterior mean is the best predictor under squared error loss regardless of whether we assume a Gaussian model and that, in general, it need not be linear; it need not be the BLUP which is the ordinary kriging predictor. However, under a Gaussian process model assumption, this posterior mean is linear, as we now clarify.

In particular, let us explicitly obtain the form of the posterior mean of $Y(\mathbf{s}_0)$. It will emerge as a function of all the population parameters ($\boldsymbol{\beta}, \sigma^2, \phi$, and τ^2). From standard multivariate normal theory we have the following general result: If

$$\begin{pmatrix} \mathbf{Y}_1 \\ \mathbf{Y}_2 \end{pmatrix} \sim N\left(\begin{pmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{pmatrix}, \begin{pmatrix} \Omega_{11} & \Omega_{12} \\ \Omega_{21} & \Omega_{22} \end{pmatrix}\right),$$

where $\Omega_{21} = \Omega_{12}^T$, then the conditional distribution $p(\mathbf{Y}_1|\mathbf{Y}_2)$ is normal with mean and variance:

$$\begin{aligned} E[\mathbf{Y}_1|\mathbf{Y}_2] &= \boldsymbol{\mu}_1 + \Omega_{12}\Omega_{22}^{-1}(\mathbf{Y}_2 - \boldsymbol{\mu}_2); \\ Var[\mathbf{Y}_1|\mathbf{Y}_2] &= \Omega_{11} - \Omega_{12}\Omega_{22}^{-1}\Omega_{21}\,. \end{aligned}$$

In our framework, we have $\mathbf{Y}_1 = Y(\mathbf{s}_0)$ and $\mathbf{Y}_2 = \mathbf{y}$. It then follows that

$$\Omega_{11} = \sigma^2 + \tau^2, \quad \Omega_{12} = \boldsymbol{\gamma}^T, \quad \text{and } \Omega_{22} = \Sigma = \sigma^2 H(\phi) + \tau^2 I\,,$$

where $\boldsymbol{\gamma}^T = \left(\sigma^2\rho(\phi; d_{01}), \ldots, \sigma^2\rho(\phi; d_{0n})\right)$. Substituting these values into the mean and variance formulae above, we obtain

$$E[Y(\mathbf{s}_0)|\mathbf{y}] = \mathbf{x}_0^T\boldsymbol{\beta} + \boldsymbol{\gamma}^T\Sigma^{-1}(\mathbf{y} - X\boldsymbol{\beta})\,, \tag{2.15}$$

$$\text{and } Var[Y(\mathbf{s}_0)|\mathbf{y}] = \sigma^2 + \tau^2 - \boldsymbol{\gamma}^T\Sigma^{-1}\boldsymbol{\gamma}\,. \tag{2.16}$$

We see that the posterior mean is a linear predictor. We remark that this solution assumes we have actually observed the covariate value $\mathbf{x}_0 = \mathbf{x}(\mathbf{s}_0)$ at the "new" site $\mathbf{s}_0$; we defer the issue of missing $\mathbf{x}_0$ for the moment.

Since, in practice, the model parameters are unknown, they must be estimated from the data. Here we would modify $h(\mathbf{y})$ to

$$\widehat{h(\mathbf{y})} = \mathbf{x}_0^T\widehat{\boldsymbol{\beta}} + \widehat{\boldsymbol{\gamma}}^T\widehat{\Sigma}^{-1}\left(\mathbf{y} - X\widehat{\boldsymbol{\beta}}\right)\,,$$

where $\widehat{\boldsymbol{\gamma}} = \left(\hat{\sigma}^2\rho(\hat{\phi}; d_{01}), \ldots, \hat{\sigma}^2\rho(\hat{\phi}; d_{0n})\right)^T$, $\widehat{\boldsymbol{\beta}} = \left(X^T\widehat{\Sigma}^{-1}X\right)^{-1}X^T\widehat{\Sigma}^{-1}\mathbf{y}$, the usual weighted least squares estimator of $\boldsymbol{\beta}$, and $\widehat{\Sigma} = \hat{\sigma}^2 H(\hat{\phi})$. Thus $\widehat{h(\mathbf{y})}$ can be written as $\boldsymbol{\lambda}^T\mathbf{y}$, where

$$\boldsymbol{\lambda} = \widehat{\Sigma}^{-1}\widehat{\boldsymbol{\gamma}} + \widehat{\Sigma}^{-1}X\left(X^T\widehat{\Sigma}^{-1}X\right)^{-1}\left(\mathbf{x}_0 - X^T\widehat{\Sigma}^{-1}\widehat{\boldsymbol{\gamma}}\right)\,. \tag{2.17}$$

The reader may be curious as to whether $\boldsymbol{\lambda}^T\mathbf{Y}$ is the same as $\boldsymbol{\ell}^T\mathbf{Y}$ assuming all parameters are known say, in the ordinary kriging case. That is, assuming we have $\gamma(\mathbf{h}) = c(\mathbf{0}) - c(\mathbf{h})$, we can write both linear predictors in terms of $c(\mathbf{h})$ or $\gamma(\mathbf{h})$ but will they agree? Immediately, we know that the answer is no because $\boldsymbol{\ell}^T\mathbf{Y}$ depends only on $\gamma(\mathbf{h})$ or $c(\mathbf{h})$ while $\boldsymbol{\lambda}^T\mathbf{Y}$ requires $\boldsymbol{\beta}$, hence μ in the constant mean case. In fact, it is a straightforward calculation to show that, if we replace μ by $\hat{\mu}$, the best linear unbiased estimator of μ, in $\boldsymbol{\lambda}$, we do obtain the ordinary kriging estimator. We leave this as an exercise.

If $\mathbf{x}_0$ is unobserved, we can estimate it and $Y(\mathbf{s}_0)$ jointly by iterating between this formula and a corresponding one for $\hat{\mathbf{x}}_0$, namely

$$\hat{\mathbf{x}}_0 = X^T\boldsymbol{\lambda}\,,$$

which arises simply by multiplying both sides of (2.17) by X^T and simplifying. This is essentially an EM (expectation-maximization) algorithm (Dempster, Laird, and Rubin, 1977), with the calculation of $\hat{\mathbf{x}}_0$ being the E step and (2.17) being the M step.

In the classical framework a lot of energy is devoted to the determination of the optimal estimates to plug into the above equations. Typically, restricted maximum likelihood (REML) estimates are selected and shown to have certain optimal properties. However, as we shall see in Chapter 6, how to perform the estimation is not an issue in the Bayesian setting. There, we instead impose prior distributions on the parameters and produce the full posterior predictive distribution $p\,(Y(\mathbf{s}_0)|\mathbf{y})$. Any desired point or interval estimate (the latter to express our uncertainty in such prediction), as well as any desired probability statements, may then be computed with respect to this distribution.

2.4.0.3 Noiseless kriging

It is natural to ask whether the kriging predictor (ordinary or posterior mean) will return the observed value at the $\mathbf{s}_i$'s where we observed the process. If so, we would refer to the predictor as an *exact* interpolator. The literature addresses this problem through detailed inspection of the kriging equations. However, the answer is immediately clear once we look at the model specification. And, it is possibly counterintuitive; we obtain exact interpolation in the case of no nugget, rather than in the seemingly more flexible case where we add a nugget.

Analytically, one can determine whether or not the predictor in (2.15) will equal the observed value at a given $\mathbf{s}_i$. We leave as a formal exercise to verify that if $\tau^2 = 0$ (i.e., the no-nugget case, or so-called noiseless prediction), then the answer is yes, while if $\tau^2 > 0$ then the answer is no.

However, we can illuminate the situation without formal calculation. There are two potential settings: (i) $Y(\mathbf{s}) = \mu(\mathbf{s}) + w(\mathbf{s})$ and (ii) $Y(\mathbf{s}) = \mu(\mathbf{s}) + w(\mathbf{s}) + \epsilon(\mathbf{s})$, the no-nugget and nugget cases, respectively. Under (i), evidently, predicting $Y(\mathbf{s}_0)$ is the same as predicting $\mu(\mathbf{s}_0) + w(\mathbf{s}_0)$. There is only one surface for the process realization. Under (ii), predicting $Y(\mathbf{s}_0)$ is different from predicting $\mu(\mathbf{s}_0) + w(\mathbf{s}_0)$; we have two random surfaces where the former is everywhere discontinuous while the latter, for customary covariance functions, is continuous. Here, terminology can be confusing. Under (i), there is no noise, so we could refer to this case as noiseless kriging. However, under (ii), predicting $\mu(\mathbf{s}_0) + w(\mathbf{s}_0)$ could be referred to as noiseless interpolation. We ignore the terminology issue and address the exact interpolation question.

Now, suppose $\mathbf{s}_0$ is, in fact, one of the $\mathbf{s}_i$. Then, we need to distinguish the observed value at $\mathbf{s}_i$ say $Y_{obs}(\mathbf{s}_i)$ from a new or replicate value at $\mathbf{s}_i$, say $Y_{rep}(\mathbf{s}_i)$. Under (i) $f(Y_{rep}(\mathbf{s}_i)|\text{Data}) = f(\mu(\mathbf{s}_i) + w(\mathbf{s}_i)|\text{Data})$ and since, given the data, $\mu(\mathbf{s}_i) + w(\mathbf{s}_i) = Y_{obs}(\mathbf{s}_i)$, $f(Y_{rep}(\mathbf{s}_i)|\text{Data})$ is degenerate at $Y_{obs}(\mathbf{s}_i)$; we have exact interpolation. Under (ii), $Y_{rep}(\mathbf{s}_i) \sim Y_{obs}(\mathbf{s}_i)$ and $E(Y_{rep}(\mathbf{s}_i)|\text{Data}) = E(\mu(\mathbf{s}_i) + w(\mathbf{s}_i)|\text{Data}) \neq Y_{obs}(\mathbf{s}_i)$; we do not have exact interpolation.

2.5 Computer tutorials

2.5.1 EDA and spatial data visualization in R

We will illustrate our approaches using the so called WEF forest inventory data from a long-term ecological research site in western Oregon. These data consist of a census of all trees in a 10 ha stand. Diameter at breast height (DBH) and tree height (HT) have been measured for all trees in the stand. For a subset of these trees, the distance from the center of the stem to the edge of the crown was measured at each of the cardinal directions.

Rarely do we have the luxury of a complete census, but rather a subset of trees are sampled using a series of inventory plots. This sample is then used to draw inferences about parameters of interest (e.g., mean stand DBH or correlation between DBH and HT). We defer these analyses to subsequent example sessions. Here, we simply use these data to demonstrate some basics of spatial data manipulation, visualization, and exploratory analysis. We will use the following three packages:

```
> library(spBayes)
> library(classInt)
> library(RColorBrewer)
```

We begin by removing rows that have NA for the variables of interest, creating a vector of DBH and HT, and plotting the tree coordinates, which appear in Figure 2.12.

```
> data(WEF.dat)
> WEF.dat <- WEF.dat[!apply(WEF.dat[,c("East_m","North_m",
+        "DBH_cm","Tree_height_m","ELEV_m")], 1, function(x)any(is.na(x))),]
> DBH <- WEF.dat$DBH_cm
> HT <- WEF.dat$Tree_height_m
> coords <- as.matrix(WEF.dat[,c("East_m","North_m")])
> plot(coords, pch=1, cex=sqrt(DBH)/10, col="darkgreen",
+      xlab="Easting (m)", ylab="Northing (m)")
> leg.vals <- round(quantile(DBH),0)
> legend("topleft", pch=1, legend=leg.vals, col="darkgreen",
+        pt.cex=sqrt(leg.vals)/10, bty="n", title="DBH (cm)")
```

To further explore spatial patterns in DBH, it is often useful to use a color gradient or ramp to construct the pallet. These can be created using colorRampPalette which creates a function that interpolates over a given set of colors. Here we make a color ramp function then create a color palette with four colors, each with five shades.

```
> col.br <- colorRampPalette(c("blue", "cyan", "yellow", "red"))
> col.pal <- col.br(5)
```

We can associate the colors with intervals that are meaningful to the analysis. For instance, in a forest inventory, trees are often identified as sapling, poletimber, sawtimber, and large sawtimber classes. The following code-block produces Figure 2.13, which displays the coordinates colored according to their class.

```
> fixed <- classIntervals(DBH, n = 4, style = "fixed",
+                         fixedBreaks = c(0, 12.7, 30.48, 60, max(DBH) + 1))
> fixed.col <- findColours(fixed, col.pal)
> plot(coords, col = fixed.col, pch = 19, cex = 0.5,
+      main = "Forestry tree size classes",
+      xlab = "Easting (m)", ylab = "Northing (m)")
> legend("topleft", fill = attr(fixed.col, "palette"),
+     legend = c("sapling", "poletimber", "sawtimber", "large sawtimber"),
+     bty = "n")
```

Figure 2.12 *Tree locations on the WEF.*

Figure 2.13 *Intervals based on previously defined tree size classes.*

Figure 2.14 *Interpolation of DBH using multilevel B-spline.*

If the data observations are well distributed over the domain, spatial patterns can often be detected by estimating a continuous surface using an interpolation function. Several packages provide suitable interpolators, include the `akima` for linear or cubic spline interpolation and `MBA` which provides efficient interpolation of large data sets with multilevel B-splines. Functions within both packages produce grids of interpolated values which can be passed to `image` or `image.plot` to produce $\Re^2$ depictions. The code below uses `MBA` for spatial interpolation and produces Figure 2.14.

```
> library(MBA)
> library(fields) ## For using the image.plot function
> x.res <- 100
> y.res <- 100
> surf <- mba.surf(cbind(coords, DBH), no.X = x.res, no.Y = y.res,
+                  h = 5, m = 2, extend = FALSE)$xyz.est
> image.plot(surf, xaxs = "r", yaxs = "r", xlab = "Easting (m)",
+            ylab = "Northing (m)", col = col.br(25))
> contour(surf, add=T) ## (Optional) Adds contour lines to the plot
```

Alternatively, 3-D graphics functions in the `rgl` package can be used to produce $\Re^3$ depictions, as in Figure 2.15, using the code

Figure 2.15 *Perspective plot of multilevel B-spline interpolation of DBH.*

```
> library(rgl)
> col <- rbind(0, cbind(matrix(drape.color(surf[[3]],
+     col = col.br(25)), x.res - 1, y.res - 1), 0))
> surface3d(surf[[1]], surf[[2]], surf[[3]], col = col)
> axes3d()
> title3d(main = "DBH", xlab = "Easting (m)", ylab = "Northing (m)",
+     zlab = "DBH (cm)")
> drape.plot(surf[[1]], surf[[2]], surf[[3]], col = col.br(150),
+     theta = 225, phi = 50, border = FALSE, add.legend = FALSE,
+     xlab = "Easting (m)", ylab = "Northing (m)",
+     zlab = "DBH (cm)")
> image.plot(zlim = range(surf[[3]], na.rm = TRUE),
+     legend.only = TRUE, horizontal = FALSE)
```

2.5.2 Variogram analysis in R

Our visual inspection of the forest inventory data suggest that there is some degree of spatial dependence in the distribution of DBH across the WEF. This encourages further exploration using a variogram analysis to quantify the range of spatial dependence and obtain an idea about the sill and the nugget.

We begin by fitting an isotropic empirical semivariogram using functions within the `geoR` package. In the code blocks provided below, we first fit an exponential variogram model to DBH, then fit a second variogram model to the residuals of a linear regression of DBH onto tree species. We first compute the inter-site distances by applying the `iDist` function to the coordinates. We then set the number of bins and use the `variog` function in `geoR` to compute an omnidirectional variogram.

```
> library(geoR)
> max.dist <- 0.25 * max(iDist(coords))
> bins <- 50
```

```
> vario.DBH <- variog(coords = coords, data = DBH,
+        uvec = (seq(0, max.dist, length = bins)))
```

Covariograms and correlograms, if desired, can be invoked using the `covariogram` and `correlogram` functions.

The computed `vario.DBH` object can then be passed to the `variofit` function in `geoR`, which uses nonlinear least squares to fit an exponential variogram model.

```
> fit.DBH <- variofit(vario.DBH, ini.cov.pars = c(600,200/-log(0.05)),
+            cov.model = "exponential", minimisation.function = "nls",
+            weights = "equal")
```

The function `variofit` can estimate the sill, the range, and the nugget parameters under a specified covariance model. A variogram object (typically an output from the `variog` function) is taken as input, together with initial values for the range and sill (in `ini.cov.pars`), and the covariance model is specified through `cov.model`. The covariance modeling options include `exponential`, `gaussian`, `spherical`, `circular`, `cubic`, `wave`, `power`, `powered.exponential`, `cauchy`, `gneiting`, `gneiting.matern`, and `pure.nugget` (no spatial covariance). Also, the initial values provided in `ini.cov.pars` do not include those for the nugget. It is concatenated with the value of the `nugget` option only if `fix.nugget=FALSE`. If the latter is `TRUE`, then the value in the `nugget` option is taken as the fixed true value.

We next fit a second variogram model to the residuals of an ordinary linear regression of DBH onto tree species. The code below produces a summary of regressing `DBH` on tree species (a categorical variable with five classes).

```
> lm.DBH <- lm(DBH ~ Species, data = WEF.dat)
> summary(lm.DBH)
Call:
lm(formula = DBH ~ Species, data = WEF.dat)

Residuals:
    Min      1Q  Median      3Q     Max
-78.423  -9.969  -3.561  10.924 118.277

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   89.423      1.303  68.629   <2e-16 ***
SpeciesGF    -51.598      4.133 -12.483   <2e-16 ***
SpeciesNF     -5.873     15.744  -0.373    0.709
SpeciesSF    -68.347      1.461 -46.784   <2e-16 ***
SpeciesWH    -48.062      1.636 -29.377   <2e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 .

Residual standard error: 22.19 on 1950 degrees of freedom
Multiple R-squared: 0.5332,Adjusted R-squared: 0.5323
F-statistic: 556.9 on 4 and 1950 DF,  p-value: < 2.2e-16
```

We now collect the residuals from the above model anf fit a variogram to the residuals.

```
> DBH.resid <- resid(lm.DBH)
> vario.DBH.resid <- variog(coords = coords, data = DBH.resid,
+                           uvec = (seq(0, max.dist, length = bins)))
> fit.DBH.resid <- variofit(vario.DBH.resid,
+      ini.cov.pars = c(300, 200/-log(0.05)), cov.model = "exponential",
```

Figure 2.16 *Isotropic semivariograms for DBH and residuals of a linear regression of DBH onto tree species.*

```
+      minimisation.function = "nls", weights = "equal")
```

Finally, we plot the two variograms – one for the response and the other for the residuals – along with fitted exponential variograms.

```
> par(mfrow = c(1, 2))
> plot(vario.DBH, ylim = c(200, 1200), main = "DBH")
> lines(fit.DBH)
> abline(h = fit.DBH$nugget, col = "blue")
> abline(h = fit.DBH$cov.pars[1] + fit.DBH$nugget, col = "green")
> abline(v = -log(0.05) * fit.DBH$cov.pars[2], col = "red3")
> plot(vario.DBH.resid, ylim = c(200, 500), main = "DBH residuals")
> lines(fit.DBH.resid)
> abline(h = fit.DBH.resid$nugget, col = "blue")
> abline(h = fit.DBH.resid$cov.pars[1] + fit.DBH.resid$nugget, col = "green")
> abline(v = -log(0.05) * fit.DBH.resid$cov.pars[2], col = "red3")
```

The resulting variograms are offered in Figure 2.16. Here the upper and lower horizontal lines are the *sill* and *nugget*, respectively, and the vertical line is the effective range (i.e., that distance at which the correlation drops to 0.05).

We can also check for possible anisotropic patterns in spatial dependence. This can also be accomplished using functions in `geoR` but let us switch to a different `R` package, `gstat`, for illustrative purposes. The `gstat` package offers a `variogram` function that can be regarded as more versatile than its `geoR` counterpart in that it allows us to specify a linear regression model *within* it and conduct the variogram analysis on the residuals without having to explicitly collect the residuals. This is demonstrated in the code below. The resulting plot appears in Figure 2.17.

```
> ELEV_m <- WEF.dat$ELEV_m
> sp.dat <- as.data.frame(cbind(DBH, HT, coords, ELEV_m))
> coordinates(sp.dat) <- c("East_m", "North_m")
> vario.DBH <- variogram(DBH ~ ELEV_m, data = sp.dat,
+     cutoff = max.dist, width = 5, alpha = (0:3) *
+         45)
> fit.DBH <- fit.variogram(vario.DBH, vgm(1000, "Exp",
+     200/-log(0.05), 600))
> print(plot(vario.DBH, fit.DBH))
```

Figure 2.17 *Directional semivariograms for DBH.*

2.6 Exercises

1. Consider the time series , $Y_t = X\sin(\omega t + \theta)$ (so X is the amplitude, ω is the frequency and θ is the phase) where X is distributed with mean 0 and variance 1 independent of $\theta \sim U(-\pi, \pi)$). Show that Y_t is weakly stationary.
2. Recalling Section 2.4, show that, under a constant mean assumption, i.e., $\mu(\mathbf{s}) = \mu$, the best linear unbiased estimator (BLUE), $\hat{\mu}$ of μ under a Gaussian model is $\frac{\mathbf{1}^T\Sigma^{-1}\mathbf{Y}}{\mathbf{1}^T\Sigma^{-1}\mathbf{1}}$. (In fact, this only requires first and second moments and not normality.) Then, show that $\boldsymbol{\ell}^T\mathbf{Y}$ is $\boldsymbol{\lambda}^T\mathbf{Y}$ with $\hat{\mu}$ replacing μ.
3. For semivariogram models #2, 4, 5, 6, 7, and 8 in Subsection 2.1.3,
   (a) identify the nugget, sill, and range (or effective range) for each;
   (b) find the covariance function $C(t)$ corresponding to each $\gamma(t)$, provided it exists.
4. Prove that for Gaussian processes, strong stationarity is equivalent to weak stationarity.
5.(a) What is the issue with regard to specifying $\widehat{c}(0)$ in the covariance function estimate (2.12)?
   (b) Show either algebraically or numerically that regardless of how $\widehat{c}(0)$ is obtained, $\widehat{\gamma}(t_k) \neq \widehat{c}(0) - \widehat{c}(t_k)$ for all t_k.
6. The scallops data can be read into `R` directly from our website by starting the program and typing

```
> myscallops <- read.table(
+ ‘‘http://www.biostat.umn.edu/~brad/data/myscallops.txt’’,
+     header=T)
```

   Carry out the steps outlined in Section 2.5.1 using `R` packages. In addition:

(a) Provide a descriptive summary of the `myscallops` data with the plots derived from the above session.

(b) Experiment with the `variog` function in `geoR` to obtain rough estimates of the nugget, sill, and range.

(c) Repeat the theoretical variogram fitting with an exponential variogram, and report your results.

7. Consider the `coalash` data frame in the `gstat` package in `R` and available from `http://www.biostat.umn.edu/~brad/data/coal.ash.txt`. This data comes from the Pittsburgh coal seam on the Robena Mine Property in Greene County, PA (Cressie, 1993, p. 32). This data frame contains 208 coal ash core samples (the variable `coal` in the data frame) collected on a grid given by x and y planar coordinates (*not* latitude and longitude).
   Carry out the following tasks in `R`:

(a) Plot the sampled sites embedded on a map of the region. Add contour lines to the plot.

(b) Provide a descriptive summary (histograms, stems, quantiles, means, range, etc.) of the variable `coal` in the data frame.

(c) Plot variograms and correlograms of the response and comment on the need for spatial analysis here.

(d) If you think that there is need for spatial analysis, arrive at your best estimates of the range, nugget, and sill.

8. Confirm expressions (2.15) and (2.16), and subsequently verify the form for $\boldsymbol{\lambda}$ given in equation (2.17).

9. Show that when using (2.15) to predict the value of the surface at one of the existing data locations $\mathbf{s}_i$, the predictor will equal the observed value at that location if and only if $\tau^2 = 0$. (That is, the usual Gaussian process is a spatial interpolator only in the "noiseless prediction" scenario.)

10. Recall that

$$\begin{aligned} \mathbf{Y} &= X\boldsymbol{\beta} + \boldsymbol{\epsilon}, \text{ where } \boldsymbol{\epsilon} \sim N(\mathbf{0}, \Sigma) \ , \\ \text{and } \Sigma &= \sigma^2 H(\phi) + \tau^2 I \ , \text{ where } (H(\phi))_{ij} = \rho(\phi; d_{ij}) \, . \end{aligned}$$

Thus the dispersion matrix of $\widehat{\boldsymbol{\beta}}$ is given as $Var(\widehat{\boldsymbol{\beta}}) = \left(X^T \Sigma^{-1} X\right)^{-1}$. Thus $\widehat{Var}(\widehat{\boldsymbol{\beta}}) = \left(X^T \widehat{\Sigma}^{-1} X\right)^{-1}$ where $\widehat{\Sigma} = \widehat{\sigma}^2 H(\widehat{\phi}) + \widehat{\tau}^2 I$ and $X = [\mathbf{1}, \mathtt{long}, \mathtt{lat}]$. Given the estimates of the sill, range, and nugget (from the `nls` function), it is possible to estimate the covariance matrix $\widehat{\Sigma}$, and thereby get $\widehat{Var}(\widehat{\boldsymbol{\beta}})$. Develop an `R` program to perform this exercise to obtain estimates of standard errors for $\widehat{\boldsymbol{\beta}}$ for the scallops data.

Hint: $\hat{\tau}^2$ is the nugget; $\hat{\sigma}^2$ is the partial sill (the sill minus the nugget). Finally, the correlation matrix $H(\hat{\phi})$ can be obtained from the spherical covariance function, part of your solution to Exercise 3.

Chapter 3

Some theory for point-referenced data models

The intent of this chapter is to provide a brief review of the basic theory of stochastic processes needed for the development of point-referenced spatial data or geostatistical models. We begin with the development of spatial stochastic processes built from independent increment processes, with particular interest in stationary spatial processes. We briefly discuss the connection between covariance functions and spectral measures. We discuss the validity of covariance functions as well as simple constructions of valid covariance functions. We then turn to smoothness of process realizations as driven by stationary covariance functions with a brief discussion of directional derivative processes (anticipating a fuller discussion in Chapter 12). Next, we expand our development of isotropic covariance functions. We conclude with a section on nonstationary covariance specifications.

3.1 Formal modeling theory for spatial processes

When we write the collection of random variables $\{Y(\mathbf{s}) : \mathbf{s} \in D\}$ for some region of interest D or more generally $\{Y(\mathbf{s}) : \mathbf{s} \in \Re^r\}$, it is evident that we are envisioning a stochastic process indexed by $\mathbf{s}$. To capture spatial association it is also evident that these variables will be pairwise dependent with strength of dependence that is specified by their locations.

So, in fact, we have to determine the joint distribution for an uncountable number of random variables. In fact, we do this through specification of arbitrary finite dimensional distributions, i.e., for an arbitrary number of and choice of locations. This characterizes the stochastic process. More precisely, for the set of locations, $\{\mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_n\}$, let the finite dimensional distribution be $P(Y(\mathbf{s}_1), Y(\mathbf{s}_2), \ldots, Y(\mathbf{s}_n) \in A)$ for A in a suitable σ-algebra of sets. In fact, without loss of generality, we can take $A = A_1 \times A_2 \times ...A_n$, a product set, and write $P(Y(\mathbf{s}_1) \in A_1, Y(\mathbf{s}_2) \in A_2, \ldots, Y(\mathbf{s}_n) \in A_n)$.

Consider the following two "consistency" conditions:

1. Under any permutation α of the indices $1, 2, \ldots, n$, say $\alpha_1, \alpha_2, \ldots, \alpha_n$,

$$P(Y(\mathbf{s}_{\alpha_1}) \in A_{\alpha_1}, Y(\mathbf{s}_{\alpha_2}) \in A_{\alpha_2}, \ldots, Y(\mathbf{s}_{\alpha_n}) \in A_n) = \\ P(Y(\mathbf{s}_1) \in A_1, Y(\mathbf{s}_2) \in A_2, \ldots, Y(\mathbf{s}_n) \in A_n) \ .$$

   That is, permutation of the indices does not change the probability of events.

2. For any set of locations, $\{\mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_n\}$ consider an additional arbitrary location $\mathbf{s}_{n+1}$. If we marginalize the $n+1$ dimensional joint distribution specified for $Y(\mathbf{s}_1), Y(\mathbf{s}_2), ..., Y(\mathbf{s}_n), Y(\mathbf{s}_{n+1})$ over $Y(\mathbf{s}_{n+1})$, we obtain the n dimensional joint distribution specified for $Y(\mathbf{s}_1), Y(\mathbf{s}_2), \ldots, Y(\mathbf{s}_n)$.

All stochastic processes on a continuous space satisfy these two conditions. Remarkably, a theorem due to Kolmogorov (see, e.g., Billingsley, 1995) informally states that the collection of finite dimensional probability measures satisfies (1) and (2) if and only if a stochastic process exists on the associated probability space having these finite dimensional distributions. Of course, characterizing the entire collection of finite dimensional distributions can

be challenging. A convenient way to do this is by confining ourselves to Gaussian processes (possibly transformations of) or to mixtures of such processes (a very rich class). That is, in this case, we can work with multivariate normal distributions and all that is required is a mean surface, $\mu(\mathbf{s})$ and a valid correlation function which provides the covariance matrix, as we discuss below.

Again, to clarify the inference setting, in practice we will only observe $Y(\mathbf{s})$ at a finite set of locations, $\mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_n$. Based upon $\{Y(\mathbf{s}_i), i = 1, \ldots, n\}$, we seek to infer about the mean, variability, and association structure of the process. We also seek to predict $Y(\mathbf{s})$ at arbitrary unobserved locations. Since our focus is on hierarchical modeling, often the spatial process is introduced through random effects at the second stage of the modeling specification. In this case, we still have the same inferential questions but now the process is never actually observed. It is latent and the data, modeled at the first stage, helps us to learn about the process.

In this sense, we can make intuitive connections with familiar dynamic models (e.g., West and Harrison, 1997) where there is a latent state space model that is temporally updated. In fact, this reminds us of a critical difference between the one-dimensional time domain and the two-dimensional spatial domain: we have full order in the former, but only partial order in two or more dimensions.

The implications of this remark are substantial. Large sample analysis for time series usually lets time go to ∞. Asymptotics envision an increasing time domain. By contrast, large sample analysis for spatial process data usually envisions a fixed region with more and more points filling in this domain (so-called infill asymptotics). When applying increasing domain asymptotic results, we can assume that, as we collect more and more data, we can learn about temporal association at increasing distances in time. When applying infill asymptotic results for a fixed domain we can learn more and more about association as distance between points tends to 0. However, with a maximum distance fixed by the domain we cannot learn about association (in terms of consistent inference) at increasing distance. The former remark indicates that we may be able to do an increasingly better job with regard to spatial prediction at a given location. However, we need not be doing better in terms of inferring about other features of the process. Learning about process parameters will be bounded; Fisher information does not go to ∞, Cramèr-Rao lower bounds and asymptotic variances do not go to 0. See the work of Stein (1999a, 1999b) as well as Loh (2005) and Zhang and Zimmerman (2005) for a full technical discussion regarding such asymptotic results. Here, we view such concerns as providing encouragement for using a Bayesian framework for inference, since then we need not rely on any asymptotic theory for inference, but rather, under the implemented model, we obtain exact inference given whatever data we have observed. Of course, the fact that information is bounded implies that the data never overwhelms the prior, as is customarily assumed. There is no free lunch and prior sensitivity analysis may be needed.

In the ensuing subsections we turn to some technical discussion regarding specification of spatial stochastic processes as well as covariance and correlation functions. However, we note that the above restriction to Gaussian processes enables several advantages. First, it allows very convenient distribution theory. Joint marginal and conditional distributions are all immediately obtained from standard theory once the mean and covariance structure have been specified. In fact, as above, this is all we need to specify in order to determine all distributions. Also, as we shall see, in the context of hierarchical modeling, a Gaussian process assumption for spatial random effects introduced at the second stage of the model is very natural in the same way that independent random effects with variance components are customarily introduced in linear or generalized linear mixed models. From a technical point of view, as noted in Subsection 2.1.1, if we work with Gaussian processes and stationary models, strong stationarity is equivalent to weak stationarity. We will clarify these notions in the next subsection. Lastly, in most applications, it is difficult to criticize a Gaussian

assumption. To argue this as simply as possible, in the absence of replication we have $\mathbf{Y} = (Y(\mathbf{s}_1), \ldots, Y(\mathbf{s}_n))^T$, a single realization from an n-dimensional distribution. With a sample size of one, how can we criticize *any* multivariate distributional specification (Gaussian or otherwise)?

Strictly speaking this last assertion is not quite true with a Gaussian process model. That is, the joint distribution is a multivariate normal with mean, say, $\mathbf{0}$, and a covariance matrix that is a parametric function of the parameters in the covariance function. As n grows large enough, the effective sample size will also grow. By linear transformation we can obtain a set of approximately uncorrelated variables through which the adequacy of the normal assumption might be studied. We omit details.

3.1.1 Some basic stochastic process theory for spatial processes

A key remark here is that, when we develop a spatial stochastic process model, we can proceed along two paths. First, we can specify the process stochastically and obtain its induced covariance function, i.e., its induced dependence structure. A second path is to start with, say, a Gaussian process and then specify a valid covariance function to overlay on the Gaussian process. This covariance function supplies the joint covariance matrix for any finite number of variables at any selected set of sites. In this subsection we follow the first path, in order to illuminate some of the rigor associated with formal stochastic process specification. In the subsequent subsections we define valid covariance functions and focus on their properties and examples.

To begin our development here, we start with independent increment processes. It will take a few paragraphs before we are able to explicitly connect them to spatial processes. A real-valued independent increment process, Z over, say, R^d (here, we are only interested in the case $d = 2$) is such that, for arbitrary disjoint sets A and B, where A and B belong to a suitable σ-algebra of sets, $Z(A)$ and $Z(B)$ are independent. In particular, we let $Z(d\mathbf{w})$ be the generator for these random variables in the sense that $Z(A) = \int_A Z(d\mathbf{w})$. Evidently, if A and B are disjoint, then $Z(A \bigcup B) = Z(A) + Z(B)$. We assume that the Z process has first and second moments and, for convenience that the first moment is 0. We set $E(Z^2(A)) = \text{var}(A) = G(A)$, i.e., $G(A) = \int_A G(dw)$. Also, we see that $E(Z(A)Z(B)) = E(Z(A \bigcap B) + Z(A \bigcap B^C))(Z(A \bigcap B)Z(A^C \bigcap B)) = E(Z^2(A \bigcap B)) = G(A \bigcap B)$ due to independence of increments.

Let $Z_f = \int f(\mathbf{w})Z(d\mathbf{w})$ for an appropriately measurable function f. Of course, $Z(A) = \int 1(\mathbf{w} \in A)Z(d\mathbf{w})$ is a special case. This suggests that we proceed to step functions as a way to study the behavior of Z_f for measurable f and the dependence between, say, Z_{f1} and Z_{f2}. If $f(\mathbf{w}) = \sum_l a_l 1(\mathbf{w} \in A_l)$, then $Z_f = \sum_l a_l Z(A_l)$. Because of the independent increments, $\text{var} Z_f = \sum a_l^2 G(A_l)$.

Next, consider $f_1(\mathbf{w}) = \sum_l a_l 1(\mathbf{w} \in A_l)$ and $f_2(\mathbf{w}) = \sum_k b_k 1(\mathbf{w} \in B_k)$. Then,

$$\begin{aligned}
\text{Cov}(Z_{f1}, Z_{f2}) &= E(Z_{f1}Z_{f2}) = \sum_l \sum_k a_l b_k E(Z(A_l)Z(B_k) \\
&= \sum_l \sum_k a_l b_k E(Z(A_l \bigcap B_k)) = \sum_l \sum_k a_l b_k G(A_l \bigcap B_k) \ .
\end{aligned}$$

But, also, $\int f_1(\mathbf{w})f_2(\mathbf{w})G(d\mathbf{w}) = \sum_l \sum_k a_l b_k G(A_l \bigcap B_k)$. So, we have shown that $E(Z_{f1}Z_{f2}) = \int f_1(\mathbf{w})f_2(\mathbf{w})G(d\mathbf{w})$.

Now, let's bring in the spatial process setting. Define

$$Y(\mathbf{s}) = \int \psi(\mathbf{s}, \mathbf{w})Z(d\mathbf{w}) \ ,$$

i.e., $Y(\mathbf{s})$ is $Z_{\psi_{\mathbf{s}}}$ in our notation above. Here, we allow ψ to be a complex valued function since we want to employ the particular form $\psi(\mathbf{s},\mathbf{w}) = e^{i\mathbf{s}^T\mathbf{w}}$. We also allow Z to be complex valued, introducing the complex conjugate using an overline. So, now $G(A) = E(Z(A)\overline{Z}(A)) = E|Z(A)|^2$. Then, we have defined a stochastic process and to calculate $\text{cov}Y(\mathbf{s}), Y(\mathbf{s}')$, we compute $E(Y(\mathbf{s})\overline{Y}(\mathbf{s}')) = \int \psi(\mathbf{s},\mathbf{w})\overline{\psi}(\mathbf{s}',\mathbf{w})G(d\mathbf{w})$, using the result of the previous paragraph.

With $\psi(\mathbf{s},\mathbf{w}) = e^{i\mathbf{s}^T\mathbf{w}}$, we obtain

$$\text{cov}(Y(\mathbf{s}), Y(\mathbf{s}')) = \int e^{i\mathbf{s}^T\mathbf{w}} e^{-i\mathbf{s}'^T\mathbf{w}} G(d\mathbf{w}) = \int e^{i(\mathbf{s}-\mathbf{s}')^T\mathbf{w}} G(d\mathbf{w}) \ .$$

In other words, the association between $Y(\mathbf{s})$ and $Y(\mathbf{s}')$ depends only upon the separation vector $\mathbf{h} = \mathbf{s} - \mathbf{s}'$, not on the individual locations; $\text{cov}(Y(\mathbf{s}), Y(\mathbf{s}')) = C(\mathbf{s} - \mathbf{s}')$. Such a stochastic process is said to be *stationary*. In fact, we have another characterization, an elegant result which says that $Y(\mathbf{s})$ is a stationary stochastic process if and only if it can be represented in the form $Y(\mathbf{s}) = \int e^{i\mathbf{s}^T\mathbf{w}} Z(d\mathbf{w})$ where Z is a possibly complex-valued, mean 0, independent increments process (Yaglom, 1987, Section 8). We note the implicit parallel structure, $Y(\mathbf{s}) = \int e^{i\mathbf{s}^T\mathbf{w}} Z(d\mathbf{w})$ and $C(\mathbf{s}) = \int e^{i\mathbf{s}^T\mathbf{w}} G(d\mathbf{w})$.

Other choices for ψ appear in the literature. For instance, ψ might be a kernel function $K(\mathbf{s} - \mathbf{w})$ which is integrable over $\Re^2$. Then, we have $Y(\mathbf{s}) = \int K(\mathbf{s} - \mathbf{w})Z(d\mathbf{w})$ and $\text{cov}(Y(\mathbf{s}), Y(\mathbf{s}')) = \int K(\mathbf{s}-\mathbf{w})K(\mathbf{s}'-\mathbf{w})G(d\mathbf{w})$. In the special case where $G(d\mathbf{w}) = \sigma^2 d\mathbf{w}$, we see that, after a simple change of variable, $\text{cov}(Y(\mathbf{s}), Y(\mathbf{s}')) = \sigma^2 \int K(\mathbf{s}-\mathbf{s}'+\mathbf{u})K(\mathbf{u})d\mathbf{u}$. Such a process construction is called *kernel convolution* and is discussed further in Section 3.2.2. A question we might ask is, "How rich is the class of stationary process obtainable under kernel convolution?" We can illuminate this in the next subsection, after we characterize valid covariance functions through Bochner's Theorem.

We might also ask when does the foregoing construction provide a Gaussian process? In particular, if $Z(\mathbf{s})$ is Brownian motion then $Y(\mathbf{s})$ is a Gaussian process. That is, Brownian motion is an independent increments process providing jointly normally distributed random variables. Informally, integration of such variables against a choice of ψ is like taking linear transformations of jointly normal variables; the resulting variable is normal, a resulting set of variables is jointly normal. Let us add a few words about Brownian motion.

Brownian motion in one dimension is, again, an independent increments process, usually defined on $\Re^+$. In fact, for $t > 0$, we let $Z(t) \equiv Z((0,t])$, i.e., we convert a set function to a point function in order to define $\{Z(t) : t \in \Re^+\}$. We assume $Z(t) \sim N(0, \sigma^2 t)$ and $\text{cov}(Z(t), Z(t') = \sigma^2 \min(t,t') = \frac{1}{2}\sigma^2(|t| + |t'| - |t - t'|)$. In other words, the spectral measure for Brownian motion in one dimension is $G(dt) = \sigma^2|dt|$. We can also calculate that $E(Z(t+h) - Z(t))^2 = \sigma^2|h|$. From this we can infer that process realizations are mean square continuous, i.e., $\lim_{h\to 0} E(Z(t+h) - Z(t))^2 = 0$. However, process realizations are not mean square differentiable. (See Section 13.2 for more discussion on mean square smoothness.)

Moving to two dimensions, we would like to define $Z(\mathbf{s})$. Paralleling the one-dimensional case, we want $Z(\mathbf{s}) \sim N(0, \sigma^2||\mathbf{s}||)$. The easiest way to envision this is through circles. That is, $Z(\mathbf{s}) \sim Z(\mathbf{s}')$ if $||\mathbf{s}|| = ||\mathbf{s}'||$; marginal univariate distributions are common on circles. Again, imitating the one-dimensional case we want $\text{cov}(Z(\mathbf{s}), Z(\mathbf{s}') = \frac{1}{2}\sigma^2(||\mathbf{s}|| + ||\mathbf{s}'|| - ||\mathbf{s} - \mathbf{s}'||)$. We see that, even if $||\mathbf{s}'|| = ||\mathbf{s}''||$, $\text{cov}(Z(\mathbf{s}), Z(\mathbf{s}') \neq (Z(\mathbf{s}), Z(\mathbf{s}''))$. Again, $E(Z(\mathbf{s}+\mathbf{h}) - Z(\mathbf{s}))^2 = \sigma^2||\mathbf{h}||$ implying mean square continuity of process realizations. Moreover, we now see that this is a linear variogram (Section 2.1.3). That is, we have a stationary (in fact, isotropic) variogram associated with a nonstationary covariance function, illuminating our comment in that subsection regarding the fact that the linear variogram does not correspond to a weakly stationary process model.

Next, we clarify the connection between Brownian motion and white noise. Recall that white noise is usually defined as the process $V(\mathbf{s})$ where the $V(\mathbf{s})$ are i.i.d. $N(0, \sigma^2)$. Consider the finite differential $\dfrac{Z(\mathbf{s}+\mathbf{h}) - Z(\mathbf{s})}{||\mathbf{h}||}$. From the above, this variable is distributed as $N(0, \sigma^2/||\mathbf{h}||)$. As $||\mathbf{h}|| \to 0$, we do not obtain a limiting distribution. So, white noise is not the *derivative* of Brownian motion (though it is sometimes referred to as a *generalized* derivative). In fact, while $Y(\mathbf{s}) = \int \phi(\mathbf{s}, \mathbf{w}) Z(d\mathbf{w})$ defines a stochastic process, $Y(\mathbf{s}) = \int \phi(\mathbf{s}, \mathbf{w}) V(\mathbf{w}) d(\mathbf{w})$ does not. Rather, we must work with finite dimensional versions of this expression, $Y(\mathbf{s}) = \sum_{j=1}^{J} \phi(\mathbf{s}, \mathbf{s}_j) V(\mathbf{s}_j)$ to create a well defined process. And, in no sense, should this finite version be viewed as an approximation to integration over $\Re^2$. The special case where $\phi(\mathbf{s}, \mathbf{w})$ is a kernel $K(\mathbf{s} - \mathbf{w})$ is discussed above and a stationary covariance function results.

Finally, we mention the notion of fractional Brownian motion. Here, we define $\text{cov}(Z(\mathbf{s}), Z(\mathbf{s}') = \frac{1}{2}\sigma^2(||\mathbf{s}||^{2H} + ||\mathbf{s}'||^{2H} - ||\mathbf{s} - \mathbf{s}'||^{2H})$. $H \in (0, 1)$ is called the Hurst index. Fractional Brownian motion generalizes the foregoing specifications and allows for process with dependent increments. In particular, $H = 1/2$ is usual Brownian motion. If H is greater than $1/2$ we have positively correlated increments while if it is less than $1/2$ we have negatively correlated increments. We leave this as an exercise.

3.1.2 Covariance functions and spectra

In order to specify a stationary process we must provide a valid covariance function. Here "valid" means that $C(\mathbf{h}) \equiv cov(Y(\mathbf{s}), Y(\mathbf{s} + \mathbf{h}))$ is such that for any finite set of sites $\mathbf{s}_1, \ldots, \mathbf{s}_n$ and for any $a_1, \ldots, a_n$,

$$Var\left[\sum_i a_i Y(s_i)\right] = \sum_{i,j} a_i a_j Cov(Y(\mathbf{s}_i), Y(\mathbf{s}_j)) = \sum_{i,j} a_i a_j C(\mathbf{s}_i - \mathbf{s}_j) \geq 0 \ ,$$

with strict inequality if not all the a_i are 0. That is, we need $C(\mathbf{h})$ to be a positive definite function.

Verifying the positive definiteness condition is evidently not routine. Fortunately, we have *Bochner's Theorem* (see, e.g., Gikhman and Skorokhod, 1974, p. 208), which provides a necessary and sufficient condition for $C(\mathbf{h})$ to be positive definite. This theorem is applicable for $\mathbf{h}$ in arbitrary r-dimensional Euclidean space, although our primary interest is in $r = 2$.

In general, for real-valued processes, Bochner's Theorem states that $C(\mathbf{h})$ is positive definite if and only if

$$C(\mathbf{h}) = \int \cos(\mathbf{w}^T \mathbf{h})\, G(d\mathbf{w}) \ , \tag{3.1}$$

where G is a bounded, positive, symmetric about 0 measure in $\Re^r$. Then $C(\mathbf{0}) = \int G d(\mathbf{w})$ becomes a normalizing constant, and $G(d\mathbf{w})/C(\mathbf{0})$ is referred to as the *spectral distribution* that induces $C(\mathbf{h})$. If $G(d\mathbf{w})$ has a density with respect to Lebesgue measure, i.e., $G(d\mathbf{w}) = g(\mathbf{w})d\mathbf{w}$, then $g(\mathbf{w})/C(\mathbf{0})$ is referred to as the *spectral density.* Evidently, (3.1) can be used to generate valid covariance functions; see Section 3.1.2.1 below. Of course, the behavioral implications associated with C arising from a given G will only be clear in special cases, and (3.1) will be integrable in closed form only in cases that are even more special.

Since $e^{i\mathbf{w}^T\mathbf{h}} = \cos(\mathbf{w}^T\mathbf{h}) + i\sin(\mathbf{w}^T\mathbf{h})$, we have $C(\mathbf{h}) = \int e^{i\mathbf{w}^T\mathbf{h}} G(d\mathbf{w})$. That is, the imaginary term disappears due to the symmetry of G around 0. In other words, $C(\mathbf{h})$ is a valid covariance function if and only if it is the characteristic function of an r-dimensional symmetric random variable (random variable with a symmetric distribution). We note that if G is not assumed to be symmetric about $\mathbf{0}$, $C(\mathbf{h}) = \int e^{i\mathbf{w}^T\mathbf{h}} G(d\mathbf{w})$ still provides a valid covariance function (i.e., positive definite) but now for a complex-valued random process on $\Re^r$.

The Fourier transform of $C(\mathbf{h})$ is

$$\widehat{c}(\mathbf{w}) = \int e^{-i\mathbf{w}^T\mathbf{h}}\, C(\mathbf{h})d\mathbf{h}\ . \tag{3.2}$$

Applying the inversion formula, $C(\mathbf{h}) = (2\pi)^{-r}\int e^{i\mathbf{w}^T\mathbf{h}}\widehat{c}(\mathbf{w})d\mathbf{w}$, we see that $(2\pi)^{-r}\widehat{c}(\mathbf{w})/C(0) = g(\mathbf{w})$, the spectral density. Explicit computation of (3.2) is usually not possible except in special cases. However, approximate calculation is available through the fast Fourier transform (FFT); see Appendix A, Section A.1. Expression (3.2) can be used to check whether a given $C(\mathbf{h})$ is valid: we simply compute $\widehat{c}(\mathbf{w})$ and check whether it is positive and integrable (so it is indeed a density up to normalization).

The one-to-one relationship between $C(\mathbf{h})$ and $g(\mathbf{w})$ enables examination of spatial processes in the spectral domain rather than in the observational domain. Computation of $g(\mathbf{w})$ can often be expedited through fast Fourier transforms; g can be estimated using the so-called *periodogram*. Likelihoods can be obtained approximately in the spectral domain enabling inference to be carried out in this domain. See, e.g., Guyon (1995) or Stein (1999a) for a full development. Likelihood evaluation is much faster in the spectral domain. However, in this book we confine ourselves to the observational domain because of concerns regarding the accuracy associated with approximation in the spectral domain (e.g., the likelihood of Whittle, 1954), and with the ad hoc creation of the periodogram (e.g., how many low frequencies are ignored). We do however note that the spectral domain may afford the best potential for handling computation associated with large data sets.

Isotropic covariance functions, i.e., $C(\|\mathbf{h}\|)$, where $\|\mathbf{h}\|$ denotes the length of $\mathbf{h}$, are the most frequently adopted choice within the stationary class. There are various direct methods for checking the permissibility of isotropic covariance and variogram specifications. See, e.g., Armstrong and Diamond (1984), Christakos (1984), and McBratney and Webster (1986). Again denoting $||\mathbf{h}||$ by t for notational simplicity, recall that Tables 2.1 and 2.2 provide the covariance function $C(t)$ and variogram $\gamma(t)$, respectively, for the widely encountered parametric istropic choices that were initially presented in Section 2.1.3.

It is noteworthy that an isotropic covariance function that is valid in dimension r need not be valid in dimension $r+1$. This intuition may be gleaned by considering $r = 1$ versus $r = 2$. For three points, in one-dimensional space, given the distances separating points 1 and 2 (d_{12}) and points 2 and 3 (d_{23}), then the distance separating points 1 and 3 d_{13} is either $d_{12}+d_{23}$ or $|d_{12}-d_{23}|$. But in two-dimensional space, given d_{12} and d_{23}, d_{13} can take any value in $\Re^+$ (subject to triangle inequality). With increasing dimension more sets of interlocation distances are possible for a given number of locations; it will be more difficult for a function to satisfy the positive definiteness condition. Armstrong and Jabin (1981) provide an explicit example that we defer to Exercise 7.

There are isotropic correlation functions that are valid in all dimensions. The Gaussian correlation function, $k\rho(\|h\|) = \exp(-\phi\,\|h\|^2)$ is an example. It is the characteristic function associated with r i.i.d. normal random variables, each with variance $1/(2\phi)$ for any r. More generally, the powered exponential, $\exp(-\phi\,\|h\|^\alpha)$, $0 < \alpha \leq 2$ (and hence the exponential correlation function) is valid for any r. Here, we note the general result that $C(||\mathbf{h}||)$ is a positive definite isotropic function on $\Re^r$ for all r if and only if it has the representation, $C(||\mathbf{h}||) = \int e^{-w||\mathbf{h}||^2}G(dw)$ where G is nondecreasing and bounded and $w \in R^+$. So, $C(||\mathbf{h}||)$ arises as a scale mixture of Gaussian correlation functions. G might be a c.d.f. on R^+ with a p.d.f., $g(w)$, i.e., $G(dw) = g(w)dw$.

Rather than seeking isotropic correlation functions that are valid in all dimensions, we might seek all valid isotropic correlation function in a particular dimension r. Matérn (1960, 1986) provides the general result. The set of $C(\|h\|)$ of the form

$$C(\|h\|) = \int_0^\infty \left(\frac{2}{w\,\|\mathbf{h}\|}\right)^\alpha \Gamma(\nu+1)J_\nu(w\,\|\mathbf{h}\|)G(dw)\ , \tag{3.3}$$

Figure 3.1 *A plot of* $J_0(d)$ *out to* $d = 100$.

where G is nondecreasing and integrable on $\Re^+$, J_ν is the Bessel function of the first kind of order ν, and $\nu = (r-2)/2$ provides all valid isotropic correlation functions on $\Re^r$.

When $r = 2$, $v = 0$ so that arbitrary correlation functions in two-dimensional space arise as scale mixtures of Bessel functions of order 0. In particular, $J_0(d) = \sum_{k=0}^{\infty} \frac{(-1)^k}{(k!)^2} \left(\frac{d}{2}\right)^{k/2}$. J_0 decreases from 1 at $d = 0$ and will oscillate above and below 0 with amplitudes and frequencies that are diminishing as d increases (see Figure 3.1). Typically, correlation functions that are monotonic and decreasing to 0 are chosen but, apparently, valid correlation functions can permit negative associations with w determining the scale in distance space. Such behavior might be appropriate in certain applications.

The form in (3.3) at $\nu = 0$ was exploited in Shapiro and Botha (1981) and Ver Hoef and Barry (1998) to develop "nonparametric" variogram models and "black box" kriging. It was employed in Ecker and Gelfand (1997) to obtain flexible spatial process models within which to do inference from a Bayesian perspective (see Section 6.1.3).

If we confine ourselves to strictly monotonic isotropic covariance functions, then we can introduce the notion of a range. As described above, the range is conceptualized as the distance beyond which association becomes negligible. If the covariance function reaches 0 in a finite distance, then we refer to this distance as the range. However, as Table 2.1 reveals, we customarily work with covariance functions that attain 0 asymptotically as $\|\mathbf{h}\| \to \infty$. In this case, it is common to define the range as the distance beyond which correlation is less than .05, and this is the definition we employ in the sequel. So if ρ is the correlation function, then writing the range as R we solve $\rho(R; \boldsymbol{\theta}) = .05$, where $\boldsymbol{\theta}$ denotes the parameters in the correlation function. Therefore, R is an implicit function of the parameter $\boldsymbol{\theta}$.

We do note that some authors define the range through the variogram, i.e., the distance at which the variogram reaches .95 of its sill. That is, we would solve $\gamma(R) = .95(\sigma^2 + \tau^2)$. Note, however, that if we rewrite this equation in terms of the correlation function we obtain

$\tau^2 + \sigma^2(1 - \rho(R; \boldsymbol{\theta})) = .95(\tau^2 + \sigma^2)$, so that $\rho(R; \boldsymbol{\theta}) = .05\left(\frac{\sigma^2+\tau^2}{\sigma^2}\right)$. Evidently, the solution to this equation is quite different from the solution to the above equation. In fact, this latter equation may not be solvable, e.g., if $\sigma^2/(\sigma^2 + \tau^2) \leq .05$, the case of very weak "spatial story" in the model. As such, one might argue that a spatial model is inappropriate in this case. However, with σ^2 and τ^2 unknown, it seems safer to work with the former definition.

3.1.2.1 More general isotropic correlation functions

From Subsection 3.1.2, a correlation function $\rho(d, \phi)$ is valid only if it is positive definite in d, $\rho(0, \phi) = 1$, and $|\rho(d, \phi)| \leq 1$ for all d. From Bochner's Theorem (3.1), the characteristic function of a symmetric distribution in R^r satisfies these constraints. From Khinchin's Theorem above (e.g., Yaglom, 1962, p. 106) as well as (3.3), the class of all valid functions $\rho(d, \phi)$ in $\Re^r$ can be expressed as

$$\rho(d, \phi) = \int_0^\infty \Omega_r(zd) dG_\phi(z)\,, \tag{3.4}$$

where G_ϕ is nondecreasing integrable and $\Omega_r(x) = \left(\frac{2}{x}\right)^{\frac{r-2}{2}} \Gamma\left(\frac{r}{2}\right) J_{\left(\frac{r-2}{2}\right)}(x)$. Here again, $J_\upsilon(\cdot)$ is the Bessel function of the first kind of order υ. For $r = 1, \Omega_1(x) = \cos(x)$; for $r = 2, \Omega_2(x) = J_0(x)$; for $r = 3, \Omega_3(x) = \sin(x)/x$; for $r = 4, \Omega_4(x) = \frac{2}{x} J_1(x)$; and for $r = \infty, \Omega_\infty(x) = \exp(-x^2)$. Specifically, $J_0(x) = \sum_{k=0}^\infty \frac{(-1)^k}{k!^2}\left(\frac{x}{2}\right)^{2k}$ and $\rho(d, \phi) = \int_0^\infty J_0(zd) dG_\phi(z)$ provides the class of all permissible correlation functions in $\Re^2$. Figure 3.1 provides a plot of $J_0(x)$ versus x, revealing that it is not monotonic. (This must be the case in order for $\rho(d, \phi)$ above to capture all correlation functions in $\Re^2$.) These more general isotropic covariance functions are revisited in much greater detail in Section 6.1.3.

3.1.3 Constructing valid covariance functions

We note that one can offer constructive strategies to build larger classes of correlation functions. Three approaches are mixing, products, and convolution. Mixing notes simply that if $C_1, \ldots, C_m$ are valid correlation functions in $\Re^r$ and if $\sum_{i=1}^m p_i = 1$, $p_i > 0$, then $C(\mathbf{h}) = \sum_{i=1}^m p_i C_i(\mathbf{h})$ is also a valid correlation function in $\Re^r$. This follows since $C(\mathbf{h})$ is the characteristic function associated with $\sum p_i f_i(\mathbf{x})$, where $f_i(\mathbf{x})$ is the symmetric about 0 density in r-dimensional space associated with $C_i(\mathbf{h})$. In fact, the sum $\sum_{i=1}^\infty a_i C_i(\mathbf{h})$ yields a valid covariance function as well, provided the a_i are all greater than 0 and $\sum_{i=1}^\infty a_i < \infty$.

Using product forms simply notes that again if $C_1, \ldots, C_m$ are valid in $\Re^r$, then $\prod_{i=1}^m C_i$ is a valid correlation function in $\Re^r$. This follows since $\prod_{i=1}^m C_i(\mathbf{h})$ is the characteristic function associated with $V = \sum_{i=1}^m V_i$ where the V_i are independent with V_i having characteristic function $C_i(\mathbf{h})$.

The use of products has attracted attention recently in the context of so-called *covariance tapering* (see Kaufman et al., 2008, and references therein) The idea here is that, with covariance matrices which do not reach 0 until $||\mathbf{h}||$ reaches ∞, the resulting covariance matrices are never sparse. With a large number of locations, handling a large covariance matrix in terms of, say, inversion and determinant calculation can be very challenging (see Chapter 12). Introducing sparsity into this matrix can facilitates this computation. A naive thought might be to set to 0 all entries in the matrix that are smaller than some specified value, arguing that these are negligible. Unfortunately, this strategy can result in a covariance matrix which is no longer positive definite. As an alternative, suppose $C(\mathbf{h})$ is the covariance matrix you start with and suppose $\tilde{C}(\mathbf{h})$ is a covariance function with bounded support, i.e., it reaches 0 for all $\mathbf{h}$ with length $||\mathbf{h}|| < d_0$ for some $d_0 < \infty$. Then $C^*(\mathbf{h}) = \tilde{C}(\mathbf{h})C(\mathbf{h})$ is valid and produces sparsity at distance greater than d_0. Of course, one might ask, why not just use $\tilde{C}$ from the start?

We note that this connects to the issue of nested variograms or covariance functions which have been suggested in the literature(see, e.g., Hohn, 1988, and, more recently, Wackernagel, 2003). The idea here would be to use different such functions in different portions of the region of interest. The problem is that, in doing this, the aggregated covariance matrix need not be positive definite; we have a model specification that cannot possibly produce the data we are observing.

Convolution simply recognizes that if C_1 and C_2 are valid correlation functions in $\Re^r$, then $C_{12}(\mathbf{h}) = \int C_1(\mathbf{h}-\mathbf{t})C_2(\mathbf{t})d\mathbf{t}$ is a valid correlation function in $\Re^r$. The argument here is to look at the Fourier transform of $C_{12}(\mathbf{h})$. That is,

$$\begin{aligned}\widehat{c}_{12}(\mathbf{w}) &= \int e^{-i\mathbf{w}^T\mathbf{h}}C_{12}(\mathbf{h})d\mathbf{h} \\ &= \int e^{-i\mathbf{w}^T\mathbf{h}}\int C_1(\mathbf{h}-\mathbf{t})C_2(\mathbf{t})d\mathbf{t}d\mathbf{h} \\ &= \widehat{c}_1(\mathbf{w})\cdot\widehat{c}_2(\mathbf{w})\,,\end{aligned}$$

where $\widehat{c}_i(\mathbf{w})$ is the Fourier transform of $C_i(\mathbf{h})$ for $i = 1, 2$. But then $C_{12}(\mathbf{h}) = (2\pi)^{-2}\int e^{i\mathbf{w}^T\mathbf{h}}\widehat{c}_1(\mathbf{w})\widehat{c}_2(\mathbf{w})d\mathbf{w}$. Now $\widehat{c}_1(\mathbf{w})$ and $\widehat{c}_2(\mathbf{w})$ are both symmetric about $\mathbf{0}$ since, up to a constant, they are the spectral densities associated with $C_1(\mathbf{h})$ and $C_2(\mathbf{h})$, respectively. Hence, $C_{12}(\mathbf{h}) = \int \cos \mathbf{w}^T\mathbf{h}G(d\mathbf{w})$ where $G(d\mathbf{w}) = (2\pi)^{-2}\widehat{c}_1(\mathbf{w})\widehat{c}_2(\mathbf{w})d\mathbf{w}$.

Thus, from (3.1), $C_{12}(\mathbf{h})$ is a valid correlation function, i.e., G is a bounded, positive, symmetric about 0 measure on $\Re^2$. In fact, if C_1 and C_2 are isotropic then C_{12} is as well; we leave this verification as Exercise 9.

3.1.4 Smoothness of process realizations

How does one select among the various choices of correlation functions? Usual model selection criteria will typically find it difficult to distinguish, say, among one-parameter isotropic scale choices such as the exponential, Gaussian, or Cauchy. Ecker and Gelfand (1997) provide some graphical illustration showing that, through suitable alignment of parameters, the correlation curves will be very close to each other. Of course, in comparing choices with parametrizations of differing dimensions (e.g., correlation functions developed using results from the previous section), we will need to employ a selection criterion that penalizes complexity and rewards parsimony (see Section 5.2.3).

An alternative perspective is to make the selection based upon theoretical considerations. This possibility arises from the powerful fact that the choice of correlation function determines the smoothness of realizations from the spatial process. More precisely, a process realization is viewed as a random surface over the region. By choice of C we can ensure that these realizations will be almost surely continuous, or mean square continuous, or mean square differentiable, and so on. Of course, at best the process is only observed at finitely many locations. (At worst, it is never observed, e.g., when the spatial process is used as a second stage model for random spatial effects.) So, it is not possible to "see" the smoothness of the process realization. Elegant theory, developed in Kent (1989) and Stein (1999a) and extended in Banerjee and Gelfand (2003), clarifies the relationship between the choice of correlation function and such smoothness. We provide a bit of this theory below, with further discussion in Section 13.2. For now, the key point is that, according to the process being modeled, we may, for instance, anticipate surfaces to not be continuous (as with digital elevation models in the presence of gorges, escarpments, or other topographic features), or to be differentiable (as in studying land value gradients or temperature gradients). We can choose a correlation function to essentially ensure such behavior.

Of particular interest in this regard is the Matérn class of covariance functions. The parameter v (see Table 2.1) is, in fact, a smoothness parameter. In two-dimensional space, the greatest integer in v indicates the number of times process realizations will be mean square differentiable. Indeed, since $v = \infty$ corresponds to the Gaussian correlation function, the

implication is that use of the Gaussian correlation function results in process realizations that are mean square analytic, which may be too smooth to be appropriate in practice. That is, it is possible to predict $Y(\mathbf{s})$ perfectly for all $\mathbf{s} \in \Re^2$ based upon observing $Y(\mathbf{s})$ in an arbitrarily small neighborhood. Expressed in a different way, use of the Matérn covariance function as a model enables the data to inform about v; we can learn about process smoothness despite observing the process at only a finite number of locations.

Hence, we follow Stein (1999a) in recommending the Matérn class as a general specification for building spatial models. The computation of this function requires evaluation of a modified Bessel function. In fact, evaluation will be done repeatedly to obtain a covariance matrix associated with n locations, and then iteratively if a model is to fit via MCMC methods. This may appear off-putting but, in fact, it is routinely available in all of the software packages we discuss in this book. In fact, such computation can be done efficiently using expansions to approximate $K_v(\cdot)$ (Abramowitz and Stegun, 1965, p. 435), or working through the inversion formula below (3.2), which in this case becomes

$$2\left(\frac{\phi\,\|\mathbf{h}\|}{2}\right)^{\nu}\frac{K_v(\phi(\|\mathbf{h}\|))}{\phi^{2\nu}\Gamma(v+\frac{r}{2})} = \int_{\Re^r} e^{i\mathbf{w}^T\mathbf{h}}(\phi^2+\|\mathbf{w}\|^2)^{-(v+r/2)}d\mathbf{w}\ , \tag{3.5}$$

where K_v is the modified Bessel function of order $\nu > 0$. We see that the Matérn covariance function arises as the characteristic function from a Cauchy spectral density. In fact, this is how Matérn came upon this covariance function when doing his thesis research (Matérn, 1960).

Computation of (3.5) is discussed further in Appendix Section A.1. In particular, the right side of (3.5) is readily approximated using fast Fourier transforms. Again, we revisit process smoothness in Section 13.2.

We conclude this subsection by returning to the question of how rich is the class of stationary processes obtained using kernel mixing. From Section 3.2, we have the induced covariance function to be $C(\mathbf{s}) = \sigma^2 \int_{\Re^2} K(\mathbf{s}-\mathbf{w})K(\mathbf{w})d\mathbf{w}$. Using (3.2), we have the Fourier transform

$$\hat{c}(\mathbf{w}) = \int_{\Re^2}\int_{\Re^2} e^{i\mathbf{w}^T\mathbf{s}}K(\mathbf{s}-\mathbf{w})K(\mathbf{w})d\mathbf{w}d\mathbf{s} = \overline{\hat{K}}(\mathbf{w})\hat{K}(\mathbf{w}) = |\hat{K}(\mathbf{w})|^2\ .$$

In other words, K induces C, $K \Leftrightarrow \hat{K}$ and $C \Leftrightarrow \hat{c}$ by the one-to-one relationship between distributions and characteristic functions. So, if we start with C, is there a $\hat{K}$ that yields $\hat{c}$? Can $\hat{c}(\mathbf{w})$ be written as $|\hat{K}(\mathbf{w})|^2$, i.e. does $\hat{c}$ admit a "square root"? We have the elegant result (Yaglom, 1987) which says that a stationary random process can be defined by kernel mixing if and only if it has a spectral density.

Also, immediately, we can create an example of a stationary random process that does not arise from kernel mixing as follows (R. Wolpert, personal communication). In $\Re^1$, let V_1 and V_2 be independent $N(0,1)$ variables and set $Y(t) = Z_1\cos(t) + Z_2\sin(t)$. Then, it is easy to see that $E(Y(t)) = 0$ and $\mathrm{cov}(Y(t), Y(t')) = \cos(t-t')$. So, $Y(t)$ is, in fact, a stationary Gaussian process. But, directly, $C(h) = \int_{-\infty}^{\infty} e^{ihw}\frac{1}{2}[\delta_{-1}(w)+\delta_1(w)]dw$ where δ is the usual delta function, $\delta_c(x) = 1$ if $x = c$, $= 0$ otherwise. So, $\hat{c}(w) = \frac{1}{2}[\delta_{-1}(w)+\delta_1(w)]$ which does not admit a square root and, therefore, there is no kernel representation.

Finally, we can consider the foregoing in the context of the Matérn class of covariance functions above. The key point here is that, in (3.5), the Matérn covariance functions are one-to-one with Cauchy-type spectral densities. In particular, the smoothness parameter, $v > 0$, appears in $\hat{c}(\mathbf{w})$ in the exponent as $-(v+\frac{r}{2})$. Hence, taking a square root yields the power $-(\frac{v}{2}+\frac{r}{4}) = -(v'+\frac{r}{2})$ with $v' = \frac{v}{2}-\frac{r}{4}$. So, in order that $v' > 0$, we need $v > \frac{r}{2}$ to have a spectral density. In two dimensions, this implies that $v > 1$, i.e., only Matérn covariance functions that produce at least mean square differentiable realizations arise from kernel convolution. Since $k = \frac{1}{2}$ for the exponential, we cannot create the exponential covariance

from kernel mixing. Also noteworthy here is the implication that one should not employ Gaussian kernels in using kernel convolution since they produce Gaussian covariance functions, again, processes realizations that are too smooth. See Paciorek and Schervish (2006) and Section 3.2.2 for further discussion in this regard.

3.1.5 Directional derivative processes

The previous section offered discussion intended to clarify, for a spatial process, the connection between the correlation function and the smoothness of process realizations. When realizations are mean square differentiable, we can think about a directional derivative process. That is, for a given direction, at each location we can define a random variable that is the directional derivative of the original process at that location in the given direction. The entire collection of random variables can again be shown to be a spatial process. We offer brief development below but note that, intuitively, such variables would be created through limits of finite differences. In other words, we can also formalize a finite difference process in a given direction. The value of formalizing such processes lies in the possibility of assessing where, in a region of interest, there are sharp gradients and in which directions. They also enable us to work at different scales of resolution. Application could involve land-value gradients away from a central business district, temperature gradients in a north-south direction as mentioned above, or perhaps the maximum gradient at a location and the direction of that gradient, in order to identify zones of rapid change (boundary analysis). Some detail on the development of directional derivative processes appears in Section 13.3.

3.2 Nonstationary spatial process models ⋆

Recognizing that isotropy is an assumption regarding spatial association that will rarely hold in practice, Section 2.2 proposed classes of covariance functions that were still stationary but anisotropic. However, we may wish to shed the stationarity assumption entirely and merely assume that $\text{cov}(Y(\mathbf{s}), Y(\mathbf{s}')) = C(\mathbf{s}, \mathbf{s}')$ where $C(\cdot, \cdot)$ is symmetric in its arguments. The choice of C must still be valid. Theoretical classes of valid nonstationary covariance functions can be developed (Rehman and Shapiro, 1996), but they are typically described through existence theorems, perhaps as functions in the complex plane.

We seek classes that are flexible but also offer attractive interpretation and are computationally tractable. To this end, we prefer constructive approaches. We first observe that nonstationarity can be immediately introduced through scaling and through marginalization of stationary processes.

For the former, suppose $w(\mathbf{s})$ is a mean 0, variance 1 stationary process with correlation function ρ. Then $v(\mathbf{s}) = \sigma(\mathbf{s})w(\mathbf{s})$ is a nonstationary process. In fact,

$$\begin{aligned} var\ v(\mathbf{s}) &= \sigma^2(\mathbf{s}) \\ \text{and } cov(v(\mathbf{s}), v(\mathbf{s}')) &= \sigma(\mathbf{s})\sigma(\mathbf{s}')\rho(\mathbf{s}-\mathbf{s}') \ , \end{aligned} \tag{3.6}$$

so $v(\mathbf{s})$ could be used as a spatial error process, replacing $w(\mathbf{s})$ in (6.1). Where would $\sigma(\mathbf{s})$ come from? Since the use of $v(\mathbf{s})$ implies heterogeneous variance for $Y(\mathbf{s})$ we could follow the familiar course in regression modeling of setting $\sigma(\mathbf{s}) = g(x(\mathbf{s}))\sigma$ where $x(\mathbf{s})$ is a suitable positive covariate and g is a strictly increasing positive function. Hence, $var\ Y(\mathbf{s})$ increases in $x(\mathbf{s})$. Customary choices for $g(\cdot)$ are $(\cdot)$ or $(\cdot)^{\frac{1}{2}}$.

Instead, suppose we set $v(\mathbf{s}) = w(\mathbf{s}) + \delta z(\mathbf{s})$ with $z(\mathbf{s}) > 0$ and with δ being random with mean 0 and variance σ_δ^2. Then $v(\mathbf{s})$ is still a mean 0 process but now unconditionally, i.e., marginalizing over δ,

$$\begin{aligned} var\ v(\mathbf{s}) &= \sigma_w^2 + z^2(\mathbf{s})\sigma_\delta^2 \\ \text{and } cov(v(\mathbf{s}), v(\mathbf{s}')) &= \sigma_w^2\rho(\mathbf{s}-\mathbf{s}') + z(\mathbf{s})z(\mathbf{s}')\sigma_\delta^2 \ . \end{aligned} \tag{3.7}$$

(There is no reason to impose $\sigma_w^2 = 1$ here.) Again, this model for $v(\mathbf{s})$ can replace that for $w(\mathbf{s})$ as above. Now where would $z(\mathbf{s})$ come from? One possibility is that $z(\mathbf{s})$ might be a function of the distance from s to some externality in the study region. (For instance, in modeling land prices, we might consider distance from the central business district.) Another possibility is that $z(\mathbf{s})$ is an explicit function of the location, e.g., of latitude or longitude, of eastings or northings (after some projection). Of course, we could introduce a vector $\boldsymbol{\delta}$ and a vector $\mathbf{z}(\mathbf{s})$ such that $\boldsymbol{\delta}^T\mathbf{z}(\mathbf{s})$ is a trend surface and then do a trend surface marginalization. In this fashion the spatial structure in the mean is converted to the association structure. And since $\mathbf{z}(\mathbf{s})$ varies with $\mathbf{s}$, the resultant association must be nonstationary.

In (3.6) the departure from stationarity is introduced in a multiplicative way, i.e., through scaling. The nonstationarity is really just in the form of a nonhomogeneous variance; the spatial correlations are still stationary. Similarly, through (3.7) it arises in an additive way. Now, the nonstationarity is really just arising by revising the mean structure with a regression term; the residual spatial dependence is still stationary. In fact, it is evident that we could create $v(\mathbf{s}) = \sigma(\mathbf{s})w(\mathbf{s}) + \delta z(\mathbf{s})$ yielding both types of departures from stationarity. But it is also evident that (3.6) and (3.7) are of limited value.

However, the foregoing suggests a simple strategy for developing nonstationary covariance structure using known functions. For instance, for a function $g(\mathbf{s})$ on $\Re^2$, $C(\mathbf{s}, \mathbf{s}') = \sigma^2 g(\mathbf{s})g(\mathbf{s}')$ is immediately seen to be a valid covariance function. Of course, it is not very interesting since, for locations $\mathbf{s}_1, \mathbf{s}_2, ..., \mathbf{s}_n$, the resulting joint covariance matrix is of rank 1 regardless of n! But then, the idea can evidently be extended by introducing more functions. This leads to the flexible class of nonstationary covariance functions introduced by Cressie and Johannesson (2008) to implement so-called fixed rank kriging for large spatial data sets. Specifically, let $C(\mathbf{s}, \mathbf{s}') = \mathbf{g}(\mathbf{s})^T K \mathbf{g}(\mathbf{s}')$ where $\mathbf{g}(\mathbf{s})$ is an $r \times 1$ vector of known functions and K is an $r \times r$ positive definite matrix. Again, the validity of this C is immediate. There is no requirement that the functions in $\mathbf{g}(\mathbf{s})$ be orthogonal and standard classes such as smoothing splines, radial basis functions, or wavelets can be used. The challenges include the choice of r and the estimation of K (setting $K = I$ is not rich enough). For the latter, Cressie and Johannesson (2008) propose to obtain an empirical estimate of K, say, $\hat{K}$ and then minimize a Frobenius norm[1] between $\hat{K}$ and $K(\boldsymbol{\theta})$ for some parametric class of positive definite matrices. Hence, this approach, when applied to kriging, will suffer the same problems regarding capturing uncertainty as noted in the context of ordinary kriging in Chapter 2. However, with the challenges regarding choice of r and specifying K as a somewhat high dimensionally unknown parametric matrix, at present there is no Bayesian version of this fixed rank kriging available.

3.2.1 Deformation

In what is regarded as a landmark paper in spatial data analysis, Sampson and Guttorp (1992) introduced an approach to nonstationarity through *deformation*. The basic idea is to transform the geographic region D to a new region G, a region such that stationarity and, in fact, isotropy holds on G. The mapping $\mathbf{g}$ from D to G is bivariate, i.e., if $\mathbf{s} = (\ell_1, \ell_2)$, $\mathbf{g}(\ell_1, \ell_2) = (g_1(\ell_1, \ell_2), g_2(\ell_1, \ell_2))$. If C denotes the isotropic covariance function on G we have

$$cov(Y(\mathbf{s}), Y(\mathbf{s}')) = C(\|\mathbf{g}(\mathbf{s}) - \mathbf{g}(\mathbf{s}')\|) \ . \tag{3.8}$$

Thus, from (3.8) there are two unknown functions to estimate, $\mathbf{g}$ and C. The latter is assumed to be a parametric choice from a standard class of covariance functions (as in Table 2.1). To determine the former is a challenging "fitting" problem. To what class of transformations shall we restrict ourselves? How shall we obtain the "best" member of this

[1]The Frobenius norm between two matrices A and B is $||A - B||^2 \equiv \mathtt{tr}(A - B)^T(A - B)$.

class? Sampson and Guttorp (1992) employ the class of thin plate splines and optimize a version of a two-dimensional nonmetric multidimensional scaling criterion (see, e.g., Mardia et al., 1979), providing an algorithmic solution. The solution is generally not well behaved, in the sense that $\mathbf{g}$ will be bijective, often folding over itself. Smith (1996) embedded this approach within a likelihood setting but worked instead with the class of radial basis functions.

Damian, Sampson, and Guttorp (2001) and Schmidt and O'Hagan (2002) have formulated fully Bayesian approaches to implement (3.8). The former still work with thin plate splines, but place priors over an identifiable parametrization (which depends upon the number of points, n being transformed). The latter elect not to model $\mathbf{g}$ directly but instead model the transformed locations. The set of n transformed locations are modeled as n realizations from a bivariate Gaussian spatial process (see Chapter 9) and a prior is placed on the process parameters. That is, the $\mathbf{g}(\mathbf{s})$ surface arises as a random realization of a bivariate process over the $\mathbf{s}$ rather than through the values over $\mathbf{s}$ of an unknown bivariate transformation.

A fundamental limitation of the deformation approach is that implementation requires independent replications of the process in order to obtain an estimated sample covariance matrix for the set of $(Y(\mathbf{s}), ..., Y(\mathbf{s}_n))$. In practice, we rarely obtain i.i.d. replications of a spatial process. If we obtain repeated measurements at a particular location, they are typically collected across time. We would prefer to incorporate a temporal aspect in the modeling rather than attempting repairs (e.g., differencing and detrending) to achieve approximately i.i.d. observations. This is the focus of Chapter 11. Moreover, even if we are prepared to assume independent replications, to adequately estimate an $n \times n$ covariance matrix, even for a moderate size n, requires a very large number of them, more than we would imagine in practice.

3.2.2 Nonstationarity through kernel mixing of process variables

Kernel mixing was described above in the context of creating stationary processes. Here, we show that it provides a strategy for introducing nonstationarity while retaining clear interpretation and permitting analytic calculation. We look at two distinct approaches, one due to Higdon (e.g., Higdon, 1998b, 2002; Higdon et al., 1999) and the other to Fuentes (e.g., Fuentes 2002a, b; Fuentes and Smith, 2001, 2003). We note that kernel mixing has a long tradition in the statistical literature, especially in density estimation and regression modeling (Silverman, 1986). Kernel mixing is often done with distributions and we will look at this idea in a later subsection. Here, we focus on kernel mixing of random variables.

Adding some detail to our earlier discussion, suppose we work with bivariate kernels starting with stationary choices of the form $K(\mathbf{s}-\mathbf{s}')$, e.g., $K(\mathbf{s}-\mathbf{s}') = \exp\{-\frac{1}{2}(\mathbf{s}-\mathbf{s}')^T V(\mathbf{s}-\mathbf{s}')\}$. A natural choice for V would be diagonal with V_{11} and V_{22} providing componentwise scaling to the separation vector $\mathbf{s}-\mathbf{s}'$. Other choices of kernel function are available; specialization to versions based on Euclidean distance is immediate; again see, e.g., Silverman (1986). Higdon et al. (1998b, 2002) let $z(\mathbf{s})$ be a white noise process, i.e., $E(z(\mathbf{s}) = 0)$, $var(z(\mathbf{s})) = \sigma^2$ and $cov(z(\mathbf{s}), z(\mathbf{s}')) = 0$ and set

$$w(\mathbf{s}) = \int_{\Re^2} K(\mathbf{s}-\mathbf{t})z(\mathbf{t})d\mathbf{t} \ . \tag{3.9}$$

Rigorously speaking, (3.9) is not defined. As we noted in Section 3.1.1, the convolution should be written as $w(\mathbf{s}) = \int K(\mathbf{s}-\mathbf{t})\mathcal{X}(d\mathbf{t})$ where $\mathcal{X}(\mathbf{t})$ is two-dimensional Brownian motion.

Reiterating earlier details, the process $w(\mathbf{s})$ is said to arise through *kernel convolution.* By change of variable, (3.9) can be written as

$$w(\mathbf{s}) = \int_{\Re^2} K(\mathbf{u})z(\mathbf{s}+\mathbf{u})d\mathbf{u} \ , \tag{3.10}$$

emphasizing that $w(\mathbf{s})$ arises as a kernel-weighted average of z's centered around $\mathbf{s}$. It is straightforward to show that $E[w(\mathbf{s})] = 0$, but also that

$$\begin{aligned} var\ w(\mathbf{s}) &= \sigma^2 \int_{\Re^2} k^2(\mathbf{s}-\mathbf{t})d\mathbf{t}\ , \\ \text{and}\quad cov(w(\mathbf{s}), w(\mathbf{s}')) &= \sigma^2 \int_{\Re^2} K(\mathbf{s}-\mathbf{t})K(\mathbf{s}'-\mathbf{t})d\mathbf{t}\ . \end{aligned} \tag{3.11}$$

A simple change of variables ($\mathbf{t} \to \mathbf{u} = \mathbf{s}' - \mathbf{t}$), shows that

$$cov(w(\mathbf{s}), w(\mathbf{s}')) = \sigma^2 \int_{\Re^2} K(\mathbf{s}-\mathbf{s}'+\mathbf{u})K(\mathbf{u})d\mathbf{u}\ , \tag{3.12}$$

i.e., $w(\mathbf{s})$ is stationary. In fact, (3.9) is a way of generating classes of stationary processes (see, e.g., Yaglom, 1962, Ch. 26) whose limitations we have considered above.

We can extend (3.9) so that $z(\mathbf{s})$ is a mean 0 stationary spatial process with covariance function $\sigma^2\rho(\cdot)$. Again $E[w(\mathbf{s})] = 0$ but now

$$\begin{aligned} var\ w(\mathbf{s}) &= \sigma^2 \int_{\Re^2}\int_{\Re^2} K(\mathbf{s}-\mathbf{t})K(\mathbf{s}'-\mathbf{t})\rho(\mathbf{t}-\mathbf{t}')d\mathbf{t}d\mathbf{t}' \\ \text{and}\quad cov(w(\mathbf{s}), w(\mathbf{s}')) &= \sigma^2 \int_{\Re^2}\int_{\Re^2} K(\mathbf{s}-\mathbf{t})K(\mathbf{s}'-\mathbf{t}')\rho(\mathbf{t}-\mathbf{t}')d\mathbf{t}d\mathbf{t}'\ . \end{aligned} \tag{3.13}$$

Interestingly, $w(\mathbf{s})$ is still stationary. We now use the change of variables ($\mathbf{t} \to \mathbf{u} = \mathbf{s}' - \mathbf{t}$, $\mathbf{t}' \to \mathbf{u}' = \mathbf{s}' - \mathbf{t}'$) to obtain

$$cov(w(\mathbf{s}), w(\mathbf{s}')) = \sigma^2 \int_{\Re^2}\int_{\Re^2} K(\mathbf{s}-\mathbf{s}'+\mathbf{u})K(\mathbf{u}')\rho(\mathbf{u}-\mathbf{u}')d\mathbf{u}d\mathbf{u}'\ . \tag{3.14}$$

Note that (3.11) and (3.13) can be proposed as covariance functions. It is straightforward to argue that they are positive definite functions and so can be attached for instance to a Gaussian process if we wish. However, the integrations in (3.12) and (3.14) will not be possible to do explicitly except in certain special cases (see, e.g., Ver Hoef and Barry, 1998). Numerical integration across $\Re^2$ for (3.12) or across $\Re^2 \times \Re^2$ for (3.14) may be difficult, requiring nonlinear transformation to a bounded set. If we work with a subset $D \subset \Re^2$, we sacrifice stationarity. Monte Carlo integration is also not attractive here: we would have to sample from the standardized density associated with K. But since $\mathbf{s} - \mathbf{s}'$ enters into the argument, we would have to do a separate Monte Carlo integration for each pair of locations $(\mathbf{s}_i, \mathbf{s}_j)$.

An alternative is to replace (3.9) with a finite sum approximation, i.e., to define

$$w(\mathbf{s}) = \sum_{j=1}^{L} K(\mathbf{s}-\mathbf{t}_j)z(\mathbf{t}_j) \tag{3.15}$$

for locations $\mathbf{t}_j$, $j = 1, \ldots, L$. In the case of a white noise assumption for the z's,

$$\begin{aligned} \text{var}w(\mathbf{s}) &= \sigma^2 \textstyle\sum_{j=1}^{L} k^2(\mathbf{s}-\mathbf{t}_j) \\ \text{and cov}(w(\mathbf{s}), w(\mathbf{s}')) &= \sigma^2 var\ w(\mathbf{s}) = \sigma^2 \textstyle\sum_{j=1}^{L} K(\mathbf{s}-\mathbf{t}_j)K(\mathbf{s}'-\mathbf{t}_j)\ . \end{aligned} \tag{3.16}$$

In the case of spatially correlated z's,

$$\begin{aligned} var\ w(\mathbf{s}) &= \sigma^2 \textstyle\sum_{j=1}^{L}\sum_{j'=1}^{L} K(\mathbf{s}-\mathbf{t}_j)K(\mathbf{s}-\mathbf{t}_{j'})\rho(\mathbf{t}_j-\mathbf{t}_{j'}) \\ \text{and } cov(w(\mathbf{s}), w(\mathbf{s}')) &= \sigma^2 \textstyle\sum_{j=1}^{L}\sum_{j'=1}^{L} K(\mathbf{s}-\mathbf{t}_j)K(\mathbf{s}'-\mathbf{t}_{j'})\rho(\mathbf{t}_j-\mathbf{t}_{j'})\ . \end{aligned} \tag{3.17}$$

Expressions (3.16) and (3.17) can be calculated directly from (3.15) and, in fact, can be used to provide a limiting argument for expressions (3.11) and (3.13); see Exercise 5. Note that (3.15) provides a dimension reduction; we express the entire stochastic process of random variables through the finite collection of $z_{\mathbf{t}_j}$. Dimension reduction has been offered as a

general way to handle the computational challenges associated with large datasets in space and time; see Chapter 12.

Note further that, while (3.16) and (3.17) are available explicitly, these forms reveal that the finite sum process in (3.15) is no longer stationary. While nonstationary specifications are the objective of this section, their creation through (3.15) is rather artificial as it arises from the arbitrary $\{\mathbf{t}_j\}$. It would be more attractive to modify (3.9) to achieve a class of nonstationary processes.

So, instead, suppose we allow the kernel in (3.9) to vary spatially. Notationally, we can write such an object as $K(\mathbf{s}-\mathbf{s}';\mathbf{s})$. Illustratively, we might take $K(\mathbf{s}-\mathbf{s}';\mathbf{s}) = \exp\{-\frac{1}{2}(\mathbf{s}-\mathbf{s}')^T V_{\mathbf{s}}(\mathbf{s}-\mathbf{s}')\}$. As above, we might take $V_{\mathbf{s}}$ to be diagonal with, if $\mathbf{s} = (\ell_1, \ell_2)$, $(V_{\mathbf{s}})_{11} = V(\ell_1)$ and $(V_{\mathbf{s}})_{22} = V(\ell_2)$. Higdon, Swall, and Kern (1999) adopt such a form with V taken to be a slowly varying function. One explicit choice would take $V_{\mathbf{s}} = V_{A_i}$ for $s \in A_i$ where the V_{A_i} are obtained through local range anisotropy. Another choice might take $V_{\mathbf{s}} = \eta(\mathbf{s})I$ where $\eta(\mathbf{s})$ might be a simple trend surface or a function of a local covariate $X(\mathbf{s})$ such as elevation.

We can insert $K(\mathbf{s}-\mathbf{s}';\mathbf{s})$ into (3.9) in place of $K(\mathbf{s}-\mathbf{s}')$ with obvious changes to (3.11), (3.12), (3.13), and (3.14). Evidently, the process is now nonstationary. In fact, the variation in V provides insight into the departure from stationarity. For computational reasons Higdon et al. (1999) implement this modified version of (3.9) through a finite sum analogous to (3.15). A particularly attractive feature of employing a finite sum approximation is dimension reduction. If $z(\mathbf{s})$ is white noise we have an approach for handling large data sets (see Section 12.1). That is, regardless of n, $\{w(\mathbf{s}_i)\}$ depends only on L latent variables z_j, $j = 1, \ldots, L$, and these variables are independent. Rather than fitting the model in the space of the $\{w(\mathbf{s}_i)\}$ we can work in the space of the z_ℓ.

Paciorek and Schervish (2006) substantially extend this work. First, they note that the general form $C(\mathbf{s},\mathbf{s}') = \int_{\Re^2} K_{\mathbf{s}}(\mathbf{u})K_{\mathbf{s}'}(\mathbf{u})d\mathbf{u}$ is a valid covariance function, in fact for $\mathbf{s},\mathbf{s}' \in \Re^r$ for an positive integer r. This is easily shown by direct calculation and is left as an exercise. More importantly, it means we do not need a process construction to justify such covariance functions. They also note that the use of Gaussian kernels in this construction (as well as the earlier stationary version) is unattractive for the same reasons noted above; the resulting covariance function will produce process realizations that are too smooth.

In fact, Paciorek and Schervish (2006) provide a much richer class of nonstationary covariance functions as follows. Let $Q(\mathbf{s},\mathbf{s}') = (\mathbf{s}-\mathbf{s}')^T(\frac{V_{\mathbf{s}}+V_{\mathbf{s}'}}{2})^{-1}(\mathbf{s}-\mathbf{s}')$ and let ρ be any positive definite function on $\Re^p$. Then

$$C(\mathbf{s},\mathbf{s}') = |V_{\mathbf{s}}|^{\frac{1}{2}}|V_{bs'}|^{\frac{1}{2}}\left|\frac{V_{\mathbf{s}}+V_{\mathbf{s}'}}{2}\right|^{-\frac{1}{2}}\rho(\sqrt{Q(\mathbf{s},\mathbf{s}')}) \tag{3.18}$$

is a valid nonstationary correlation function on $\Re^r$. We can use the Matérn or other choices for ρ with choices for $V_{\mathbf{s}}$ as above. We can multiply by σ^2 to obtain a covariance function. Under conditions given in the Ph.D. thesis of Paciorek, process realizations under $C(\mathbf{s},\mathbf{s}')$ inherit the smoothness properties of $\rho(\cdot)$. We leave as an exercise to show that if, in the general kernel convolution form above, we take $K_{\mathbf{s}}(\mathbf{u}) = |V_{\mathbf{s}}|^{-\frac{1}{2}}\mathtt{exp}(-(\mathbf{s}-\mathbf{u})^T V_{\mathbf{s}}(\mathbf{s}-\mathbf{u}))$, then $C(\mathbf{s},\mathbf{s}') = |V_{\mathbf{s}}|^{\frac{1}{2}}|V_{bs'}|^{\frac{1}{2}}|\frac{V_{\mathbf{s}}+V_{\mathbf{s}'}}{2}|^{-\frac{1}{2}}\mathtt{exp}(-Q(\mathbf{s},\mathbf{s}'))$.

Fuentes (2002a, b) offers a kernel mixing form that initially appears similar to (3.9) but is fundamentally different. Let

$$w(\mathbf{s}) = \int K(\mathbf{s}-\mathbf{t})z_{\boldsymbol{\theta}(\mathbf{t})}(\mathbf{s})d\mathbf{t}\ . \tag{3.19}$$

In (3.19), $K(\cdot)$ is as in (3.9) but $z_{\boldsymbol{\theta}}(\mathbf{s})$ denotes a mean 0 stationary spatial process with covariance function that is parametrized by $\boldsymbol{\theta}$. For instance $C(\cdot;\boldsymbol{\theta})$ might be $\sigma^2\exp(-\phi\,\|\cdot\|^\alpha)$,

a power exponential family with $\boldsymbol{\theta} = (\sigma^2, \phi, \alpha)$. In (3.19) $\boldsymbol{\theta}(\mathbf{t})$ indexes an uncountable number of processes. These processes are assumed independent across $\mathbf{t}$. Note that (3.19) is mixing an uncountable number of stationary spatial processes each at $\mathbf{s}$ while (3.9) is mixing a single process across all locations.

Formally, $w(\mathbf{s})$ has mean 0 and

$$\begin{array}{c} var(w(\mathbf{s})) = \int_{\Re^2} k^2(\mathbf{s}-\mathbf{t})C(0;\boldsymbol{\theta}(\mathbf{t}))d\mathbf{t} \\ \text{and } cov(w(\mathbf{s}), w(\mathbf{s}')) = \int_{\Re^2} K(\mathbf{s}-\mathbf{t})K(\mathbf{s}'-\mathbf{t})C(\mathbf{s}-\mathbf{s}';\boldsymbol{\theta}(\mathbf{t}))d\mathbf{t}\ . \end{array} \tag{3.20}$$

Expression (3.20) reveals that (3.19) defines a nonstationary process. Suppose k is very rapidly decreasing and $\boldsymbol{\theta}(\mathbf{t})$ varies slowly. Then $w(\mathbf{s}) \approx K(0)z_{\boldsymbol{\theta}(\mathbf{s})}(\mathbf{s})$. But also, if $\mathbf{s}-\mathbf{s}'$ is small, $w(\mathbf{s})$ and $w(\mathbf{s}')$ will behave like observations from a stationary process with parameter $\boldsymbol{\theta}(\mathbf{s})$. Hence, Fuentes refers to the class of models in (3.19) as a nonstationary class that exhibits *local* stationarity.

In practice, one cannot work with (3.19) directly. Again, finite sum approximation is employed. Again, a finite set of locations $\mathbf{t}_1, \ldots, \mathbf{t}_L$ is selected and we set

$$w(\mathbf{s}) = \sum_j K(\mathbf{s}-\mathbf{t}_j)z_j(\mathbf{s})\ , \tag{3.21}$$

writing $\boldsymbol{\theta}(\mathbf{t}_j)$ as j. Straightforwardly,

$$\begin{array}{c} var(w(\mathbf{s}) = \sum_{j=1}^L k^2(\mathbf{s}-\mathbf{t}_j)C_j(0) \\ \text{and } cov(w(\mathbf{s}), w(\mathbf{s}') = \sum_{j=1}^L K(\mathbf{s}-\mathbf{t}_j)K(\mathbf{s}'-\mathbf{t}_j)C_j(\mathbf{s}-\mathbf{s}')\ . \end{array} \tag{3.22}$$

It is worth noting that the discretization in (3.21) does not provide a dimension reduction. In fact, it is a dimension *explosion*; we need L $z_j(\mathbf{s})$'s for each $w(\mathbf{s})$.

In (3.21) it can happen that some $\mathbf{s}$'s may be far enough from each of the $\mathbf{t}_j$'s so that each $K(\mathbf{s}-\mathbf{t}_j) \approx 0$, whence $w(\mathbf{s}) \approx 0$. Of course, this cannot happen in (3.19) but we cannot work with this expression. A possible remedy was proposed in Banerjee et al. (2004). Replace (3.21) with

$$w(\mathbf{s}) = \sum_{j=1}^L \alpha(\mathbf{s}, \mathbf{t}_j)z_j(\mathbf{s})\ . \tag{3.23}$$

In (3.23), the $z_j(\mathbf{s})$ are as above, but $\alpha(\mathbf{s}, \mathbf{t}_j) = \gamma(\mathbf{s}, \mathbf{t}_j)/\sqrt{\sum_{j=1}^L \gamma^2(\mathbf{s}, \mathbf{t}_j)}$, where $\gamma(\mathbf{s}, \mathbf{t})$ is a decreasing function of the distance between $\mathbf{s}$ and $\mathbf{t}$, which may change with $\mathbf{s}$, i.e., $\gamma(\mathbf{s}, \mathbf{t}) = k_{\mathbf{s}}(\|\mathbf{s}-\mathbf{t}\|)$. (In the terminology of Higdon et al., 1999, $k_{\mathbf{s}}$ would be a spatially varying kernel function.) As a result, $\sum_{j=1}^L \alpha^2(\mathbf{s}, \mathbf{t}_j) = 1$, so regardless of where $\mathbf{s}$ is, not all of the weights in (3.23) can be approximately 0. Other standardizations for γ are possible; we have proposed this one because if all σ_j^2 are equal, then $var(w(\mathbf{s})) = \sigma^2$. That is, if each local process has the same variance, then this variance should be attached to $w(\mathbf{s})$. Furthermore, suppose $\mathbf{s}$ and $\mathbf{s}'$ are near to each other, whence $\gamma(\mathbf{s}, \mathbf{t}_j) \approx \gamma(\mathbf{s}', \mathbf{t}_j)$ and thus $\alpha(\mathbf{s}, \mathbf{t}_j) \approx \alpha(\mathbf{s}', \mathbf{t}_j)$. So, if in addition all $\phi_j = \phi$, then $cov(w(\mathbf{s}), w(\mathbf{s}')) \approx \sigma^2\rho(\mathbf{s}-\mathbf{s}';\phi)$. So, if the process is in fact stationary over the entire region, we obtain essentially the second-order behavior of this process.

The alternative scaling $\widetilde{\alpha}(\mathbf{s}, \mathbf{t}_j) = \gamma(\mathbf{s}, \mathbf{t}_j)/\sum_{j'} \gamma(\mathbf{s}, \mathbf{t}_{j'})$ gives a weighted average of the component processes. Such weights would preserve an arbitrary constant mean. However, since, in our context, we are modeling a mean 0 process, such preservation is not a relevant feature.

Useful properties of the process in (3.23) are

$$E\left(w\left(\mathbf{s}\right)\right) \quad = \quad 0\ ,$$

$$\begin{aligned} Var\left(w\left(\mathbf{s}\right)\right) &= \sum_{j=1}^{L} \alpha^2(\mathbf{s},\mathbf{t}_j)\sigma_j^2 \ , \\ \text{and } \ cov(w(\mathbf{s}),w(\mathbf{s}')) &= \sum_{j=1}^{L} \alpha(\mathbf{s},\mathbf{t}_j)\alpha(\mathbf{s}',\mathbf{t}_j)\sigma_j^2\rho(\mathbf{s}-\mathbf{s}';\phi_j) \ . \end{aligned}$$

We have clearly defined a proper spatial process through (3.23). In fact, for arbitrary locations $\mathbf{s}_1,\ldots,\mathbf{s}_n$, let $\mathbf{w}_\ell^T = (w_\ell(\mathbf{s}_1)),\ldots,w_\ell(\mathbf{s}_n))$, $\mathbf{w}^T = (w(\mathbf{s}_1),\ldots,w(\mathbf{s}_n))$, and let A_ℓ be diagonal with $(A_\ell)_{ii} = \alpha(\mathbf{s}_i,\mathbf{t}_\ell)$. Then $\mathbf{w} \sim N(\mathbf{0}, \sum_{\ell=1}^{L}\boldsymbol{\sigma}_\ell^2 A_\ell R(\phi_\ell)A_\ell)$ where $(\Sigma(\phi_\ell))_{ii'} = \rho\left(\mathbf{s}_i-\mathbf{s}_{i'},\phi_\ell\right)$. Note that $L=1$ is permissible in (3.23); $w\left(\mathbf{s}\right)$ is still a nonstationary process. Finally, Fuentes and Smith (2003) and Banerjee et al. (2004) offer some discussion regarding precise number of and locations for the $\mathbf{t}_j$.

We conclude this subsection by noting that for a general nonstationary spatial process there is no sensible notion of a range. However, for the class of processes in (3.23) we can define a meaningful range. Under (3.23),

$$corr(w(\mathbf{s}),w(\mathbf{s}')) = \frac{\sum_{j=1}^{L}\alpha(\mathbf{s},\mathbf{t}_j)\alpha(\mathbf{s}',\mathbf{t}_j)\sigma_j^2\rho(\mathbf{s}-\mathbf{s}';\phi_j)}{\sqrt{\left(\sum_{j=1}^{L}\alpha^2(\mathbf{s},\mathbf{t}_j)\sigma_j^2\right)\left(\sum_{j=1}^{L}\alpha^2(\mathbf{s}',\mathbf{t}_j)\sigma_j^2\right)}} \ . \tag{3.24}$$

Suppose ρ is positive and strictly decreasing asymptotically to 0 as distance tends to ∞, as is usually assumed. If ρ is, in fact, isotropic, let d_ℓ be the range for the ℓth component process, i.e., $\rho(d_\ell,\phi_\ell) = .05$, and let $\widetilde{d} = \max_\ell d_\ell$. Then (3.24) immediately shows that, at distance $\widetilde{d}$ between $\mathbf{s}$ and $\mathbf{s}'$, we have $corr(w(\mathbf{s}),w(\mathbf{s}')) \leq .05$. So $\widetilde{d}$ can be interpreted as a conservative range for $w(\mathbf{s})$. Normalized weights are not required in this definition. If ρ is only assumed stationary, we can similarly define the range in an arbitrary direction $\boldsymbol{\mu}$. Specifically, if $\boldsymbol{\mu}/\left|\left|\boldsymbol{\mu}\right|\right|$ denotes a unit vector in $\boldsymbol{\mu}$'s direction and if $d_{\boldsymbol{\mu},\ell}$ satisfies $\rho(d_{\boldsymbol{\mu},\ell}\boldsymbol{\mu}/\left\|\boldsymbol{\mu}\right\|;\phi_\ell) = .05$, we can take $\widetilde{d}_{\boldsymbol{\mu}} = \max_\ell \widetilde{d}_{\boldsymbol{\mu},\ell}$.

3.2.3 Mixing of process distributions

If a kernel $K(\cdot)$ is integrable and standardized to a density function and if f is also a density function, then

$$f_K(y) = \int K(y-x)f(x)dx \tag{3.25}$$

is a density function. (It is, of course, the distribution of $X+Y-X$ where $X \sim f$, $Y-X \sim k$, and X and $Y-X$ are independent). In (3.25), we can extend to allow $\mathbf{Y}$, a vector of dimension n. But recall that we have specified the distribution for a spatial process through arbitrary finite dimensional distributions (see Section 3.1.1). This suggests that we can use (3.25) to build a process distribution.

Operating formally, let V_D be the set of all $V(\mathbf{s})$, $\mathbf{s} \in D$. Write $V_D = V_{0,D} + V_0 - V_{0,D}$ where $V_{0,D}$ is a realization of a mean 0 stationary Gaussian process over D, and $V_D - V_{0,D}$ is a realization of a white noise process with variance σ^2 over D. Write

$$f_K(V_D \mid \tau) = \int \frac{1}{\sigma}\, K\left(\frac{1}{\tau}(V_D - V_{0,D})\right) f(V_{0,D})dV_{0,D} \ . \tag{3.26}$$

Formally, f_K is the distribution of the spatial process $v(\mathbf{s})$. In fact, $v(\mathbf{s})$ is just the customary model for the residuals in a spatial regression, i.e., of the collection $v(\mathbf{s}) = w(\mathbf{s}) + \epsilon(\mathbf{s})$ where $w(\mathbf{s})$ is a spatial process and $\epsilon(\mathbf{s})$ is a noise or nugget process.

Of course, in this familiar case there is no reason to employ the form (3.26). However it does reveal how, more generally, a spatial process can be developed through "kernel mixing" of a process distribution. More importantly, it suggests that we might introduce an alternative specification for $V_{0,D}$. For example, suppose $f(V_{0,D})$ is a discrete distribution, say, of the form $\sum_\ell p_\ell \delta(v^*_{\ell,D})$ where $p_\ell \geq 0$, $\sum p_\ell = 1$, $\delta(\cdot)$ is the Dirac delta function, and $V^*_{\ell,D}$ is a surface over D. The sum may be finite or infinite. An illustration of the former arises when $f(V_{0,D})$ is a realization from a finite discrete mixture (Duan and Gelfand, 2003) or from a finite Dirichlet process; the latter arises under a general Dirichlet process and can be extend to more general Pitman-Yor processesed; see Gelfand, Kottas, and MacEachern (2005) for further details in this regard.

But then, if $K(\cdot)$ is Gaussian white noise, given $\{p_\ell\}$ and $\{v^*_{\ell,D}\}$, for any set of locations $\mathbf{s}_1, \ldots, \mathbf{s}_n$, if $\mathbf{V} = (v(\mathbf{s}_1), \ldots, v(\mathbf{s}_n)$,

$$f_K(\mathbf{V}) = \sum_\ell p_\ell N(\mathbf{v}^*_\ell, \sigma^2 I) \ , \tag{3.27}$$

where $\mathbf{v}^*_\ell = (v^*_\ell(\mathbf{s}_1), \ldots, v^*_\ell(\mathbf{s}_n))^T$. So $v(\mathbf{s})$ is a continuous process that is non-Gaussian. But also, $E[v(\mathbf{s}_i)] = \sum_\ell p_\ell v^*_\ell(\mathbf{s}_i)$ and

$$\begin{aligned} var\ v(\mathbf{s}_i) &= \textstyle\sum_\ell p_\ell v^{2*}_\ell(\mathbf{s}_i) - (\sum_\ell p_\ell v^*_\ell(\mathbf{s}_i))^2 \\ \text{and } cov(v(\mathbf{s}_i), v(\mathbf{s}_j)) &= \textstyle\sum_\ell p_\ell v^*_\ell(\mathbf{s}_i) v^*_\ell(\mathbf{s}_j) - (\sum_\ell p_\ell v^*_\ell(\mathbf{s}_i))(\sum_\ell p_\ell v^*_\ell(\mathbf{s}_j)) \ . \end{aligned} \tag{3.28}$$

This last expression shows that $v(\mathbf{s})$ is *not* a stationary process. However, a routine calculation shows that if the $v^*_{\ell,D}$ are continuous surfaces, the $v(\mathbf{s})$ process is mean square continuous and almost surely continuous.

3.3 Exercises

1. Show that, if $\rho(\mathbf{s}-\mathbf{s}')$ is a valid correlation function, then $e^{\sigma^2 \rho(\mathbf{s}-\mathbf{s}')}$ and $\sinh(\sigma^2 \rho(\mathbf{s}-\mathbf{s}'))$ are valid covariance functions.
2. If $c(\mathbf{h};\gamma)$ is valid for $\gamma \in \Gamma$ and ν is a positive measure on Γ, then $c_\nu(\mathbf{h} = \int_\Gamma c(\mathbf{h};\gamma)\nu(d\gamma)$ is valid provided the integral exists for all $\mathbf{h}$
3. Suppose $Y(\mathbf{s})$ is a Gaussian process with mean surface $\mu(\mathbf{s})$ and covariance function $c(\mathbf{s},\mathbf{s}')$. Let $Z(\mathbf{s})$ be the induced log Gaussian process, i.e., $Y(\mathbf{s}) = \log Z(\mathbf{s})$. Find the mean surface and covariance function for the $Z(\mathbf{s})$ process. If $Y(\mathbf{s})$ is stationary, is $Z(\mathbf{s})$ necessarily stationary?
4. Suppose $W(\mathbf{s})$ is a mean 0 stationary Gaussian process with correlation function $\rho(\mathbf{h})$ Let $Y(\mathbf{s}) = \frac{\sigma W(\mathbf{s})}{\sqrt{\lambda}}$. If $\lambda \sim Ga(\nu/2, \nu/2)$, when will the marginal process have a covariance function? If it does, what is the covariance function?
5. Some mixing results:
   (a) If $Z \sim N(0,1)$ and $V \sim Ga(r/2, r/2)$ independent of Z, show that $Y = \sigma z/\sqrt{V} \sim \sigma t_r$ where t_r is a t-distribution with r d.f. So, if $Z(\mathbf{s})$ is a mean 0 Gaussian process with covariance function/correlation function $\rho(\mathbf{s}-\mathbf{s}')$, then we will define $Y(\mathbf{s}) = \sigma Z(\mathbf{s})/sqrtV$ as a t-process with r d.f. Obtain the joint distribution of $Y(\mathbf{s})$ and $Y(\mathbf{s}')$. Obtain the covariance function of the process, provided $r > 2$.
   (b) If $Z \sim N(0,1)$ and $V \sim \exp(1)$ independent of Z, show that $Y = \sigma Z/\sqrt{2V} \sim$ Laplace$(0,\sigma)$ where Laplace(μ,σ) is the Laplace (double exponential) distribution with location μ and scale σ. So, if $Z(\mathbf{s})$ is a mean 0 Gaussian process with covariance function/correlation function $\rho(\mathbf{s}-\mathbf{s}')$, then we will define $Y(\mathbf{s}) = \sigma Z(\mathbf{s})/\sqrt{2V}$ as a Laplace process. Obtain the joint distribution of $Y(\mathbf{s})$ and $Y(\mathbf{s}')$. Obtain the covariance function of the process.

6. Show, for fractional Brownian motion in two dimensions, if the Hurst index is greater than 1/2 we have positively correlated increments while if it is less than 1/2 we have negatively correlated increments.

7. Consider the *triangular* (or "tent") covariance function,

$$C(\|h\|) = \begin{cases} \sigma^2(1 - \|h\|/\delta) & \text{if } \|h\| \le \delta,\ \sigma^2 > 0,\ \delta > 0, \\ 0 & \text{if } \|h\| > \delta \end{cases} .$$

It is valid in one dimension. (The reader can verify that it is the characteristic function of the density function $f(x)$ proportional to $[1 - \cos(\delta x)]/\delta x^2$.) Now in two dimensions, consider a 6×8 grid with locations $\mathbf{s}_{jk} = (j\delta/\sqrt{2}, k\delta/\sqrt{2})$, $j = 1, \ldots, 6$, $k = 1, \ldots, 8$. Assign a_{jk} to $\mathbf{s}_{jk}$ such that $a_{jk} = 1$ if $j + k$ is even, $a_{jk} = -1$ if $j + k$ is odd. Show that $Var[\Sigma a_{jk} Y(\mathbf{s}_{jk})] < 0$, and hence that the triangular covariance function is *invalid* in two dimensions.

8. The *turning bands method* (Christakos, 1984; Stein, 1999a) is a technique for creating stationary covariance functions on $\Re^r$. Let $\mathbf{u}$ be a random unit vector on $\Re^r$ (by random we mean that the coordinate vector that defines $\mathbf{u}$ is randomly chosen on the surface of the unit sphere in $\Re^r$). Let $c(\cdot)$ be a valid stationary covariance function on $\Re^1$, and let $W(t)$ be a mean 0 process on $\Re^1$ having $c(\cdot)$ as its covariance function. Then for any location $\mathbf{s} \in \Re^r$, define

$$Y(\mathbf{s}) = W(\mathbf{s}^T\mathbf{u}) .$$

Note that we can think of the process either conditionally given $\mathbf{u}$, or marginally by integrating with respect to the uniform distribution for $\mathbf{u}$. Note also that $Y(\mathbf{s})$ has the possibly undesirable property that it is constant on planes (i.e., on $\mathbf{s}^T\mathbf{u} = k$).

   (a) If W is a Gaussian process, show that, given $\mathbf{u}$, $Y(\mathbf{s})$ is also a Gaussian process and is stationary.

   (b) Show that marginally $Y(\mathbf{s})$ is *not* a Gaussian process, but is stationary. [*Hint:* Show that $\mathrm{Cov}(Y(\mathbf{s}), Y(\mathbf{s}')) = \mathrm{E}_{\mathbf{u}} c((\mathbf{s} - \mathbf{s}')^T\mathbf{u})$.]

   (c) If $c(\cdot)$ is isotropic, then so is $Cov(Y(\mathbf{s}), Y(\mathbf{s}'))$.

9.(a) Based on (3.1), show that $c_{12}(\mathbf{h})$ is a valid correlation function; i.e., that G is a bounded, positive, symmetric about 0 measure on $\Re^2$.

   (b) Show further that if c_1 and c_2 are isotropic, then c_{12} is.

10. Show by direct calculation that $C(\mathbf{s}, \mathbf{s}') = \int_{Re^2} K_{\mathbf{s}}(\mathbf{u}) K_{\mathbf{s}'}(\mathbf{u}) d\mathbf{u}$ is a valid covariance function for $\mathbf{s}, \mathbf{s}' \in Re^p$ for any positive integer p.

11. Suppose we take $K_{\mathbf{s}}(\mathbf{u}) = |V_{\mathbf{s}}|^{-\frac{1}{2}} \exp(-(\mathbf{s} - \mathbf{u})^T V_{\mathbf{s}}(\mathbf{s} - \mathbf{u}))$, then, if $C(\mathbf{s}, \mathbf{s}') = \int_{Re^2} K_{\mathbf{s}}(\mathbf{u}) K_{\mathbf{s}'}(\mathbf{u}) d\mathbf{u}$, show that

$$C(\mathbf{s}, \mathbf{s}') = |V_{\mathbf{s}}|^{\frac{1}{2}} |V_{\mathbf{s}'}|^{\frac{1}{2}} \left| \frac{V_{\mathbf{s}} + V_{\mathbf{s}'}}{2} \right|^{-\frac{1}{2}} \exp(-Q(\mathbf{s}, \mathbf{s}')).$$

Chapter 4

Basics of areal data models

We now present a development of exploratory tools and modeling approaches that are customarily applied to data collected for areal units. Again, this literature is sometimes referred to as discrete spatial modeling to reflect the fact that we are only specifying a joint model for a finite set of random variables. We have in mind general, possibly irregular geographic units, but of course include the special case of regular grids of cells (pixels). Indeed, many of the ensuing models have been proposed for regular lattices of points and parameters, and sometimes even for point-referenced data (see Chapter 12 on the problem of inverting very large matrices).

In the context of areal units the general inferential issues are the following:

(i) Is there spatial pattern? If so, how strong is it? Intuitively, "spatial pattern" suggests that measurements for areal units which are near to each other will tend to take more similar values than those for units far from each other. Though you might "know it when you see it," this notion is evidently vague and in need of quantification. Indeed, with independent measurements for the units, we expect to see *no pattern*, i.e., a completely random arrangement of larger and smaller values. But again, randomness will inevitably produce some patches of similar values.

(ii) Do we want to smooth the data? If so, how much? Suppose, for example, that the measurement for each areal unit is a count, say, a number of cancers. Even if the counts were independent, and perhaps even after population adjustment, there would still be extreme values, as in any sample. Are the observed high counts more elevated than would be expected by chance? If we sought to present a surface of expected counts we might naturally expect that the high values would tend to be pulled down, the low values to be pushed up. This is the notion of smoothing. No smoothing would present a display using simply the observed counts. Maximal smoothing would result in a single common value for all units, clearly excessive. Suitable smoothing would fall somewhere in between, and take the spatial arrangement of the units into account.

Of course, how much smoothing is appropriate is not readily defined. In particular, for model-based smoothers such as we describe below, it is not evident what the extent of smoothing is, or how to control it. Specification of a utility function for smoothing (as attempted in Stern and Cressie, 1999) would help to address these questions but does not seem to be considered in practice..

(iii) For a new areal unit or set of units, how can we infer about what data values we expect to be associated with these units? That is, if we modify the areal units to new units, e.g., from zip codes to census block groups, what can we say about the cancer counts we expect for the latter, given those for the former? This is the so-called *modifiable areal unit problem (MAUP)*, which historically (and in most GIS software packages) is handled by crude areal allocation. Sections 7.2 and 7.3 propose model-based methodology for handling this problem.

As a matter of fact, in order to facilitate interpretation and better assess uncertainty, we will suggest model-based approaches to treat the above issues, as opposed to the more descriptive or algorithmic methods that have dominated the literature and are by now widely available in GIS software packages. We will also introduce further flexibility into these models by examining them in the context of regression. That is, we assume interest in explaining the areal unit responses and that we have available potential covariates to do this. These covariates may be available at the same or at different scales from the responses, but, regardless, we will now question whether there remains any spatial structure adjusted for these explanatory variables. This suggests that we may not try to model the data in a spatial way directly, but instead introduce spatial association through random effects. This will lead to versions of generalized linear mixed models (Breslow and Clayton, 1993). We will generally view such models in the hierarchical fashion that is the primary theme of this text.

4.1 Exploratory approaches for areal data

We begin with the presentation of some tools that can be useful in the initial exploration of areal unit data. The primary concept here is a *proximity matrix*, W. Given measurements $Y_1, \ldots, Y_n$ associated with areal units $1, 2, \ldots, n$, the entries w_{ij} in W spatially connect units i and j in some fashion. (Customarily w_{ii} is set to 0.) Possibilities include binary choices, i.e., $w_{ij} = 1$ if i and j share some common boundary, perhaps a vertex (as in a regular grid). Alternatively, w_{ij} could reflect "distance" between units, e.g., a decreasing function of intercentroidal distance between the units (as in a county or other regional map). But distance can be returned to a binary determination. For example, we could set $w_{ij} = 1$ for all i and j within a specified distance. Or, for a given i, we could get $w_{ij} = 1$ if j is one of the K nearest (in distance) neighbors of i. The preceding choices suggest that W would be symmetric. However, for irregular areal units, this last example provides a setting where this need not be the case. Also, the w_{ij}'s may be standardized by $\sum_j w_{ij} = w_{i+}$. If $\widetilde{W}$ has entries $\widetilde{w}_{ij} = w_{ij}/w_{i+}$, then evidently $\widetilde{W}$ is row stochastic, i.e., $\widetilde{W}\mathbf{1} = \mathbf{1}$, but now $\widetilde{W}$ need not be symmetric.

As the notation suggests, the entries in W can be viewed as weights. More weight will be associated with j's closer (in some sense) to i than those farther away from i. In this exploratory context (but, as we shall see, more generally) W provides the mechanism for introducing spatial structure into our formal modeling.

Lastly, working with distance suggests that we can define distance bins, say, $(0, d_1], (d_1, d_2], (d_2, d_3]$, and so on. This enables the notion of *first-order neighbors* of unit i, i.e., all units within distance d_1 of i, *second-order neighbors*, i.e., all units more than d_1 but at most d_2 from i, *third-order neighbors*, and so on. Analogous to W we can define $W^{(1)}$ as the proximity matrix for first-order neighbors. That is, $w_{ij}^{(1)} = 1$ if i and j are first-order neighbors, and equal to 0 otherwise. Similarly we define $W^{(2)}$ as the proximity matrix for second-order neighbors; $w_{ij}^{(2)} = 1$ if i and j are second-order neighbors, and 0 otherwise, and so on to create $W^{(3)}$, $W^{(4)}$, etc.

Of course, the most obvious exploratory data analysis tool for lattice data is a map of the data values. Figure 4.1 gives the statewide average verbal SAT exam scores as reported by the College Board and initially analyzed by Wall (2004). Clearly these data exhibit strong spatial pattern, with midwestern states and Utah performing best, and coastal states and Indiana performing less well. Of course, before jumping to conclusions, we must realize there are any number of spatial covariates that may help to explain this pattern; for instance, the percentage of eligible students taking the exam (Midwestern colleges have historically relied on the ACT exam, not the SAT, and only the best and brightest students in these states typically take the latter exam). Still, the map of these raw data shows significant spatial pattern.

Figure 4.1 *Choropleth map of 1999 average verbal SAT scores, lower 48 U.S. states and the district of Columbia.*

4.1.1 Measures of spatial association

Two standard statistics that are used to measure strength of spatial association among areal units are Moran's I and Geary's C (see, e.g., Ripley, 1981, Sec. 5.4). These are spatial analogues of statistics for measuring association in time series, the lagged autocorrelation coefficient and the Durbin-Watson statistic, respectively. They can also be seen to be areal unit analogues of the empirical estimates for the correlation function and the variogram, respectively. Recall that, for point-referenced data, the empirical covariance function (2.12) and semivariogram (2.9), respectively, provide customary nonparametric estimates of these measures of association.

Moran's I takes the form

$$I = \frac{n \sum_i \sum_j w_{ij}(Y_i - \overline{Y})(Y_j - \overline{Y})}{\left(\sum_{i \neq j} w_{ij}\right) \sum_i (Y_i - \overline{Y})^2} \,. \tag{4.1}$$

I is not strictly supported on the interval $[-1, 1]$. It is evidently a ratio of quadratic forms in $\mathbf{Y}$, which provides the idea for obtaining approximate first and second moments through the delta method (see, e.g., Agresti, 2002, Ch. 14). Moran shows under the null model where the Y_i are i.i.d., I is asymptotically normally distributed with mean $-1/(n-1)$ and a rather unattractive variance of the form

$$\mathrm{Var}(I) = \frac{n^2(n-1)S_1 - n(n-1)S_2 - 2S_0^2}{(n+1)(n-1)^2 S_0^2} \,. \tag{4.2}$$

In (4.2), $S_0 = \sum_{i \neq j} w_{ij}$, $S_1 = \frac{1}{2}\sum_{i \neq j}(w_{ij} + w_{ji})^2$, and $S_2 = \sum_k (\sum_j w_{kj} + \sum_i w_{ik})^2$. We recommend the use of Moran's I as an exploratory measure of spatial association, rather than as a "test of spatial significance."

For the data mapped in Figure 4.1, we used the `moran.test` function in the `spdep` package in `R` (see Section 4.5.2) to obtain a value for Moran's I of 0.6125, a reasonably large value. The associated standard error estimate of 0.0979 suggests very strong evidence against the null hypothesis of no spatial correlation in these data.

Geary's C takes the form

$$C = \frac{(n-1)\sum_i \sum_j w_{ij}(Y_i - Y_j)^2}{2\left(\sum_{i \neq j} w_{ij}\right)\sum_i (Y_i - \overline{Y})^2} \ . \tag{4.3}$$

C is never negative, and has mean 1 for the null model; *small* values (i.e., between 0 and 1) indicate *positive* spatial association. Also, C is a ratio of quadratic forms in $\mathbf{Y}$ and, like I, is asymptotically normal if the Y_i are i.i.d. We omit details of the distribution theory, recommending the interested reader to Cliff and Ord (1973), or Ripley (1981, p. 99).

Using the `geary.test` function on the SAT verbal data in Figure 4.1, we obtained a value of 0.3577 for Geary's C, with an associated standard error estimate of 0.0984. Again, the marked departure from the mean of 1 indicates strong positive spatial correlation in the data.

Convergence to asymptotic normality for a ration of quadratic forms is extremely slow. We may believe the significant rejection of independence using the asymptotic theory for the example above because the results are so extreme. However, if one truly seeks to run a significance test using (4.1) or (4.3), our recommendation is a Monte Carlo approach. Under the null model the distribution of I (or C) is invariant to permutation of the Y_i's. The exact null distribution of I (or C) requires computing its value under all $n!$ permutations of the Y_i's, infeasible for n in practice. However, a Monte Carlo sample of say 1000 permutations, including the observed one, will position the observed I (or C) relative to the remaining 999, to determine whether it is extreme (perhaps via an empirical p-value). Again using `spatial.cor` function on our SAT verbal data, we obtained empirical p-values of 0 using both Moran's I and Geary's C; *no* random permutation achieved I or C scores as extreme as those obtained for the actual data itself.

A further display that can be created in this spirit is the *correlogram*. Working with say, I, in (4.1) we can replace w_{ij} with the previously defined $w_{ij}^{(1)}$ and compute, say $I^{(1)}$. Similarly, we can replace w_{ij} with $w_{ij}^{(2)}$ and obtain $I^{(2)}$. A plot of $I^{(r)}$ vs. r is called a correlogram and, if spatial pattern is present, is expected to decline in r initially and then perhaps vary about 0. Evidently, this display is a spatial analogue of a temporal lag autocorrelation plot (e.g., see Carlin and Louis, 2000, p. 181). In practice, the correlogram tends to be very erratic and its information context is often not clear.

With large, regular grids of cells as we often obtain from remotely sensed imagery, it may be of interest to study spatial association in a particular direction (e.g., east-west, north-south, southwest-northeast, etc.). Now the spatial component reduces to one dimension and we can compute lagged autocorrelations (lagged appropriately to the size of the grid cells) in the specific direction. An analogue of this was proposed for the case where the Y_i are binary responses (e.g., presence or absence of forest in the cell) by Agarwal, Gelfand, and Silander (2002). In particular, Figure 4.2 shows rasterized maps of binary land use classifications for roughly 25,000 1 km $\times$ 1 km pixels in eastern Madagascar; see Agarwal et al. (2002) as well as Section 7.5 for further discussion.

While the binary map in Figure 4.2 shows spatial pattern in land use, we develop an additional display to provide quantification. For data on a regular grid or lattice, we calculate binary analogues of the sample autocovariances, using the 1 km $\times$ 1 km resolution with four illustrative directions: East (E), Northeast (NE), North (N), and Northwest (NW). In particular for any pair of pixels, we can identify, say, a Euclidean distance and direction between them by labeling one as X and the other as Y, creating a correlated binary pair. Then, we can go to the lattice and identify all pairs which share the same distance and direction. The collection of all such (X,Y) pairs yields a 2×2 table of counts (with table cells labeled as X=0, Y=0; X=0, Y=1; X=1, Y=0, X=1, Y=1). The resultant log-odds ratio measures the association between pairs in that direction at that distance. (Note that

Figure 4.2 *Rasterized north and south regions (1 km × 1 km) with binary land use classification overlaid.*

if we followed the same procedure but reversed direction, e.g., changed from E to W, the corresponding log odds ratio would be unchanged.)

In Figure 4.3, we plot log odds ratio against direction for each of the four directions. Note that the spatial association is quite strong, requiring a distance of at least 40 km before it drops to essentially 0. This suggests that we would not lose much spatial information if we work with the lower (4 km × 4 km) resolution. In exchange we obtain a richer response variable (17 ordered levels, indicating number of forested cells from 0 to 16) and a substantial reduction in number of pixels (from 26,432 to $1,652$ in the north region, from 24,544 to $1,534$ in the south region) to facilitate model fitting.

4.1.2 Spatial smoothers

Recall from the beginning of this chapter that often a goal for, say, a choropleth map of the Y_i's is *smoothing*. Depending upon the number of classes used to make the map, there is already some implicit smoothing in such a display (although this is not *spatial* smoothing, of course).

The W matrix directly provides a spatial smoother; that is, we can replace Y_i by $\widehat{Y}_i = \sum_j w_{ij} Y_j / w_{i+}$. This ensures that the value for areal unit i "looks like" its neighbors, and that the more neighbors we use in computing $\widehat{Y}_i$, the more smoothing we will achieve. In fact, $\widehat{Y}_i$ may be viewed as an unusual smoother in that it ignores the value actually observed for unit i. As such, we might revise the smoother to

$$\widehat{Y}_i^* = (1-\alpha) Y_i + \alpha \widehat{Y}_i \ , \tag{4.4}$$

where $\alpha \in (0,1)$. Working in an exploratory mode, various choices may be tried for α, but for any of these, (4.4) is a familiar *shrinkage* form. Thus, under a specific model with a suitable loss function, an optimal α could be sought. Finally, the form (4.4), viewed generally as

Figure 4.3 *Land use log-odds ratio versus distance in four directions.*

a linear combination of the Y_j, is customarily referred to as a *filter* in the GIS literature. In fact, such software will typically provide choices of filters, and even a default filter to automatically smooth maps.

In Section 5.1 we will present a general discussion revealing how smoothing (shrinkage) emerges as a byproduct of the hierarchical models we propose to use to explain the Y_i. In particular, when W is used in conjunction with a stochastic model (as in Section 4.3), the $\widehat{Y}_i$ are updated across i and across Monte Carlo iterations as well. So the observed Y_i will affect the eventual $\widehat{Y}_i$; we achieve model-driven smoothing and a "manual" inclusion of Y_i as in (4.4) with some choice of α is unnecessary.

4.2 Brook's Lemma and Markov random fields

A useful technical result for obtaining the joint distribution of the Y_i in some of the models we discuss below is *Brook's Lemma* (Brook, 1964). The usefulness of this lemma is exposed in Besag's (1974) seminal paper on conditionally autoregressive models.

It is clear that given $p(y_1, \ldots, y_n)$, the so-called *full conditional* distributions, $p(y_i|y_j, j \neq i)$, $i = 1, \ldots, n$, are uniquely determined. Brook's Lemma demonstrates the converse and, in fact, enables us to constructively retrieve the unique joint distribution determined by these full conditionals. But first, it is also clear that we cannot write down an arbitrary set of full conditional distributions and assert that they uniquely determine the joint distribution. To see this, let $Y_1|Y_2 \sim N(\alpha_0 + \alpha_1 Y_2, \sigma_1^2)$ and let $Y_2|Y_1 \sim N(\beta_0 + \beta_1 Y_1^3, \sigma_2^2)$, where N denotes the normal (Gaussian) distribution. Intuitively, it seems that a mean for Y_1 given Y_2 which is linear in Y_2 is incompatible with a mean for Y_2 given Y_1 which is linear in Y_1^3. More formally, we see that

$$E(Y_1) = E[E(Y_1|Y_2)] = E[\alpha_0 + \alpha_1 Y_2] = \alpha_0 + \alpha_1 E(Y_2) \ , \tag{4.5}$$

i.e., $E(Y_1)$ and $E(Y_2)$ are linearly related. But in fact, it must also be the case that

$$E(Y_2) = E[E(Y_2|Y_1)] = E[\beta_0 + \beta_1 Y_1] = \beta_0 + \beta_1 E(Y_1^3) \ . \tag{4.6}$$

Equations (4.5) and (4.6) could simultaneously hold only in trivial cases, so the two mean specifications are *incompatible*. Thus we can say that $f(y_1|y_2)$ and $f(y_2|y_1)$ are incompatible with regard to determining $p(y_1, y_2)$. We do not examine conditions for compatibility of conditional distributions here, although there has been considerable work in this area (see, e.g., Arnold and Strauss, 1991, and references therein).

Another point is that $p(y_1 \ldots, y_n)$ may be improper even if $p(y_i|y_j, j \neq i)$ is proper for all i. As an elementary illustration, consider $p(y_1, y_2) \propto \exp[-\frac{1}{2}(y_1 - y_2)^2]$. Evidently $p(y_1|y_2)$ is $N(y_2, 1)$ and $p(y_2|y_1)$ is $N(y_1, 1)$, but $p(y_1, y_2)$ is improper. Casella and George (1992) provide a similar example in a bivariate exponential (instead of normal) setting.

Brook's Lemma notes that

$$\begin{aligned} p(y_1, \ldots, y_n) \quad = \quad & \frac{p(y_1|y_2, \ldots, y_n)}{p(y_{10}|y_2, \ldots, y_n)} \cdot \frac{p(y_2|y_{10}, y_3, \ldots, y_n)}{p(y_{20}|y_{10}, y_3, \ldots, y_n)} \\ & \cdots \frac{p(y_n|y_{10}, \ldots, y_{n-1,0})}{p(y_{n0}|y_{10}, \ldots, y_{n-1,0})} \cdot p(y_{10}, \ldots, y_{n0}) \ , \end{aligned} \tag{4.7}$$

an identity which is easily checked (Exercise 1). Here, $\mathbf{y}_0 = (y_{10}, \ldots, y_{n0})'$ is any fixed point in the support of $p(y_1, \ldots, y_n)$. Hence $p(y_1, \ldots, y_n)$ is determined by the full conditional distributions, since apart from the constant $p(y_{10}, \ldots, y_{n0})$ they are the only objects appearing on the right-hand side of (4.7). Hence the joint distribution is determined up to a proportionality constant. If $p(y_1, \ldots, y_n)$ is improper then this is, of course, the best we can do; if $p(y_1, \ldots, y_n)$ is proper then the fact that it integrates to 1 determines the constant. Perhaps most important is the constructive nature of (4.7): we can create $p(y_1, \ldots, y_n)$ simply by calculating the product of ratios. For more on this point, see Exercise 2.

When the number of areal units is very large (say, a regular grid of pixels associated with an image or a large number of small geographic regions), we do not seek to write down the joint distribution of the Y_i. Rather we prefer to work (and model) exclusively with the n corresponding full conditional distributions. In fact, from a spatial perspective, we would imagine that the full conditional distribution for Y_i would be more "local," that is, it should really depend only upon the neighbors of cell i. Adopting some definition of a neighbor structure (e.g., the one setting $W_{ij} = 1$ or 0 depending on whether i and j are adjacent or not), let ∂_i denote the set of neighbors of cell i.

Next suppose we specify a set of full conditional distributions for the Y_i such that

$$p(y_i|y_j, j \neq i) = p(y_i|y_j, j \in \partial_i) \tag{4.8}$$

A critical question to ask is whether a specification such as (4.8) uniquely determines a joint distribution for $Y_1, \ldots Y_n$. That is, we do not need to see the explicit form of this distribution. We merely want to be assured that if, for example, we implement a Gibbs sampler (see Subsection 5.3.1) to simulate realizations from the joint distribution, that there is indeed a unique stationary distribution for this sampler.

The notion of using *local* specification to determine a joint (or global) distribution in the form (4.8) is referred to as a *Markov random field* (MRF). There is by now a substantial literature in this area, with Besag (1974) being a good place to start. Geman and Geman (1984) provide the next critical step in the evolution, while Kaiser and Cressie (2000) offer a current view and provide further references. See also Rue and Held (2005) and references therein.

A critical definition in this regard is that of a *clique*. A clique is a set of cells (equivalently, indices) such that each element is a neighbor of every other element. With n cells, depending upon the definition of the neighbor structure, cliques can possibly be of size 1, 2, and so on up to size n. A *potential function* (or simply *potential*) of order k is a function of k arguments that is exchangeable in these arguments. The arguments of the potential would be the values taken by variables associated with the cells for a clique of size k. For continuous

Y_i, a customary potential on cliques of size $k = 2$ is $(Y_i - Y_j)^2$ when $i \sim j$. (We use the notation $i \sim j$ if i is a neighbor of j and j is a neighbor of i.) In fact, we may also view this potential as a sum of a potential on cliques of size $k = 1$, i.e., Y_i^2 with a potential on cliques of size $k = 2$, i.e., Y_iY_j. For, say, binary Y_i, a common potential on cliques of size $k = 2$ is

$$I(Y_i = Y_j) = Y_iY_j + (1 - Y_i)(1 - Y_j) \ ,$$

where again $i \sim j$ and I denotes the indicator function. Next, we define a *Gibbs distribution* as follows: $p(y_1, \ldots, y_n)$ is a Gibbs distribution if it is a function of the Y_i only through potentials on cliques. That is,

$$p(y_1, \ldots, y_n) \propto \exp\left\{\gamma \sum_k \sum_{\boldsymbol{\alpha} \in \mathcal{M}_k} \phi^{(k)}(y_{\alpha_1}, y_{\alpha_2}, \ldots, y_{\alpha_k})\right\} \ . \tag{4.9}$$

Here, $\phi^{(k)}$ is a potential of order k, $\mathcal{M}_k$ is the collection of all subsets of size k from $\{1, 2, \ldots, n\}$, $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_k)'$ indexes this set, and $\gamma > 0$ is a scale (or "temperature") parameter. If we only use cliques of size 1, we see that we obtain an independence model, evidently not of interest. When $k = 2$, we achieve spatial structure. In practice, cliques with $k = 3$ or more are rarely used, introducing complexity with little benefit. So, throughout this book, only cliques of order less than or equal to 2 are considered.

Informally, the *Hammersley-Clifford Theorem* (see Besag, 1974; also Clifford, 1990) demonstrates that if we have an MRF, i.e., if (4.8) defines a unique joint distribution, then this joint distribution is a Gibbs distribution. That is, it is of the form (4.9), with all of its "action" coming in the form of potentials on cliques. Cressie (1993, pp. 417–18) offers a proof of this theorem, and mentions that its importance for spatial modeling lies in its limiting the complexity of the conditional distributions required, i.e., full conditional distributions can be specified locally.

Geman and Geman (1984) provided essentially the converse of the Hammersley-Clifford Theorem. If we begin with (4.9) we have determined an MRF. As a result, they argued that to sample a Markov random field, one could sample from its associated Gibbs distribution, hence coining the term "Gibbs sampler."

For continuous data on $\Re^1$, a common choice for the joint distribution is a pairwise difference form

$$p(y_1, \ldots, y_n) \propto \exp\left\{-\frac{1}{2\tau^2} \sum_{i,j} (y_i - y_j)^2 I(i \sim j)\right\} \ . \tag{4.10}$$

Distributions such as (4.10) will be the focus of the next section. For the moment, we merely note that it is a Gibbs distribution on potentials of order 1 and 2 and that

$$p(y_i \,\Big|\, y_j, j \neq i) = N\left(\sum_{j \in \partial_i} y_i/m_i \ , \ \tau^2/m_i\right) \ , \tag{4.11}$$

where m_i is the number of neighbors of cell i. The distribution in (4.11) is clearly of the form (4.8) and shows that the mean of Y_i is the average of its neighbors, exactly the sort of local smoother we discussed in the section on spatial smoothers.

4.3 Conditionally autoregressive (CAR) models

Although they were introduced by Besag (1974) approximately 30 years ago, conditionally autoregressive (CAR) models have enjoyed a dramatic increase in usage only in the past decade or so. This resurgence arises from their convenient employment in the context of Gibbs sampling and more general Markov chain Monte Carlo (MCMC) methods for fitting certain classes of hierarchical spatial models (seen, e.g., in Section 6.4.3).

4.3.1 The Gaussian case

We begin with the Gaussian (or *autonormal*) case. Suppose we set

$$Y_i \mid y_j, j \neq i \sim N\left(\sum_j b_{ij} y_j \ , \ \tau_i^2\right) \ , \ i = 1, \ldots, n \ . \tag{4.12}$$

These full conditionals are compatible, so through Brook's Lemma we can obtain

$$p(y_1, \ldots, y_n) \propto \exp\left\{-\frac{1}{2}\mathbf{y}' D^{-1}(I - B)\mathbf{y}\right\} \ , \tag{4.13}$$

where $B = \{b_{ij}\}$ and D is diagonal with $D_{ii} = \tau_i^2$. Expression (4.13) suggests a joint multivariate normal distribution for $\mathbf{Y}$ with mean $\mathbf{0}$ and variance matrix $\Sigma_{\mathbf{y}} = (I - B)^{-1} D$.

But we are getting ahead of ourselves. First, we need to ensure that $D^{-1}(I - B)$ is symmetric. The resulting conditions are

$$\frac{b_{ij}}{\tau_i^2} = \frac{b_{ji}}{\tau_j^2} \quad \text{for all} \ \ i, j \ . \tag{4.14}$$

Evidently, from (4.14), B need not be symmetric. Returning to our proximity matrix W (which we assume to be symmetric), suppose we set $b_{ij} = w_{ij}/w_{i+}$ and $\tau_i^2 = \tau^2/w_{i+}$. Then (4.14) is satisfied and (4.12) yields $p(y_i|y_j, j \neq i) = N\left(\sum_j w_{ij} y_j / w_{i+} \ , \ \tau^2/w_{i+}\right)$. Also, (4.13) becomes

$$p(y_1, \ldots, y_n) \propto \exp\left\{-\frac{1}{2\tau^2}\mathbf{y}'(D_w - W)\mathbf{y}\right\} \ , \tag{4.15}$$

where D_w is diagonal with $(D_w)_{ii} = w_{i+}$.

Now a second aspect is noticed. $(D_w - W)\mathbf{1} = \mathbf{0}$, i.e., $\Sigma_{\mathbf{y}}^{-1}$ is singular, so that $\Sigma_{\mathbf{y}}$ does not exist and the distribution in (4.15) is improper. (The reader is encouraged to note the difference between the case of $\Sigma_{\mathbf{y}}^{-1}$ singular and the case of $\Sigma_{\mathbf{y}}$ singular. With the former we have a density function but one that is not integrable; effectively we have too many variables and we need a constraint on them to restore propriety. With the latter we have no density function but a proper distribution that resides in a lower dimensional space; effectively we have too *few* variables.) With a little algebra (4.15) can be rewritten as

$$p(y_1, \ldots, y_n) \propto \exp\left\{-\frac{1}{2\tau^2}\sum_{i \neq j} w_{ij}(y_i - y_j)^2\right\} \ . \tag{4.16}$$

This is a pairwise difference specification slightly more general than (4.10). But the impropriety of $p(\mathbf{y})$ is also evident from (4.16) since we can add any constant to all of the Y_i and (4.16) is unaffected; the Y_i are not "centered." A constraint such as $\sum_i Y_i = 0$ would provide the needed centering. Thus we have a more general illustration of a joint distribution that is improper, but has all full conditionals proper. The specification in (4.15) or (4.16) is often referred to as an *intrinsically autoregressive* (IAR) model.

As a result, $p(\mathbf{y})$ in (4.15) cannot be used as a model for data; data could not arise under an improper stochastic mechanism, and we cannot impose a constant center on randomly realized measurements. Hence, the use of an improper autonormal model must be relegated to a *prior* distributional specification. That is, it will be attached to random spatial effects introduced at the second stage of a hierarchical specification (again, see, e.g., Section 6.4.3).

The impropriety in (4.15) can be remedied in an obvious way. Redefine $\Sigma_{\mathbf{y}}^{-1} = D_w - \rho W$ and choose ρ to make $\Sigma_{\mathbf{y}}^{-1}$ nonsingular. This is guaranteed if $\rho \in \left(1/\lambda_{(1)}, 1/\lambda_{(n)}\right)$, where

$\lambda_{(1)} < \lambda_{(2)} < \cdots < \lambda_{(n)}$ are the ordered eigenvalues of $D_w^{-1/2} W D_w^{-1/2}$; see Exercise 5; Moreover, since $tr(D_w^{-1/2} W D_w^{-1/2}) = 0 = \sum_{i=1}^n \lambda_{(i)}$, $\lambda_{(1)} < 0$, $\lambda_{(n)} > 0$, and 0 belongs to $\left(1/\lambda_{(1)}, 1/\lambda_{(n)}\right)$.

Simpler bounds than those given above for the propriety parameter ρ may be obtained if we replace the adjacency matrix W by the scaled adjacency matrix $\widetilde{W} \equiv Diag(1/w_{i+})W$; recall $\widetilde{W}$ is not symmetric, but it will be row stochastic (i.e., all of its rows sum to 1). $\Sigma_{\mathbf{y}}^{-1}$ can then be written as $M^{-1}(I - \alpha\widetilde{W})$ where M is diagonal. Then if $|\alpha| < 1$, $I - \alpha\widetilde{W}$ is nonsingular. (See the SAR model of the next section, as well as Exercise 9;) Carlin and Banerjee (2003) show that $\Sigma_{\mathbf{y}}^{-1}$ is diagonally dominant and symmetric. But diagonally dominant symmetric matrices are positive definite (Harville, 1997), providing an alternative argument for the propriety of the joint distribution.

An elegant way to look at the propriety of CAR models is through the Gershgorin disk theorem (Golub and Van Loan, 2012; Horn and Johnson, 2012). This famous theorem of linear algebra focuses on so-called *diagonal dominance* and, in its simplest form, asserts that, for any symmetric matrix A, if all $a_{ii} > 0$ and $a_{ii} > \sum_{j \neq i} |a_{ij}|$, then A is positive definite. For instance, if $D_w^{-1}(I - B)$ is symmetric, then it is positive definite if, for each i, $\sum_{j \neq i} |B_{ij}| < 1$. With $D_w - \rho W$, a sufficient condition is that $|\rho| < 1$, weaker than the conditions created above. However, since we have positive definiteness for $\rho < 1$, impropriety at $\rho = 1$, this motivates us to examine the behavior of CAR models for ρ near 1, as we do below.

Returning to the unscaled situation, ρ can be viewed as an additional parameter in the CAR specification, enriching this class of spatial models. Furthermore, $\rho = 0$ has an immediate interpretation: the Y_i become independent $N(0, \tau^2/w_{i+})$. If ρ is not included, independence cannot emerge as a limit of (4.15). (Incidentally, this suggests a clarification of the role of τ^2, the variance parameter associated with the full conditional distributions: the magnitude of τ^2 should *not* be viewed as, in any way, *quantifying* the strength of spatial association. Indeed if all Y_i are multiplied by c, τ^2 becomes $c\tau^2$ but the strength of spatial association among the Y_i is clearly unaffected.) Lastly, $\rho \sum_j w_{ij} Y_j / w_{i+}$ can be viewed as a *reaction function*, i.e., ρ is the expected proportional "reaction" of Y_i to $\sum_j w_{ij} Y_j / w_{i+}$. (This interpretation is more common in the SAR literature (Section 4.4).)

With these advantages plus the fact that $p(\mathbf{y})$ (or the Bayesian posterior distribution, if the CAR specification is used to model constrained random effects) is now proper, is there any reason not to introduce the ρ parameter? In fact, the answer may be yes. Under $\Sigma_{\mathbf{y}}^{-1} = D_w - \rho W$, the full conditional $p(y_i | y_j, j \neq i)$ becomes $N\left(\rho \sum_j w_{ij} y_j / w_{i+}\ ,\ \tau^2/w_{i+}\right)$. Hence we are modeling Y_i not to have mean that is an average of its neighbors, but some *proportion* of this average. Does this enable any sensible spatial interpretation for the CAR model? Moreover, does ρ calibrate very well with any familiar interpretation of "strength of spatial association"? Fixing $\tau^2 = 1$ without loss of generality, we can simulate CAR realizations for a given n, W, and ρ. We can also compute for these realizations a descriptive association measure such as Moran's I or Geary's C. Here we do not present explicit details of the range of simulations we have conducted. However, for a 10×10 grid using a first-order neighbor system, when $\rho = 0.8$, I is typically 0.1 to 0.15; when $\rho = 0.9$, I is typically 0.2 to 0.25; and even when $\rho = 0.99$, I is typically at most 0.5. It thus appears that ρ can mislead with regard to strength of association. Expressed in a different way, within a Bayesian framework, a prior on ρ that encourages a consequential amount of spatial association would place most of its mass near 1.

A related point is that if $p(\mathbf{y})$ is proper, the breadth of spatial pattern may be too limited. In the case where a CAR model is applied to random effects, an improper choice may actually enable wider scope for posterior spatial pattern. As a result, we do not take a position with regard to propriety or impropriety in employing CAR specifications (though

in the remainder of this text we do sometimes attempt to illuminate relative advantages and disadvantages).

Referring to (4.12), we may write the entire system of random variables as

$$\mathbf{Y} = B\mathbf{Y} + \boldsymbol{\epsilon}\,, \quad \text{or equivalently,} \tag{4.17}$$

$$(I - B)\mathbf{Y} = \boldsymbol{\epsilon}\,. \tag{4.18}$$

In particular, the distribution for $\mathbf{Y}$ induces a distribution for $\boldsymbol{\epsilon}$. If $p(\mathbf{y})$ is proper then $\mathbf{Y} \sim N(\mathbf{0}, (I-B)^{-1}D)$ whence $\boldsymbol{\epsilon} \sim N(\mathbf{0}, D(I-B)^T)$, i.e., the components of $\boldsymbol{\epsilon}$ are not independent. Also, $Cov(\boldsymbol{\epsilon}, \mathbf{Y}) = D$. The SAR specification in Section 4.4 reverses this specification, supplying a distribution for $\boldsymbol{\epsilon}$ which induces a distribution for $\mathbf{Y}$.

When $p(\mathbf{y})$ is proper we can appeal to standard multivariate normal distribution theory to interpret the entries in $\Sigma_{\mathbf{y}}^{-1}$. For example, $1/(\Sigma_{\mathbf{y}}^{-1})_{ii} = Var(Y_i|Y_j, j \neq i)$. Of course with $\Sigma_{\mathbf{y}}^{-1} = D^{-1}(I-B)$, $(\Sigma_{\mathbf{y}}^{-1})_{ii} = 1/\tau_i^2$ providing immediate agreement with (4.12). But also, if $(\Sigma_{\mathbf{y}}^{-1})_{ij} = 0$, then Y_i and Y_j are conditionally independent given $Y_k, k \neq i, j$, a fact you are asked to show in Exercise 10. Hence if any $b_{ij} = 0$, we have conditional independence for that pair of variables. Connecting b_{ij} to w_{ij} shows that the choice of neighbor structure implies an associated collection of conditional independences. With first-order neighbor structure, all we are asserting is a spatial illustration of the local Markov property (Whittaker, 1990, p. 68).

We conclude this subsection with four remarks. First, one can directly introduce a regression component into (4.12), e.g., a term of the form $\mathbf{x}_i'\boldsymbol{\beta}$. Conditional on $\boldsymbol{\beta}$, this does not affect the association structure that ensues from (4.12); it only revises the mean structure. However, we omit details here (the interested reader can consult Besag, 1974), since we will only use the autonormal CAR as a distribution for spatial random effects. These effects are added onto the regression structure for the mean on some transformed scale (again, see Section 6.4.3).

We also note that in suitable contexts it may be appropriate to think of $\mathbf{Y}_i$ as a vector of dependent areal unit measurements or, in the context of random effects, as a vector of dependent random effects associated with an areal unit. This leads to the specification of multivariate conditionally autoregressive (MCAR) models, which is the subject of Section 10.1. From a somewhat different perspective, $\mathbf{Y}_i$ might arise as $(Y_{i1}, \ldots, Y_{iT})^T$ where Y_{it} is the measurement associated with areal unit i at time t, $t = 1, \ldots, T$. Now we would of course think in terms of spatiotemporal modeling for Y_{it}. This is the subject of Section 11.7.

Thirdly, a (proper) CAR model can in principle be used for point-level data, taking w_{ij} to be, say, an inverse distance between points i and j. However, unlike the spatial prediction described in Section 2.4, now spatial prediction becomes *ad hoc*. That is, to predict at a new site Y_0, we might specify the distribution of Y_0 given $Y_1, \ldots, Y_n$ to be a normal distribution, such as a $N\left(\rho \sum_j w_{0j} y_j / w_{0+}\,,\ \tau^2/w_{0+}\right)$. Note that this determines the joint distribution of $Y_0, Y_1, \ldots, Y_n$. However, this joint distribution is *not* the CAR distribution that would arise by specifying the full conditionals for $Y_0, Y_1, \ldots, Y_n$ and using Brook's Lemma, as in constructing (4.15). In this regard, we cannot "marginalize" a CAR model. That is, suppose we specify a CAR model for, say, n areal units and we want a CAR model for a subset of them, say the first m. If we consider the multivariate normal distribution with upper left $m \times m$ block $(D^{-1}(I-B))_m^{-1}$, the inverse of this matrix need not look anything like the CAR model for these m units.

Finally, Gaussian Markov random fields can introduce proximities more general than those that we have discussed here. In particular, working with a regular lattice, there is much scope for further theoretical development. For instance, Rue and Held (2005, p. 114) describe the derivation of the following model weights based on the forward difference analogue of penalizing the derivatives of a surface used to specify the thin plate spline. They consider 12 neighbors of a given point. The north, east, south, and west neighbors

each receive a weight of +8, the northeast, southeast, southwest, and northwest neighbors, each receive a weight of −2 and the "two away" north, east, south, and west neighbors, each receive a weight of −1. Thus, the $w_{i+} = 20$. These weights would possibly viewed as unusual with regard to spatial smoothing, in particular the negative values, but, again, they do have a probabilistic justification through the two-dimensional random walk on the lattice. Moreover, they do play a role in Markov random field approximation to Gaussian processes. Some care needs to be taken with regard to edge specifications. See further discussion in Rue and Held (2005).

4.3.2 The non-Gaussian case

If one seeks to model the data directly using a CAR specification, then in many cases a normal distribution would not be appropriate. Binary response data and sparse count data are two examples. In fact, one might imagine any exponential family model as a first-stage distribution for the data. Here, we focus on the case where the Y_i are binary variables and present the so-called autologistic CAR model (historically, the Ising model; see Brush, 1967). This model has received attention in the literature; see, e.g., Heikkinen and Hogmander (1994), Hogmander and Møller (1995), and Hoeting et al. (2000). Ignoring covariates for the moment, as we did with the CAR models above, consider the joint distribution

$$\begin{aligned} p(y_1, y_2, ..., y_n; \psi) &\propto \exp(\psi \sum_{i,j} w_{ij} 1(y_i = y_j)) \\ &= \exp(\psi \sum_{i,j} w_{ij}(y_i y_j + (1 - y_i)(1 - y_j))). \end{aligned} \tag{4.19}$$

We immediately recognize this specification as a Gibbs distribution with a potential on cliques of order $k = 2$. Moreover, this distribution is always proper since it can take on only 2^n values. However, we will assume that ψ is an unknown parameter (how would we know it in practice?) and hence we will need to calculate the normalizing constant $c(\psi)$ in order to infer about ψ. But, computation of this constant requires summation over all of the 2^n possible values that $(Y_1, Y_2, ..., Y_n)$ can take on. Even for moderate sample sizes this will present computational challenges. Hoeting et al. (2000) propose approximations to the likelihood using a pseudo-likelihood and a normal approximation.

From (4.19) we can obtain the full conditional distributions for the Y_i's. In fact, $P(Y_i = 1|y_j, j \neq i) = e^{\psi S_{i,1}}/(e^{\psi S_{i,0}} + e^{\psi(S_{i,1})})$ where $S_{i,1} = \sum_{j\sim i} 1(y_j = 1)$ and $S_{i,0} = \sum_{j\sim i} 1(y_j = 0)$ and $P(Y_i = 0|y_j, j \neq i) = 1 - P(Y_i = 1|y_j, j \neq i)$. That is, $S_{i,1}$ is the number of neighbors of i that are equal to 1 and $S_{i,0}$ is the number of neighbors of i that are equal to 0. We can see the role that ψ plays; larger values of ψ place more weight on matching. This is most easily seen through $\log \frac{P(Y_i=1|y_j,j\neq i)}{P(Y_i=0|y_j,j\neq i)} = \psi(S_{i,1} - S_{i,0})$. Since the full conditional distributions take on only two values, there are no normalizing issues with them.

Bringing in covariates is natural on the log scale, i.e.,

$$\log \frac{P(Y_i = 1|y_j, j \neq i)}{P(Y_i = 0|y_j, j \neq i)} = \psi(S_{i,1} - S_{i,0}) + \mathbf{X}_i^T \boldsymbol{\beta}. \tag{4.20}$$

Solving for $P(Y_i = 1|y_j, j \neq i)$, we obtain

$$P(Y_i = 1 \mid y_j, j \neq i) = \frac{\exp\{\psi(S_{i,1} - S_{i,0}) + \mathbf{X}_i^T \boldsymbol{\beta}\}}{1 + \exp\{(S_{i,1} - S_{i,0}) + \mathbf{X}_i^T \boldsymbol{\beta}\}} .$$

Now, to update both ψ and $\boldsymbol{\beta}$, we will again need the normalizing constant, now $c(\psi, \boldsymbol{\beta})$. In fact, we leave as an exercise, the joint distribution of $(Y_1, Y_2, ..., Y_n)$ up to a constant.

The case where Y_i can take on one of several categorical values presents a natural extension to the autologistic model. If we label the (say) L possible outcomes as simply $1, 2, ..., L$, then we can define the joint distribution for $(Y_1, Y_2, ..., Y_n)$ exactly as in (4.19), i.e.,

$$p(y_1, y_2, ..., y_n; \psi) \propto \exp(\psi \sum_{i,j} w_{ij} 1(y_i = y_j)) \tag{4.21}$$

with w_{ij} as above. The distribution in (4.21) is referred to as a *Potts model* (Potts, 1952). Now the distribution takes on L^n values; now, calculation of the normalizing constant is even more difficult. Because of this challenge, fitting Potts models to data is rare in the literature; rather, it is customary to run a forward simulation using the Potts model since this only requires implementing a routine Gibbs sampler, updating the Y_i's (see Chapter 5) for a fixed ψ. However, one nice data analysis example is the allocation model in Green and Richardson (2002). There, the Potts model is employed as a random effects specification, in a disease mapping context (see Chapter 6), as an alternative to a CAR model.

4.4 Simultaneous autoregressive (SAR) models

Returning to (4.17), suppose that instead of letting $\mathbf{Y}$ induce a distribution for $\boldsymbol{\epsilon}$, we let $\boldsymbol{\epsilon}$ induce a distribution for $\mathbf{Y}$. Imitating usual autoregressive time series modeling, suppose we take the ϵ_i to be independent innovations. For a little added generality, assume that $\boldsymbol{\epsilon} \sim N\left(0, \tilde{D}\right)$ where $\tilde{D}$ is diagonal with $\left(\tilde{D}\right)_{ii} = \sigma_i^2$. (Note $\tilde{D}$ has no connection with D in Section 4.3; the B we use below may or may not be the same as the one we used in that section.) Analogous to (4.12), now $Y_i = \sum_j b_{ij} Y_j + \epsilon_i$, $i = 1, 2, ..., n$, with $\epsilon_i \sim N\left(0, \sigma_i^2\right)$ or, equivalently, $(I - B)\mathbf{Y} = \boldsymbol{\epsilon}$ with $\boldsymbol{\epsilon}$ distributed as above. Therefore, if $(I - B)$ is full rank,

$$\mathbf{Y} \sim N\left(\mathbf{0}\,,\, (I - B)^{-1} \tilde{D}\, ((I - B)^{-1})^T\right)\,. \tag{4.22}$$

Also, $Cov(\boldsymbol{\epsilon}, \mathbf{Y}) = \tilde{D}(I - B)^{-1}$. If $\tilde{D} = \sigma^2 I$ then (4.22) simplifies to $\mathbf{Y} \sim N\left(\mathbf{0}\,,\, \sigma^2 \left[(I - B)(I - B)^T\right]^{-1}\right)$. In order that (4.22) be proper, $I - B$ must be full rank but not necessarily symmetric. Two choices are most frequently discussed in the literature (e.g., Griffith, 1988). The first assumes $B = \rho W$, where W is a so-called contiguity matrix, i.e., W has entries that are 1 or 0 according to whether or not unit i and unit j are direct neighbors (with $w_{ii} = 0$). So W is our familiar first-order neighbor proximity matrix. Here ρ is called a *spatial autoregression parameter* and, evidently, $Y_i = \rho \sum_j Y_j I(j \in \partial_i) + \epsilon_i$, where ∂_i denotes the set of neighbors of i. In fact, any symmetric proximity matrix can be used and, paralleling the discussion below (4.15), $I - \rho W$ will be nonsingular if $\rho \in \left(\frac{1}{\lambda_{(1)}}, \frac{1}{\lambda_{(n)}}\right)$ where now $\lambda_{(1)} < \cdots < \lambda_{(n)}$ are the ordered eigenvalues of W. As a weaker conclusion, if W is symmetric, we can apply the diagonal dominance result from Section 4.3.1. Now, if $\rho \sum_{j \neq i} w_{ij} < 1$ for each i, i.e., $\rho < \min \frac{1}{w_{i+}}$, we have positive definiteness, hence nonsingularity.

Alternatively, W can be replaced by $\widetilde{W}$ where now, for each i, the ith row has been normalized to sum to 1. That is, $\left(\tilde{W}\right)_{ij} = w_{ij}/w_{i+}$. Again, $\widetilde{W}$ is not symmetric, but it is row stochastic, i.e., $\widetilde{W}\mathbf{1} = \mathbf{1}$. If we set $B = \alpha\widetilde{W}$, α is called a *spatial autocorrelation parameter* and, were W a contiguity matrix, now $Y_i = \alpha \sum_j Y_i I(j \in \partial_i)/w_{i+} + \epsilon_i$. With a very regular grid the w_{i+} will all be essentially the same and thus α will be a multiple of ρ. But, perhaps more importantly, with $\widetilde{W}$ row stochastic the eigenvalues of $\widetilde{W}$ are all less than or equal to 1 (i.e., $\max |\lambda_i| = 1$). Thus $I - \alpha\widetilde{W}$ will be nonsingular if $\alpha \in (-1, 1)$, justifying referring to α as an autocorrelation parameter; see Exercise 9.

A SAR model is customarily introduced in a regression context, i.e., the *residuals* $\mathbf{U} = \mathbf{Y} - X\boldsymbol{\beta}$ are assumed to follow a SAR model, rather than $\mathbf{Y}$ itself. But then, following (4.17), if $\mathbf{U} = B\mathbf{U} + \boldsymbol{\epsilon}$, we obtain the attractive form

$$\mathbf{Y} = B\mathbf{Y} + (I - B)\, X\boldsymbol{\beta} + \boldsymbol{\epsilon}\,. \tag{4.23}$$

Expression (4.23) shows that $\mathbf{Y}$ is modeled through a component that provides a spatial weighting of neighbors and a component that is a usual linear regression. If B is the zero matrix we obtain an OLS regression; if $B = I$ we obtain a purely spatial model.

We note that from (4.23) the SAR model does not introduce any spatial effects; the errors in (4.23) are independent. Expressed in a different way, if we modeled $\mathbf{Y} - X\boldsymbol{\beta}$ as $\mathbf{U} + \mathbf{e}$ with $\mathbf{e}$ independent errors, we would have $\mathbf{U} + \mathbf{e} = B\mathbf{U} + \boldsymbol{\epsilon} + \mathbf{e}$ and $\boldsymbol{\epsilon} + \mathbf{e}$ would result in a redundancy. Equivalently, if we write $\mathbf{U} = B\mathbf{U} + \boldsymbol{\epsilon}$, we see, from the distribution of $B\mathbf{U}$, that both terms on the right side are driven by the same variance component, σ_ϵ^2. As a result, in practice a SAR specification is not used in conjunction with a GLM. To introduce $\mathbf{U}$ as a vector of spatial adjustments to the mean vector, a transformed scale creates redundancy between the independent Gaussian error in the definition of the U_i and the stochastic mechanism associated with the conditionally independent Y_i.

We briefly note the somewhat related spatial modeling approach of Langford et al. (1999). Rather than modeling the residual vector $\mathbf{U} = B\mathbf{U} + \boldsymbol{\epsilon}$, they propose that $\mathbf{U} = \tilde{B}\boldsymbol{\epsilon}$ where $\boldsymbol{\epsilon} \sim N\left(\mathbf{0}, \sigma^2 I\right)$, i.e., that $\mathbf{U}$ be modeled as a spatially motivated linear combination of independent variables. This induces $\Sigma_U = \sigma^2 \tilde{B}\tilde{B}^T$. Thus, the U_i and hence the Y_i will be dependent and given $\tilde{B}$, $cov\left(Y_i, Y_{i'}\right) = \sigma^2 \sum_j \tilde{b}_{ij}\tilde{b}_{i'j}$. If $\tilde{B}$ arises through some proximity matrix W, the more similar rows i and i' of W are, the stronger the association between Y_i and $Y_{i'}$. However, the difference in nature between this specification and that in (4.23) is evident. To align the two, we would set $(I - B)^{-1} = \tilde{B}$, i.e. $B = I - \tilde{B}^{-1}$ (assuming $\tilde{B}$ is of full rank). $I - \tilde{B}^{-1}$ would not appear to have any interpretation through a proximity matrix.

Perhaps the most important point to note with respect to SAR models is that they are well suited to maximum likelihood estimation but not at all for MCMC fitting of Bayesian models. That is, the log likelihood associated with (4.23) (assuming $\tilde{D} = \sigma^2 I$) is

$$\frac{1}{2}\log\left|\sigma^{-1}\left(I - B\right)\right| - \frac{1}{2\sigma^2}\left(\mathbf{Y} - X\boldsymbol{\beta}\right)^T \left(I - B\right)\left(I - B\right)^T \left(\mathbf{Y} - X\boldsymbol{\beta}\right)\,. \tag{4.24}$$

Though B will introduce a regression or autocorrelation parameter, the quadratic form in (4.24) is quick to calculate (requiring no inverse) and the determinant can usually be calculated rapidly using diagonally dominant, sparse matrix approximations (see, e.g., Pace and Barry, 1997a,b). Thus maximization of (4.24) can be done iteratively but, in general, efficiently.

Also, note that while the form in (4.24) can certainly be extended to a full Bayesian model through appropriate prior specifications, the absence of a hierarchical form with random effects implies straightforward Bayesian model fitting as well. Indeed, the general spatial slice Gibbs sampler (see Appendix Section A.2, or Agarwal and Gelfand, 2002) can easily handle this model. However, suppose we attempt to introduce SAR random effects in some fashion. Unlike CAR random effects that are defined through full conditional distributions, the full conditional distributions for the SAR effects have no convenient form. For large n, computation of such distributions using a form such as (4.22) will be expensive.

SAR models as in (4.23) are frequently employed in the spatial econometrics literature. With point-referenced data, B is taken to be ρW where W is the matrix of inter-point distances. Likelihood-based inference can be implemented in the `spdep` package in `R` as well as more specialized software, such as that from the Spatial Analysis Laboratory (`sal.agecon.uiuc.edu`). Software for large data sets is supplied there, as well as through the website of Prof. Kelley Pace, `www.spatial-statistics.com`. An illustrative example is provided in Exercise 12.

4.4.1 CAR versus SAR models

Cressie (1993, pp. 408–10) credits Brook (1964) with being the first to make a distinction between the CAR and SAR models, and offers a comparison of the two. To begin with, we may note from (4.13) and (4.22) that, under propriety, the two forms are equivalent if and only if

$$(I-B)^{-1}D = (I-\tilde{B})^{-1}\tilde{D}((I-\tilde{B})^{-1})^T \ ,$$

where we use the tilde to indicate matrices in the SAR model. Cressie then shows that any SAR model can be represented as a CAR model (since D is diagonal, we can straightforwardly solve for B), but gives a counterexample to prove that the converse is not true. Since all SAR models are proper while we routinely employ improper CAR models, it is not surprising that the latter is a larger class.

For the "proper" CAR and SAR models that include spatial correlation parameters ρ, Wall (2004) shows that the correlations between neighboring regions implied by these two models can be rather different; in particular, the first-order neighbor correlations increase at a slower rate as a function of ρ in the CAR model than they do for the SAR model. (As an aside, she notes that these correlations are not even monotone for $\rho < 0$, another reason to avoid negative spatial correlation parameters.) Also, correlations among pairs can switch in nonintuitive ways. For example, when working with the adjacency relationships generated by the lower 48 contiguous U.S. states, she finds that when $\rho = .49$ in the CAR model, $Corr(Alabama, Florida) = .20$ and $Corr(Alabama, Georgia) = .16$. But when ρ increases to .975, we instead get $Corr(Alabama, Florida) = .65$ and $Corr(Alabama, Georgia) = .67$, a slight reversal in ordering.

4.4.2 STAR models

In the literature SAR models have frequently been extended to handle spatiotemporal data. The idea is that in working with proximity matrices, we can define neighbors in time as well as in space. Figure 4.4 shows a simple illustration with 9 areal units, 3 temporal units for each areal unit yielding $i = 1, \ldots, 9$, $t = 1, 2, 3$, labeled as indicated.

The measurements Y_{it} are spatially associated at each fixed t. But also, we might seek to associate, say, Y_{i2} with Y_{i1} and Y_{i3}. Suppose we write Y as the 27×1 vector with the first nine entries at $t = 1$, the second nine at $t = 2$, and the last nine at $t = 3$. Also let $W_S = BlockDiag(W_1, W_1, W_1)$, where

$$W_1 = \begin{pmatrix} 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 \end{pmatrix} .$$

Then W_S provides a spatial contiguity matrix for the Y's. Similarly, let $W_T = \begin{pmatrix} 0 & W_2 & 0 \\ W_2 & 0 & W_2 \\ 0 & W_2 & 0 \end{pmatrix}$, where $W_2 = I_{9\times 9}$. Then W_T provides a *temporal* contiguity matrix for the Y's. But then, in our SAR model we can define $B = \rho_s W_S + \rho_t W_T$. In fact, we can

Figure 4.4 *Illustration of spatiotemporal areal unit setting for STAR model.*

also introduce $\rho_{ST} W_S W_T$ into B and note that

$$W_S W_T = \begin{pmatrix} 0 & W_1 & 0 \\ W_1 & 0 & W_1 \\ 0 & W_1 & 0 \end{pmatrix}.$$

In this way, we introduce association across both space and time. For instance Y_{21} and Y_{41} affect the mean of Y_{12} (as well as affecting Y_{11}) from W_S by itself. Many more possibilities exist. Models formulated through such more general definitions of B are referred to as *spatiotemporal autoregressive* (STAR) models. See Pace et al. (2000) for a full discussion and development. The interpretation of the ρ's in the above example measures the relative importance of first-order spatial neighbors, first-order temporal neighbors, and first-order spatiotemporal neighbors.

4.5 Computer tutorials

In this section we outline the use of some GIS functions in `R` for obtaining neighborhood (adjacency) matrices, computing Moran's and Geary's statistic and fitting CAR and SAR models using traditional maximum likelihood techniques, and mapping the results for certain classes of problems. Here we confine ourselves to the modeling of Gaussian data on areal units. For the following illustrations we will load the following libraries and use functions therein:

```
> library(maps)
> library(maptools)
> library(spdep)
> library(classInt) ## Will be used for plotting maps later
> library(RColorBrewer) ## Will be used for colored maps
```

4.5.1 Adjacency matrices from maps using `spdep`

Consider, for example, the SAT exam scores data from the lower 48 contiguous states of the U.S. and the District of Columbia. We can construct this map using the `maps` and `maptools` packages. The idea is to create a sequence of data structures that will eventually produce an adjacency matrix. We execute the following commands

```
> usa.state = map(database="state", fill=TRUE, plot=FALSE)
> state.ID <- sapply(strsplit(usa.state$names, ":"), function(x) x[1])
> usa.poly = map2SpatialPolygons(usa.state, IDs=state.ID)
> usa.nb = poly2nb(usa.poly)
> usa.adj.mat = nb2mat(usa.nb, style="B")
```

The object `"usa.state"` is returned by the `"map()"` function. Next, we extract the state ID's from the `"state"` database and use them to construct a `"SpatialPolygon"` object called `"usa.poly"`, which is then coverted to a neighborhood object `"usa.nb"` using the `spdep` function `poly2nb`. Finally the `nb2mat` function (also in `spdep`) produces the adjacency matrix. The option `style="B"` produces the basic binary coding. Therefore, `usa.adj.mat` produced above is a 49×49 matrix whose (i, j)-th entry is equal to 1 if i is a neighbor of j and 0 otherwise. All diagonal entries are 0. The option `style="W"` produces a row-normalized adjacency matrix. Other options are available and left to the reader to explore.

For constructing adjacency matrices for counties in a U.S. state, the above code needs to be modified slightly because of the way the `"state"` and `"county"` databases list their entries. In the `"state"` database, the states are listed by simply their names (e.g. `"maryland"`) when they make up a single region or are separated by a `":"` in cases when they are split into subregions (e.g. `"michigan:north"` and `"michigan:south"`). This can be easily checked by typing `"usa.state$names"` after the `"usa.state"` object has been created as above. Counties within a state, on the other hand, are listed by the name of the state followed by the name of the county (e.g. `"minnesota,hennepin"`). Therefore, to produce the county neighborhood matrix for the State of Minnesota, we execute

```
> mn.county = map("county", "minnesota", fill=TRUE, plot=FALSE)
> county.ID <- sapply(strsplit(mn.county$names, ","), function(x) x[2])
> mn.poly = map2SpatialPolygons(mn.county, IDs=county.ID)
> mn.nb = poly2nb(mn.poly)
> mn.adj.mat = nb2mat(mn.nb, style="B")
```

Note the different specification in the way the `"strsplit()"` function is implemented for getting the county identifiers. The rest of the code is fairly similar to that for the state adjacencies. Neighbors of any given county can be easily found from the adjacency matrix. For example, the neighbors of Winona county in Minnesota can be found as

```
> mn.region.id <- attr(mn.nb, "region.id")
> winona.neighbors.index = mn.nb[[match("winona", mn.region.id)]]
> winona.neighbors = rownames(mn.adj.mat[winona.neighbors.index,])
> winona.neighbors
[1] "fillmore" "houston"  "olmsted"  "wabasha"
```

Note: Since the region is restricted to Minnesota, this lists Winona's adjacent counties in Minnesota only. Winona has three other adjacent counties in Wisconsin: Buffalo, Trempealeau and La Crosse.

One could also create adjacency matrices from external shapefiles by executing

```
> mn.map.shp = readShapeSpatial("minnesota.shp")
> mn.nb.shp = poly2nb(mn.map.shp)
> mn.adj.mat.shp = nb2mat(mn.nb, style="B")
```

However, the adjacency matrices obtained from external shapefiles need not be identical to those obtained from R's map databases. In fact, the rows and columns of the neighborhood matrix mn.adj.mat.shp obtained from the ESRI shapefiles for Minnesota will not correspond to mn.adj.mat from R's maps. Nevertheless, these can be easily brought to the same order:

```
> ordered.index =  order(as.character(mn.map$NAME))
> mn.adj.mat.shp = mn.adj.mat.shp[ordered.index, ordered.index]
```

These are now in the same order. However, they are still not quite identical with 99.87% of the entries in agreement. This happens because the polygons in the ESRI shapefiles slightly differ from those in the maps package in R.

4.5.2 Moran's I and Geary's C in spdep

We can subsequently invoke the moran.test and geary.test functions in the spdep package to obtain Moran's I and Geary's C statistics. Let us illustrate using the SAT scores data presented in Figure 4.1. We first use the nb object usa.nb and convert it to a listw object in R.

```
> usa.listw = nb2listw(usa.nb, style="W")
```

The option style="W" takes a 0-1 neighbors list, where regions are either listed as neighbors or are absent, and creates row-normalized weights. We read the SAT scores using the dataset available in http://www.biostat.umn.edu/~brad/data2.html.

```
state.sat.scores = read.table("state-sat.dat", header=T)
```

Next, we use the moran.test function to obtain

```
> moran.test(state.sat.scores$VERBAL, listw=usa.listw),
```

which gives us the calculated Moran's I and the associated variance of the estimate. This yields a sample estimate of 0.6125 with an associated standard error of 0.0979. Geary's C can be computed analogously using the geary.test function:

```
geary.test(state.sat.scores$VERBAL, listw=usa.listw),
```

which yields the sample estimate and standard error of 0.3577 and 0.0984, respectively. For maps with "islands" or regions without neighbors, we need to set zero.policy=TRUE in the nb2listw() function. This ensures that weight vectors of zero length are applied to islands, i.e., regions without any neighbor in the neighbors list. Otherwise, the program terminates with an error.

4.5.3 SAR and CAR model fitting using spdep *in* R

We now turn to fitting spatial autoregression models using available functions in the spdep package. A convenient illustration is offered by the SIDS (sudden infant death syndrome) data, analyzed by Cressie and Read (1985), Cressie and Chan (1989), Cressie (1993, Sec. 6.2) and Kaluzny et al. (1998, Sec. 5.3), and already loaded into the spdep package. This dataset contains counts of SIDS deaths from 1974 to 1978 and counts from 1979 to 1983 along with related covariate information for the 100 counties in the U.S. State of North Carolina.

The dataset can be read from a shapefile sids.shp and an nb object can be constructed from a GAL file ncCC89 containing the data analyzed by Cressie and Chan (1989). This data relates to counts aggregated from 1979 to 1983. Both these files are included in spdep. We execute the following steps:

```
> nc.sids <- readShapePoly(system.file("etc/shapes/sids.shp", package=
  "spdep")[1],
```

```
+       ID="FIPSNO", proj4string=CRS("+proj=longlat +ellps=clrk66"))
> rn <- sapply(slot(nc.sids, "polygons"), function(x) slot(x, "ID"))
> ncCC89.nb <- read.gal(system.file("etc/weights/ncCC89.gal", package=
  "spdep")[1],
+ region.id=rn)
```

The first step produces a `SpatialPolygonsDataFrame` object `nc.sids`, while the second step produces the region IDs and stores them in `rn`. The third step uses these region IDs to produce an `nb` object by directly reading from the GAL file. We next use a Freeman-Tukey transformation to produce the transformed rates and append them to the `nc.sids` object.

```
> nc.sids.rates.FT = sqrt(1000)*(sqrt(nc.sids$SID79/nc.sids$BIR79)
+                        + sqrt((nc.sids$SID79 + 1)/nc.sids$BIR79))
> nc.sids$rates.FT = nc.sids.rates.FT
```

We wish to regress these rates on the non-white birth rates over the same period. This variable is available as `NWBIR79` in the `nc.sids` object. We will use the Freeman-Tukey transformed birth rates:

```
> nc.sids.nwbir.FT = sqrt(1000)*(sqrt(nc.sids$NWBIR79/nc.sids$BIR79)
+                        + sqrt((nc.sids$NWBIR79 + 1)/nc.sids$BIR79))
> nc.sids$nwbir.FT = nc.sids.nwbir.FT
```

Maximum likelihood estimation of (4.23), which has the likelihood in (4.24), can be carried out using the `errorsarlm()` or, equivalently, the `spautolm()` function in `spdep`. These functions produce the same output. Below we demonstrate the latter. We first create a `listw` object using the 0-1 adjacency structure

```
> ncCC89.listw = nb2listw(ncCC89.nb, style="B", zero.policy=TRUE)
```

Note that the `zero.policy=TRUE` is required here because the county shapefile in `spdep` lists two counties, Dare and Hyde, as having zero neighbors. These counties are situated in coastal North Carolina and are adjacent to substantial bodies of water. These two counties can be identified as

```
> nc.county.id = attr(ncCC89.nb, "region.id")
> nc.no.neighbors = card(ncCC89.nb)
> nc.islands = as.character(nc.sids[card(ncCC89.nb) == 0, ]$NAME)
> nc.islands
[1] "Dare" "Hyde"
```

We now estimate the SAR model in (4.23) using

```
> nc.sids.sar.out = spautolm(rates.FT~ nwbir.FT, data=nc.sids, family="SAR",
+                            listw=ncCC89.listw, zero.policy=TRUE)
> summary(nc.sids.sar.out)
Call:
spautolm(formula = rates.FT ~ nwbir.FT, data = nc.sids, listw = ncCC89.listw,
    family = "SAR", zero.policy = TRUE)

Residuals:
       Min         1Q     Median         3Q        Max
-2.2660489 -0.4281394  0.0043310  0.4978178  2.5164979

Regions with no neighbours included:
 37055 37095

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) 2.5216462  0.2570334  9.8106  < 2e-16
nwbir.FT    0.0124001  0.0071867  1.7254  0.08445
```

```
Lambda: 0.05206 LR test value: 2.218 p-value: 0.13641

Log likelihood: -123.6898
ML residual variance (sigma squared): 0.68697, (sigma: 0.82883)
Number of observations: 100
Number of parameters estimated: 4
AIC: 255.38
```

A CAR model can also be estimated by setting `family=CAR` in the above. We see the presence of a significantly positive intercept. The regression coefficient corresponding to the non-white birth rates seems to indicate a positive correlation but is not statistically significant. These results seem to be consistent with the estimates from ordinary least squares. We point out that an analysis of the 1974–1978 data by Kaluzny et al. (1998, Sec. 5.3) using the `S+ Spatial Stats` software, which was presented in the earlier edition of this book, returned a significant coefficient for birth rates. However, the dataset used there was different, the adjacency matrix of the state was modified to split a county into three regions and the weights used for the adjacency matrix also were different. The spatial autocorrelation parameter, denoted by `Lambda` in the output summary, is also not significant.

Once the estimates from the model have been obtained, we may wish to plot the fitted rates and compare them with the raw data. We first compute the fitted rates and append them to the `nc.sids` object. We then plot the fitted rates on a map.

```
> nc.sids$fitted.sar = fitted(nc.sids.sar.out)
> brks = c(0, 2.0, 3.0, 3.5, 6.0)
> color.pallete = rev(brewer.pal(4,"RdBu"))
> class.fitted = classIntervals(var=nc.sids$fitted.sar, n=4, style="fixed",
+                               fixedBreaks=brks, dataPrecision=4)
> color.code.fitted = findColours(class.fitted, color.pallete)
> plot(nc.sids, col=color.code.fitted)
```

The raw rates available as `nc.sids$rates.FT` can be plotted analogously. The resulting maps, with an added legend, are displayed in Figure 4.5. Although the spatial autocorrelation in the data was found to be modest, the fitted values from the SAR model clearly show the smoothing. Both the maps have the same color scheme.

Instead of defining a neighborhood structure completely in terms of spatial adjacency on the map, we may want to construct neighbors using a distance function. For example, given centroids of the various regions, we could identify regions as neighbors if and only if their intercentroidal distance is below a particular threshold. We illustrate using a dataset offering neighborhood-level information on crime, mean home value, mean income, and other variables for 49 neighborhoods in Columbus, OH, during 1980. More information on these data is available from Anselin (1988, p.189), or in Exercise 12. The data can be downloaded from `www.biostat.umn.edu/~brad/data/Columbus.dat` but that is not needed as it is available within the `spdep` package.

We begin by reading a Columbus shapefile and creating a `"SpatialPolygonsDataFrame"` object.

```
> library(spdep)
> columbus.poly <- readShapePoly(system.file("etc/shapes/columbus.shp",
+                                 package="spdep")[1])
```

Suppose we would like to have regions with intercentroidal distances less than half the maximum intercentroidal distance as neighbors. We first construct an object, `columbus.coords`, that contain the centroids of the different regions.

Figure 4.5 *Unsmoothed raw (a) and spatially smoothed fitted (b) rates, North Carolina SIDS data.*

```
> columbus.coords = coordinates(columbus.poly)
```

A particularly useful function provided by `spdep` is the `dnearneigh()` function. This function can take a matrix of coordinates along with specified lower and upper distance bounds as inputs and returns a list of neighbors categorized by all points that are within the specified distance threshold from each other. Note that the function uses Euclidean distances, which means that the coordinates must be projected onto a plane (and cannot be in terms of latitude and longitude) before the function is applied. There is, however, a potential pitfall of directly using the `dnearneigh()` function in that it may generate "islands" (i.e., regions with no neighbors) unless we are judicious about our choice of the upper distance bound (the lower bound is usually set to zero). One way to circumvent this problem is to first apply

the k-nearest neighbors algorithm to the coordinates and then to create an neighborhood list from these k-nearest neighbors:

```
> columbus.knn = knearneigh(columbus.coords)
> columbus.knn2nb = knn2nb(columbus.knn)
```

Next, the Euclidean distances between the neighbors are constructed by applying the `nbdists()` function to the k-nearest neighbor list. This returns the nearest-neighbor distances.

```
> columbus.dist.list = nbdists(columbus.knn2nb, columbus.coords)
> columbus.dist.vec = unlist(columbus.dist.list)
```

where the second step converts the list data structure into a vector. Next, we find the maximum of the nearest neighbor distances and pass it to the `dnearneigh()` function as an input for the upper bound.

```
> columbus.dist.max = max(columbus.dist.vec)
> columbus.dnn.nb = dnearneigh(columbus.coords, 0, columbus.dist.max)
```

This ensures that there are no islands. Next, we create a `listw` object from the returned `dnearneigh()` function and estimate the SAR model.

```
> columbus.dnn.listw = nb2listw(columbus.dnn.nb, style="B", zero.policy=TRUE)
> columbus.dnn.sar.out = spautolm(CRIME~HOVAL+INC, data=columbus.poly,
+                  family="SAR", listw=columbus.dnn.listw, zero.policy=TRUE)
> summary(columbus.dnn.sar.out)
```

```
Call: spautolm(formula = CRIME ~ HOVAL + INC, data = columbus.poly,
    listw = columbus.dnn.listw, family = "SAR", zero.policy = TRUE)

Residuals:
      Min        1Q    Median        3Q       Max
-31.18714  -4.18864  -0.24961   6.10122  22.69041

Coefficients:
             Estimate Std. Error z value  Pr(>|z|)
(Intercept) 50.229861   5.422987  9.2624 < 2.2e-16
HOVAL       -0.251366   0.079881 -3.1468  0.001651
INC         -0.826528   0.298414 -2.7697  0.005610

Lambda: 0.11302 LR test value: 19.536 p-value: 9.8716e-06

Log likelihood: -177.6092
ML residual variance (sigma squared): 73.306, (sigma: 8.5619)
Number of observations: 49
Number of parameters estimated: 5
AIC: 365.22
```

We see that both house value (HOVAL) and income (INC) have a significant negative impact on crime, which is not unexpected. The spatial autocorrelation parameter `Lambda` is also significant, which indicates strong spatial dependence in the data. Note, however, that this significance of the spatial dependence may be an artefact of an undesirably dense connectedness structure imposed by setting the upper distance bound to the maximum intercentroidal distance. To mitigate this effect one can set some proportion of the maximum distance as the distance upper bound. For example, we can set the upper bound to be `0.25*columbus.dist.max` in the above code and repeat the subsequent steps. This specification generates islands and `zero.policy=TRUE` option is required to obtain estimates from the SAR model. We do not show this output but it is similar to the above. While

the regression coefficients do not change substantially and are still significant, the spatial autocorrelation parameter is no longer statistically significant (p-value is approximately 0.25257). The result from the CAR model (obtained by setting `family="CAR"`) is also very similar.

4.6 Exercises

1. Verify Brook's Lemma, equation (4.7).

2.(a) To appreciate how Brook's Lemma works, suppose Y_1 and Y_2 are both binary variables, and that their joint distribution is defined through conditional logit models. That is,

$$\log \frac{P(Y_1 = 1|Y_2)}{P(Y_1 = 0|Y_2)} = \alpha_0 + \alpha_1 Y_2 \quad \text{and} \quad \log \frac{P(Y_2 = 1|Y_1)}{P(Y_2 = 0|Y_1)} = \beta_0 + \beta_1 Y_1 \ .$$

Obtain the joint distribution of Y_1 and Y_2.

(b) This result can be straightforwardly extended to the case of more than two variables, but the details become increasingly clumsy. Illustrate this issue in the case of *three* binary variables, Y_1, Y_2, and Y_3.

3. Returning to (4.13) and (4.14), let $B = ((b_{ij}))$ be an $n \times n$ matrix with positive elements; that is, $b_{ij} > 0$, $\sum_j b_{ij} \leq 1$ for all i, and $\sum_j b_{ij} < 1$ for at least one i. Let $D = Diag\left(\tau_i^2\right)$ be a diagonal matrix with positive elements τ_i^2 such that $D^{-1}(I - B)$ is symmetric; that is, $b_{ij}/\tau_i^2 = b_{ji}/\tau_j^2$, for all i, j. Show that $D^{-1}(I - B)$ is positive definite.

4. Looking again at (4.13), obtain a simple sufficient condition on B such that the CAR specification with precision matrix $D^{-1}(I - B)$ is a pairwise difference specification, as in (4.16).

5. Show that, for W symmetric, $\Sigma_{\mathbf{y}}^{-1} = D_w - \rho W$ is positive definite (thus resolving the impropriety in (4.15)) if $\rho \in \left(1/\lambda_{(1)}, 1/\lambda_{(n)}\right)$, where $\lambda_{(1)} < \lambda_{(2)} < \cdots < \lambda_{(n)}$ are the ordered eigenvalues of $D_w^{-1/2} W D_w^{-1/2}$.

6. Show that if all entries in W are nonnegative and $D_w - \rho W$ is positive definite with $0 < \rho < 1$, then all entries in $(D_w - \rho W)^{-1}$ are nonnegative.

7. Under a proper CAR model for $\mathbf{Y}$, i.e., with $\Sigma_{\mathbf{y}} = D_w - \rho W$, obtain the correlation and covariance between Y_i and Y_j.

8. Obtain the joint distribution, up to normalizing constant, for $(Y_1, Y_2, ..., Y_n)$ under (4.20). Hint: You might try to guess it but Brook's Lemma can be used as well.

9. Recalling the SAR formulation using the scaled adjacency matrix $\widetilde{W}$ just below (4.22), prove that $I - \alpha\widetilde{W}$ will be nonsingular if $\alpha \in (-1, 1)$, so that α may be sensibly referred to as an "autocorrelation parameter."

10. In the setting of Subsection 4.3.1, if $(\Sigma_{\mathbf{y}}^{-1})_{ij} = 0$, then show that Y_i and Y_j are conditionally independent given $Y_k, k \neq i, j$.

11. The file `www.biostat.umn.edu/~brad/data/state-sat.dat` gives the 1999 state average SAT data (part of which is mapped in Figure 4.1).

(a) Use the `spautolm` function to fit the SAR model of Section 4.4, taking the verbal SAT score as the response Y and the percent of eligible students taking the exam in each state as the covariate X. Do this analysis twice: first using binary weights and then using row-normalized weights. Is the analysis sensitive to these choices of weights? Is knowing X helpful in explaining Y?

(b) Using the `maps` library in `R`, draw choropleth maps similar to Figure 4.1 of both the fitted verbal SAT scores and the spatial residuals from the SAR model. Is there

evidence of spatial correlation in the response Y once the covariate X is accounted for?

(c) Repeat your SAR model analysis above, again using `spautolm` but now assuming the CAR model of Section 4.3. Compare your estimates with those from the SAR model and interpret any changes.

(d) One might imagine that the percentage of eligible students taking the exam should perhaps affect the variance of our model, not just the mean structure. To check this, refit the SAR model replacing your row-normalized weights with weights equal to the reciprocal of the percentage of students taking the SAT. Is this model sensible?

12. Consider the data `www.biostat.umn.edu/~brad/data/Columbus.dat`, taken from Anselin (1988, p. 189) and also available within the `spdep` `R` package (but with possibly different variable names). These data record crime information for 49 neighborhoods in Columbus, OH, during 1980. Variables measured include NEIG, the neighborhood id value (1–49); HOVAL, its mean housing value (in \$1,000); INC, its mean household income (in \$1,000); CRIME, its number of residential burglaries and vehicle thefts per thousand households; OPEN, a measure of the neighborhood's open space; PLUMB, the percentage of housing units without plumbing; DISCBD, the neighborhood centroid's distance from the central business district; X, an x-coordinate for the neighborhood centroid (in arbitrary digitizing units, not polygon coordinates); Y, the same as X for the y-coordinate; AREA, the neighborhood's area; and PERIM, the perimeter of the polygon describing the neighborhood.

(a) Use `spdep` in `R` to construct adjacency matrices for the neighborhoods of Columbus based upon centroid distances less than

i. 25% of the maximum intercentroidal distances;
ii. 50% of the maximum intercentroidal distances;
iii. 75% of the maximum intercentroidal distances.

(b) For each of the three spatial neighborhoods constructed above, use the `spautolm` function to fit SAR models with CRIME as the dependent variable, and HOVAL, INC, OPEN, PLUMB, and DISCBD as the covariates. Compare your results and interpret your parameter estimates in each case.

(c) Repeat your analysis by setting $B = \rho W$ in equation (4.23) with W_{ij} the Euclidean distance between location i and location j.

(d) Repeat part (b) for CAR models. Compare your estimates with those from the SAR model and interpret them.

Chapter 5

Basics of Bayesian inference

In this chapter we provide a brief review of hierarchical Bayesian modeling and computing for readers not already familiar with these topics. Of course, in one chapter we can only scratch the surface of this rapidly expanding field, and readers may well wish to consult one of the many recent textbooks on the subject, either as preliminary work or on an as-needed basis. It should come as little surprise that the book we most highly recommend for this purpose is the one by Carlin and Louis (2008); the Bayesian methodology and computing material below roughly follows Chapters 2 and 3, respectively, in that text.

However, a great many other good Bayesian books are available, and we list a few of them and their characteristics. First we mention the texts stressing Bayesian theory, including DeGroot (1970), Box and Tiao (1992), Berger (1985), Bernardo and Smith (1994), and Robert (2007). These books tend to focus on foundations and decision theory, rather than computation or data analysis. On the more methodological side, a nice introductory book is that of Lee (1997), with O'Hagan (1994) and Gelman et al. (2013) offering more general Bayesian modeling treatments. More recent texts, with greater emphasis on Bayesian modeling using `R`, include Albert (2008) and Hoff (2009).

5.1 Introduction to hierarchical modeling and Bayes' Theorem

By modeling both the observed data and any unknowns as random variables, the Bayesian approach to statistical analysis provides a coherent framework for combining complex data models and external knowledge or expert opinion. In this approach, in addition to specifying the distributional model $f(\mathbf{y}|\boldsymbol{\theta})$ for the observed data $\mathbf{y} = (y_1, \ldots, y_n)$ given a vector of unknown parameters $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)$, we suppose that $\boldsymbol{\theta}$ is a random quantity sampled from a *prior* distribution $\pi(\boldsymbol{\theta}|\boldsymbol{\lambda})$, where $\boldsymbol{\lambda}$ is a vector of hyperparameters. For instance, y_i might be the empirical mammography rate in a sample of women aged 40 and over from county i, θ_i the underlying true mammography rate for all such women in this county, and $\boldsymbol{\lambda}$ a parameter controlling how these true rates vary across counties. If $\boldsymbol{\lambda}$ is known, inference concerning $\boldsymbol{\theta}$ is based on its *posterior* distribution,

$$p(\boldsymbol{\theta}|\mathbf{y}, \boldsymbol{\lambda}) = \frac{p(\mathbf{y}, \boldsymbol{\theta}|\boldsymbol{\lambda})}{p(\mathbf{y}|\boldsymbol{\lambda})} = \frac{p(\mathbf{y}, \boldsymbol{\theta}|\boldsymbol{\lambda})}{\int p(\mathbf{y}, \boldsymbol{\theta}|\boldsymbol{\lambda})\, d\boldsymbol{\theta}} = \frac{f(\mathbf{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta}|\boldsymbol{\lambda})}{\int f(\mathbf{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta}|\boldsymbol{\lambda})\, d\boldsymbol{\theta}}\,. \tag{5.1}$$

Notice the contribution of both the data (in the form of the likelihood f) and the external knowledge or opinion (in the form of the prior π) to the posterior. Since, in practice, $\boldsymbol{\lambda}$ will not be known, a second stage (or *hyperprior*) distribution $h(\boldsymbol{\lambda})$ will often be required, and (5.1) will be replaced with

$$p(\boldsymbol{\theta}|\mathbf{y}) = \frac{p(\mathbf{y}, \boldsymbol{\theta})}{p(\mathbf{y})} = \frac{\int f(\mathbf{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta}|\boldsymbol{\lambda})h(\boldsymbol{\lambda})\, d\boldsymbol{\lambda}}{\int f(\mathbf{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta}|\boldsymbol{\lambda})h(\boldsymbol{\lambda})\, d\boldsymbol{\theta} d\boldsymbol{\lambda}}. \tag{5.2}$$

It is worth noting the *hierarchical* structure implicit in (5.2), i.e., three levels of distributional specification, typically with primary interest in the $\boldsymbol{\theta}$ level. As the title of this book

suggests, such hierarchical specification is the focus of the inference methodology in this volume.

Rather than integrating over $\boldsymbol{\lambda}$, we might replace $\boldsymbol{\lambda}$ by an estimate $\hat{\boldsymbol{\lambda}}$ obtained by maximizing the marginal distribution $p(\mathbf{y}|\boldsymbol{\lambda}) = \int f(\mathbf{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta}|\boldsymbol{\lambda})d\boldsymbol{\theta}$, viewed as a function of $\boldsymbol{\lambda}$. Inference could then proceed based on the *estimated* posterior distribution $p(\theta|\mathbf{y}, \hat{\boldsymbol{\lambda}})$, obtained by plugging $\hat{\boldsymbol{\lambda}}$ into equation (5.1). This approach is referred to as *empirical Bayes* analysis; see Berger (1985), Maritz and Lwin (1989), Carlin and Louis (2008) and, more recently, Efron (2010) for details regarding empirical Bayes methodology and applications.

The Bayesian inferential paradigm offers potentially attractive advantages over the classical, frequentist statistical approach through its more philosophically sound foundation, its unified approach to data analysis, and its ability to formally incorporate prior opinion or external empirical evidence into the results via the prior distribution π. Data analysts, formerly reluctant to adopt the Bayesian approach due to skepticism concerning its philosophy and a lack of necessary computational tools, are now turning to it with increasing regularity since classical methods emerge as both theoretically and practically inadequate to handle the challenges of today's complex modeling landscape. In our context, modeling the θ_i as random (instead of fixed) effects allows us to induce specific (e.g., spatial) correlation structures among them, hence among the observed data y_i as well. Hierarchical Bayesian methods now enjoy broad application in the analysis of spatial data, as the remainder of this book reveals.

A computational challenge in applying Bayesian methods is that for most realistic problems, the integrations required to do inference under (5.1) are generally not tractable in closed form, and thus must be approximated numerically. Forms for π and h (called *conjugate* priors) that enable at least partial analytic evaluation of these integrals may often be found, but in the presense of nuisance parameters (typically unknown variances), some intractable integrations remain. Here the emergence of inexpensive, high-speed computing equipment and software comes to the rescue, enabling the application of Markov chain Monte Carlo (MCMC) integration methods, such as the Metropolis-Hastings algorithm (Metropolis et al., 1953; Hastings, 1970) and the Gibbs sampler (Geman and Geman, 1984; Gelfand and Smith, 1990). This is the subject of Section 5.3.

5.1.1 Illustrations of Bayes' Theorem

Equation (5.1) is a generic version of what is referred to as *Bayes' Theorem* or *Bayes' Rule*. It is attributed to the Reverend Thomas Bayes, an 18th-century nonconformist minister and part-time mathematician; a version of the result was published (posthumously) in Bayes (1763). In this subsection we consider a few basic examples of its use.

Example 5.1 Suppose we have observed a single normal (Gaussian) observation $Y \sim N\left(\theta, \sigma^2\right)$ with σ^2 known, so that the likelihood $f\left(y|\theta\right) = N\left(y|\theta, \sigma^2\right) \equiv \frac{1}{\sigma\sqrt{2\pi}} \exp(-\frac{(y-\theta)^2}{2\sigma^2})$, $y \in \Re$, $\theta \in \Re$, and $\sigma > 0$. If we specify the prior distribution as $\pi(\theta) = N\left(y \,\middle|\, \mu, \tau^2\right)$ with $\boldsymbol{\lambda} = (\mu, \tau^2)'$ fixed, then from (5.1) we can compute the posterior as

$$\begin{aligned} p\left(\theta|y\right) &= \frac{N\left(\theta|\mu, \tau^2\right) N\left(y|\theta, \sigma^2\right)}{p\left(y\right)} \\ &\propto N\left(\theta|\mu, \tau^2\right) N\left(y|\theta, \sigma^2\right) \\ &= N\left(\theta \,\middle|\, \frac{\sigma^2}{\sigma^2+\tau^2}\mu + \frac{\tau^2}{\sigma^2+\tau^2}y\,,\, \frac{\sigma^2\tau^2}{\sigma^2+\tau^2}\right) . \end{aligned} \tag{5.3}$$

That is, the posterior distribution of θ given y is also normal with mean and variance as given. The proportionality in the second row arises since the marginal distribution $p(y)$ does

not depend on θ, and is thus constant with respect to the Bayes' Theorem calculation. The final equality in the third row results from collecting like (θ^2 and θ) terms in the exponential and then completing the square.

Note that the posterior mean $E(\theta|y)$ is a weighted average of the prior mean μ and the data value y, with the weights depending on our relative uncertainty with respect to the prior and the likelihood. Also, the posterior *precision* (reciprocal of the variance) is equal to $1/\sigma^2 + 1/\tau^2$, which is the sum of the likelihood and prior precisions. Thus, thinking of precision as "information," we see that in the normal/normal model, the information in the posterior is the total of the information in the prior and the likelihood.

Suppose next that instead of a single datum we have a set of n observations $\mathbf{y} = (y_1, y_2, \ldots, y_n)'$. From basic normal theory we know that $f(\bar{y}|\theta) = N(\theta, \sigma^2/n)$. Since $\overline{y}$ is sufficient for θ, from (5.3) we have

$$\begin{aligned} p(\theta|\mathbf{y}) = p\left(\theta|\bar{y}\right) &= N\left(\theta \,\middle|\, \frac{(\sigma^2/n)}{(\sigma^2/n)+\tau^2}\mu + \frac{\tau^2}{(\sigma^2/n)+\tau^2}\overline{y}\,,\ \frac{(\sigma^2/n)\tau^2}{(\sigma^2/n)+\tau^2}\right) \\ &= N\left(\theta \,\middle|\, \frac{\sigma^2}{\sigma^2+n\tau^2}\mu + \frac{n\tau^2}{\sigma^2+n\tau^2}\overline{y}\,,\ \frac{\sigma^2\tau^2}{\sigma^2+n\tau^2}\right)\ . \end{aligned}$$

Again we obtain a posterior mean that is a weighted average of the prior (μ) and data (in this case summarized through $\bar{y}$).

In these two examples, the prior chosen leads to a posterior distribution for θ that is available in closed form, and is a member of the same distributional family as the prior. Such a prior is referred to as a *conjugate* prior. We will often use such priors in our work, since, when they are available, conjugate families are convenient and still allow a variety of shapes wide enough to capture our prior beliefs.

Note that setting $\tau^2 = \infty$ in the previous example corresponds to a prior that is arbitrarily vague, or *noninformative*. This then leads to a posterior of $p\left(\theta|y\right) = N\left(\theta|\overline{y}, \sigma^2/n\right)$, exactly the same as the likelihood for this problem. This arises since the limit of the conjugate (normal) prior here is actually a uniform, or "flat" prior, and thus the posterior becomes the likelihood (possibly renormalized to integrate to 1 as a function of θ). Of course, the flat prior is *improper* here, since the uniform does not integrate to anything finite over the entire real line; however, the posterior is still well defined since the likelihood can be integrated with respect to θ. Bayesians often use flat or otherwise improper noninformative priors, since prior feelings are often rather vague relative to the information in the likelihood. Such inference is broadly referred to as *objective* Bayes analysis. In any case, we generally want the data (and not the prior) to dominate the determination of the posterior.

Example 5.2 *(the general linear model).* Let $\mathbf{Y}$ be an $n \times 1$ data vector, X an $n \times p$ matrix of covariates, and adopt the likelihood and prior structure,

$$\begin{aligned} \mathbf{Y}|\boldsymbol{\beta} &\sim N_n\left(X\boldsymbol{\beta}, \Sigma\right), \text{ i.e. } f\left(\mathbf{Y}|\boldsymbol{\beta}\right) \equiv N_n\left(\mathbf{Y}|X\boldsymbol{\beta}, \Sigma\right), \\ \boldsymbol{\beta} &\sim N_p\left(A\boldsymbol{\alpha}, V\right), \text{ i.e. } \pi\left(\boldsymbol{\beta}\right) \equiv N\left(\boldsymbol{\beta}|A\boldsymbol{\alpha}, V\right)\ . \end{aligned}$$

Here $\boldsymbol{\beta}$ is a $p \times 1$ vector of regression coefficients and Σ is a $p \times p$ covariance matrix. Then it can be shown (now a classic result, first published by Lindley and Smith, 1972), that the marginal distribution of $\mathbf{Y}$ is

$$\mathbf{Y} \sim N\left(XA\boldsymbol{\alpha}\,,\ \Sigma + XVX^T\right)\ ,$$

and the posterior distribution of $\boldsymbol{\beta}|\mathbf{Y}$ is

$$\begin{aligned} \boldsymbol{\beta}|Y &\sim N\left(D\mathbf{d}, D\right), \\ \text{where } D^{-1} &= X^T\Sigma^{-1}X + V^{-1} \\ \text{and } \mathbf{d} &= X^T\Sigma^{-1}\mathbf{Y} + V^{-1}A\boldsymbol{\alpha}\ . \end{aligned}$$

Thus $E(\boldsymbol{\beta}|\mathbf{Y}) = D\mathbf{d}$ provides a point estimate for $\boldsymbol{\beta}$, with variability captured by the associated variance matrix D.

In particular, note that for a vague prior we may set $V^{-1} = 0$, so that $D^{-1} = X\Sigma^{-1}X$ and $\mathbf{d} = X^T\Sigma^{-1}\mathbf{Y}$. In the simple case where $\Sigma = \sigma^2 I_p$, the posterior becomes

$$\boldsymbol{\beta}|Y \sim N\left(\hat{\boldsymbol{\beta}}\,,\, \sigma^2(X'X)^{-1}\right) ,$$

where $\hat{\boldsymbol{\beta}} = (X'X)^{-1}X'\mathbf{y}$. Since the usual likelihood approach produces

$$\hat{\boldsymbol{\beta}} \sim N\left(\boldsymbol{\beta}\,,\, \sigma^2(X'X)^{-1}\right) ,$$

we once again we see "flat prior" Bayesian results that are formally equivalent to the usual likelihood approach.

5.2 Bayesian inference

While the computing associated with Bayesian methods can be daunting, the subsequent inference is relatively straightforward, especially in the case of estimation. This is because once we have computed (or obtained an estimate of) the posterior, inference comes down merely to extracting features of this distribution, since by Bayes' Rule the posterior summarizes everything we know about the model parameters in the light of the data. In the remainder of this section, we shall assume for simplicity that the posterior $p(\boldsymbol{\theta}|\mathbf{y})$ itself (and not merely an estimate of it) is available for summarization.

Bayesian methods for estimation are also reminiscent of corresponding maximum likelihood methods. This should not be surprising, since likelihoods form part of the Bayesian specification; we have even seen that a normalized (i.e., standardized) likelihood can be thought of a posterior when this is possible. However, when we turn to hypothesis testing (Bayesians prefer the term *model comparison*), the approaches have little in common. Bayesians (and many likelihoodists) have a deep and abiding antipathy toward p-values, for a long list of reasons we shall not go into here; the interested reader may consult Berger (1985, Sec. 4.3.3), Kass and Raftery (1995, Sec. 8.2), or Carlin and Louis (2000, Sec. 2.3.3).

5.2.1 Point estimation

Suppose for the moment that θ is univariate. Given the posterior $p(\theta|\mathbf{y})$, a natural Bayesian point estimate of θ would be some measure of centrality. Three familiar choices are the posterior mean,

$$\hat{\theta} = E(\theta|\mathbf{y}) ,$$

the posterior median,

$$\hat{\theta} \;:\; \int_{-\infty}^{\hat{\theta}} p(\theta|\mathbf{y})d\theta = 0.5 ,$$

and the posterior mode,

$$\hat{\theta} \;:\; p(\hat{\theta}|\mathbf{y}) = \sup_{\theta} p(\theta|\mathbf{y}) .$$

Notice that the lattermost estimate is typically easiest to compute, since it does not require any integration: we can replace $p(\theta|\mathbf{y})$ by its unstandardized form, $f(\mathbf{y}|\theta)p(\theta)$, and get the same answer (since these two differ only by a multiplicative factor of $m(\mathbf{y})$, which does not depend on θ). Indeed, if the posterior exists under a flat prior $p(\theta) = 1$, then the posterior mode is nothing but the maximum likelihood estimate (MLE).

Note that for symmetric unimodal posteriors (e.g., a normal distribution), the posterior mean, median, and mode will all be equal. However, for multimodal or otherwise nonnormal

posteriors, the mode will often be an unsatisfying choice of centrality measure (consider for example the case of a steadily decreasing, one-tailed posterior; the mode will be the very first value in the support of the distribution — hardly central!). By contrast, the posterior mean, though arguably the most commonly used, will sometimes be overly influenced by heavy tails (just as the sample mean $\bar{y}$ is often nonrobust against outlying observations). As a result, the posterior median will often be the best and safest point estimate. It is also the most difficult to compute (since it requires both an integration and a rootfinder), but this difficulty is somewhat mitigated for posterior estimates obtained via MCMC; see Section 5.3 below.

5.2.2 Interval estimation

The posterior allows us to make any desired probability statements about θ. By inversion, we can infer about any quantile. For example, suppose we find the $\alpha/2$- and $(1-\alpha/2)$-quantiles of $p(\theta|\mathbf{y})$, that is, the points q_L and q_U such that

$$\int_{-\infty}^{q_L} p(\theta|\mathbf{y})d\theta = \alpha/2 \quad \text{and} \quad \int_{q_U}^{\infty} p(\theta|\mathbf{y})d\theta = 1 - \alpha/2 \; .$$

Then clearly $P(q_L < \theta < q_U|\mathbf{y}) = 1-\alpha$; our confidence that θ lies in (q_L, q_U) is $100\times(1-\alpha)\%$. Thus this interval is a $100 \times (1-\alpha)\%$ *credible set* (or simply *Bayesian confidence interval*) for θ. This interval is relatively easy to compute, and enjoys a direct interpretation ("the probability that θ lies in (q_L, q_U) is $(1-\alpha)$") which the usual frequentist interval does not.

The interval just described is often called the *equal tail* credible set, for the obvious reason that is obtained by chopping an equal amount of support $(\alpha/2)$ off the top and bottom of $p(\theta|\mathbf{y})$. Note that for symmetric unimodal posteriors, this equal tail interval will be symmetric about this mode (which we recall equals the mean and median in this case). It will also be optimal in the sense that it will have shortest length among sets C satisfying

$$1 - \alpha \leq P(C|\mathbf{y}) = \int_C p(\theta|\mathbf{y})d\theta \; . \tag{5.4}$$

Note that any such set C could be thought of as a $100 \times (1-\alpha)\%$ credible set for θ. For posteriors that are not symmetric and unimodal, a better (shorter) credible set can be obtained by taking only those values of θ having posterior density greater than some cutoff $k(\alpha)$, where this cutoff is chosen to be as large as possible while C continues to satisfy equation (5.4). This *highest posterior density* (HPD) confidence set will always be of optimal length, but will typically be significantly more difficult to compute. The equal tail interval emerges as HPD in the symmetric unimodal case since there too it captures the "most likely" values of θ. Fortunately, many of the posteriors we will be interested in will be (at least approximately) symmetric unimodal, so the much simpler equal tail interval will often suffice. In fact, it is the routine choice in practice.

5.2.3 Hypothesis testing and model choice

We have seen that Bayesian inference (point or interval) is quite straightforward given the posterior distribution, or an estimate thereof. By contrast, hypothesis testing is less straightforward, for two reasons. First, there is less agreement among Bayesians as to the proper approach to the problem. For years, posterior probabilities and Bayes factors were considered the only appropriate method. But these methods are only suitable with fully proper priors, and for relatively low-dimensional models. With the recent proliferation of very complex models, employing at least partly improper priors, other methods have come to the fore. Second, solutions to hypothesis testing questions often involve not just the

posterior $p(\boldsymbol{\theta}|\mathbf{y})$, but also the *marginal* distribution, $m(\mathbf{y})$. Unlike the case of posterior and the predictive distributions, samples from the marginal distribution do not naturally emerge from most MCMC algorithms. Thus, the sampler must often be "tricked" into producing the necessary samples.

Model choice essentially requires specification of the *utility* for a model. This is a challenging exercise and, in practice, is not often considered. Off the shelf criteria are typically adopted. For instance, do we see our primary goal for the model to be explanation or, alternatively, is it prediction? Utility for the former places emphasis on the parameters and in fact, calculates a criterion over the parameter space. An approximate yet very easy-to-use model choice tool of this type, known as the Deviance Information Criterion (DIC), has gained popularity, as well as implementation in the `WinBUGS` software package. Utility for the latter places us in the space of the observations, considering the performance of *predictive* distributions. One such choice is the posterior predictive criterion due to Gelfand and Ghosh (1998). More recent attention has been paid to proper scoring rules (Gneiting and Raftery, 2007). In any event, Bayesian model selection (and model selection in general) is always a contentious issue; there is rarely a unanimously agreed upon criterion. In this subsection, we limit our attention to Bayes factors, the DIC, the Gelfand/Ghosh criterion and the (continuous) rank probability score (CRPS). A further important point in this regard is the notion of cross-validation or hold out data, i.e., data that is not used to fit the model but rather only for model validation or comparison. See the discussion on the use of hold out samples at the end of this section.

5.2.3.1 Bayes factors

We begin by setting up the hypothesis testing problem as a model choice problem, replacing the customary two hypotheses H_0 and H_A by two candidate parametric models M_1 and M_2 having respective parameter vectors $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$. Under prior densities $\pi_i(\boldsymbol{\theta}_i)$, $i = 1, 2$, the marginal distributions of $\mathbf{Y}$ are found by integrating out the parameters,

$$p(\mathbf{y}|M_i) = \int f(\mathbf{y}|\boldsymbol{\theta}_i, M_i)\pi_i(\boldsymbol{\theta}_i)d\boldsymbol{\theta}_i \ , \ i = 1, 2 \ . \tag{5.5}$$

Bayes' Theorem (5.1) may then be applied to obtain the posterior probabilities $P(M_1|\mathbf{y})$ and $P(M_2|\mathbf{y}) = 1 - P(M_1|\mathbf{y})$ for the two models. The quantity commonly used to summarize these results is the *Bayes factor*, BF, which is the ratio of the posterior odds of M_1 to the prior odds of M_1, given by Bayes' Theorem as

$$\begin{aligned} BF &= \frac{P(M_1|\mathbf{y})/P(M_2|\mathbf{y})}{P(M_1)/P(M_2)} && (5.6) \\ &= \frac{\left[\frac{p(\mathbf{y}|M_1)P(M_1)}{p(\mathbf{y})}\right] / \left[\frac{p(\mathbf{y}|M_2)P(M_2)}{p(\mathbf{y})}\right]}{P(M_1)/P(M_2)} \\ &= \frac{p(\mathbf{y} \mid M_1)}{p(\mathbf{y} \mid M_2)} \ , && (5.7) \end{aligned}$$

the ratio of the observed marginal densities for the two models. Assuming the two models are *a priori* equally probable (i.e., $P(M_1) = P(M_2) = 0.5$), we have that $BF = P(M_1|\mathbf{y})/P(M_2|\mathbf{y})$, the posterior odds of M_1.

Consider the case where both models share the same parametrization (i.e., $\boldsymbol{\theta}_1 = \boldsymbol{\theta}_2 = \boldsymbol{\theta}$), and both hypotheses are simple (i.e., $M_1 : \boldsymbol{\theta} = \boldsymbol{\theta}^{(1)}$ and $M_2 : \boldsymbol{\theta} = \boldsymbol{\theta}^{(2)}$). Then $\pi_i(\boldsymbol{\theta})$ consists of a point mass at $\boldsymbol{\theta}^{(i)}$ for $i = 1, 2$, and so from (5.5) and (5.7) we have

$$BF = \frac{f(\mathbf{y}|\boldsymbol{\theta}^{(1)})}{f(\mathbf{y}|\boldsymbol{\theta}^{(2)})} \ ,$$

which is nothing but the likelihood ratio between the two models. Hence, in the simple-versus-simple setting, the Bayes factor is precisely the odds in favor of M_1 over M_2 *given solely by the data*.

A popular "shortcut" method is the *Bayesian Information Criterion* (BIC) (also known as the *Schwarz Criterion*), the change in which across the two models is given by

$$\Delta BIC = W - (p_2 - p_1) \log n \ , \tag{5.8}$$

where p_i is the number of parameters in model $M_i, i = 1, 2$, and

$$W = -2 \log \left[\frac{\sup_{M_1} f(\mathbf{y}|\boldsymbol{\theta})}{\sup_{M_2} f(\mathbf{y}|\boldsymbol{\theta})} \right] \ ,$$

the usual likelihood ratio test statistic. Schwarz (1978) showed that for nonhierarchical (two-stage) models and large sample sizes n, BIC approximates $-2 \log BF$. An alternative to BIC is the *Akaike Information Criterion* (AIC), which alters (5.8) slightly to

$$\Delta AIC = W - 2(p_2 - p_1) \ . \tag{5.9}$$

Both AIC and BIC are *penalized likelihood ratio* model choice criteria, since both have second terms that act as a penalty, correcting for differences in size between the models (to see this, think of M_2 as the "full" model and M_1 as the "reduced" model). Evidently, using a penalty which depends upon the sample size, BIC criticizes difference in model dimension more strongly than AIC does.

The more serious (and aforementioned) limitation in using Bayes factors or their approximations is that they are not appropriate under noninformative priors. To see this, note that if $\pi_i(\boldsymbol{\theta}_i)$ is improper, then $p(\mathbf{y}|M_i) = \int f(\mathbf{y}|\boldsymbol{\theta}_i, M_i)\pi_i(\boldsymbol{\theta}_i)d\boldsymbol{\theta}_i$ necessarily is as well, and so BF as given in (5.7) is not well defined. While several authors (see, e.g., Berger and Pericchi, 1996; O'Hagan, 1995) have attempted to modify the definition of BF to repair this deficiency, we suggest more informal yet general approaches described below.

5.2.3.2 The DIC criterion

Spiegelhalter et al. (2002) propose a generalization of the AIC, whose asymptotic justification is not appropriate for hierarchical (3 or more level) models. The generalization is based on the posterior distribution of the *deviance* statistic,

$$D(\boldsymbol{\theta}) = -2 \log f(\mathbf{y}|\boldsymbol{\theta}) + 2 \log h(\mathbf{y}) \ , \tag{5.10}$$

where $f(\mathbf{y}|\boldsymbol{\theta})$ is the likelihood function and $h(\mathbf{y})$ is some standardizing function of the data alone. These authors suggest summarizing the *fit* of a model by the posterior expectation of the deviance, $\overline{D} = E_{\theta|y}[D]$, and the *complexity* of a model by the effective number of parameters p_D (which may well be less than the total number of model parameters, due to the borrowing of strength across random effects). In the case of Gaussian models, one can show that a reasonable definition of p_D is the expected deviance minus the deviance evaluated at the posterior expectations,

$$p_D = E_{\theta|y}[D] - D(E_{\theta|y}[\boldsymbol{\theta}]) = \overline{D} - D(\bar{\boldsymbol{\theta}}) \ . \tag{5.11}$$

The *Deviance Information Criterion* (DIC) is then defined as

$$DIC = \overline{D} + p_D = 2\overline{D} - D(\bar{\boldsymbol{\theta}}) \ , \tag{5.12}$$

with smaller values of DIC indicating a better-fitting model. Both building blocks of DIC and p_D, $E_{\theta|y}[D]$ and $D(E_{\theta|y}[\boldsymbol{\theta}])$, are easily estimated via MCMC methods (see below),

enhancing the approach's appeal. Indeed, DIC may be computed automatically for any model in `WinBUGS`.

While the p_D portion of this expression does have meaning in its own right as an effective model size, DIC itself does not, since it has no absolute scale (due to the arbitrariness of the scaling constant $h(\mathbf{y})$, which is often simply set equal to zero). Thus only *differences* in DIC across models are meaningful. Relatedly, when DIC is used to compare nested models in standard exponential family settings, the unnormalized likelihood $L(\boldsymbol{\theta};\mathbf{y})$ is often used in place of the normalized form $f(\mathbf{y}|\boldsymbol{\theta})$ in (5.10), since in this case the normalizing function $m(\boldsymbol{\theta}) = \int L(\boldsymbol{\theta};\mathbf{y})d\mathbf{y}$ will be free of $\boldsymbol{\theta}$ and constant across models, hence contribute equally to the DIC scores of each (and thus have no impact on model selection). However, in settings where we require comparisons across different likelihood distributional forms, generally one must be careful to use the properly scaled joint density $f(\mathbf{y}|\boldsymbol{\theta})$ for each model.

Identification of what constitutes a *significant* difference is also a bit awkward; delta method approximations to $Var(DIC)$ have to date met with little success (Zhu and Carlin, 2000). In practice one typically adopts the informal approach of simply recomputing DIC a few times using different random number seeds, to get a rough idea of the variability in the estimates. With a large number of independent DIC replicates $\{DIC_l,\ l = 1, \ldots, N\}$, one could of course estimate $Var(DIC)$ by its sample variance,

$$\widehat{Var}(DIC) = \frac{1}{N-1}\sum_{l=1}^{N}(DIC_l - \overline{DIC})^2 \ .$$

But in any case, DIC is not intended for formal identification of the "correct" model, but rather merely as a method of comparing a collection of alternative formulations (all of which may be incorrect). This informal outlook (and DIC's approximate nature in markedly nonnormal models) suggests informal measures of its variability will often be sufficient. The p_D statistic is also helpful in its own right, since how close it is to the actual parameter count provides information about how many parameters are actually "needed" to adequately explain the data. For instance, a relatively low p_D may indicate collinear fixed effects or overshrunk random effects; see Exercise 1.

DIC is general, and trivially computed as part of an MCMC run, without any need for extra sampling, reprogramming, or complicated loss function determination. Moreover, experience with DIC to date suggests it works quite well, despite the fact that no formal justification for it is yet available outside of posteriors that can be well approximated by a Gaussian distribution (a condition that typically occurs asymptotically, but perhaps not without a moderate to large sample size for many models). Still, DIC is by no means universally accepted by Bayesians as a suitable all-purpose model choice tool, as the discussion in Spiegelhalter et al. (2002) almost immediately indicates. Model comparison using DIC is not invariant to parametrization, so (as with prior elicitation) the most sensible parametrization must be carefully chosen beforehand. Unknown scale parameters and other innocuous restructuring of the model can also lead to subtle changes in the computed DIC value.

Finally, DIC will obviously depend on what part of the model specification is considered to be part of the likelihood, and what is not. Spiegelhalter et al. (2002) refer to this as the *focus* issue, i.e., determining which parameters are of primary interest, and which should "count" in p_D. For instance, in a hierarchical model with data distribution $f(\mathbf{y}|\boldsymbol{\theta})$, prior $p(\boldsymbol{\theta}|\eta)$ and hyperprior $p(\eta)$, one might choose as the likelihood either the obvious conditional expression $f(\mathbf{y}|\boldsymbol{\theta})$, or the *marginal* expression,

$$p(\mathbf{y}|\eta) = \int f(\mathbf{y}|\boldsymbol{\theta})p(\boldsymbol{\theta}|\eta)d\boldsymbol{\theta} \ . \tag{5.13}$$

We refer to the former case as "focused on $\boldsymbol{\theta}$," and the latter case as "focused on η." Spiegelhalter et al. (2002) defend the dependence of p_D and DIC on the choice of focus as perfectly natural, since while the two foci give rise to the same marginal density $m(y)$, the integration in (5.13) clearly suggests a different model complexity than the unintegrated version (having been integrated out, the θ parameters no longer "count" in the total). They thus argue that it is up to the user to think carefully about which parameters ought to be in focus before using DIC. Perhaps the one difficulty with this advice is that, in cases where the integration in (5.13) is not possible in closed form, the unintegrated version is really the only feasible choice. Indeed, the DIC tool in `WinBUGS` always focuses on the lowest level parameters in a model (in order to sidestep the integration issue), even when the user intends otherwise.

5.2.3.3 Posterior predictive loss criterion

An alternative to DIC that is also easily implemented using output from posterior simulation is the *posterior predictive loss* (performance) approach of Gelfand and Ghosh (1998). Focusing on prediction, in particular, with regard to replicates of the observed data, $Y_{\ell,rep}$, $\ell = 1, \ldots, n$, the selected models are those that perform well under a so-called *balanced* loss function. Roughly speaking, this loss function penalizes actions both for departure from the corresponding observed value ("fit") as well as for departure from what we expect the replicate to be ("smoothness"). The loss puts weights k and 1 on these two components, respectively, to allow for adjustment of relative regret for the two types of departure.

We avoid details here, but note that for squared error loss, the resulting criterion becomes

$$D_k = \frac{k}{k+1}\, G + P \;, \tag{5.14}$$

$$\text{where } G = \sum_{\ell=1}^{n} (\mu_\ell - y_{\ell,obs})^2 \;\text{ and }\; P = \sum_{\ell=1}^{n} \sigma_\ell^2 \;.$$

In (5.14), $\mu_\ell = E(Y_{\ell,rep}|\mathbf{y})$ and $\sigma_\ell^2 = Var(Y_{\ell,rep}|\mathbf{y})$, i.e., the mean and variance of the predictive distribution of $Y_{\ell,rep}$ given the observed data $\mathbf{y}$.

The components of D_k have natural interpretations. G is a goodness-of-fit term, while P is a penalty term. To clarify, we are seeking to penalize complexity and reward parsimony, just as DIC and other penalized likelihood criteria do. For a poor model we expect large predictive variance and poor fit. As the model improves, we expect to do better on both terms. But as we start to overfit, we will continue to do better with regard to goodness of fit, but also begin to inflate the variance (as we introduce multicollinearity). Eventually the resulting increased predictive variance penalty will exceed the gains in goodness of fit. So as with DIC, as we sort through a collection of models, the one with the smallest D_k is preferred. When $k = \infty$ (so that $D_k = D_\infty = G + P$), we will sometimes write D_∞ simply as D for brevity.

Two remarks are appropriate. First, we may report the first and second terms (excluding $k/(k+1)$) on the right side of (5.14), rather than reducing to the single number D_k. Second, in practice, ordering of models is typically insensitive to the particular choice of k.

The quantities μ_ℓ and σ_ℓ^2 can be readily computed from posterior samples. If under model m we have parameters $\boldsymbol{\theta}^{(m)}$, then

$$p(y_{\ell,rep}|\mathbf{y}) = \int p(y_{\ell,rep}|\boldsymbol{\theta}^{(m)})\, p(\boldsymbol{\theta}^{(m)}|\mathbf{y})\, d\boldsymbol{\theta}^{(m)} \;. \tag{5.15}$$

Hence each posterior realization (say, $\boldsymbol{\theta}^*$) can be used to draw a corresponding $y_{\ell,rep}$ from $p(y_{\ell,rep}|\boldsymbol{\theta}^{(m)} = \boldsymbol{\theta}^*)$. The resulting $y^*_{\ell,rep}$ has marginal distribution $p(y_{\ell,rep}|\mathbf{y})$. With samples

from this distribution we can obtain μ_ℓ and σ_ℓ^2. Hence development of D_k requires an extra level of simulation, one for one with the posterior samples.

More general loss functions can be used, including the so-called deviance loss (based upon $p(y_\ell|\boldsymbol{\theta}^{(m)})$), again yielding two terms for D_k with corresponding interpretation and predictive calculation. This enables application to, say, binomial or Poisson likelihoods. We omit details here since in this book, only (5.14) is used for examples that employ this criterion.

5.2.3.4 Model assessment using hold out data

An important point is that all of the foregoing criteria evaluate model performance based upon the data used to fit the model. That is, they use the data twice with regard to model comparison. Arguably, a more attractive way to compare models is to partition the dataset into a fitting (or learning) set and a validation or "hold out" set and apply the criterion to the hold out data after fitting the model to the fitting dataset. This enables us to see how a model will perform with *new* data; using the fitted data to compare models will provide too optimistic an assessment of model performance. Of course, how much data to retain for fitting and how much to use for hold out is not well defined; it depends upon the modeling and the amount of available data. Rough suggestions in the literature suggest holding out as much as $20 - 30\%$. We don't think the amount is a critical issue.

When working with point referenced spatial data, prediction is the primary goal. This is the whole intent of the kriging development in Chapter 6. If we are concerned with the predictive performance of a model, then holding out data becomes natural. In this regard, we have two potential perspectives. The first makes a comparison between an observed value, e.g., an observation at a spatial location and an estimate of this observation, obtained through, say, kriging. Applied to hold out data, it leads to criteria such as predicted mean square or absolute error, i.e., the sum of squared differences or absolute differences across the hold out data.

The second perspective compares an observed value with its associated posterior predictive distribution (see Chapter 6). For instance, using this distribution, we can create a posterior $1 - \alpha$ predictive interval for a held out observation. If we do this across many observations, we can obtain an *empirical* coverage probability (proportion of times the predictive interval contained the observed value) which can be compared with the *nominal* $1-\alpha$ coverage probability. Rough agreement supports the model. Empirical under-coverage suggests that the model is too optimistic with regard to uncertainty, intervals are too short. However, over-coverage is also not desirable. It indicates that if predictive intervals are wider than need be, uncertainty is overestimated.

For model choice, comparing a predictive distribution with an observation takes us into the realm of probabilistic calibration or forecasting, making forecasts for the future and providing suitable measures of the uncertainty associated with them. Probabilistic forecasting has become routine in such applications as weather and climate prediction, computational finance, and macroeconomic forecasting. In our context, the goodness of a predictive distribution relative to an observation is measured by how concentrated the distribution is around the observation. Bypassing all of the elegant theory regarding proper scoring rules (see Gneiting and Raftery, 2007), the proposed measure is the continuous rank probability score (CRPS), the squared integrated distance between the predictive distribution and the degenerate distribution at the observed value,

$$CRPS(F, y) = \int_{-\infty}^{\infty} (F(u) - 1(u \geq y))^2 du \ , \tag{5.16}$$

where F is the predictive distribution and y is the observed value. For us, $Y(\mathbf{s}_0)$ is the observation and F is the posterior predictive distribution for $Y(\mathbf{s}_0)$. With a collection of

such hold out observations and associated predictive distributions, we would sum the CRPS over these observations to create the model comparison criterion. Recall that, under MCMC model fitting, we will not have F explicitly but, rather, a sample from F. Fortunately, Gneiting and Raftery (2007) present a convenient alternative form for (5.16), providing F has a first moment:

$$CRPS(F, y) = \frac{1}{2}E_F|Y - Y'| + E_F|Y - y| \tag{5.17}$$

where Y and Y' are independent replicates from F. With samples from F, we have immediate Monte Carlo integrations to compute (5.17).

Finally, we do not recommend a choice among the model comparison approaches discussed. Again, DIC works in the parameter space with the likelihood, while the predictive approaches work in the space of the data with posterior predictive distributions. The former addresses comparative explanatory performance, while the latter addresses comparative predictive performance. So, if the objective is to use the model for explanation, we may prefer DIC; if instead the objective is prediction, we may prefer a predictive criterion. In different terms, we can argue that, since, with areal unit data, we are most interested in explanation, DIC will be attractive. In fact, the notion of hold out data is not clear for this setting. By contrast, with point referenced data, we can easily hold out some locations; predictive validation/comparison seems the more attractive path.

5.3 Bayesian computation

As mentioned above, in this section we provide a brief introduction to Bayesian computing, following the development in Chapter 3 of Carlin and Louis (2008). The explosion in Bayesian activity and computing power in the past 20 or so years has caused a similar explosion in the number of books in this area. The earliest comprehensive treatment was by Tanner (1996), with books by Gilks et al. (1996), Gamerman (1997), and Chen et al. (2000) offering updated and expanded discussions that are primarily Bayesian in focus. Also significant are the computing books by Liu (2001) and Robert and Casella (2004, 2009), which, while not specifically Bayesian, still emphasize Markov chain Monte Carlo methods typically used in modern Bayesian analysis. Perhaps the most current and comprehensive summary of this activity appears in Brooks, Gelman, Jones, and Meng (2010).

Without doubt, the most widely used computing tools in Bayesian practice today are Markov chain Monte Carlo (MCMC) methods. This is due to their ability (in principle) to enable inference from posterior distributions of arbitrarily large dimension, essentially by reducing the problem to one of recursively addressing a series of lower-dimensional (often unidimensional) problems. Like traditional Monte Carlo methods, MCMC methods work by producing not a closed form for the posterior (of a feature of interest) in (5.1), but a *sample* of values $\{\boldsymbol{\theta}^{(g)}, g = 1, \ldots, G\}$ from this distribution. In this sense, we revert to the most basic of statistical ideas in order to learn about a distribution/population, sample from it. While sampling obviously does not carry as much information as the closed form itself, a histogram or kernel density estimate based on such a sample is typically sufficient for reliable inference; moreover such an estimate can be made arbitrarily accurate merely by increasing the Monte Carlo sample size G. However, unlike traditional Monte Carlo methods, MCMC algorithms produce *correlated* samples from this posterior, since they arise as iterative draws from a particular Markov chain, the stationary distribution of which is the same as the posterior.

The convergence of the Markov chain to the correct stationary distribution can be guaranteed for an enormously broad class of posteriors, explaining MCMC's popularity. But this convergence is also the source of most of the difficulty in actually implementing MCMC procedures, for two reasons. First, it forces us to make a decision about when it is safe to stop the sampling algorithm and summarize its output, an area known in the business as *convergence diagnosis*. Second, it clouds the determination of the quality of the estimates

produced (since they are based not on i.i.d. draws from the posterior, but on correlated samples. This is sometimes called the *variance estimation* problem, since a common goal here is to estimate the Monte Carlo variances (equivalently standard errors) associated with our MCMC-based posterior estimates.

In the remainder of this section, we introduce the three most popular MCMC algorithms, the Gibbs sampler, the Metropolis-Hastings algorithm and the slice sampler. We then return to the convergence diagnosis and variance estimation problems.

5.3.1 The Gibbs sampler

Suppose our model features k parameters, $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)'$. To implement the Gibbs sampler, we must assume that samples can be generated from each of the *full* or *complete* conditional distributions $\{p(\theta_i \mid \boldsymbol{\theta}_{j\neq i}, \mathbf{y}),\ i = 1, \ldots, k\}$ in the model. Such samples might be available directly (say, if the full conditionals were familiar forms, like normals and gammas) or indirectly (say, via a rejection sampling approach). In this latter case two popular alternatives are the adaptive rejection sampling (ARS) algorithm of Gilks and Wild (1992), and the Metropolis algorithm described in the next subsection. In either case, under mild conditions, the collection of full conditional distributions uniquely determine the joint posterior distribution, $p(\boldsymbol{\theta}|\mathbf{y})$, and hence all marginal posterior distributions $p(\theta_i|\mathbf{y}),\ i = 1, \ldots, k$.

Given an arbitrary set of starting values $\{\theta_2^{(0)}, \ldots, \theta_k^{(0)}\}$, the algorithm proceeds as follows:

Gibbs Sampler: For $(t \in 1 : T)$, repeat:

Step 1: Draw $\theta_1^{(t)}$ from $p\left(\theta_1 \mid \theta_2^{(t-1)}, \theta_3^{(t-1)}, \ldots, \theta_k^{(t-1)}, \mathbf{y}\right)$

Step 2: Draw $\theta_2^{(t)}$ from $p\left(\theta_2 \mid \theta_1^{(t)}, \theta_3^{(t-1)}, \ldots, \theta_k^{(t-1)}, \mathbf{y}\right)$

$\vdots$

Step k: Draw $\theta_k^{(t)}$ from $p\left(\theta_k \mid \theta_1^{(t)}, \theta_2^{(t)}, \ldots, \theta_{k-1}^{(t)}, \mathbf{y}\right)$

Under mild regularity conditions that are satisfied for most statistical models (see, e.g., Geman and Geman, 1984, or Roberts and Smith, 1993), one can show that the k-tuple obtained at iteration t, $(\theta_1^{(t)}, \ldots, \theta_k^{(t)})$, converges in distribution to a draw from the true joint posterior distribution $p(\theta_1, \ldots, \theta_k|\mathbf{y})$. This means that for t sufficiently large (say, bigger than some t_0), $\{\boldsymbol{\theta}^{(t)},\ t = t_0 + 1, \ldots, T\}$ is essentially a (correlated) sample from the true posterior, from which any posterior quantities of interest may be estimated. For example, a histogram of the $\{\theta_i^{(t)},\ t = t_0 + 1, \ldots, T\}$ themselves provides a simulation-consistent estimator of the marginal posterior distribution for θ_i, $p(\theta_i \mid \mathbf{y})$. We might also use a sample mean to estimate the posterior mean, i.e.,

$$\widehat{E}(\theta_i|\mathbf{y}) = \frac{1}{T - t_0} \sum_{t=t_0+1}^{T} \theta_i^{(t)} \ . \tag{5.18}$$

Similarly, the sample variance provides an estimate of the posterior variance. Quantiles of the sample provide quantiles of the posterior distribution. The time from $t = 0$ to $t = t_0$ is commonly known as the *burn-in* period; popular methods for selection of an appropriate t_0 are discussed below.

In practice, we may actually run m *parallel* Gibbs sampling chains, instead of only 1, for some modest m (say, $m = 5$). We will see below that such parallel chains may be useful in assessing convergence of the sampler, and, in any event, can be produced with no extra time on a multiprocessor computer. By analogy with (5.18), in this case, we would again

discard all samples from the burn-in period, obtaining the posterior mean estimate,

$$\widehat{E}(\theta_i|\mathbf{y}) = \frac{1}{m(T-t_0)} \sum_{j=1}^{m} \sum_{t=t_0+1}^{T} \theta_{i,j}^{(t)} \ , \tag{5.19}$$

where now the second subscript on $\theta_{i,j}$ indicates chain number. Again we defer comment on the issues of how to choose t_0 and how to assess the quality of (5.19) and related estimators for the moment.

As a historical footnote, we add that Geman and Geman (1984) chose the name "Gibbs sampler" because the distributions used in their context (image restoration, where the parameters were actually the colors of pixels on a screen) were Gibbs distributions (as previously seen in equation (4.9)). These were in turn named after J.W. Gibbs, a 19th-century American physicist and mathematician generally regarded as one of the founders of modern thermodynamics and statistical mechanics. While Gibbs distributions form an exponential family on potentials that includes most standard statistical models as special cases, most Bayesian applications do not require anywhere near this level of generality, typically dealing solely with standard statistical distributions (normal, gamma, etc.). Yet, despite a few attempts by some Bayesians to choose a more descriptive name (e.g., the "successive substitution sampling" (SSS) moniker due to Schervish and Carlin, 1992), the Gibbs sampler name has stuck. As such the Gibbs sampler is yet another example of Stigler's Law of Eponymy, which states that no scientific discovery is named for the person(s) who actually thought of it. (In fact, Stigler's Law of Eponymy is not due to Stigler (1999), meaning that it is an example of itself!)

5.3.2 The Metropolis-Hastings algorithm

The Gibbs sampler is easy to understand and implement, but requires the ability to readily sample from each of the full conditional distributions, $p(\theta_i \mid \boldsymbol{\theta}_{j\neq i}, \mathbf{y})$. Unfortunately, when the prior distribution $p(\boldsymbol{\theta})$ and the likelihood $f(\mathbf{y}|\boldsymbol{\theta})$ are not a conjugate pair, one or more of these full conditionals may not be available in closed form. Even in this setting, however, $p(\theta_i \mid \boldsymbol{\theta}_{j\neq i}, \mathbf{y})$ *will* be available up to a proportionality constant, since it is proportional to the portion of $f(\mathbf{y}|\boldsymbol{\theta}) \times p(\boldsymbol{\theta})$ that involves θ_i.

The *Metropolis algorithm* (or *Metropolis-Hastings algorithm*) is a rejection algorithm that attacks precisely this problem, since it requires only a function proportional to the distribution to be sampled, at the cost of requiring a rejection step from a particular *candidate* density. Like the Gibbs sampler, this algorithm was not developed by statistical data analysts for this purpose, but by statistical physicists working on the Manhattan Project in the 1940s seeking to understand the particle movement theory underlying the first atomic bomb (one of the coauthors on the original Metropolis et al. (1953) paper was Edward Teller, who is often referred to as "the father of the hydrogen bomb"). In this regard, it was used to implement forward simulation of realizations (scenarios) under a complex, high-dimensional model. Fortunately, in this context, it was never fitted to real data!

While as mentioned above our main interest in the algorithm is for generation from (typically univariate) full conditionals, it is most easily described (and theoretically supported) for the full multivariate θ vector. Thus, suppose for now that we wish to generate from a joint posterior distribution distribution $p(\boldsymbol{\theta}|\mathbf{y}) \propto h(\boldsymbol{\theta}) \equiv f(\mathbf{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})$. We begin by specifying a candidate density $q(\boldsymbol{\theta}^*|\boldsymbol{\theta}^{(t-1)})$ that is a valid density function for every possible value of the conditioning variable $\boldsymbol{\theta}^{(t-1)}$, and satisfies $q(\boldsymbol{\theta}^*|\boldsymbol{\theta}^{(t-1)}) = q(\boldsymbol{\theta}^{(t-1)}|\boldsymbol{\theta}^*)$, i.e., q is *symmetric* in its arguments. Given a starting value $\boldsymbol{\theta}^{(0)}$ at iteration $t = 0$, the algorithm proceeds as follows:

Metropolis Algorithm: For $(t \in 1:T)$, repeat:

1. Draw $\boldsymbol{\theta}^*$ from $q(\cdot|\boldsymbol{\theta}^{(t-1)})$
2. Compute the ratio $r = h(\boldsymbol{\theta}^*)/h(\boldsymbol{\theta}^{(t-1)}) = \exp[\log h(\boldsymbol{\theta}^*) - \log h(\boldsymbol{\theta}^{(t-1)})]$
3. If $r \geq 1$, set $\boldsymbol{\theta}^{(t)} = \boldsymbol{\theta}^*$;
   If $r < 1$, set $\boldsymbol{\theta}^{(t)} = \begin{cases} \boldsymbol{\theta}^* \text{ with probability } r \\ \boldsymbol{\theta}^{(t-1)} \text{ with probability } 1-r \end{cases}$.

Then under generally the same mild conditions as those supporting the Gibbs sampler, a draw $\boldsymbol{\theta}^{(t)}$ converges in distribution to a draw from the true posterior density $p(\boldsymbol{\theta}|\mathbf{y})$. Note however that when the Metropolis algorithm (or the Metropolis-Hastings algorithm below) is used to update within a Gibbs sampler, it never samples from the full conditional distribution. Convergence using Metropolis steps, then, would be expected to be slower than that for a regular Gibbs sampler.

Recall that the steps of the Gibbs sampler were fully determined by the statistical model under consideration (since full conditional distributions for well-defined models are unique). By contrast, the Metropolis algorithm affords substantial flexibility through the selection of the candidate density q. This flexibility can be a blessing and a curse: while theoretically we are free to pick almost any candidate density, in practice only a "good" choice will result in sufficiently many candidate acceptances. The usual approach (after $\boldsymbol{\theta}$ has been transformed to have support $\Re^k$, if necessary) is to set

$$q(\boldsymbol{\theta}^*|\boldsymbol{\theta}^{(t-1)}) = N(\boldsymbol{\theta}^*|\boldsymbol{\theta}^{(t-1)}, \widetilde{\Sigma}) \ , \tag{5.20}$$

since this distribution obviously satisfies the symmetry property, and is "self correcting" (candidates are always centered around the current value of the chain). Equation (5.20) is usually referred to as a "random walk" proposal density. Specification of q then comes down to specification of $\widetilde{\Sigma}$. Here we might try to mimic the posterior variance by setting $\widetilde{\Sigma}$ equal to an empirical estimate of the true posterior variance, derived from a preliminary sampling run.

The reader might well imagine an optimal choice of q would produce an empirical acceptance ratio of 1, the same as the Gibbs sampler (and with no apparent "waste" of candidates). However, the issue is rather more subtle than this: accepting all or nearly all of the candidates is often the result of an overly narrow candidate density. Such a density will "baby-step" around the parameter space, leading to high acceptance but also high autocorrelation in the sampled chain and slow exploration of the support for the distribution. An overly dispersed candidate density will also struggle, proposing leaps to places far from the bulk of the support of the posterior, leading to high rejection and, again, high autocorrelation. Thus the "folklore" here is to choose $\widetilde{\Sigma}$ so that roughly 50% of the candidates are accepted. Subsequent theoretical work (e.g., Gelman et al., 1996) indicates even lower acceptance rates (25 to 40%) are optimal, but this result varies with the dimension and true posterior correlation structure of $\boldsymbol{\theta}$.

As a result, choice of $\widetilde{\Sigma}$ is often done *adaptively*. For instance, in one dimension (setting $\widetilde{\Sigma} = \widetilde{\sigma}$, and thus avoiding the issue of correlations among the elements of $\boldsymbol{\theta}$), a common trick is to simply pick some initial value of $\widetilde{\sigma}$, and then keep track of the empirical proportion of candidates that are accepted. If this fraction is too high (75 to 100%), we simply increase $\widetilde{\sigma}$; if it is too low (0 to 20%), we decrease it. Since adaptation infinitely often can actually disturb the chain's convergence to the desired stationary distribution, the simplest approach is to allow this adaptation only during the burn-in period, a practice sometimes referred to as *pilot adaptation*. This is in fact the approach currently used by `WinBUGS`, where the default pilot period is 4000 iterations. A more involved alternative is to allow adaptation at *regeneration points* which, once defined and identified, break the Markov chain into independent sections. See, e.g., Mykland, Tierney and Yu (1995), Mira and Sargent (2000), and Hobert et al. (2002) for discussions of the use of regeneration in practical MCMC settings.

As mentioned above, in practice the Metropolis algorithm is often employed as a substep in a larger Gibbs sampling algorithm, used to generate from awkward full conditionals. Such hybrid Gibbs-Metropolis applications were once known as "Metropolis within Gibbs" or "Metropolis substeps," and users would worry about how many such substeps should be used. Fortunately, it was soon realized that a single substep was sufficient to ensure convergence of the overall algorithm, and so this is now standard practice: when we encounter an awkward full conditional (say, for θ_i), we simply draw one Metropolis candidate, accept or reject it, and move on to θ_{i+1}. Further discussion of convergence properties and implementation of hybrid MCMC algorithms can be found in Tierney (1994) and Carlin and Louis (2000, Sec. 5.4.4).

We end this subsection with the important generalization of the Metropolis algorithm devised by Hastings (1970). In this variant we drop the requirement that q be symmetric in its arguments, which is often useful for bounded parameter spaces (say, $\theta > 0$) where Gaussian proposals as in (5.20) are not natural.

Metropolis-Hastings Algorithm: In Step 2 of the Metropolis algorithm above, replace the acceptance ratio r by

$$r = \frac{h(\boldsymbol{\theta}^*)q(\boldsymbol{\theta}^{(t-1)} \mid \boldsymbol{\theta}^*)}{h(\boldsymbol{\theta}^{(t-1)})q(\boldsymbol{\theta}^* \mid \boldsymbol{\theta}^{(t-1)})} \,. \tag{5.21}$$

Then again under mild conditions, a draw $\boldsymbol{\theta}^{(t)}$ converges in distribution to a draw from the true posterior density $p(\boldsymbol{\theta}|\mathbf{y})$ as $t \to \infty$.

In practice we often set $q(\boldsymbol{\theta}^* \,|\, \boldsymbol{\theta}^{(t-1)}) = q(\boldsymbol{\theta}^*)$, i.e., we use a proposal density that ignores the current value of the variable. This algorithm is sometimes referred to as a *Hastings independence chain*, so named because the proposals (though not the final $\boldsymbol{\theta}^{(t)}$ values) form an independent sequence. While easy to implement, this algorithm can be difficult to tune; it will converge slowly unless the chosen q is rather close to the true posterior (which is of course unknown in advance).

5.3.3 Slice sampling

An alternative to the Metropolis-Hastings algorithm that is still quite general is *slice sampling* (Neal, 2003). In its most basic form, suppose we seek to sample a univariate $\theta \sim f(\theta) \equiv h(\theta)/\int h(\theta)d\theta$, where $h(\theta)$ is known. Suppose we add a so-called *auxiliary variable* U such that $U|\theta \sim Unif(0, h(\theta))$. Then the joint distribution of θ and U is $p(\theta, u) \propto 1 \cdot I(U < h(\theta))$, where I denotes the indicator function. If we run a Gibbs sampler drawing from $U|\theta$ followed by $\theta|U$ at each iteration, we can obtain samples from $p(\theta, u)$, and hence from the marginal distribution of θ, $f(\theta)$. Sampling from $\theta|u$ requires a draw from a uniform distribution for θ over the set $S_U = \{\theta : U < h(\theta)\}$.

Figure 5.1 reveals why this approach is referred to as slice sampling. U "slices" the nonnormalized density, and the resulting "footprint" on the axis provides S_U. If we can enclose S_U in an interval, we can draw θ uniformly on this interval and simply retain it only if $U < h(\theta)$ (i.e., if $\theta \in S_U$). If $\boldsymbol{\theta}$ is instead multivariate, S_U is more complicated and now we would need a bounding rectangle.

Note that if $h(\theta) = h_1(\theta)h_2(\theta)$ where, say, h_1 is a standard density that is easy to sample, while h_2 is nonstandard and difficult to sample, then we can introduce an auxiliary variable U such that $U|\theta \sim U(0, h_2(\theta))$. Now $p(\theta, u) = h_1(\theta)I(U < h_2(\theta))$. Again $U|\theta$ is routine to sample, while to sample $\theta|U$ we would now draw θ from $h_1(\theta)$ and retain it only if θ is such that $U < h_2(\theta)$.

Slice sampling incurs problems similar to rejection sampling in that we may have to draw many θ's from h_1 before we are able to retain one. On the other hand, it has an advantage over the Metropolis-Hastings algorithm in that it always samples from the exact full conditional $p(\theta|u)$. As noted above, Metropolis-Hastings does not, and thus slice sampling would

Figure 5.1 *Illustration of slice sampling. For this bimodal distribution, S_U is the union of two disjoint intervals.*

be expected to converge more rapidly. Nonetheless, overall comparison of computation time may make one method a winner for some cases, and the other a winner in other cases. We do remark that slice sampling is attractive for fitting a large range of point-referenced spatial data models, as we detail in Appendix Section A.2 following Agarwal and Gelfand (2005). In fact, it has become useful in the context of spatial point pattern data for fitting so-called log Gaussian Cox processes. See Section 8.4.3 and the papers Murray et al. (2010) and Murray and Adams (2010).

5.3.4 Convergence diagnosis

As mentioned above, the most problematic part of MCMC computation is deciding when it is safe to stop the algorithm and summarize the output. This means we must make a guess as to the iteration t_0 after which all output may be thought of as coming from the true stationary distribution of the Markov chain (i.e., the true posterior distribution). The most common approach here is to run a few (say, $m = 3$ or 5) *parallel* sampling chains, initialized at widely disparate starting locations that are overdispersed with respect to the true posterior. These chains are then plotted on a common set of axes, and these *trace plots* are then viewed to see if there is an identifiable point t_0 after which all m chains seem to be "overlapping" (traversing the same part of $\boldsymbol{\theta}$-space).

Sadly, there are obvious problems with this approach. First, since the posterior is unknown at the outset, there is no reliable way to ensure that the m chains are "initially overdispersed," as required for a convincing diagnostic. We might use extreme quantiles of the prior $p(\boldsymbol{\theta})$ and rely on the fact that the support of the posterior is typically a subset of that of the prior, but this requires a proper prior and in any event is perhaps doubtful in high-dimensional or otherwise difficult problems. Second, it is hard to see how to automate such a diagnosis procedure, since it requires a subjective judgment call by a human viewer. A great many papers have been written on various convergence diagnostic statistics that summarize MCMC output from one or many chains that may be useful when associated

with various stopping rules; see Cowles and Carlin (1996) and Mengersen et al. (1999) for reviews of many such diagnostics.

Among the most popular diagnostic is that of Gelman and Rubin (1992). Here, we run a small number (m) of parallel chains with different starting points that are "initially overdispersed" with respect to the true posterior. (Of course, since we don't know the true posterior before beginning there is technically no way to ensure this; still, the rough location of the bulk of the posterior may be discernible from known ranges, the support of the (proper) prior, or perhaps a preliminary posterior mode-finding algorithm.) Running the m chains for $2N$ iterations each, we then try to see whether the variation within the chains for a given parameter of interest λ approximately equals the total variation across the chains during the latter N iterations. Specifically, we monitor convergence by the estimated *scale reduction factor*,

$$\sqrt{\hat{R}} = \sqrt{\left(\frac{N-1}{N} + \frac{m+1}{mN}\frac{B}{W}\right)\frac{df}{df-2}}\,, \tag{5.22}$$

where B/N is the variance between the means from the m parallel chains, W is the average of the m within-chain variances, and df is the degrees of freedom of an approximating t density to the posterior distribution. Equation (5.22) is the factor by which the scale parameter of the t density might shrink if sampling were continued indefinitely; the authors show it must approach 1 as $N \to \infty$.

The approach is fairly intuitive and is applicable to output from any MCMC algorithm. However, it focuses only on detecting bias in the MCMC estimator; no information about the *accuracy* of the resulting posterior estimate is produced. It is also an inherently univariate quantity, meaning it must be applied to each parameter (or parametric function) of interest in turn, although Brooks and Gelman (1998) extend the Gelman and Rubin approach in three important ways, one of which is a multivariate generalization for simultaneous convergence diagnosis of every parameter in a model.

While the Gelman-Rubin-Brooks and other formal diagnostic approaches remain popular, in practice very simple checks often work just as well and may even be more robust against "pathologies" (e.g., multiple modes) in the posterior surface that may easily fool some diagnostics. For instance, sample autocorrelations in any of the observed chains can inform about whether slow traversing of the posterior surface is likely to impede convergence. Sample cross-correlations (i.e., correlations between two different parameters in the model) may identify ridges in the surface (say, due to collinearity between two predictors) that will again slow convergence; such parameters may need to be updated in multivariate blocks, or one of the parameters dropped from the model altogether. Combined with a visual inspection of a few sample trace plots, the user can at least get a good feeling for whether posterior estimates produced by the sampler are likely to be reliable. However, as a caveat, all convergence diagnostics explore solely MCMC output. They never compare proximity of the output to the truth since the truth is not known (if it was, we wouldn't be implementing an MCMC algorithm).

5.3.5 Variance estimation

An obvious criticism of Monte Carlo methods generally is that no two analysts will obtain the same answer, since the components of the estimator are random. This makes assessment of the variance of these estimators an important point. Combined with a central limit theorem, the result would be an ability to test whether two Monte Carlo estimates were significantly different. For example, suppose we have a single chain of N post-burn-in samples of a parameter of interest λ, so that our basic posterior mean estimator (5.18) becomes $\hat{E}(\lambda|\mathbf{y}) = \hat{\lambda}_N = \frac{1}{N}\sum_{t=1}^{N}\lambda^{(t)}$. Assuming the samples comprising this estimator are

independent, a variance estimate for it would be given by

$$\widehat{Var}_{iid}(\hat{\lambda}_N) = s_\lambda^2/N = \frac{1}{N(N-1)} \sum_{t=1}^{N} (\lambda^{(t)} - \hat{\lambda}_N)^2 \ , \tag{5.23}$$

i.e., the sample variance, $s_\lambda^2 = \frac{1}{N-1} \sum_{t=1}^{N} (\lambda^{(t)} - \hat{\lambda}_N)^2$, divided by N. But while this estimate is easy to compute, it would very likely be an *underestimate* due to positive autocorrelation in the MCMC samples. One can resort to *thinning*, which is simply retaining only every kth sampled value, where k is the approximate lag at which the autocorrelations in the chain become insignificant. However, MacEachern and Berliner (1994) show that such thinning from a stationary Markov chain always increases the variance of sample mean estimators, and is thus suboptimal. This is intuitively reminiscent of Fisher's view of sufficiency: it is never a good idea to throw away information (in this case, $(k-1)/k$ of our MCMC samples) just to achieve approximate independence among those that remain.

A better alternative is to use all the samples, but in a more sophisticated way. One such alternative uses the notion of *effective sample size*, or ESS (Kass et al., 1998, p. 99). ESS is defined as

$$ESS = N/\kappa(\lambda) \ ,$$

where $\kappa(\lambda)$ is the *autocorrelation time* for λ, given by

$$\kappa(\lambda) = 1 + 2 \sum_{k=1}^{\infty} \rho_k(\lambda) \ , \tag{5.24}$$

where $\rho_k(\lambda)$ is the autocorrelation at lag k for the parameter of interest λ. We may estimate $\kappa(\lambda)$ using sample autocorrelations estimated from the MCMC chain. The variance estimate for $\hat{\lambda}_N$ is then

$$\widehat{Var}_{ESS}(\hat{\lambda}_N) = s_\lambda^2/ESS(\lambda) = \frac{\kappa(\lambda)}{N(N-1)} \sum_{t=1}^{N} (\lambda^{(t)} - \hat{\lambda}_N)^2 \ .$$

Note that unless the $\lambda^{(t)}$ are uncorrelated, $\kappa(\lambda) > 1$ and $ESS(\lambda) < N$, so that $\widehat{Var}_{ESS}(\hat{\lambda}_N) > \widehat{Var}_{iid}(\hat{\lambda}_N)$, in concert with intuition. That is, since we have fewer than N effective samples, we expect some inflation in the variance of our estimate.

In practice, the autocorrelation time $\kappa(\lambda)$ in (5.24) is often estimated simply by cutting off the summation when the magnitude of the terms first drops below some "small" value (say, 0.1). This procedure is simple but may lead to a biased estimate of $\kappa(\lambda)$. Gilks et al. (1996, pp. 50–51) recommend an *initial convex sequence estimator* mentioned by Geyer (1992) which, while still output-dependent and slightly more complicated, actually yields a consistent (asymptotically unbiased) estimate here.

A final and somewhat simpler (though also more naive) method of estimating $Var(\hat{\lambda}_N)$ is through *batching*. Here we divide our single long run of length N into m successive batches of length k (i.e., $N = mk$), with batch means $B_1, \ldots, B_m$. Clearly $\hat{\lambda}_N = \bar{B} = \frac{1}{m} \sum_{i=1}^{m} B_i$. We then have the variance estimate

$$\widehat{Var}_{batch}(\hat{\lambda}_N) = \frac{1}{m(m-1)} \sum_{i=1}^{m} (B_i - \hat{\lambda}_N)^2 \ , \tag{5.25}$$

provided that k is large enough so that the correlation between batches is negligible, and m is large enough to reliably estimate $Var(B_i)$. It is important to verify that the batch means are indeed roughly independent, say, by checking whether the lag 1 autocorrelation of the B_i is less than 0.1. If this is not the case, we must increase k (hence N, unless the current m is already quite large), and repeat the procedure.

Regardless of which of the above estimates, $\hat{V}$, is used to approximate $Var(\hat{\lambda}_N)$, a 95% confidence interval for $E(\lambda|\mathbf{y})$ is then given by

$$\hat{\lambda}_N \pm z_{.025}\sqrt{\hat{V}}\ ,$$

where $z_{.025} = 1.96$, the upper .025 point of a standard normal distribution. If the batching method is used with fewer than 30 batches, it is a good idea to replace $z_{.025}$ by $t_{m-1,.025}$, the upper .025 point of a t distribution with $m-1$ degrees of freedom. `WinBUGS` offers both naive (5.23) and batched (5.25) variance estimates; this software is the subject of the next section.

5.4 Computer tutorials

5.4.1 Basic Bayesian modeling in `R`

In this subsection we merely point out that for simple (typically low-dimensional) Bayesian calculations employing standard likelihoods paired with conjugate priors, the built-in density, quantile, and plotting functions in standard statistical packages may well offer sufficient power; there is no need to use a "Bayesian" package per se. In such cases, statisticians might naturally turn to `R` due to its broad array of libraries and special functions (especially those offering summaries of standard distributions), graphics, interactive environments, and easy extendability.

As a concrete example, suppose we are observing a data value Y from a $Bin(n, \theta)$ distribution, with density proportional to

$$p(y|\theta) \propto \theta^y (1-\theta)^{n-y}\ . \tag{5.26}$$

The $Beta(\alpha, \beta)$ distribution offers a conjugate prior for this likelihood, since its density is proportional to (5.26) as a function of θ, namely

$$p(\theta) \propto \theta^{\alpha-1}(1-\theta)^{\beta-1}\ . \tag{5.27}$$

Using Bayes' Rule (5.1), it is clear that

$$\begin{aligned} p(\theta|y) &\propto \theta^{y+\alpha-1}(1-\theta)^{n-y+\beta-1} \\ &\propto Beta(y+\alpha,\ n-y+\beta)\ , \end{aligned} \tag{5.28}$$

another Beta distribution.

Now consider a setting where $n = 10$ and we observe $Y = y_{obs} = 7$. Choosing $\alpha = \beta = 1$ (i.e., a uniform prior for θ), the posterior is a $Beta(y_{obs}+1, n-y_{obs}+1) = Beta(8, 4)$ distribution. In `R` we can obtain a plot of this distribution by typing

```
> theta <- seq(from=0, to=1, length=101)
> yobs <- 7; n <- 10
> plot(theta, dbeta(theta, yobs+1, n-yobs+1), type="l",
    ylab="posterior density",xlab="")
```

The posterior median may be obtained as

```
> qbeta(.5, yobs+1, n-yobs+1)
```

while the endpoints of a 95% equal-tail credible interval are

```
> qbeta(c(.025, .975), yobs+1, n-yobs+1)
```

In fact, these points may be easily added to our posterior plot (see Figure 5.2) by typing

```
> abline(v=qbeta(.5, yobs+1, n-yobs+1))
> abline(v=qbeta(c(.025, .975), yobs+1, n-yobs+1), lty=2)
```

Figure 5.2 *Illustrative beta posterior, with vertical reference lines added at the .025, .5, and .975 quantiles.*

The `pbeta` and `rbeta` functions may be used similarly to obtain prespecified posterior probabilities (say, $Pr(\theta < 0.8|y_{obs})$) and random draws from the posterior, respectively.

Indeed, similar density, quantile, cumulative probability, and random generation routines are available in `R` for a wide array of standard distributional families that often emerge as posteriors (gamma, normal, multivariate normal, Dirichlet, etc.). Thus in settings where MCMC techniques are unnecessary, these languages may offer the most sensible approach. They are especially useful in situations requiring code to be wrapped around statements like those above so that repeated posterior calculations may be performed. For example, when designing an experiment to be analyzed at some later date using a Bayesian procedure, we would likely want to simulate the procedure's performance in repeated sampling (the Bayesian analog of a power or "sample size" calculation). Such repeated sampling might be of the data for fixed parameters, or over both the data and the parameters. (We hasten to add that `WinBUGS` can be called from `R`, albeit in a special way; see `www.stat.columbia.edu/~gelman/bugsR/`. Future releases of `WinBUGS` may be available directly within `R` itself.)

5.4.2 Advanced Bayesian modeling in `WinBUGS`

In this subsection we provide a introduction to Bayesian data analysis in `WinBUGS`, the most well-developed and general Bayesian software package available to date. `WinBUGS` is the `Windows` successor to `BUGS`, a UNIX package whose name originally arose as a humorous acronym for `Bayesian inference Using Gibbs Sampling`. The package is freely available from the website `http://www.mrc-bsu.cam.ac.uk/bugs/welcome.shtml`. The software comes with a user manual, as well as two examples manuals that are enormously helpful for learning the language and various strategies for Bayesian data analysis.

We remark that for further examples of good applied Bayesian work, in addition to the fine book by Gilks et al. (1996), there are the series of "Bayesian case studies" books by Gatsonis et al. (1993, 1995, 1997, 1999, 2002, 2003) and the Bayesian modeling book by Congdon (2001). While this lattermost text assumes a walking familiarity with the Bayesian approach, it also includes a great many examples and corresponding computer code for their implementation in `WinBUGS`.

WinBUGS has an interactive environment that enables the user to specify models (hierarchical) and it actually performs Gibbs sampling to generate posterior samples. Convergence diagnostics, model checks and comparisons, and other helpful plots and displays are also available. We will now look at some WinBUGS code for greater insight into its modeling language.

Example 5.3 The line example from the main WinBUGS manual will be considered in stages, in order to both check the installation and to illustrate the use of WinBUGS.

Consider a set of 5 (obviously artificial) (X, Y) pairs: (1, 1), (2, 3), (3, 3), (4, 3), (5, 5). We shall fit a simple linear regression of Y on X using the notation,

$$\begin{aligned} Y_i &\sim N\left(\mu_i, \sigma^2\right) , \\ \text{where } \mu_i &= \alpha + \beta x_i . \end{aligned}$$

As the WinBUGS code below illustrates, the language allows a concise expression of the model, where dnorm(a,b) denotes a normal distribution with mean a and *precision* (reciprocal of the variance) b, and dgamma(c,d) denotes a gamma distribution with mean c/d and variance c/d^2. The data means mu[i] are specified using a *logical* link (denoted by <-), instead of a *stochastic* one (denoted by ∼). The second logical expression allows the standard deviation σ to be estimated.

```
model
  {
  for(i in 1:N){
      Y[i] ~ dnorm(mu[i], tau)
        mu[i] <- alpha + beta * x[i]
    }
  sigma <- 1/sqrt(tau)
  alpha ~ dnorm(0, 1.0E-6)
  beta ~ dnorm(0, 1.0E-6)
  tau ~ dgamma(1.0E-3, 1.0E-3)
  }
```

The parameters in the Gibbs sampling order here will be α, β, and $\tau \equiv 1/\sigma^2$; note all are given proper but minimally informative prior distributions.

We next need to load in the data. The data can be represented using R object notation as: list(x = c(1, 2, 3, 4, 5), Y = c(1, 3, 3, 3, 5), N = 5), or as a combination of an R object and a rectangular array with labels at the head of the columns:

```
list(N=5)
x[ ]   Y[ ]
  1     1
  2     3
  3     3
  4     3
  5     5
```

Implementation of this code in WinBUGS is most easily accomplished by pointing and clicking through the menu on the Model/Specification, Inference/Samples, and Inference/Update tools.

Example 5.4 Consider a basic kriging model of the form

$$\begin{aligned} \mathbf{Y} &\sim MVN\left(\boldsymbol{\mu}, w^2 H(\phi) + v^2 I\right) , \\ \text{where } \boldsymbol{\mu} &= X\boldsymbol{\beta} . \end{aligned}$$

Here I is an $N \times N$ identity matrix, while $\Sigma = w^2 H(\phi)$, an $N \times N$ correlation matrix of the form $H(\phi)_{ij} = \exp(-\phi d_{ij})$ where as usual d_{ij} is the distance between locations i and j.

What follows is some WinBUGS code to do this problem directly, i.e., using the multivariate normal distribution dnorm and constructing the H matrix directly using the exponential (exp) and power (pow) functions.

```
model
{
for(i in 1:N) {
  Y[i] ~ dnorm(mu[i], tauv)
  mu[i] <- inprod(X[i,],beta[]) + W[i]
  muW[i] <- 0
  }
for(i in 1:p) {beta[i] ~ dnorm(0.0, 0.0001)}
W[1:N] ~ dmnorm(muW[], Omega[,])
tauv ~ dgamma(0.001,0.001)
v <- 1/sqrt(tauv)
tauw ~ dgamma(0.001,0.001)
w <- 1/sqrt(tauw)
phi~ dgamma(0.01,0.01)

for (i in 1:N) {
  for(j in 1:N) {
    H[i,j] <- (1/tauw)*exp(-phi*pow(d[i,j],2)) } }
Omega[1:N,1:N] <- inverse(H[1:N,1:N])
}
```

We can also fit this model using the spatial.exp function now available in WinBUGS releases 1.4 and later:

```
model
{
for(i in 1:N) {
  Y[i] ~ dnorm(mu[i], tauv
  mu[i] <- inprod(X[i,],beta[]) + W[i]
  muW[i] <- 0
  }
for(i in 1:p) {beta[i] ~ dnorm(0.0, 0.0001)}
W[1:N] ~ spatial.exp(muW[], x[], y[], tauw, phi, 1)
tauv ~ dgamma(0.001,0.001)
v <- 1/sqrt(tauv)
tauw ~ dgamma(0.001,0.001)
w <- 1/sqrt(tauw)
phi ~ dgamma(0.01,0.01)
}
```

You are asked to compare the results of these two approaches using a "toy" ($N = 10$) data set in Exercise 4.

5.5 Exercises

1. During her senior year in high school, Minnesota basketball sensation Carolyn Kieger scored at least 30 points in 9 consecutive games, helping her team win 7 of those games.

Game	Points scored by Kieger	Points scored by Rest of team	Game outcome
1	31	31	W, 62–49
2	31	16	W, 47–39
3	36	35	W, 71–64
4	30	42	W, 72–48
5	32	19	L, 64–51
6	33	37	W, 70–49
7	31	29	W, 60–37
8	33	23	W, 56–45
9	32	15	L, 57–47

Table 5.1 *Carolyn Kieger prep basketball data.*

The data for this remarkable streak are shown in Table 5.1. Notice that the rest of the team *combined* managed to outscore Kieger on only 2 of the 9 occasions.
A local press report on the streak concluded (apparently quite sensibly) that Kieger was primarily responsible for the team's relatively good win-loss record during this period. A natural statistical model for testing this statement would be the *logistic regression* model,

$$\begin{aligned} Y_i &\overset{ind}{\sim} Bernoulli(p_i)\ , \\ \text{where}\quad \text{logit}(p_i) &= \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i}\ . \end{aligned}$$

Here, Y_i is 1 if the team won game i and 0 if not, x_{1i} and x_{2i} are the corresponding points scored by Kieger and the rest of the team, respectively, and the logit transformation is defined as $\text{logit}(p_i) \equiv \log(p_i/(1-p_i))$, so that

$$p_i = \frac{\exp(\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i})}{1 + \exp(\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i})}\ .$$

(a) Using vague (or even flat) priors for the β_j, $j = 0, 1, 2$, fit this model to the data using the `WinBUGS` package. After downloading the program from `http://www.mrc-bsu.cam.ac.uk/bugs/` you may wish to follow the models provided by the similar `Surgical` or `Beetles` examples (click on "Help" and pull down to "Examples Vol I" or "Examples Vol II"). Obtain posterior summaries for the β_j parameters, as well as a DIC score and effective number of parameters p_D. Also investigate MCMC convergence using trace plots, autocorrelations, and crosscorrelations (the latter from the "Correlations" tool under the "Inference" menu). Is this model acceptable, numerically or statistically?

(b) Fit an appropriate two-parameter reduction of the model in part (a). Center the remaining covariate(s) around their own mean to reduce crosscorrelations in the parameter space, and thus speed MCMC convergence. Is this model an improvement?

(c) Fit one additional two-parameter model, namely,

$$\text{logit}(p_i) = \beta_0 + \beta_1 z_i\ ,$$

where $z_i = x_{1i}/(x_{1i} + x_{2i})$, the *proportion* of points scored by Kieger in game i. Again investigate convergence behavior, the β_j posteriors, and model fit relative to those in parts (a) and (b).

(d) For this final model, look at the estimated posteriors for the p_i themselves, and interpret the striking differences among them. What does this suggest might still be missing from our model?

2. Show that (5.20) is indeed a symmetric proposal density, as required by the conditions of the Metropolis algorithm.
3. Suppose now that θ is univariate but confined to the range $(0, \infty)$, with density proportional to $h(\theta)$.
   (a) Find the Metropolis acceptance ratio r assuming a Gaussian proposal density (5.20). Is this an efficient generation method?
   (b) Find the Metropolis acceptance ratio r assuming a Gaussian proposal density for $\eta \equiv \log \theta$. (Hint: Don't forget the Jacobian of this transformation!)
   (c) Finally, find the Metropolis-Hastings acceptance ratio r assuming a $Gamma(a, b)$ proposal density for θ.
4. Using the `WinBUGS` code and corresponding data set available from the web at `www.biostat.umn.edu/~brad/data/direct.bug`, attempt to fit the Bayesian kriging model in Example 5.4.
   (a) Using the "direct" code (which builds the $H(\phi)$ matrix explicitly).
   (b) Using the intrinsic `spatial.exp` function in `WinBUGS 1.4`.
   (c) Do your results in (a) and (b) agree? How do the runtimes compare?
   (d) Check to see if `WinBUGS` can handle the $N = 100$ case using the simulated data set `www.biostat.umn.edu/~brad/data/direct.bigdat` with a suitably modified version of your code.
5. Guo and Carlin (2004) consider a joint analysis of the AIDS longitudinal and survival data originally analyzed separately by Goldman et al. (1996) and Carlin and Louis (2000, Sec. 8.1). These data compare the effectiveness of two drugs, didanosine (ddI) and zalcitabine (ddC), in both preventing death and improving the longitudinal CD4 count trajectories in patients with late-stage HIV infection. The joint model used is one due to Henderson, Diggle, and Dobson (2000), which links the two submodels using bivariate Gaussian random effects. Specifically,

   Longitudinal model: For data $y_{i1}, y_{i2}, \ldots, y_{in_i}$ from the ith subject at times $s_{i1}, s_{i2}, \ldots, s_{i,n_i}$, let
   $$y_{ij} = \mu_i(s_{ij}) + W_{1i}(s_{ij}) + \epsilon_{ij} \ , \tag{5.29}$$
   where $\mu_i(s) = \mathbf{x}_{1i}^T(s)\boldsymbol{\beta}_1$ is the mean response, $W_{1i}(s) = \mathbf{d}_{1i}^T(s)\mathbf{U}_i$ incorporates subject-specific random effects (adjusting the main trajectory for any subject), and $\epsilon_{ij} \sim N(0, \sigma_\epsilon^2)$ is a sequence of mutually independent measurement errors. This is the classic longitudinal random effects setting of Laird and Ware (1982).

   Survival model: Letting t_i is time to death for subject i, we assume the parametric model,
   $$t_i \sim \text{Weibull}\,(p, \mu_i(t)) \ ,$$
   where $p > 0$ and
   $$\log(\mu_i(t)) = \mathbf{x}_{2i}^T(t)\boldsymbol{\beta}_2 + W_{2i}(t) \ .$$
   Here, $\boldsymbol{\beta}_2$ is the vector of fixed effects corresponding to the (possibly time-dependent) explanatory variables $\mathbf{x}_{2i}(t)$ (which may have elements in common with $\mathbf{x}_{1i}$), and $W_{2i}(t)$ is similar to $W_{1i}(s)$, including subject-specific covariate effects and an intercept (often called a *frailty*).

   The specific joint model studied by Guo and Carlin (2004) assumes
   $$W_{1i}(s) = U_{1i} + U_{2i}\, s \ , \quad \text{and} \tag{5.30}$$
   $$W_{2i}(t) = \gamma_1 U_{1i} + \gamma_2 U_{2i} + \gamma_3(U_{1i} + U_{2i}\, t) + U_{3i} \ , \tag{5.31}$$
   where $(U_{1i}, U_{2i})^T \overset{iid}{\sim} N(\mathbf{0}, \boldsymbol{\Sigma})$ and $U_{3i} \overset{iid}{\sim} N(0, \sigma_3^2)$, independent of the $(U_{1i}, U_{2i})^T$. The γ_1, γ_2, and γ_3 parameters in model (5.31) measurethe association between the two submodels

induced by the random intercepts, slopes, and fitted longitudinal value at the event time $W_{1i}(t)$, respectively.

(a) Use the code at `www.biostat.umn.edu/~brad/software.html` to fit the version of this model with $U_{3i} = 0$ for all i ("Model XII") in `WinBUGS`, as well as the further simplied version that sets $\gamma_3 = 0$ ("Model XI"). Which model fits better according to the DIC criterion?

(b) For your chosen model, investigate and comment on the posterior distributions of γ_1, γ_2, $\beta_{1,3}$ (the relative effect of ddI on the overall CD4 slope), and $\beta_{2,2}$ (the relative effect of ddI on survival).

(c) For each drug group separately, estimate the posterior distribution of the median survival time of a hypothetical patient with covariate values corresponding to a male who is AIDS-negative and intolerant of AZT at study entry. Do your answers change if you fit only the survival portion of the model (i.e., ignoring the longitudinal information)?

(d) Use the code at `www.biostat.umn.edu/~brad/software.html` to fit the `SAS Proc NLMIXED` code (originally written by Dr. Oliver Schabenberger) for Models XI and XII above. Are the answers consistent with those you obtained from `WinBUGS` above? How do the computer runtimes compare? What is your overall conclusion about Bayesian versus classical estimation in this setting?

Chapter 6

Hierarchical modeling for univariate spatial data

Having reviewed the basics of inference and computing under the hierarchical Bayesian modeling paradigm, we now turn our attention to its application in the setting of univariate point-referenced and areal unit data. Many of the models discussed in Chapter 3 and Chapter 4 will be of interest, but now they may be introduced in either the first-stage specification, to directly model the data in a spatial fashion, *or* in the second-stage specification, to model spatial structure in the random effects. We begin with models for point-level data, and proceed on to areal data models.

There is a substantial body of literature focusing on spatial prediction from a Bayesian perspective. Early work includes Le and Zidek (1992), Handcock and Stein (1993), Brown, Le, and Zidek (1994), Handcock and Wallis (1994), DeOliveira, Kedem, and Short (1997), Ecker and Gelfand (1997), Diggle, Tawn, and Moyeed (1998), and Karson et al. (1999). The work of Woodbury (1989), Abrahamsen (1993), and Omre and colleagues (Omre, 1987; Omre, 1988; Omre and Halvorsen, 1989; Omre, Halvorsen, and Berteig, 1989; and Hjort and Omre, 1994) is partially Bayesian in the sense that prior specification of the mean parameters and covariance function are elicited; however, no distributional assumption is made for the $Y(\mathbf{s})$. Since the publication of the first edition of this book, there has been an explosion of Bayesian spatial work; too many papers to cite here. Rather, the reader will find references to them sprinkled throughout the remaining chapters of this new edition.

6.1 Stationary spatial process models

The basic model we will work with is

$$Y(\mathbf{s}) = \mu(\mathbf{s}) + w(\mathbf{s}) + \epsilon(\mathbf{s}) \ , \tag{6.1}$$

where the mean structure $\mu(\mathbf{s}) = \mathbf{x}^T(\mathbf{s})\boldsymbol{\beta}$. The residual is partitioned into two pieces, one spatial and one nonspatial. That is, the $w(\mathbf{s})$ are assumed to be realizations from a zero-centered stationary Gaussian spatial process (see Section 3.1), capturing residual spatial association, while the $\epsilon(\mathbf{s})$ are uncorrelated pure error terms. Thus, for the covariance functions presented in Chapter 2, the $w(\mathbf{s})$ introduce the partial sill (σ^2) and range (ϕ) parameters, while the $\epsilon(\mathbf{s})$ add the nugget effect (τ^2). Valid correlation functions were discussed in Section 2.1. More specifically, supplying the correlation function as a function of the separation between sites yields a *stationary* model. If this dependence is captured only through the distance $||s_i - s_j||$, we obtain *isotropy*. Again, the most common such forms (exponential, Matérn, etc.) were presented in Subsection 2.1.3 and Tables 2.1 and 2.2.

Several interpretations can be attached to $\epsilon(\mathbf{s})$ and its associated variance τ^2. For instance, $\epsilon(\mathbf{s})$ can be viewed as a pure error term, as opposed to the spatial error term $w(\mathbf{s})$. That is, we would not necessarily insist that the residual error be entirely spatially structured. Correspondingly, the nugget τ^2 is a variance component of $Y(\mathbf{s})$, as is σ^2. In other

words, while $w(\mathbf{s}+\mathbf{h}) - w(\mathbf{s}) \to 0$ as $\mathbf{h} \to 0$ (if process realizations are continuous; see Subsection 3.1.4 and Section 13.2), $[w(\mathbf{s}+\mathbf{h}) + \epsilon(\mathbf{s}+\mathbf{h})] - [w(\mathbf{s}) - \epsilon(\mathbf{s})]$ will not. We are proposing residuals that are not spatially continuous, but not because the spatial process is not smooth. Instead, it is because we envision additional variabilty associated with the observed process, $Y(\mathbf{s})$. This could be viewed as measurement error (as might be the case with data from a monitoring device) or more generally as "noise" associated with replication of measurement at location $\mathbf{s}$ (as might be the case with the sale of a single-family home at $\mathbf{s}$, in which case $\epsilon(\mathbf{s})$ would capture the effect of the particular seller, buyer, realtors, and so on).

Another view of τ^2 is that it represents *microscale* variability, i.e., variability at distances smaller than the smallest interlocation distance in the data. In this sense, arguably, $\epsilon(\mathbf{s})$ could also be viewed as a spatial process, but with very rapid decay in association, i.e., with very small range. The dependence between the $\epsilon(\mathbf{s})$ would only matter at very high resolution. In this regard, Cressie (1993, pp. 112–113) suggests that $\epsilon(\mathbf{s})$ and τ^2 may themselves be partitioned into two pieces, one reflecting pure error and the other reflecting microscale error. In practice, we will rarely know much about the latter (and the data can not inform about it since we never observe the process at finer spatial resolution), so in this book we employ $\epsilon(\mathbf{s})$ to represent only the former.

6.1.1 Isotropic models

Suppose we have data $Y(\mathbf{s}_i)$, $i = 1, \ldots, n$, and let $\mathbf{Y} = (Y(\mathbf{s}_1), \ldots, Y(\mathbf{s}_n))^T$. The basic Gaussian isotropic kriging models of Section 2.4 are a special case of the general linear model, and therefore their Bayesian analysis can be viewed as a special case of Example 5.2. The problem just boils down to the appropriate definition of the Σ matrix. For example, in the case with a nugget effect,

$$\Sigma = \sigma^2 H(\phi) + \tau^2 I \ ,$$

where H is a correlation matrix with $H_{ij} = \rho(\mathbf{s}_i - \mathbf{s}_j; \phi)$ and ρ is a valid isotropic correlation function on $\Re^2$ indexed by a parameter (or parameters) ϕ. Collecting the entire collection of model parameters into a vector $\boldsymbol{\theta} = \left(\boldsymbol{\beta}, \sigma^2, \tau^2, \phi\right)^T$, a Bayesian solution requires an appropriate prior distribution $p(\boldsymbol{\theta})$. Parameter estimates may then be obtained from the posterior distribution, which by (5.1) is

$$p(\boldsymbol{\theta}|\mathbf{y}) \propto f(\mathbf{y}|\boldsymbol{\theta})\, p(\boldsymbol{\theta}) \ , \tag{6.2}$$

where

$$\mathbf{Y} \mid \boldsymbol{\theta} \sim N\left(X\boldsymbol{\beta}\,,\ \sigma^2 H(\phi) + \tau^2 I\right) \ . \tag{6.3}$$

6.1.1.1 Prior specification

Typically, independent priors are chosen for the different parameters, i.e.,

$$p(\boldsymbol{\theta}) = p(\boldsymbol{\beta})p(\sigma^2)p(\tau^2)p(\phi) \ ,$$

and natural candidates are multivariate normal for $\boldsymbol{\beta}$ and inverse gamma for σ^2 and τ^2. Specification for ϕ of course depends upon the choice of ρ function; in the simple exponential case where $\rho(\mathbf{s}_i - \mathbf{s}_j; \phi) = \exp(-\phi||\mathbf{s}_i - \mathbf{s}_j||)$ (and ϕ is thus univariate), a gamma might seem sensible. However, at this point, we need to devote a few paragraphs to a more careful development of prior specifications. First, let us consider the setting with no nugget so $\Sigma = \sigma^2 H(\phi)$. Let's focus on the Matérn class of covariance function. An elegant result due to Zhang (2004), which requires a somewhat sophisticated analysis of equivalent measures for stochastic processes, tells us that for the Matérn covariance function with smoothness parameter ν, the product $\sigma^2\phi^{2\nu}$ can be identified but not the individual parameters. For

instance, with the exponential, we can identify $\sigma^2\phi$ but not the range or the variances themselves. Only if we fix one can we identify the other. Implications for inference become apparent. Kriging will only be sensitive to the product. Which parameter are we more interested in learning about? Likely, it is the spatial variance, especially in the interests of comparison with the pure error variance. The two variance components inform about the relative strength of the spatial story vs. the pure error story, suggesting greater interest in σ^2. Furthermore, generally, learning regarding the spatial range is weak. So, practically, we recommend a very informative prior for ϕ and a relatively vague prior for σ^2. For the former, rather than a Gamma distribution, we often employ a uniform over a specified interval or a discretized uniform over a finite set of points.

Still care must be taken with regard to the latter. Here, we note the work of Berger et al. (2001). Again, in the case of a spatial model with no nugget, with the exponential covariance function, they consider the class of *objective* priors of the form $p(\boldsymbol{\beta}, \sigma^2, \phi) \propto \frac{p(\phi)}{(\sigma^2)^\alpha}$ which implies a flat prior for the regression coefficients, $\boldsymbol{\beta}$. They demonstrate that, with, say, a uniform prior for ϕ, an improper posterior arises for $\alpha < 2$. The implication for us is that if we adopt an inverse Gamma distribution prior $IG(\epsilon, \epsilon)$ prior for σ^2, this corresponds to the case of $\alpha = 1 + \epsilon$. For small ϵ, we have a specification that yields a posterior which is close to improper. While we may not explicitly see problems with our MCMC chains, we know that sampling from a nearly improper posterior can yield poorly behaved MCMC inference. Hence, our recommendation is to always use $IG(a, b)$ priors with $a \geq 1$ (implying $\alpha \geq 2$). Again, these priors are still quite vague since, with $a = 1$, we have no mean, hence no variance, and with $a = 2$, we have no variance. The foregoing theory has not been extended to the case when we bring in the nugget, τ^2. However, practical experience suggests that the foregoing problems only worsen and thus the same cautions should be taken.

Since we will often want to make inferential statements about the parameters separately, we will need to obtain *marginal* posterior distributions. For example, a point estimate or credible interval for $\boldsymbol{\beta}$ arises from

$$\begin{aligned} p(\boldsymbol{\beta}|\mathbf{y}) &= \int\int\int p\left(\boldsymbol{\beta}, \sigma^2, \tau^2, \phi|\mathbf{y}\right) d\sigma^2 d\tau^2 d\phi \\ &\propto p(\boldsymbol{\beta}) \int\int\int f(\mathbf{y}|\boldsymbol{\theta})\, p(\sigma^2) p(\tau^2) p(\phi) d\sigma^2 d\tau^2 d\phi\,. \end{aligned}$$

In principle this is simple, but in practice there will be no closed form for the above integrations. As such, we will often resort to MCMC or other numerical integration techniques, as described in Section 5.3.

We emphasize the role of hierarchical modeling throughout this book. In particular, we emphasize the generic model

$$[\text{data} \mid \text{process}, \text{parameters}][\text{process} \mid \text{parameters}][\text{parameters}]$$

which is more flexible than might first appear since the nature of the data and the nature of the process are not specified. This rich framework emphasizes our goal of learning about a process (typically, fairly complex) that is driving the data we are observing, a process that we seek to better understand. This framework occurs throughout the duration of this book. Here, we note that Expression (6.3) can be recast as a hierarchical model of this form by writing the first-stage specification as $\mathbf{Y}$ conditional not only on $\boldsymbol{\theta}$, but also on the vector of spatial random effects $\mathbf{W} = (w(\mathbf{s}_1), \ldots, w(\mathbf{s}_n))^T$. That is,

$$\mathbf{Y} \mid \boldsymbol{\theta}, \mathbf{W} \sim N(X\boldsymbol{\beta} + \mathbf{W}, \tau^2 I)\,. \tag{6.4}$$

The $Y(\mathbf{s}_i)$ are conditionally independent given the $w(\mathbf{s}_i)$. The second-stage specification is for $\mathbf{W}$, namely, $\mathbf{W}|\sigma^2, \phi \sim N(\mathbf{0}, \sigma^2 H(\phi))$ where $H(\phi)$ is as above. This is the *process* model.

Here, it is quite simple, a process merely introduced to capture spatial dependence. In later examples it will become richer as we incorporate more process features into its specification. Lastly, the model specification is completed by adding priors for $\boldsymbol{\beta}$ and τ^2 as well as for σ^2 and ϕ, the latter two of which may be viewed as hyperparameters. The parameter space is now augmented from $\boldsymbol{\theta}$ to $(\boldsymbol{\theta}, \mathbf{W})$, and its dimension is increased by n.

Regardless, the resulting $p(\boldsymbol{\theta}|\mathbf{y})$ is the same, but we have the choice of using Gibbs sampling (or some other MCMC method) to fit the model either as $f(\mathbf{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})$, or as $f(\mathbf{y}|\boldsymbol{\theta}, \mathbf{W})p(\mathbf{W}|\boldsymbol{\theta})p(\boldsymbol{\theta})$. The former is the result of marginalizing the latter over $\mathbf{W}$. Generally, we would prefer to work with the former (see Appendix Section A.2). Apart from the conventional wisdom that we should do as much marginalization in closed form as possible before implementing an MCMC algorithm (i.e., in as low a dimension as possible), the matrix $\sigma^2 H(\phi) + \tau^2 I$ is typically better behaved than $\sigma^2 H(\phi)$. To see this, note that if, say, $\mathbf{s}_i$ and $\mathbf{s}_j$ are very close to each other, $\sigma^2 H(\phi)$ will be close to singular while $\sigma^2 H(\phi) + \tau^2 I$ will not. Determinant and inversion calculation will also tend to be better behaved for the marginal model form than the conditional model form.

Interest is often in the spatial surface that involves $\mathbf{W}|\mathbf{y}$, as well as prediction for $W(\mathbf{s}_0)|\mathbf{y}$ for various choices of $\mathbf{s}_0$. At first glance it would appear that fitting the conditional model here would have an advantage, since realizations essentially from $p(\mathbf{W}|\mathbf{y})$ are directly produced in the process of fitting the model. However, since $p(\mathbf{W}|\mathbf{y}) = \int p(\mathbf{W}|\boldsymbol{\theta}, \mathbf{y})p(\boldsymbol{\theta}|\mathbf{y})d\boldsymbol{\theta}$, posterior realizations of $\mathbf{W}$ can be obtained one for one via *composition* sampling using posterior realizations of $\boldsymbol{\theta}$. Specifically, if the values $\boldsymbol{\theta}^{(g)}$ are draws from an MCMC algorithm with stationary distribution $p(\boldsymbol{\theta}|\mathbf{y})$, then corresponding draws $\mathbf{W}^{(g)}$ from $p(\mathbf{W}|\boldsymbol{\theta}^{(g)}, \mathbf{y})$ will have marginal distribution $p(\mathbf{W}|\mathbf{y})$, as desired. Thus we need not generate the $\mathbf{W}^{(g)}$ within the Gibbs sampler itself, but instead obtain them immediately given the output of the smaller, marginal sampler. Note that marginalization over $\mathbf{W}$ is only possible if the hierarchical form has a first-stage Gaussian specification, as in (6.4). We consider this matter in Section 3.2.

Next we turn to prediction of the response Y at a new value s_0 with associated covariate vector $\mathbf{x}(\mathbf{s}_0)$; this predictive step is the Bayesian "kriging" operation. Denoting the unknown value at that point by $Y(\mathbf{s}_0)$ and using the notations $Y_0 \equiv Y(\mathbf{s}_0)$ and $\mathbf{x}_0 \equiv \mathbf{x}(\mathbf{s}_0)$ for convenience, the solution in the Bayesian framework simply amounts to finding the predictive distribution,

$$p\left(y_0|\mathbf{y}, X, \mathbf{x}_0\right) = \int p\left(y_0, \boldsymbol{\theta}|\mathbf{y}, X, \mathbf{x}_0\right) d\boldsymbol{\theta} = \int p\left(y_0|\mathbf{y}, \boldsymbol{\theta}, \mathbf{x}_0\right) p\left(\boldsymbol{\theta}|\mathbf{y}, X\right) d\boldsymbol{\theta} \ , \tag{6.5}$$

where $p\left(y_0|\mathbf{y}, \boldsymbol{\theta}, \mathbf{x}_0\right)$ has a conditional normal distribution arising from the joint multivariate normal distribution of Y_0 and the original data $\mathbf{Y}$; see (2.15) and (2.16).

In practice, MCMC methods may again be readily used to obtain estimates of (6.5). Suppose we draw (after burn-in, etc.) our posterior sample $\boldsymbol{\theta}^{(1)}, \boldsymbol{\theta}^{(2)}, \ldots, \boldsymbol{\theta}^{(G)}$ from the posterior distribution $p\left(\boldsymbol{\theta}|\mathbf{y}, X\right)$. Then the above predictive integral may be computed as a Monte Carlo mixture of the form

$$\widehat{p}\left(y_0|\mathbf{y}, X, \mathbf{x}_0\right) = \frac{1}{G}\sum_{g=1}^{G} p\left(y_0|\mathbf{y}, \boldsymbol{\theta}^{(g)}, \mathbf{x}_0\right) \ . \tag{6.6}$$

In practice we typically use composition sampling to draw, one for one for each $\boldsymbol{\theta}^{(g)}$, a $y_0^{(g)} \sim p\left(y_0|\mathbf{y}, \boldsymbol{\theta}^{(g)}, \mathbf{x}_0\right)$. The collection $\left\{y_0^{(1)}, y_0^{(2)}, \ldots, y_0^{(G)}\right\}$ is a sample from the posterior predictive density, and so can be fed into a histogram or kernel density smoother to obtain an approximate plot of the density, bypassing the mixture calculation (6.6). A point estimate and credible interval for the predicted Y_0 may be computed in the same manner as in the

estimation case above. This is all routinely done in the `spBayes` package in `R`, or in `WinBUGS`; see Subsection 6.1.2 for more details on using the latter package.

Next suppose that we want to predict at a *set* of m sites, denoted, say, by $S_0 = \{\mathbf{s}_{01}, \mathbf{s}_{02}, \ldots, \mathbf{s}_{0m}\}$. We could individually predict at each of these points "independently" using the above method. But *joint* prediction may also be of interest, since it enables realizations from the same random spatial surface. As a result it allows estimation of posterior associations among the m predictions. We may form an unobserved vector $\mathbf{Y}_0 = (Y(\mathbf{s}_{01}), \ldots, Y(\mathbf{s}_{0m}))^T$ with associated design matrix X_0 having rows $\mathbf{x}(\mathbf{s}_{0j})^T$, and compute its joint predictive density as

$$\begin{aligned} p\left(\mathbf{y}_0|\mathbf{y}, X, X_0\right) &= \int p\left(\mathbf{y}_0|\mathbf{y}, \boldsymbol{\theta}, X_0\right) p\left(\boldsymbol{\theta}|\mathbf{y}, X\right) d\boldsymbol{\theta} \\ &\approx \frac{1}{G}\sum_{g=1}^{G} p\left(\mathbf{y}_0|\mathbf{y}, \boldsymbol{\theta}^{(g)}, X_0\right), \end{aligned}$$

where again $p\left(\mathbf{y}_0|\mathbf{y}, \boldsymbol{\theta}^{(j)}, X_0\right)$ is available from standard conditional normal formulae. We could also use composition to obtain, one for one for each $\boldsymbol{\theta}^{(g)}$, a collection of $\mathbf{y}_0^{(g)}$ and make any inferences we like based on this sample, either jointly or componentwise.

Often we are interested in not only the variables $Y(\mathbf{s})$, but also in functions of them, e.g., $\log Y(\mathbf{s})$ (if $Y(\mathbf{s}) > 0$), $I(Y(\mathbf{s}) > c)$, and so on. These functions are random variables as well. More generally we might be interested in functions $g(\mathbf{Y}_D)$ where $\mathbf{Y}_D = \{Y(\mathbf{s}) : \mathbf{s} \in D\}$. These include, for example, $(Y(\mathbf{s}_i) - Y(\mathbf{s}_j))^2$, which enter into the variogram, linear transformations $\sum_i \ell_i Y(\mathbf{s}_i)$, which include filters for spatial prediction at some location, and finite differences in specified directions, $[Y(\mathbf{s} + h\mathbf{u}) - Y(\mathbf{s})]/\mathbf{h}$, where $\mathbf{u}$ is a particular unit vector (see Subsection 13.3).

Functions of the form $g(\mathbf{Y}_D)$ also include block averages, i.e., $Y(A) = \frac{1}{|A|}\int_A g(Y(\mathbf{s}))d\mathbf{s}$. Block averages are developed in much more detail in Chapter 7. The case where $g(\mathbf{Y}_D) = I(Y(\mathbf{s}) \leq c)$ leads to the definition of the spatial CDF (SCDF) as in Section 15.3. Integration of a process or of a function of a process yields a new random variable, i.e., the integral is random and is usually referred to as a stochastic integral. An obvious but important point is that $E_A g(Y(\mathbf{s})) \neq g(E_A Y(\mathbf{s}))$ if g is not linear. Hence modeling $g(Y(\mathbf{s}))$ is not the same as modeling $g(Y(A))$. See Wakefield and Salway (2001) for further discussion.

6.1.2 Bayesian kriging in `WinBUGS`

Both the prediction techniques mentioned in the previous section (univariate and joint) are automated in `WinBUGS` (versions 1.3.1 and later). As usual we illustrate in the context of an example.

Example 6.1 Here we revisit the basic kriging model first considered in Example 5.4. Recall in that example we showed how to specify the model in the `WinBUGS` language directly, without making use of any special functions. Unfortunately, `WinBUGS`' standard matrix inversion routines are too slow for this approach to work for any but the smallest geostatistical data sets. However, the language does offer several special functions for Bayesian kriging, which we now describe.

First consider the pure spatial (no-nugget effect) model that is compatible with `WinBUGS` 1.4. In this model (with corresponding computer code available at `http://www.biostat.umn.edu/~brad/data2.html`), Y are the observed responses with covariate data X, N is the number of observed data sites with spatial coordinates (x[],y[]), M is the number of missing data sites with spatial coordinates (x0[],y0[]), and we seek to predict the response $Y0$ given the observed covariate data $X0$.

In the example given at the website, the data were actually simulated from a (purely spatial) Gaussian field with a univariate mean structure. Specifically, the true parameter values are $\beta = 5.0$, $\phi = 1.05$, and spatial variance $\sigma^2 = 2.0$.

```
model
{
for (i in 1:N) { mu[i] <- inprod(X[i,],beta[]) }

for (i in 1:p) {beta[i] ~ dnorm(0.0, 0.0001)}
Y[1:N] ~ spatial.exp(mu[], x[], y[], spat.prec, phi, 1)
phi~dgamma(0.1,0.1)
spat.prec ~ dgamma(0.10, 0.10)
sigmasq <- 1/spat.prec

# Predictions Joint
Y0[1:M] ~ spatial.pred(mu0[], x0[], y0[], Y[])
for(j in 1:M) { mu0[j] <- inprod(X0[j,], beta[]) }
}
```

In this code, the `spatial.exp` command fits the exponential kriging model directly to the observed data $\mathbf{Y}$, meaning that we are forgoing the nugget effect here. The final argument "1" in this command indicates an ordinary exponential model; another option is "2," corresponding to a powered exponential model where the power used is 2 (i.e., spatial dependence between two observations varies as the *square* of the distance between them). The `spatial.pred` command handles the joint prediction (kriging) at the new sites $X0$.

The following modification handles the modification where we add the nugget to the spatial model in `WinBUGS1.4`:

```
model
{
    for (i in 1:N) {
        Y[i] ~dnorm(mu[i], error.prec)
        mu[i] <- inprod(X[i,],beta[]) + W[i]
        muW[i] <- 0
      }

    for (i in 1:p) {beta[i] ~dnorm(0.0, 0.0001)}
    tausq <- 1/error.prec
    W[1:N] ~ spatial.exp(muW[], x[], y[], spat.prec, phi, 1)
    phi~dgamma(0.1,0.1)
    spat.prec ~dgamma(0.10, 0.10)
    sigmasq <- 1/spat.prec

  # Predictions Joint
    W0[1:M] ~spatial.pred(muW0[], x0[], y0[], W[])
    for(j in 1:M) {
        muW0[j] <- 0
        Y0[j] <- inprod(X0[j,], beta[]) + W0[j]
    }
}
```

Here, the `spatial.exp` command is used not with the observed data $\mathbf{Y}$, but with the random effects vector $\mathbf{W}$. Adding $\mathbf{W}$ into the mean structure and placing an ordinary normal error structure on $\mathbf{Y}$ conditional on $\mathbf{W}$ produces the "spatial plus nugget" error total structure we desire (see (6.3) above). ■

6.1.3 More general isotropic correlation functions, revisited

In Section 3.1.2.1 we noted the general characterization result for all valid isotropic correlation functions in R^r. That is, from Khinchin's Theorem (e.g., Yaglom, 1962, p. 106) as well as (3.3), this class of functions $\rho(d,\phi)$ in $\Re^r$ can be expressed as

$$\rho(d,\phi) = \int_0^\infty \Omega_r(zd)dG_\phi(z)\,, \tag{6.7}$$

where G_ϕ is nondecreasing integrable and $\Omega_r(x) = \left(\frac{2}{x}\right)^{\frac{r-2}{2}} \Gamma\left(\frac{r}{2}\right) J_{\left(\frac{r-2}{2}\right)}(x)$. Repeating, $J_\upsilon(\cdot)$ is the Bessel function of the first kind of order υ. For $r = 1, \Omega_1(x) = \cos(x)$; for $r = 2, \Omega_2(x) = J_0(x)$; for $r = 3, \Omega_3(x) = \sin(x)/x$; for $r = 4, \Omega_4(x) = \frac{2}{x}J_1(x)$; and for $r = \infty, \Omega_\infty(x) = \exp(-x^2)$. Specifically, $J_0(x) = \sum_{k=0}^\infty \frac{(-1)^k}{k!^2}\left(\frac{x}{2}\right)^{2k}$ and $\rho(d,\phi) = \int_0^\infty J_0(zd)dG_\phi(z)$ provides the class of all permissible correlation functions in $\Re^2$. Figure 3.1 provides a plot of $J_0(x)$ versus x, revealing that it is not monotonic. (This must be the case in order for $\rho(d,\phi)$ above to capture all correlation functions in $\Re^2$.)

In practice, a convenient simple choice for $G_\phi(z)$ is a step function that assigns positive mass (jumps or weights) w_ℓ at points (nodes) ϕ_ℓ, $\ell = 1,\ldots,p$ yielding, with $\mathbf{w} = (w_1, w_2, \ldots, w_p)$,

$$\rho(d,\phi,\mathbf{w}) = \sum_{\ell=1}^p w_\ell \Omega_n(\phi_\ell d)\,. \tag{6.8}$$

The forms in (6.8) are referred to as *nonparametric* variogram models in the literature to distinguish them from standard or parametric forms for $\rho(d,\phi)$, such as those given in Table 2.2. This is a separate issue from selecting a parametric or nonparametric methodology for parameter estimation. Sampson and Guttorp (1992), Shapiro and Botha (1991), and Cherry, Banfield, and Quimby (1996) use a step function for G_ϕ. Barry and Ver Hoef (1996) employ a mixture of piecewise linear variograms in R^1 and piecewise-planar models for sites in $\Re^2$. Hall, Fisher, and Hoffmann (1994) transform the problem from choosing ϕ_ℓ's and w_ℓ's in (6.8) to determining a kernel function and its associated bandwidth. Lele (1995) proposes iterative spline smoothing of the variogram yielding a ρ which is not obviously of the form (6.7). Most of these *nonparametric* models are fit to some version of the empirical semivariogram (2.9).

Sampson and Guttorp (1992) fit their model, using $\Omega_\infty(x)$ in (6.8), to the semivariogram cloud rather than to the smoothed Matheron semivariogram estimate. Their example involves a data set with 12 sites yielding only 66 points in the semivariogram cloud, making this feasible. Application of their method to a much larger (hence "noisier") data set would be expected to produce a variogram mixing hundreds and perhaps thousands of Gaussian forms. The resulting variogram will follow the semivariogram cloud too closely to be plausible.

Working in $\Re^2$, where again $\Omega_2(x) = J_0(x)$, under the Bayesian paradigm we can introduce (6.8) directly into the likelihood but keep p small (at most 5), allowing random w_ℓ or random ϕ_ℓ. This offers a compromise between the rather limiting standard parametric forms (Table 2.1) that specify two or three parameters for the covariance structure, and above nonparametric methods that are based upon a practically implausible (and potentially overfitting) mixture of hundreds of components. Moreover, by working with the likelihood, inference is conditioned upon the observed $\mathbf{y}$, rather than on a summary such as a smoothed version of the semivariogram cloud.

Returning to (6.7), when $n = 2$ we obtain

$$\rho(d,\phi) = \int_0^\infty \sum_{k=0}^\infty \frac{(-1)^k}{k!^2}\left(\frac{zd}{2}\right)^{2k} dG_\phi(z)\,. \tag{6.9}$$

Only if z is bounded, i.e., if G_ϕ places no mass on say $z > \phi_{max}$, can we interchange summation and integration to obtain

$$\rho(d,\phi) = \sum_{k=0}^{\infty} \frac{(-1)^k}{k!^2} \left(\frac{d}{2}\right)^{2k} \delta_{2k} \ , \tag{6.10}$$

where $\delta_{2k} = \int_0^{\phi_{\max}} z^{2k} dG_\phi(z)$. The simplest such choice for G_ϕ puts discrete mass w_ℓ at a finite set of values $\phi_\ell \in (0, \phi_{\max})$, $\ell = 1, \ldots, p$ resulting in a finite mixture of Bessels model for $\rho(d,\phi)$, which in turn yields

$$\gamma(d_{ij}) = \tau^2 + \sigma^2 \left(1 - \sum_{\ell=1}^{p} w_\ell J_0(\phi_\ell d_{ij})\right) \ . \tag{6.11}$$

Under a Bayesian framework for a given p, if the w_ℓ's are each fixed to be $\frac{1}{p}$ with ϕ_ℓ's unknown (hence random), they are constrained by $0 < \phi_1 < \phi_2 < \cdots < \phi_p < \phi_{max}$ for identifiability. The result is an equally weighted mixture of random curves. If a random mixture of fixed curves is desired, then the w_ℓ's are random and the ϕ_ℓ's are systematically chosen to be $\phi_\ell = \left(\frac{\ell}{p+1}\right)\phi_{\max}$. We examine $p = 2, 3, 4, 5$ for fixed nodes and $p = 1, 2, 3, 4, 5$ for fixed weights. Mixture models using random w_ℓ's and random ϕ_ℓ's might be considered but, in our limited experience, the posteriors have exhibited weak identifiability in the parameters and thus are not recommended.

In choosing ϕ_{max}, we essentially determine the maximum number of sign changes we allow for the dampened sinusoidal Bessel correlation function over the range of d's of interest. For, say, $0 \le d \le d^{\max}$ where $d^{\max}$ is the maximum of the $d_{ij} = ||\mathbf{s}_i - \mathbf{s}_j||$, the larger ϕ is, the more sign changes $J_0(\phi d)$ will have over this range. This suggests making ϕ_{max} very large. However, as noted earlier in this section, we seek to avoid practically implausible ρ and γ, which would arise from an implausible $J_0(\phi d)$. For illustration, the plot in Figure 3.1 above allows several sign changes, to show the longer term stability of its oscillation. Letting κ be the value of x where $J_0(x) = 0$ attains its kth sign change (completes its $\frac{k-1}{2}$ period) we set $\kappa = \phi_{\max} d^{\max}$, thus determining $\phi_{\max}$. We reduce the choice of $\phi_{\max}$ to choosing the maximum number of Bessel periods allowable. For a given p, when the ϕ's are random, the posterior distribution for ϕ_p will reveal how close to $\phi_{\max}$ the data encourages ϕ_p to be.

Example 6.2 We return to the 1990 log-transformed scallop data, originally presented in Subsection 2.3.2. In 1990, 148 sites were sampled in the New York Bight region of the Atlantic Ocean, which encompasses the area from the tip of Long Island to the mouth of the Delaware River. These data have been analyzed by Ecker and Heltshe (1994), Ecker and Gelfand (1997, 1999), Kaluzny et al. (1998), and others. Figure 6.1 shows the semivariogram cloud (panel a) together with boxplots (panel b) formed from the cloud using the arbitrary lag $\delta = 0.05$. The 10,731 pairs of points that produce the semivariogram cloud do not reveal any distinct pattern. In a sense, this shows the folly of fitting a curve to this data: we have a weak signal, and a great deal of noise.

However, the boxplots and the Matheron empirical semivariograms each based on lag $\delta = 0.05$ (Figure 6.2) clearly exhibit spatial dependence, in the sense that when separation distances are small, the spatial variability tends to be less. Here the attempt is to remove the noise to see whatever signal there may be. Of course, the severe skewness revealed by the boxplots (and expected from squared differences) raises the question of whether the bin averages are an appropriate summary (Expression (2.9)); see Ecker and Gelfand (1997) in this regard. Clearly such displays and attempts to fit an empirical variogram must be viewed as part of the exploratory phase of our data analysis.

For the choice of ϕ_{max} in the nonparametric setup, we selected seven sign changes, or three Bessel periods. With $d_{ij}^{max} = 2.83$ degrees, ϕ_{max} becomes 7.5. A sensitivity analysis

Figure 6.1 *Semivariogram cloud (a) and boxplot produced from 0.05 lag (b), 1993 scallop data.*

Figure 6.2 *Matheron empirical semivariograms for lag* $\delta = 0.05$.

with two Bessel mixtures ($p = 2$) having a fixed weight w_1 and random nodes was undertaken. Two, four, and five Bessel periods revealed little difference in results as compared with three. However, when one Bessel period was examined ($\phi_{max} = 3$), the model fit poorly and in fact ϕ_p was just smaller than 3. This is an indication that more flexibility (i.e., a larger value of ϕ_{max}) is required.

Several of the parametric models from Tables 2.1 and 2.2 and several nonparametric Bessel mixtures with different combinations of fixed and random parameters were fit to the 1990 scallop data. (Our analysis here parallels that of Ecker and Gelfand, 1997, although our results are not identical to theirs since they worked with the 1993 version of the data

Figure 6.3 *Posterior means for various semivariogram models.*

set.) Figure 6.3 shows the posterior mean of each respective semivariogram, while Table 6.1 provides the value of model choice criteria for each model along with the independence model, $\Sigma_{\mathbf{Y}} = (\tau^2 + \sigma^2)I$. Here we use the Gelfand and Ghosh (1998) model selection criterion (5.14), as described in Subsection 5.2.3. However, since we are fitting variograms, we work somewhat less formally using $Z_{ij,obs} = (Y(\mathbf{s}_i) - Y(\mathbf{s}_j))^2/2$. Since Z_{ij} is distributed as a multiple of a χ^2_1 random variable, we use a loss associated with a gamma family of distributions, obtaining a $D_{k,m}$ value of

$$(k+1)\sum_{i,j}\left\{\log\left(\frac{\lambda_{ij}^{(m)} + kz_{ij,obs}}{k+1}\right) - \frac{\log(\lambda_{ij}^{(m)}) + k\log(z_{ij,obs})}{k+1}\right\} + \sum_{i,j}\Big(\log(\lambda_{ij}^{(m)}) - E(\log(z_{ij,rep}) \mid \mathbf{y}, m)\Big) \tag{6.12}$$

for model m, where $\lambda_{ij}^{(m)} = E(z_{ij,rep} \mid \mathbf{y}, m)$. The concavity of the log function ensures that both summations on the right-hand side of (6.12) are positive. (As an aside, in theory $z_{ij,obs} > 0$ almost surely, but in practice we may observe some $z_{ij} = 0$ as, for example, with the log counts in the scallop data example. A correction is needed and can be achieved by adding ϵ to $z_{ij,obs}$ where ϵ is, say, one half of the smallest possible positive $z_{ij,obs}$.)

Setting $k = 1$ in (6.12), we note that of the Bessel mixtures, the five-component model with fixed ϕ's and random weights is best according to the $D_{1,m}$ statistic. Here, given $\phi_{max} = 7.5$, the nodes are fixed to be $\phi_1 = 1.25, \phi_2 = 2.5, \phi_3 = 3.75, \phi_4 = 5.0$, and $\phi_5 = 6.25$. One would expect that the fit measured by the $G_{1,m}$ criterion should improve with increasing p. However, the models do not form a nested sequence in p, except in some instances (e.g., the $p = 2$ model is a special case of the $p = 5$ model). Thus, the apparent

Model	$G_{1,m}$	P_m	$D_{1,m}$
Parametric			
exponential	10959	13898	24857
Gaussian	10861	13843	24704
Cauchy	10683	13811	24494
spherical	11447	13959	25406
Bessel	11044	14037	25081
independent	11578	16159	27737
Semiparametric			
fixed ϕ_ℓ, random w_ℓ:			
two	11071	13968	25039
three	10588	13818	24406
four	10934	13872	24806
five	10567	13818	24385
random ϕ_ℓ, fixed w_ℓ:			
two	10673	13907	24580
three	10677	13959	24636
four	10636	13913	24549
five	10601	13891	24492

Table 6.1 *Model choice for fitted variogram models, 1993 scallop data.*

poorer fit of the four-component fixed ϕ model relative to the three-component model is indeed possible. The random ϕ Bessel mixture models were all very close and, as a class, these models fit as well or better than the best parametric model. Hence, modeling mixtures of Bessel functions appears more sensitive to the choice of fixed ϕ's than to fixed weights. ■

6.1.4 Modeling geometric anisotropy

Anistropy was introduced in Section 2.2, in the form of geometric, sill, and nugget anisotropy, to refer to to particular cases of stationarity. In any event, we have $Cov\left(Y\left(\mathbf{s}+\mathbf{h}\right), Y\left(\mathbf{s}\right)\right) = C\left(\mathbf{h};\phi\right)$. The most prominent, tractable, and interesting case in applications is *geometric anisotropy*. This refers to the situation where the coordinate space can be linearly transformed to an isotropic space. A linear transformation may correspond to rotation or stretching of the coordinate axes. Thus in general,

$$\rho\left(\mathbf{h};\phi\right) = \rho_0\left(||L\mathbf{h}||\,;\phi\right) \ ,$$

where l is a $d \times d$ matrix describing the linear transformation. Of course, if L is the identity matrix, this reduces to the isotropic case.

We assume a second-order stationary normal model for $\mathbf{Y}$, arising from the customary model, $Y(\mathbf{s}) = \mu + w(\mathbf{s}) + \epsilon(\mathbf{s})$ as in (6.1). This yields $\mathbf{Y} \sim N(\mu\mathbf{1}\ ,\ \Sigma(\boldsymbol{\alpha}))$, where $\boldsymbol{\alpha} = (\tau^2, \sigma^2, B)^T$, $B = L^T L$, and

$$\Sigma(\boldsymbol{\alpha}) = \tau^2 I + \sigma^2 H((\mathbf{h}'B\mathbf{h})^{\frac{1}{2}}) \ . \tag{6.13}$$

In (6.13), the matrix H has (i,j)th entry $\rho((\mathbf{h}'_{ij}B\mathbf{h}_{ij})^{\frac{1}{2}})$ where ρ is a valid correlation function and $\mathbf{h}_{ij} = \mathbf{s}_i - \mathbf{s}_j$. Common forms for ρ would be those in Table 2.2. In (6.13), τ^2 is the semiovariogram nugget and $\tau^2 + \sigma^2$ is the sill. The variogram is $2\gamma(\tau^2, \sigma^2, (\mathbf{h}'B\mathbf{h})^{\frac{1}{2}}) = 2(\tau^2 + \sigma^2(1 - \rho((\mathbf{h}'B\mathbf{h})^{\frac{1}{2}}))$.

Turning to $\Re^2$, B is 2×2 and the orientation of the associated ellipse, ω, is related to B by (see, e.g., Anton, 1984, p. 691)

$$\cot(2\omega) = \frac{b_{11} - b_{22}}{2b_{12}} \ . \tag{6.14}$$

The range in the direction η, where η is the angle $\mathbf{h}$ makes with the x-axis and which we denote as r_η, is determined by the relationship

$$\rho(r_\eta(\widetilde{\mathbf{h}}'_\eta B \widetilde{\mathbf{h}}_\eta)^{\frac{1}{2}}) = 0.05 \ , \tag{6.15}$$

where $\widetilde{\mathbf{h}}_\eta = (\cos\eta, \sin\eta)$ is a unit vector in direction η.

The *ratio of anisotropy* (Journel and Huijbregts, 1978, pp. 178–181), also called the *ratio of affinity* (Journel and Froidevaux, 1982, p. 228), which here we denote as λ, is the ratio of the major axis of the ellipse to the minor axis, and is related to B by

$$\lambda = \frac{r_\omega}{r_{(\pi-\omega)}} = \left(\frac{\widetilde{\mathbf{h}}'_{(\pi-\omega)} B \widetilde{\mathbf{h}}_{(\pi-\omega)}}{\widetilde{\mathbf{h}}'_\omega B \widetilde{\mathbf{h}}_\omega} \right)^{\frac{1}{2}} \ , \tag{6.16}$$

where again $\widetilde{\mathbf{h}}_\eta$ is the unit vector in direction η. Since (6.14), (6.15), and (6.16) are functions of B, posterior samples (hence inference) for them is straightforward given posterior samples of $\boldsymbol{\alpha}$.

A customary prior distribution for a positive definite matrix such as B is Wishart(R, p), where

$$\pi(b) \propto |B|^{\frac{p-n-1}{2}} \exp\left(-\frac{1}{2} tr(pBR^{-1}) \right) \ , \tag{6.17}$$

so that $E(B) = R$ and $p \geq n$ is a precision parameter in the sense that $Var(B)$ increases as p decreases. In $\Re^2$, the matrix $R = \begin{bmatrix} R_{11} & R_{12} \\ R_{12} & R_{22} \end{bmatrix}$. Prior knowledge is used to choose R, but we choose the prior precision parameter, p, to be as small as possible, i.e., $p = 2$.

A priori, it is perhaps easiest to assume that the process is isotropic, so we set $R = \delta I$ and then treat δ as fixed or random. For δ random, we model $p(B, \delta) = p(B|\delta)p(\delta)$, where $p(B|\delta)$ is the Wishart density given by (6.17) and $p(\delta)$ is an inverse gamma distribution with mean obtained from a rough estimate of the range and infinite variance (i.e., shape paramater equal to 2).

However, if we have prior evidence suggesting geometric anisotropy, we could attempt to capture it using (6.14), (6.15), or (6.16) with $\widetilde{\mathbf{h}}'_\eta R \widetilde{\mathbf{h}}_\eta$ replacing $\widetilde{\mathbf{h}}'_\eta B \widetilde{\mathbf{h}}_\eta$. For example, with a prior guess for ω, the angle of orientation of the major axis of the ellipse, a prior guess for λ, the ratio of major to minor axis (say, from a rose diagram), and a guess for the range in a specific direction (say, from a directional semivariogram), then (6.14), (6.15), and (6.16) provide a system of three linear equations in three unknowns to solve for R_{11}, R_{12}, and R_{22}. Alternatively, from three previous directional semivariograms, we might guess the range in three given directions, say, r_{η_1}, r_{η_2}, and r_{η_3}. Now, using (6.15), we again arrive at three linear equations with three unknowns in R_{11}, R_{12}, and R_{22}. One can also use an empirical semivariogram in $\Re^2$ constructed from prior data to provide guesses for R_{11}, R_{12}, and R_{22}. By computing a $0°$ and $90°$ directional semivariogram based on the ESC plot with rows where $h_y \approx 0$ for the former and columns where $h_x \approx 0$ in the latter, we obtain guesses for $R_{11}$and R_{22}, respectively. Finally, R_{12} can be estimated by examining a bin where neither $h_x \approx 0$ nor $h_y \approx 0$. Equating the empirical semivariogram to the theoretical semivariogram at the associated (x_i, y_j), with R_{11} and R_{22} already determined, yields a single equation to solve for R_{12}.

	Isotropic prior	Geometrically anisotropic prior		
	fixed $\widehat{\psi} = 0.0003$	ω, λ and r_{50°	three ranges	ESC plot
τ^2	1.29	1.43	1.20	1.33
	(1.00, 1.64)	(1.03, 1.70)	(1.01, 1.61)	(0.97, 1.73)
σ^2	2.43	2.35	2.67	2.58
	(1.05, 5.94)	(1.27, 5.47)	(1.41, 5.37)	(1.26, 5.67)
sill	3.72	3.80	3.87	3.91
	(2.32, 7.17)	(2.62, 6.69)	(2.66, 6.76)	(2.39, 7.09)
μ	2.87	2.55	3.14	2.90
	(2.16, 3.94)	(1.73, 3.91)	(2.24, 3.99)	(2.14, 4.02)
ω	55.3	64.4	57.2	60.7
	(26.7, 80.7)	(31.9, 77.6)	(24.5, 70.7)	(46.7, 75.2)
λ	2.92	3.09	3.47	3.85
	(1.59, 4.31)	(1.77, 4.69)	(1.92, 4.73)	(2.37, 4.93)

Table 6.2 *Posterior means and 95% interval estimates for a stationary Gaussian model with Gaussian correlation structure under various prior specifications.*

Example 6.3 Here we return again to the log-transformed sea scallop data of Subsection 2.3.2, and reexamine it for geometric anisotropy. Previous analyses (e.g., Ecker and Heltshe, 1994) have detected geometric anisotropy with the major axes of the ellipse oriented parallel to the coastline ($\approx 50^\circ$ referenced counterclockwise from the x-axis). Kaluzny et al. (1998, p. 90) suggest that λ, the ratio of major axis to minor axis, is approximately 3. The 1993 scallop catches with 147 sites were analyzed in Ecker and Gelfand (1997) under isotropy. Referring back to the ESC plot in Figure 2.10, a geometrically anisotropic model seems reasonable. Here we follow Ecker and Gelfand (1999) and illustrate with a Gaussian correlation form, $\rho((\mathbf{h}'B\mathbf{h})^{\frac{1}{2}}) = \exp(-\mathbf{h}'B\mathbf{h})$.

We can use the 1990 scallop data to formulate isotropic and geometrically anisotropic prior specifications for R, the prior mean for B. The first has $R = \delta I$ with fixed $\widehat{\delta} = 0.0003$, i.e., a prior isotropic range of 100 km. Another has $\widehat{\delta} = 0.000192$, corresponding to a 125-km isotropic prior range to assess the sensitivity of choice of $\widehat{\delta}$, and a third has δ random. Under prior geometric anisotropy, we can use $\omega = 50^\circ$, $\lambda = 3$, and $r_{50^\circ} = 125$ km to obtain a guess for R. Solving (6.14), (6.15), and (6.16) gives $R_{11} = 0.00047$, $R_{12} = -0.00023$, and $R_{22} = 0.00039$. Using the customary directional semivariograms with the 1990 data, another prior guess for R can be built from the three prior ranges $r_{0^\circ} = 50$ km, $r_{45^\circ} = 125$ km, and $r_{135^\circ} = 30$ km. Via (6.15), we obtain $R_{11} = 0.012$, $R_{12} = -0.00157$, and $R_{22} = 0.00233$. Using the ESC plot for the 1990 data, we use all bins where $h_x = h_{long} \approx 0$ (90° semivariogram) to provide $R_{22} = 0.0012$, and bins where $h_y = h_{lat} \approx 0$ (0° semivariogram) to provide $R_{11} = 0.00053$. Finally, we pick three bins with large bin counts (328, 285, 262) and along with the results of the 0° and 90° ESC plot directional semivariograms, we average the estimated R_{12} for each of these three bins to arrive at $R_{12} = -0.00076$.

The mean and 95% interval estimates for the isotropic prior specification with $\widehat{\delta} = 0.0003$, and the three geometrically anisotropic specifications are presented in Table 6.2. Little sensitivity to the prior specifications is observed as expected, given that we use the smallest allowable prior precision. The posterior mean for the angle of orientation, ω, is about 60° and the ratio of the major ellipse axis to minor axis, λ, has a posterior mean of

Figure 6.4 *Posterior distributions under the geometrically anisotropic prior formed from the ESC plot.*

about 3 to 3.5. Furthermore, the value 1 is not in any of the three 95% interval estimates for λ, indicating that isotropy is inappropriate.

We next present posterior inference associated with the ESC plot-based prior specification. Figure 6.4 shows the posteriors for the nugget in panel (a), sill in panel (b), angle of orientation in panel (c), and the ratio of major axis to minor axis in panel (d). Figure 6.5 shows the mean posterior range plotted as a function of angle with associated individual 95% intervals. This plot is much more informative in revealing departure from isotropy than merely examining whether the 95% interval for λ contains 1. Finally, Figure 6.6 is a plot of the contours of the posterior mean surface of the semivariogram. Note that it agrees with the contours of the ESC plot given in Figure 2.10 reasonably well.

6.2 Generalized linear spatial process modeling

In some point-referenced data sets we obtain measurements $Y(\mathbf{s})$ that would not naturally be modeled using a normal distribution; indeed, they need not even be continuous. For instance, $Y(\mathbf{s})$ might be a binary variable indicating whether or not measurable rain fell at location $\mathbf{s}$ in the past 24 hours, or a count variable indicating the number of insurance claims over the past five years by the residents of a single-family home at location $\mathbf{s}$. In an aggregate data context examining species range and richness, $Y(\mathbf{s})$ might indicate presence or absence of a particular species at $\mathbf{s}$ (although here, strictly speaking $\mathbf{s}$ is not a point, but really an area that is sufficiently small to be thought of as a point within the overall study area).

Following Diggle, Tawn, and Moyeed (1998), we formulate a hierarchical model analogous to those in Section 6.1, but with the Gaussian model for $Y(\mathbf{s})$ replaced by another suitable member of the class of exponential family models. Assume the observations $Y(\mathbf{s}_i)$ are conditionally independent given $\boldsymbol{\beta}$ and $w(\mathbf{s}_i)$ with distribution,

$$f(y(\mathbf{s}_i)|\boldsymbol{\beta}, w(\mathbf{s}_i), \gamma) = h(y(\mathbf{s}_i), \gamma) \exp\{\gamma[y(\mathbf{s}_i)\eta(\mathbf{s}_i) - \psi(\eta(\mathbf{s}_i))]\} , \tag{6.18}$$

Figure 6.5 *Posterior range as a function of angle for the geometrically anisotropic prior formed from the ESC plot.*

where $g(\eta(\mathbf{s}_i)) = \mathbf{x}^T(\mathbf{s}_i)\boldsymbol{\beta} + w(\mathbf{s}_i)$ for some link function g, and γ is a dispersion parameter. We presume the $w(\mathbf{s}_i)$ to be spatial random effects coming from a Gaussian process, as in Section 6.1. The second-stage specification is $\mathbf{W} \sim N(\mathbf{0}, \sigma^2 H(\boldsymbol{\phi}))$ as before. Were the $w(\mathbf{s}_i)$ i.i.d., we would have a customary generalized linear mixed effects model (Breslow and Clayton, 1993). Hence (6.18) is still a generalized linear mixed model, but now with spatial structure in the random effects.

Note that, although we have defined a process for $w(\mathbf{s})$ we have not created a process for $Y(\mathbf{s})$. That is, using conditional independence, what we have done is to create a joint distribution $f(y(\mathbf{s}_1), \ldots, y(\mathbf{s}_n)|\boldsymbol{\beta}, \sigma^2, \boldsymbol{\phi}, \gamma)$, namely,

$$\int \left(\prod_{i=1}^{n} f(y(\mathbf{s}_i)|\boldsymbol{\beta}, w(\mathbf{s}_i), \gamma) \right) p(\mathbf{W}|\sigma^2, \boldsymbol{\phi}) d\mathbf{W} \ . \tag{6.19}$$

The class of distributions that can support a stochastic process is limited, characterized through mixtures of elliptical distributions which, of course, includes Gaussian processes.

We have an opportunity to make another important point here. A frequent first stage spatial specification is a binary response model. That is, at every location $\mathbf{s}$, there is a binary variable $Y(\mathbf{s})$. The resulting surface is frequently referred to as a binary map (DeOliveira, 2000). In this case, we can usefully consider two hierarchical specifications to model the binary map. The first sets $Y(\mathbf{s}) = 1$ or 0 according whether $Z(\mathbf{s}) \geq 0$ or < 0. Then, we model $Z(\mathbf{s}) = \mathbf{x}(\mathbf{s})^T\boldsymbol{\beta} + w(\mathbf{s}) + \epsilon(\mathbf{s})$. That is $Z(\mathbf{s})$ is our usual geostatistical model, (6.1). So, $Z(\mathbf{s})$ is a Gaussian process and determines $Y(\mathbf{s})$. In particular, $P(Y(\mathbf{s}) = 1) = P(Z(\mathbf{s}) \geq 0) = \Phi(\mathbf{x}(\mathbf{s})^T\boldsymbol{\beta} + w(\mathbf{s}))$. As an alternative, resembling (6.18), let $P(Y(\mathbf{s}) = 1) \equiv p(\mathbf{s})$. Now, adopt a link function to take $p(\mathbf{s})$ to $\Re^1$, say $\Phi^{-1}(\cdot)$ and set $\Phi^{-1}(p(\mathbf{s})) = \mathbf{x}(\mathbf{s})^T\boldsymbol{\beta} + w(\mathbf{s})$. Though they appear different, these two models are equivalent; rather, now $Y(\mathbf{s})|p(\mathbf{s})$ is a random mechanism while $Y(\mathbf{s})|Z(\mathbf{s})$ is a deterministic mechanism. We have exchanged the binary first stage stochastic specification for a pure error Gaussian specification. But, this also clarifies why it is not sensible to add a pure error term to the specification for

Figure 6.6 *Contours of the posterior mean semivariogram surface for the geometrically anisotropic prior formed from the ESC plot.*

$p(\mathbf{s})$. Such an error term would be redundant, as is immediately evident were we to add a corresponding additional pure error term to the specification for $Z(\mathbf{s})$. In fact, with such an additional error term, we would obtain unidentified MCMC model fitting which we would see with poorly behaved convergence. This point is evidently true for whatever first stage generalized linear model specification is used in (6.18).

We also note an important consequence of modeling with spatial random effects (which incidentally is relevant for Sections 6.4 and 6.5 as well). Introducing these effects in the (transformed) mean, as below (6.18), encourages the means of the spatial variables at proximate locations to be close to each other (adjusted for covariates). Though marginal spatial dependence is induced between, say, $Y(\mathbf{s})$ and $Y(\mathbf{s}')$, the observed $Y(\mathbf{s})$ and $Y(\mathbf{s}')$ need *not* be close to each other. This would be the case even if $Y(\mathbf{s})$ and $Y(\mathbf{s}')$ had the same mean. As a result, second-stage spatial modeling is attractive when spatial explanation in the *mean* is of interest. Direct (first-stage) spatial modeling is appropriate to encourage proximate *observations* to be close.

Turning to computational issues, note that (6.19) cannot be integrated in closed form; we cannot marginalize over $\mathbf{W}$. Unlike the Gaussian case, a MCMC algorithm will have to update $\mathbf{W}$ as well as $\boldsymbol{\beta}, \sigma^2, \boldsymbol{\phi}$, and γ. This same difficulty occurs with simulation-based model fitting of standard generalized linear mixed models (again see, e.g., Breslow and Clayton, 1993). In fact, the $w(\mathbf{s}_i)$ would likely be updated using a Metropolis step with a Gaussian proposal, or through adaptive rejection sampling (since their full conditional distributions will typically be log-concave); see Exercise 4:

Example 6.4 Non-Gaussian point-referenced spatial model. Here we consider a real estate data set, with observations at 50 locations in Baton Rouge, LA. The response $Y(\mathbf{s})$ is a binary variable, with $Y(\mathbf{s}) = 1$ indicating that the price of the property at location $\mathbf{s}$ is "high" (above the median price for the region), and $Y(\mathbf{s}) = 0$ indicating that the price is "low". Observed covariates include the house's age, total living area, and other area in the property. We fit the model given in (6.18) where $Y(\mathbf{s}) \sim Bernoulli(p(\mathbf{s}))$ and g is the logit link. The `WinBUGS` code and data for this example are at `www.biostat.umn.edu/~brad/data2.html`.

Parameter	50%	(2.5%, 97.5%)
intercept	–1.096	(–4.198, 0.4305)
living area	0.659	(–0.091, 2.254)
age	0.009615	(–0.8653, 0.7235)
ϕ	5.79	(1.236, 9.765)
σ^2	1.38	(0.1821, 6.889)

Table 6.3 *Parameter estimates (posterior medians and upper and lower .025 points) for the binary spatial model.*

Figure 6.7 *Image plot of the posterior median surface of the latent spatial process $w(\mathbf{s})$, binary spatial model.*

Table 6.3 provides the parameter estimates and Figure 6.7 shows the image plot with overlaid contour lines for the posterior mean surface of the latent $w(\mathbf{s})$ process. These are obtained by assuming vague priors for $\boldsymbol{\beta}$, a Uniform$(0, 10)$ prior for ϕ, and an inverse gamma$(0.1, 0.1)$ prior for σ^2. The image plot reveals negative residuals (i.e., lower prices) in the northern region, and generally positive residuals (higher prices) in the south-central region, although the southeast shows some lower price zones. The distribution of the contour lines indicate smooth flat stretches across the central parts, with downward slopes toward the north and southeast. The covariate effects are generally uninteresting, though living area seems to have a marginally significant effect on price class.

6.3 Fitting hierarchical models for point-referenced data in `spBayes`

6.3.1 Gaussian spatial regression models

We will use the forest inventory data from the U.S. Department of Agriculture Forest Service, Bartlett Experimental Forest (BEF), Bartlett, NH. This dataset is a part of the

spBayes package in R and holds 1991 and 2002 forest inventory data for 437 plots. In our illustration below, we use log-transformed total tree biomass as the outcome and regress it on five predictors: slope, elevation, and tasseled cap brightness (TC1), greenness (TC2), and wetness (TC3) components from spring, summer, and fall 2002 Landsat images. We use these data to demonstrate some basics of univariate spatial regression analysis for Gaussian outcomes. The regression model will subsequently be used to make prediction of biomass for every image pixel across the BEF.

For the illustrations and graphics below, we will load the following packages

```
> library(spBayes)
> library(MBA)
> library(geoR)
> library(fields)
> library(sp)
> library(maptools)
> library(rgdal)
> library(classInt)
> library(lattice)
```

We obtained estimates of the partial sill, σ^2, nugget, τ^2, and decay parameter ϕ based upon some empirical semivariogram plots and used them as starting values in the spBayes univariate spatial regression functions bayesGeostatExact and spLM. The bayesGeostatExact function assumes that ϕ and the nugget to partial sill ratio τ^2/σ^2 are fixed. Assuming a normal prior on the regression coefficients $\boldsymbol{\beta}$ and an inverse gamma prior on σ^2 reduces the spatial regression problem to a conjugate Bayesian linear regression model and carry out exact inference by sampling directly from the posterior; no MCMC algorithm is needed. The block of code below describes the above steps after invoking spBayes.

```
> data(BEF.dat)
> BEF.dat <- BEF.dat[BEF.dat$ALLBIO02_KGH>0,]
> bio <- BEF.dat$ALLBIO02_KGH*0.001;
> log.bio <- log(bio)
> coords <- as.matrix(BEF.dat[,c("XUTM","YUTM")])
> p <- 6
> beta.prior.mean <- as.matrix(rep(0, times=p))
> beta.prior.precision <- matrix(0, nrow=p, ncol=p)
> phi <- 0.014
> alpha <- 0.016/0.08
> sigma.sq.prior.shape <- 2.0
> sigma.sq.prior.rate <- 0.08
> sp.exact <- bayesGeostatExact(
+ log.bio~ELEV+SLOPE+SUM_02_TC1+SUM_02_TC2+SUM_02_TC3,
+               data=BEF.dat, coords=coords, n.samples=1000,
+               beta.prior.mean=beta.prior.mean,
+               beta.prior.precision=beta.prior.precision,
+ cov.model="exponential",
+               phi=phi, alpha=alpha,
+               sigma.sq.prior.shape=sigma.sq.prior.shape,
+               sigma.sq.prior.rate=sigma.sq.prior.rate,
+ sp.effects=FALSE)
```

In the above code, data specifies the data frame containing the variables, coords denotes the objects containing the coordinates, nsamples specifies the number of posterior samples to be drawn, the two beta.prior arguments set the mean and precision matrix for the normal prior on the regression coefficients, cov.model specifies the covariance function,

`phi` denotes the spatial decay parameter, and the two `sigma.sq.prior` arguments set the inverse gamma prior on σ^2. The `sp.effects` argument indicates whether we want to sample the spatial effects from their posterior distribution. Currently we set it to false.

The above code fits the spatial regression model with an exponential covariance function 415 observations and produces the following estimates based upon 1000 posterior samples. Again, no burn-in is needed as this is exact sampling from the true posterior. A summary of the results are presented below (rounded up to 3 significant digits).

```
> round(summary(sp.exact$p.samples)$quantiles,3)
              2.5%    25%    50%    75%  97.5%
(Intercept) -0.372  0.636  1.161  1.644  2.701
ELEV         0.000  0.000  0.000  0.001  0.001
SLOPE       -0.016 -0.011 -0.009 -0.006 -0.002
SUM_02_TC1  -0.002  0.007  0.010  0.015  0.023
SUM_02_TC2  -0.002  0.003  0.006  0.008  0.013
SUM_02_TC3   0.010  0.017  0.021  0.025  0.032
sigma.sq     0.072  0.078  0.082  0.086  0.095
tau.sq       0.014  0.016  0.016  0.017  0.019
```

A more flexible alternative to `bayesGeostatExact` is the `spLM` function. The latter does not assume that ϕ is fixed, nor is it assumed that the ratio τ^2/σ^2 is fixed. Now, we can assign individual priors on σ^2, τ^2 and ϕ. In addition, we will now implement MCMC. The `spLM` function fits the marginalized model, where the spatial effects as well as the regression coefficients have been integrated out. We will see later how these spatial effects can be recovered using the `spRecover` function. The regression coefficients are updated from their normal full conditional distrbutions, while the ϕ, σ^2 and τ^2 will be updated using Metropolis steps. We need starting values and tuning parameters for these parameters. These are passed as two `lists` named `starting` and `tuning` respectively. A third `list`, named `priors`, assigns prior distributions to the parameters. The rest of the arguments and their specifications are similar to `bayesGeostatExact`. The chunk of code below illustrates the use of `spLM` to draw $10,000$ MCMC samples.

```
> n.samples <- 10000
> bef.sp <- spLM(log.bio~ELEV+SLOPE+SUM_02_TC1+SUM_02_TC2
+ +SUM_02_TC3,
+               data=BEF.dat, coords=coords,
+               starting=list("phi"=3/200,"sigma.sq"=0.08,
+                             "tau.sq"=0.02),
+                tuning=list("phi"=0.1, "sigma.sq"=0.05,
+                            "tau.sq"=0.05),
+                priors=list("phi.Unif"=c(3/1500, 3/50),
+                            "sigma.sq.IG"=c(2, 0.08),
+                            "tau.sq.IG"=c(2, 0.02)),
+                cov.model="exponential",n.samples=n.samples)
```

Note that we have not specified a prior distribution for the regression coefficients $\boldsymbol{\beta}$; a flat prior is used by default. Their posterior estimates are not very different from those obtained in `bayesGeostatExact` so we do not present them again. Below is a summary of the estimates of σ^2, τ^2 and ϕ:

```
> round(summary(mcmc(bef.sp$p.theta.samples))$quantiles,3)
          2.5%   25%   50%   75% 97.5%
sigma.sq 0.041 0.064 0.081 0.094 0.111
tau.sq   0.004 0.010 0.023 0.039 0.062
phi      0.005 0.008 0.010 0.012 0.017
```

Figure 6.8 *MCMC trace plots of* $\boldsymbol{\beta}$

The `spRecover` function uses composition sampling to obtain the posterior samples of the marginalized regression coefficients and the spatial effects. This is easily implemented by passing the `bef.sp` object computed above to the `spRecover` function:

```
> burn.in <- floor(0.75*n.samples)
> bef.sp <- spRecover(bef.sp, start=burn.in)
```

The posterior samples of the regression coefficients and the spatial effects can then be obtained as

```
> beta.samples = bef.sp$p.beta.recover.samples
> w.samples = bef.sp$p.w.recover.samples
```

respectively. The output from `spLM` is easily exported to the `CODA` package in `R` for convergence diagnostics. For example, if we wish to generate trace plots of the six regression coefficients (including the intercept), we execute

```
> par(mfrow=c(3,2))
> plot(beta.samples, auto.layout=TRUE, density=FALSE)
```

Figure 6.8 shows the resulting trace plots. We could also obtain the posterior mean and standard deviation for the spatial effects as below. Using the `apply` function in `R` helps avoid the undesirable programming practice of `for loops`.

```
> w.hat.mu <- apply(w.samples,1,mean)
> w.hat.sd <- apply(w.samples,1,sd)
```

These posterior means can then be interpolated across the domain to produce "maps" of spatial variables. Assuming that we have already obtained the residuals from a simple

Figure 6.9 *Interpolated surface of the OLS model residuals and the mean of the random spatial effects posterior distribution.*

ordinary least squares (OLS) model, and stored them in the object `bio.resid`, we plot side by side interpolated surfaces for residuals from the OLS model and for the posterior means of the spatial effects from the spatial regression model. The code to do this supplied below.

```
> par(mfrow=c(1,2))
> surf <- mba.surf(cbind(coords, bio.resid),
+ no.X=x.res, no.Y=y.res,
+ extend=FALSE)$xyz.est
> z.lim <- range(surf[[3]], na.rm=TRUE)
> image.plot(surf, xaxs = "r", yaxs = "r",
+ zlim=z.lim, main="LM residuals")
> surf <- mba.surf(cbind(coords, w.hat.mu),
+ no.X=x.res, no.Y=y.res,
+ extend=FALSE)$xyz.est
> image.plot(surf, xaxs = "r", yaxs = "r",
+ zlim=z.lim, main="Mean spatial effects")
```

6.3.1.1 Prediction

Having obtained samples from the parameters' posterior distribution, we now turn to prediction or Bayesian kriging. Using the `spLM` object and predictor variables from *new* locations, the function `spPredict` allows us to sample from the posterior predictive distribution of every pixel across the BEF. We are only interested in predictions within the BEF; however, the predictor variable grid extends well beyond the BEF bounds. Therefore, we would like to *clip* the predictor grid to the BEF bounding polygon. The code block below makes use of the `readShapePoly` function from the `maptools` package and `readGDAL` function from the `rgdal` package to read the bounding polygon and predictor variable grid stack, respectively.

```
> library(maptools)
> library(rgdal)
> BEF.shp <- readShapePoly("BEF-data/BEF_bound.shp")
> shp2poly <- BEF.shp@polygons[[1]]@Polygons[[1]]@coords
> BEF.poly <- as.matrix(shp2poly)
> BEF.grids <- readGDAL("dem_slope_lolosptc_clip_60.img")
```

We then construct the prediction design matrix for the entire grid extent. Then extract the coordinates of the BEF bounding polygon vertices and use the `pointsInPoly` `spBayes` function to obtain the desired subset of the prediction design matrix and associated prediction coordinates (i.e., pixel centroids). Finally, the `spPredict` function is called and posterior predictive samples are stored in `bef.bio.pred`. The code below implements these steps.

```
> pred.covars <- cbind(BEF.grids[["band1"]],
+                      BEF.grids[["band2"]],
+                      BEF.grids[["band3"]],
+                      BEF.grids[["band4"]],
+                      BEF.grids[["band5"]])
> pred.covars <- cbind(rep(1, nrow(pred.covars)), pred.covars)
> pred.coords <- SpatialPoints(BEF.grids)@coords
> pointsInPolyOut <- pointsInPoly(BEF.poly, pred.coords)
> pred.covars <- pred.covars[pointsInPolyOut,]
> pred.coords <- pred.coords[pointsInPolyOut,]
> bef.bio.pred <- spPredict(bef.sp, start=burn.in,
+                           pred.coords=pred.coords,
+                           pred.covars=pred.covars)
```

With access to each pixel's posterior predictive distribution we can map any summary statistics of interest. In Figure 6.10 we compare the log metric tons of biomass interpolated over the observed plots to that of the pixel-level prediction. The generation of this image plot requires some additional code to clip the interpolation grid produced by `mba.surf` to the BEF polygon. Here we also demonstrate the `sp` function `overlay` to subset the grid (that is an alternative approach to using `pointsInPoly`). The code below demonstrates this.

```
> bef.bio.pred.mu = apply(bef.bio.pred$p.y.predictive.samples,
+                         1,mean)
> bef.bio.pred.sd = apply(bef.bio.pred$p.y.predictive.samples,
+                         1,sd)
> surf <- mba.surf(cbind(coords, log.bio), no.X=x.res, no.Y=x.res,
+                  extend=TRUE, sp=TRUE)$xyz.est
> surf  <- surf [!is.na(overlay(surf, BEF.shp)),]
> surf  <- as.image.SpatialGridDataFrame(surf)
> z.lim <- range(surf[["z"]], na.rm=TRUE)
> pred.grid <- as.data.frame(list(pred.coords,
+                                 pred.mu=bef.bio.pred.mu,
+                                 pred.sd=bef.bio.pred.sd))
> coordinates(pred.grid) = c("x", "y")
> gridded(pred.grid) <- TRUE
> pred.mu.image <-
+    as.image.SpatialGridDataFrame(pred.grid["pred.mu"])
```

This completes the construction of the prediction grid. The next block of code creates the image plots.

```
> par(mfrow=c(1,2))
> image.plot(surf, axes=TRUE, zlim=z.lim, col=tim.colors(25),
+            xaxs = "r", yaxs = "r",
+            main="Log metric tons of biomass")
> plot(BEF.shp, add=TRUE)
> image.plot(pred.mu.image, zlim=z.lim, col=tim.colors(25),
+            xaxs = "r", yaxs = "r",
```

Figure 6.10 *Interpolated surface of observed log metric tons of biomass and the posterior predictive mean of each pixels.*

```
+             main="Mean predicted log metric tons of biomass")
> plot(BEF.shp, add=TRUE)
```

Finally, `sp` grid data objects can be exported using functions within `rgdal` as described below.

```
> writeGDAL(pred.grid["pred.mu"], "BEF_Pred_mu_biomass.tif")
> writeGDAL(pred.grid["pred.sd"], "BEF_Pred_sd_biomass.tif")
```

This is helpful if we want to combine our prediction surface with other spatial data within a full featured Geographic Information system (e.g., GRASS, QGIS, ArcGIS, etc.).

6.3.1.2 Model selection

To compare several alternative models with varying degrees of richness, we might use the GPD criterion (Gelfand and Ghosh, 1998) or the Deviance Information Criterion (DIC) (Spiegelhalter et al., 2002). Both these methods can be implemented in `spBayes`. The code below estimates and computes these two criterion for the three models in Tables 6.4 and 6.5.

First, we store the output from `spLM` for the OLS model.

```
> set.seed(1)
> ##Candidate model 1
> lm.obj <- lm(log.bio~ELEV+SLOPE+SUM_02_TC1+
+                    +SUM_02_TC2+SUM_02_TC3, data=BEF.dat)
> cand.1 <- bayesLMRef(lm.obj, n.samples)
```

The code below runs and stores the output for simple spatial model with only an intercept.

```
> ##Candidate model 2
> cand.2 <- spLM(log.bio~1, data=BEF.dat, coords=coords,
+               starting=list("phi"=3/200,"sigma.sq"=0.08,
+                             "tau.sq"=0.02),
+               tuning=list("phi"=0.1, "sigma.sq"=0.05,
+                           "tau.sq"=0.05),
+               priors=list("phi.Unif"=c(3/1500, 3/50),
+                           "sigma.sq.IG"=c(2, 0.08),
+                           "tau.sq.IG"=c(2, 0.02)),
```

	bar.D	D.bar.Omega	pD	DIC
Non spatial	-538.50	-545.40	6.90	-531.60
Spatial intercept only	-1202.50	-1416.80	214.30	-988.20
Spatial with predictors	-1084.90	-1266.80	182.00	-902.90

Table 6.4 *Candidate model comparison using DIC.*

	G	P	D
Non spatial	41.00	42.40	83.40
Spatial intercept only	2.80	18.20	21.00
Spatial with predictors	5.00	22.30	27.30

Table 6.5 *Candidate model comparison using GPD.*

```
+                cov.model="exponential",
+                n.samples=n.samples)
```

Next, we store the output from `spLM` for a spatial model with the intercept and the five predictors included.

```
> ##Candidate model 3
> cand.3 <- spLM(log.bio~ELEV+SLOPE+SUM_02_TC1+SUM_02_TC2
+                       +SUM_02_TC3, data=BEF.dat, coords=coords,
+                starting=list("phi"=3/200,"sigma.sq"=0.08,
+                              "tau.sq"=0.02),
+                tuning=list("phi"=0.1, "sigma.sq"=0.05,
+                            "tau.sq"=0.05),
+                priors=list("phi.Unif"=c(3/1500, 3/50),
+                            "sigma.sq.IG"=c(2, 0.08),
+                            "tau.sq.IG"=c(2, 0.02)),
+                cov.model="exponential",
+                n.samples=n.samples)
```

Finally, we compute the DIC and GPD scores for the three models using the `spDiag` function. Note that for both the second and third model, DIC computes the likelihood with the estimated spatial effects included. Therefore, we use the `spRecover` function to recover the spatial effects for these two models.

```
> cand.1.DIC <- spDiag(cand.1, start=burn.in, verbose=FALSE)
> cand.2 <- spRecover(cand.2, start=burn.in, verbose=FALSE)
> cand.2.DIC <- spDiag(cand.2, verbose=FALSE)
> cand.3 <- spRecover(cand.3, start=burn.in, verbose=FALSE)
> cand.3.DIC <- spDiag(cand.3, start=burn.in, verbose=FALSE)
```

The results from this code are displayed in Tables 6.4 and 6.5.

6.3.2 Non-Gaussian spatial GLM

The function `spGLM` fits the Poisson and binomial model using the log and logit link function, respectively. Here we illustrate the use of `spGLM` to fit a Poisson generalized linear mixed model with spatially dependent random effects. We consider a simulated dataset with 50 locations inside the unit square. We generate a latent Gaussian spatial random field $w(\mathbf{s})$ using an exponential covariance function with $\sigma^2 = 2$ and $\phi = 3/0.5$ (so the spatial range is 0.5). Finally, the outcome in each location is generated from a Poisson distribution with intensity $\exp(\beta_0 + w(\mathbf{s}_i))$. The dataset is generated in the code block below.

```
> n <- 50
> coords <- cbind(runif(n, 0, 1), runif(n, 0, 1))
> phi <- 3/0.5
> sigma.sq <- 2
> R <- exp(-phi * iDist(coords))
> w <- mvrnorm(1, rep(0, n), sigma.sq * R)
> beta.0 <- 0.1
> y <- rpois(n, exp(beta.0 + w))
```

Assuming there is no spatial dependence we might fit a simple non-spatial GLM using

```
> pois.nonsp <- glm(y ~ 1, family = "poisson")
> beta.starting <- coefficients(pois.nonsp)
> beta.tuning <- t(chol(vcov(pois.nonsp)))
```

These coefficients and the Cholesky square root of the parameters' estimated covariances will be used as starting values and Metropolis sampler tuning values in the subsequent call to `spGLM`. In addition to the regression coefficients we specify starting values for the spatial range `phi` and variance `sigma.sq` as well as the random spatial effects `w`.

Here posterior inference is based on three MCMC chains each of length 15,000. The code to generate the first of these chains is given below.

```
> n.batch <- 300
> batch.length <- 50
> n.samples <- n.batch * batch.length
> pois.sp.chain.1 <-
+       spGLM(y ~ 1,family = "poisson",coords = coords,
+             starting = list(beta = beta.starting,
+                             phi = 3/0.5,
+                             sigma.sq = 1,w = 0),
+             tuning = list(beta = 0.1, phi = 0.5,
+                           sigma.sq = 0.1,w = 0.1),
+             priors = list("beta.Flat",
+                           phi.Unif = c(3/1, 3/0.1),
+                           sigma.sq.IG = c(2, 1)),
+             amcmc = list(n.batch = n.batch,
+                          batch.length=batch.length,
+                          accept.rate = 0.43),
+                          cov.model = "exponential")
```

The `coda` package's `plot` function can be used to plot chain trace plots using the code below.

```
> samps <- mcmc.list(pois.sp.chain.1$p.beta.theta.samples,
+                    pois.sp.chain.2$p.beta.theta.samples,
+                    pois.sp.chain.3$p.beta.theta.samples)
> plot(samps)
```

The convergence of multiple chains can be assessed using diagnostics detailed in Gelman and Rubin (1992). The `gelman.diag` function in the `coda` package calculates the "potential scale reduction factor" for each each parameter, along with the associated upper and lower confidence limits. Approximate convergence is diagnosed when the upper confidence limit is close to 1. We can also plot the Gelman and Rubin's shrink factor versus the number of MCMC samples; here again convergence is diagnosed when the upper confidence limit remains close to 1. Figure 6.11 suggests we should discard the first ∼10,000 samples as burn-in prior to summarizing the parameters' posterior distributions.

```
> print(gelman.diag(samps))
```

Figure 6.11 *MCMC chain convergence diagnostics.*

```
Potential scale reduction factors:

            Point est. Upper C.I.
(Intercept)       1.09       1.20
sigma.sq          1.02       1.05
phi               1.01       1.02

Multivariate psrf

1.03
> gelman.plot(samps)
> burn.in <- 10000
> print(round(summary(window(samps, start = burn.in))$quantiles[,
+        c(3, 1, 5)], 2))
             50%  2.5% 97.5%
(Intercept) 0.18 -1.44  0.82
sigma.sq    1.58  0.87  3.41
phi         9.89  3.69 23.98
```

Figure 6.12 *Observed and estimated counts.*

Given the post burn-in samples, we can also generate surfaces of the estimated counts. The code below produces Figure 6.12.

```
> samps <- as.matrix(window(samps, start = burn.in))
> w <- cbind(pois.sp.chain.1$p.w.samples[, burn.in:n.samples],
+            pois.sp.chain.2$p.w.samples[, burn.in:n.samples],
+            pois.sp.chain.3$p.w.samples[, burn.in:n.samples])
> beta.0.hat <- mean(samps[, "(Intercept)"])
> w.hat <- apply(w, 1, mean)
> y.hat <- exp(beta.0.hat + w.hat)
> par(mfrow = c(1, 2))
> surf <- mba.surf(cbind(coords, y), no.X = 100, no.Y = 100,
+                  extend = TRUE)$xyz.est
> image.plot(surf, main = "Observed counts")
> points(coords)
> surf <- mba.surf(cbind(coords, y.hat), no.X = 100, no.Y = 100,
+                  extend = TRUE)$xyz.est
> image.plot(surf, main = "Fitted counts")
> points(coords)
```

Given the posterior samples of the model parameters, we use composition sampling to draw from the posterior predictive distribution of any new location. The code block below constructs a grid to define the prediction locations then calls `spPredict`. For each new location's posterior predictive samples we can draw a corresponding vector of realizations from `rpois`.

```
> pred.coords <- as.matrix(expand.grid(seq(0.01, 0.99,
+                          length.out = 20),
+                          seq(0.01, 0.99,
+                          length.out = 20)))
> pred.covars <- as.matrix(rep(1, nrow(pred.coords)))
> pois.pred <- spPredict(pois.sp.chain.1,
+                        start = 10000,
+                        thin = 10,
+                        pred.coords = pred.coords,
+                        pred.covars = pred.covars)
```

Figure 6.13 *Observed and predicted counts.*

```
> y.pred <- apply(pois.pred$p.y.predictive.samples,1,median)
```

These realization can then be summarized to assess predictive performance. For example, Figure 6.13 illustrates the median prediction over the domain.

```
> par(mfrow = c(1, 2))
> surf <- mba.surf(cbind(coords, y), no.X = 100, no.Y = 100,
+                  extend = TRUE)$xyz.est
> image.plot(surf, main = "Observed counts")
> points(coords)
> surf <- mba.surf(cbind(pred.coords, y.pred),
+                  no.X = 100, no.Y = 100,
+                  extend = TRUE)$xyz.est
> image.plot(surf, main = "Predicted counts")
> points(pred.coords, pch = "x", cex = 1)
> points(coords, pch = 19, cex = 1)
```

6.4 Areal data models

6.4.1 Disease mapping

An area of strong biostatistical and epidemiological interest is that of *disease mapping*. Here we typically have count data of the following sort:

$$\begin{aligned} Y_i &= \text{observed number of cases of disease in county } i,\ i = 1, \ldots, I \\ E_i &= \text{expected number of cases of disease in county } i,\ i = 1, \ldots, I \end{aligned}$$

The Y_i are thought of as random variables, while the E_i are thought of as fixed and known functions of n_i, the number of persons at risk for the disease in county i. As a simple starting point, we might assume that

$$E_i = n_i \bar{r} \equiv n_i \left(\frac{\sum_i y_i}{\sum_i n_i} \right) \equiv \sum_i y_i \frac{\sum_i n_i}{\sum_i n_i},$$

i.e., $\bar{r}$ is the overall disease rate in the entire study region. The second equivalence interprets E_i as scaling the total number of cases by the proportion of the population at risk in county

i. These E_i thus correspond to a kind of "null hypothesis," where we expect a constant disease rate in every county. This process is called *internal standardization*, since it centers the data (some counties will have observed rates higher than expected, and some less) but uses only the observed data to do so.

Internal standardization is "cheating" (or at least "empirical Bayes") in some sense, since, evidently, the E_i are not fixed but are functions of the data. A better approach might be to make reference to an existing standard table of age-adjusted rates for the disease (as might be available for many types of cancer). Then, after stratifying the population by age group, the E_i emerge as

$$E_i = \sum_j n_{ij} r_j \ ,$$

where n_{ij} is the person-years at risk in area i for age group j (i.e., the number of persons in age group j who live in area i times the number of years in the study), and r_j is the disease rate in age group j (taken from the standard table). This process is called *external standardization.* In either case, in its simplest form a disease map is just a display (in color or greyscale) of the raw disease rates overlaid on the areal units.

6.4.2 Traditional models and frequentist methods

If E_i is not too large (i.e, the disease is rare or the regions i are sufficiently small), the usual model for the Y_i is the Poisson model,

$$Y_i|\eta_i \sim Po(E_i\eta_i) \ ,$$

where η_i is the true *relative risk* of disease in region i. The maximum likelihood estimate (MLE) of η_i is readily shown to be

$$\hat{\eta}_i \equiv SMR_i = \frac{Y_i}{E_i} \ ,$$

the *standardized morbidity (or mortality) ratio* (SMR), i.e., the ratio of observed to expected disease cases (or deaths). Note that $Var(SMR_i) = Var(Y_i)/E_i^2 = \eta_i/E_i$, and so we might take $\widehat{Var}(SMR_i) = \hat{\eta}_i/E_i = Y_i/E_i^2$. This in turn permits calculation of traditional confidence intervals for η_i (although this is a bit awkward since the data are discrete), as well as hypothesis tests.

Example 6.5 To find a confidence interval for η_i, arguably, it would be preferable to work on the log scale, i.e., to assume that $\log SMR_i$ is roughly *normally* distributed. Using the delta method (Taylor series expansion), one can find that

$$Var[\log(SMR_i)] \approx \frac{1}{SMR_i^2} Var(SMR_i) = \frac{E_i^2}{Y_i^2} \times \frac{Y_i}{E_i^2} = \frac{1}{Y_i} \ .$$

An approximate 95% CI for $\log \eta_i$ is thus $\log SMR_i \pm 1.96/\sqrt{Y_i}$, and so (transforming back) an approximate 95% CI for η_i is

$$\left(SMR_i \exp(-1.96/\sqrt{Y_i}) \ , \ SMR_i \exp(1.96/\sqrt{Y_i})\right) \ .$$

■

Example 6.6 Suppose we wish to test whether the true relative risk in county i is elevated or not, i.e.,

$$H_0 : \eta_i = 1 \ \text{ versus } \ H_A : \eta_i > 1 \ .$$

Under the null hypothesis, $Y_i \sim Po(E_i)$, so the p-value for this test is

$$p = Pr(X \geq Y_i|E_i) = 1 - Pr(X < Y_i|E_i) = 1 - \sum_{x=0}^{Y_i-1} \frac{\exp(-E_i)E_i^x}{x!} \ .$$

This is the (one-sided) p-value; if it is less than 0.05 we would typically reject H_0 and conclude that there is a statistically significant excess risk in county i. ■

6.4.3 Hierarchical Bayesian methods

The methods of the previous section are fine for detecting extra-Poisson variability (overdispersion) in the observed rates, but what if we seek to *estimate* and *map* the underlying relative risk surface $\{\eta_i, i = 1, \ldots, I\}$? In this case we might naturally think of a *random effects* model for the η_i, since we would likely want to assume that all the true risks come from a common underlying distribution. Random effects models also allow the procedure to "borrow strength" across the various counties in order to come up with an improved estimate for the relative risk in each.

The random effects here, however, can be high dimensional, and are couched in a nonnormal (Poisson) likelihood. Thus, as in the previous sections of this chapter, the most natural way of handling this more complex model is through hierarchical Bayesian modeling, as we now describe.

6.4.3.1 Poisson-gamma model

As a simple initial model, consider

$$\begin{aligned} Y_i \mid \eta_i &\overset{ind}{\sim} Po(E_i\eta_i) \ , i = 1, \ldots, I, \\ \text{and } \eta_i &\overset{iid}{\sim} G(a, b) \ , \end{aligned}$$

where $G(a, b)$ denotes the *gamma* distribution with mean $\mu = a/b$ and variance $\sigma^2 = a/b^2$; note that this is the gamma parametrization used by the `WinBUGS` package. Solving these two equations for a and b we get $a = \mu^2/\sigma^2$ and $b = \mu/\sigma^2$. Suppose we set $\mu = 1$ (the "null" value) and $\sigma^2 = (0.5)^2$ (a fairly large variance for this scale). Figure 6.14(a) shows a sample of size 1000 from the resulting $G(4, 4)$ prior; note the vertical reference line at $\eta_i = \mu = 1$.

Inference about $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_I)'$ is now based on the resulting posterior distribution, which in the Poisson-gamma emerges in closed form (thanks to the conjugacy of the gamma prior with the Poisson likelihood) as $\prod_i p(\eta_i|y_i)$, where $p(\eta_i|y_i)$ is $G(y_i + a, E_i + b)$. Thus a suitable point estimate of η_i might be the posterior mean,

$$E(\eta_i|\mathbf{y}) = E(\eta_i|y_i) = \frac{y_i + a}{E_i + b} = \frac{y_i + \frac{\mu^2}{\sigma^2}}{E_i + \frac{\mu}{\sigma^2}} \tag{6.20}$$

$$\begin{aligned} &= \frac{E_i\left(\frac{y_i}{E_i}\right)}{E_i + \frac{\mu}{\sigma^2}} + \frac{\left(\frac{\mu}{\sigma^2}\right)\mu}{E_i + \frac{\mu}{\sigma^2}} \\ &= w_i \, SMR_i + (1 - w_i)\mu \ , \end{aligned} \tag{6.21}$$

where $w_i = E_i/[E_i + (\mu/\sigma^2)]$, so that $0 \leq w_i \leq 1$. Thus the Bayesian point estimate (6.21) is a *weighted average* of the the data-based SMR for region i, and the prior mean μ. This estimate is approximately equal to SMR_i when w_i is close to 1 (i.e., when E_i is big, so the data are strongly informative, or when σ^2 is big, so the prior is weakly informative). On the other hand, (6.21) will be approximately equal to μ when w_i is close to 0 (i.e., when E_i is small, so the data are sparse, or when σ^2 is small, so that the prior is highly informative).

Figure 6.14 *Samples of size 1000 from a gamma(4,4) prior (a) and a gamma(27+4, 21+4) posterior (b) for η_i.*

As an example, suppose in county i we observe $y_i = 27$ disease cases, when only $E_i = 21$ were expected. Under our $G(4,4)$ prior we obtain a $G(27+4, 21+4) = G(31, 25)$ posterior distribution; Figure 6.14(b) shows a sample of size 1000 drawn from this distribution. From (6.20) this distribution has mean $31/25 = 1.24$ (consistent with the figure), indicating slightly elevated risk (24%). However, the posterior probability that the true risk is bigger than 1 is $P(\eta_i > 1 \,|\, y_i) = .863$, which we can derive exactly (say, using `1 - pgamma(25,31)` in `R`), or estimate empirically as the proportion of samples in Figure 6.14(b) that are greater than 1. In either case, we see substantial but not overwhelming evidence of risk elevation in this county.

If we desired a $100 \times (1-\alpha)\%$ confidence interval for η_i, the easiest approach would be to simply take the upper and lower $\alpha/2$-points of the $G(31, 25)$ posterior, since the resulting interval $\left(\eta_i^{(L)}, \eta_i^{(U)}\right)$, would be such that $P\left[\eta_i \in \left(\eta_i^{(L)}, \eta_i^{(U)}\right) \mid y_i\right] = 1-\alpha$, by definition of the posterior distribution. This is the so-called *equal-tail credible interval* mentioned in Subsection 5.2.2. In our case, taking $\alpha = .05$ we obtain $(\eta_i^{(L)}, \eta_i^{(U)}) = (.842, 1.713)$, again indicating no "significant" elevation in risk for this county. (In `R` the appropriate commands here are `qgamma(.025, 31)/25` and `qgamma(.975, 31)/25`.)

Finally, in a "real" data setting we would obtain not 1 but I point estimates, interval estimates, and posterior distributions, one for each county. Such estimates would often be summarized in a choropleth map, say, in `R` or `ArcView`. Full posteriors are obviously difficult to summarize spatially, but posterior means, variances, or confidence limits are easily mapped in this way. We shall explore this issue in the next subsection.

6.4.3.2 Poisson-lognormal models

The gamma prior of the preceding section is very convenient computationally, but suffers from a serious defect: it fails to allow for spatial correlation among the η_i. To do this we would need a *multivariate* version of the gamma distribution; such structures exist but are awkward both conceptually and computationally. Instead, the usual approach is to place some sort of multivariate *normal* distribution on the $\psi_i \equiv \log \eta_i$, the *log*-relative risks.

Thus, consider the following augmentation of our basic Poisson model:

$$\begin{aligned} Y_i \,|\, \psi_i &\overset{ind}{\sim} Po\left(E_i\, e^{\psi_i}\right) , \\ \text{where } \psi_i &= \mathbf{x}_i'\boldsymbol{\beta} + \theta_i + \phi_i \, . \end{aligned} \tag{6.22}$$

The $\mathbf{x}_i$ are explanatory spatial covariates, having parameter coefficients $\boldsymbol{\beta}$. The covariates are *ecological*, or county (not individual) level, which may lead to problems of ecological bias (to be discussed later). However, the hope is that they will explain some (perhaps all) of the spatial patterns in the Y_i.

Before, we discuss modeling for the θ_i and ϕ_i, we digress to return to the internal standardization problem mentioned at the beginning of this section. It is evident in (6.22) that, with internal standardization, we do not have a valid model. That is, the data appears on both sides of the Poisson specification. This issue is generally ignored in the literature but the reader might appreciate the potential benefit of an alternative parametrization which writes $E(Y_i) = n_i p_i$. Here, again, n_i is the number of people at risk in, say, county i and p_i is the incidence rate in county i. Now, we would model p_i and avoid the internal standardization problem. In fact, this is the customary Poisson parametrization where we expect large n_i and small p_i. And, if interest is in the η_i's, the SMR's, after posterior inference on p_i, we get posterior inference for eta_i for free since, given the data, $\eta_i = n_i p_i / E_i$. Why isn't this model more widely used? We suspect it is the fact that, in principle, we would have to introduce, say, a logit or probit link to model the p_i which makes the model more difficult to fit. (The difficulty with the log link is that, upon inversion, we might find probabilities greater than 1. However, when the p_i are small this need not be a problem.) In the sequel we stay with the customary parametrization but encourage the more ambitious reader to consider the alternative.

Returning to (6.22), the θ_i capture region-wide *heterogeneity* via an ordinary, exchangeable normal prior,

$$\theta_i \overset{iid}{\sim} N(0\ ,\ 1/\tau_h)\ , \tag{6.23}$$

where τ_h is a precision (reciprocal of the variance) term that controls the magnitude of the θ_i. These random effects capture extra-Poisson variability in the log-relative risks that varies "globally," i.e., over the entire study region.

Finally, the ϕ_i are the parameters that make this a spatial model by capturing regional *clustering*. That is, they model extra-Poisson variability in the log-relative risks that varies "locally," so that nearby regions will have more similar rates. A plausible way to attempt this might be to try a point-referenced model on the parameters ϕ_i. For instance, writing $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_I)'$, we might assume that

$$\boldsymbol{\phi} \mid \mu, \boldsymbol{\lambda} \sim N_I(\mu, H(\boldsymbol{\lambda}))\ ,$$

where N_I denotes the I-dimensional normal distribution, μ is the (stationary) mean level, and $(H(\boldsymbol{\lambda}))_{ii'}$ gives the covariance between ϕ_i and $\phi_{i'}$ as a function of some hyperparameters $\boldsymbol{\lambda}$. The standard forms given in Table 2.1 remain natural candidates for this purpose. The issue here is that it is not clear what are appropriate inter-areal unit distances. Should we use centroid to centroid? Does this make sense with units of quite differing sizes and irregular shapes? A further challenge arises when we have many areal units. Then, the so-called big N problem emerges (see Chapter 11) with regard to high-dimensional matrix inversion. Moreover, with areal unit data, we are not interested in interpolation, that is, kriging is the usual benefit for introducing a covariance function. As a result, it is customary in hierarchical analyses of areal data to return to neighbor-based notions of proximity, and ultimately, to return to CAR specifications for $\boldsymbol{\phi}$ (Section 4.3). In the present context (with the CAR model placed on the elements of $\boldsymbol{\phi}$ rather than the elements of $\mathbf{Y}$), we will write

$$\boldsymbol{\phi} \sim CAR(\tau_c)\ , \tag{6.24}$$

where by this notation we mean the improper CAR (IAR) model in (4.16) with y_i replaced by ϕ_i, τ^2 replaced by $1/\tau_c$, and using the 0-1 (adjacency) weights w_{ij}. Thus τ_c is a *precision* (not variance) parameter in the CAR prior (6.24), just as τ_h is a precision parameter in the heterogeneity prior (6.23).

6.4.3.3 CAR models and their difficulties

Compared to point-level (geostatistical) models, CAR models are very convenient computationally, since our method of finding the posterior of $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$ is itself a conditional algorithm, the Gibbs sampler. Recall from Subsection 5.3.1 that this algorithm operates by successively sampling from the *full conditional* distribution of each parameter (i.e., the distribution of each parameter given the data and every other parameter in the model). So for example, the full conditional of ϕ_i is

$$p(\phi_i|\phi_{j\neq i}, \boldsymbol{\theta}, \boldsymbol{\beta}, \mathbf{y}) \propto Po(y_i \mid E_i e^{\mathbf{x}_i'\boldsymbol{\beta}+\theta_i+\phi_i}) \times N(\phi_i \mid \bar{\phi}_i \ , \ 1/(\tau_c m_i)) \ , \tag{6.25}$$

meaning that we do not need to work with the joint distribution of $\boldsymbol{\phi}$ at all. The conditional approach also eliminates the need for any matrix inversion.

While computationally convenient, CAR models have various theoretical and computational challenges, some of which have already been noted in Section 4.3, and others of which emerge now that we are using the CAR as a distribution for the random effects $\boldsymbol{\phi}$, rather than the data $\mathbf{Y}$ itself. We consider two of these.

1. Impropriety: Recall from the discussion surrounding (4.15) that the IAR prior we selected in (6.24) above is *improper*, meaning that it does not determine a legitimate probability distribution (one that integrates to 1). That is, the matrix $\Sigma^{-1} = (D_w - W)$ is singular, and thus its inverse does not exist.

As mentioned in that earlier discussion, one possible fix for this situation is to include a "propriety parameter" ρ in the precision matrix, i.e., $\Sigma^{-1} = (D_w - \rho W)$. Taking $\rho \in \big(1/\lambda_{(1)}, 1/\lambda_{(n)}\big)$, where $\lambda_{(1)}$ and $\lambda_{(n)}$ are the smallest and largest eigenvalues of $D_w^{-1/2} W D_w^{-1/2}$, respectively, ensures the existence of $\Sigma_{\mathbf{y}}$. Alternatively, using the scaled adjacency matrix $\widetilde{W} \equiv Diag(1/w_{i+})W$, Σ^{-1} can be written as $M^{-1}(I - \alpha\widetilde{W})$ where M is diagonal. One can then show (Carlin and Banerjee, 2003; Gelfand and Vounatsou, 2002) that if $|\alpha| < 1$, then $(I - \alpha\widetilde{W})$ will be positive definite, resolving the impropriety problem without eigenvalue calculation.

This immediately raises the question of whether or not to include a propriety parameter in the CAR model. On the one hand, adding, say, ρ, in addition to supplying propriety, enriches the spatial model. Moreover, it provides a natural interpretation and, when $\rho = 0$. we obtain an independence model. So, why would we not include ρ? One reason is that the interpretation is not quite what we want. Recognizing that the inclusion of ρ makes $E(\phi_i = \rho \sum_{j\sim i} w_{ij}\phi_j$, we see that it isn't necessarily sensible that we expect ϕ_i to be only a portion of the average of its neighbors. Why is that an appropriate smoother? A further difficulty with this fix (discussed in more detail near the end of Subsection 4.3.1) is that this new prior typically does not deliver enough spatial similarity unless ρ is quite close to 1, thus getting us very close to the same problem again! Indeed in model fitting, ρ always is very close to 1; it is as though, as a spatial random effects model, the data wants ρ to be 1. Some authors (e.g., Carlin and Banerjee, 2003) recommend an informative prior that insists on larger ρ's or α's (say, a $beta(18, 2)$), but this is controversial since there will typically be little true prior information available regarding the magnitude of α.

Our recommendation is to work with the intrinsic CAR specification, i.e., ignore the impropriety of the standard CAR model (6.24) and continue! After all, we are only using the CAR model as a *prior*; the *posterior* will generally still emerge as proper, so Bayesian inference may still proceed. This is the usual approach, but it also requires some care, as follows: this improper CAR prior is a *pairwise difference prior* (Besag et al., 1995) that is identified only up to an additive constant. Thus to identify an intercept term β_0 in the log-relative risk, we must add the constraint $\sum_{i=1}^{I} \phi_i = 0$. Note that, in implementing a Gibbs sampler to fit (6.22), this constraint can be imposed *numerically* by recentering each sampled $\boldsymbol{\phi}$ vector around its own mean following each Gibbs iteration, so-called centering

on the fly. Note that we will prefer to implement this centering during the course of model fitting rather than transforming to the lower dimensional proper distribution, in order to retain the convenient conditional distributions associated with the pairwise difference CAR.

As a brief digression here, we note that the simple model, $Y_i = \phi_i + \theta_i$ where, again, the θ_i are i.i.d. error terms and the ϕ's are an intrinsic CAR is a legitimate data model. That is, $Y_i|\phi_i \sim N(\phi_i, \sigma_h^2)$. (This setting is analogous to placing an improper prior on μ in the model, $Y_i = \mu + \theta_i$.) In fact, we do not need an intercept in this model. If we were to include one, we would have to center the ϕ's at each model fitting iteration. Contrast this specification with $Y_i = \mathbf{X}_i^T\boldsymbol{\beta} + \phi_i$ which cannot be a data model as we see from the conditional distribution of $Y_i|\boldsymbol{\beta}$.

2. Selection of τ_c and τ_h: Clearly the values of these two prior precision parameters will control the amount of extra-Poisson variability allocated to "heterogeneity" (the θ_i) and "clustering" (the ϕ_i). But they cannot simply be chosen to be arbitrarily large, since then the ϕ_i and θ_i would be *unidentifiable*: note that we see only a single Y_i in each county, yet we are attempting to estimate *two* random effects for each i! Eberly and Carlin (2000) investigate convergence and Bayesian learning for this data set and model, using fixed values for τ_h and τ_c.

Suppose we decide to introduce third-stage priors (*hyperpriors*) on τ_c and τ_h. Under the intrinsic CAR specification, the distribution for the ϕ_i's is improper. What power of τ_c should be introduced into the multivariate "Gaussian distribution" for the ϕ's to yield, with the hyperprior, the full conditional distribution for τ_c? Evidently, with an improper distribution, there is no *correct* answer; the power is not determined. This problem has been considered in the literature; see, e.g., Hodges, Carlin and Fan (2003). A natural way to think about this issue is to recognize the effective dimension of the intrinsic CAR model. For instance, with a connected set of areal units, we have one more ϕ than we need; the intrinsic CAR is proper in the $n-1$ dimensional space where we set one of the ϕ's equal to the sum of the remainder. Thus, the centering on the fly constraint makes the distribution proper so the suggestion is to use the power $(n - 1/2$ for τ_c. However, in some areal unit settings, we have singleton units, units with no neighbors (perhaps, say, two units not connected to the rest). Here, the recommendation is to deduce from n the number of "islands" in the data. With no singletons, we have one island, hence $n-1$, with one singleton, we have two islands, hence, $n-2$, etc. Of course, with a proper CAR, this problem disappears.

These hyperpriors cannot be arbitrarily vague due to the identifiability challenge above. Still, the gamma offers a conjugate family here, so we might simply take

$$\tau_h \sim G(a_h, b_h) \text{ and } \tau_c \sim G(a_c, b_c) \ .$$

To make this prior "fair" (i.e., equal prior emphasis on heterogeneity and clustering), it is tempting to simply set $a_h = a_c$ and $b_h = b_c$, but this would be incorrect for two reasons. First, the τ_h prior (6.23) uses the usual *marginal* specification, while the τ_c prior (6.24) is specified *conditionally*. Second, τ_c is multiplied by the number of neighbors m_i before playing the role of the (conditional) prior precision. Bernardinelli et al. (1995) note that the prior marginal standard deviation of ϕ_i is roughly equal to the prior conditional standard deviation divided by 0.7. Thus a scale that delivers

$$sd(\theta_i) = \frac{1}{\sqrt{\tau_h}} \approx \frac{1}{0.7\sqrt{\bar{m}\tau_c}} \approx sd(\phi_i) \tag{6.26}$$

where $\bar{m}$ is the average number of neighbors may offer a reasonably "fair" specification. Of course, it is fundamentally unclear how to relate the marginal variance of a proper joint distribution with the conditional variance of an improper joint distribution.

Figure 6.15 *Scotland lip cancer data: (a) crude standardized mortality ratios (observed/expected × 100); (b) AFF covariate values.*

Example 6.7 As an illustration of the Poisson-lognormal model (6.22), consider the data displayed in Figure 6.15. These data from Clayton and Kaldor (1987) are the observed (Y_i) and expected (E_i) cases of lip cancer for the $I = 56$ districts of Scotland during the period 1975–1980. One county-level covariate x_i, the percentage of the population engaged in agriculture, fishing or forestry (AFF), is also available (and also mapped in Figure 6.15). Modeling the log-relative risk as

$$\psi_i = \beta_0 + \beta_1 x_i + \theta_i + \phi_i \ , \tag{6.27}$$

we wish to investigate a variety of vague, proper, and arguably "fair" priors for τ_c and τ_h, find the estimated posterior of β_1 (the AFF effect), and find and map the fitted relative risks $E(\psi_i|\mathbf{y})$.

Recall that Y_i cannot inform about θ_i or ϕ_i, but only about their sum $\xi_i = \theta_i + \phi_i$. Making the reparameterization from $(\boldsymbol{\theta}, \boldsymbol{\phi})$ to $(\boldsymbol{\theta}, \boldsymbol{\xi})$, we have the joint posterior,

$$p(\boldsymbol{\theta}, \boldsymbol{\xi}|\mathbf{y}) \propto L(\boldsymbol{\xi}; \mathbf{y}) p(\boldsymbol{\theta}) p(\boldsymbol{\xi} - \boldsymbol{\theta}).$$

This means that

$$p(\theta_i \mid \theta_{j \neq i}, \boldsymbol{\xi}, \mathbf{y}) \propto p(\theta_i)\, p(\xi_i - \theta_i \mid \{\xi_j - \theta_j\}_{j \neq i}) \ .$$

Since this distribution is free of the data $\mathbf{y}$, the θ_i are *Bayesianly unidentified* (and so are the ϕ_i). But this does not preclude *Bayesian learning* (i.e., prior to posterior movement)

```
model
{
  for (i in 1 : regions) {
     O[i] ~ dpois(mu[i])
     log(mu[i]) <- log(E[i]) + beta0 + beta1*aff[i]/10
       + phi[i] + theta[i]
     theta[i] ~ dnorm(0.0,tau.h)
     xi[i] <- theta[i] + phi[i]
     SMRhat[i] <- 100 * mu[i]/E[i]
     SMRraw[i] <- 100 * O[i]/E[i]
  }
  phi[1:regions] ~ car.normal(adj[], weights[], num[], tau.c)

  beta0 ~ dnorm(0.0, 1.0E-5) # vague prior on grand intercept
  beta1 ~ dnorm(0.0, 1.0E-5) # vague prior on AFF effect

  tau.h ~ dgamma(1.0E-3,1.0E-3)  # ``fair'' prior from
  tau.c ~ dgamma(1.0E-1,1.0E-1)  #   Best et al. (1999)

  sd.h <- sd(theta[]) # marginal SD of heterogeneity effects
  sd.c <- sd(phi[])   # marginal SD of clustering effects
  alpha <- sd.c / (sd.h + sd.c)
}
```

Figure 6.16 `WinBUGS` *code to fit the Poisson-normal-CAR model to the Scottish lip cancer data.*

about θ_i. No Bayesian learning would instead require

$$p(\theta_i|\mathbf{y}) = p(\theta_i) \ , \tag{6.28}$$

in the case where both sides are proper (a condition not satisfied by the CAR prior). Note that (6.28) is a much stronger condition than Bayesian unidentifiability, since the data have no impact on the *marginal* (not merely the conditional) posterior distribution.

Recall that, though they are unidentified, the θ_i and ϕ_i are interesting in their own right, as is

$$\alpha = \frac{sd(\boldsymbol{\phi})}{sd(\boldsymbol{\theta}) + sd(\boldsymbol{\phi})} \ ,$$

where $sd(\cdot)$ is the empirical marginal standard deviation function. That is, α is the proportion of the variability in the random effects that is due to clustering (hence $1 - \alpha$ is the proportion due to unstructured heterogeneity). Recall we need to specify vague but proper prior values τ_h and τ_c that lead to acceptable convergence behavior, yet still allow Bayesian learning. This prior should also be "fair," i.e., lead to $\alpha \approx 1/2$ *a priori*.

Figure 6.16 contains the `WinBUGS` code for this problem, which is also available at `http://www.biostat.umn.edu/~brad/data2.html`. Note the use of vague priors for τ_c and τ_h as suggested by Best et al. (1999), and the use of the `sd` function in `WinBUGS` to greatly facilitate computation of α. In fact, `WinBUGS` permits the user to edit the adjacency matrix — for example, to permit the three island counties (the Shetland, Hebrides, and Ornkey) to be considered neighbors of the nearest mainland counties, or not.

Posterior moments (mean, sd) and convergence (lag 1 autocorrelation) summaries for α, β_1, ξ_1, and ξ_{56} are given in Table 6.6. Besides the Best et al. (1999) prior, two priors inspired by equation (6.26) are also reported; see Carlin and Pérez (2000). The AFF covariate appears significantly different from 0 under all 3 priors, although convergence is *very* slow

	Posterior for α			Posterior for β		
Priors for τ_c, τ_h	mean	sd	l1acf	mean	sd	l1acf
G(1.0, 1.0), G(3.2761, 1.81)	.57	.058	.80	.43	.17	.94
G(.1, .1), G(.32761, .181)	.65	.073	.89	.41	.14	.92
G(.1, .1), G(.001, .001)	.82	.10	.98	.38	.13	.91
	Posterior for ξ_1			Posterior for ξ_{56}		
Priors for τ_c, τ_h	mean	sd	l1acf	mean	sd	l1acf
G(1.0, 1.0), G(3.2761, 1.81)	.92	.40	.33	−.96	.52	.12
G(.1, .1), G(.32761, .181)	.89	.36	.28	−.79	.41	.17
G(.1, .1), G(.001, .001)	.90	.34	.31	−.70	.35	.21

Table 6.6 *Posterior summaries for the spatial model with Gamma hyperpriors for τ_c and τ_h, Scotland lip cancer data; "sd" denotes standard deviation while "l1acf" denotes lag 1 sample autocorrelation.*

(very high values for l1acf). The excess variability in the data seems mostly due to clustering ($E(\alpha|\mathbf{y}) > .50$), but the posterior distribution for α does *not* seem robust to changes in the prior. Finally, convergence for the ξ_i (reasonably well identified) is rapid; convergence for the ψ_i (not shown) is virtually immediate.

Of course, a full analysis of these data would also involve a map of the posterior means of the raw and estimated SMR's, which we can do directly in `GeoBUGS`, the spatial statistics module supplied with `WinBUGS` Versions 1.4 and later. While this spatial module comes with its own database of maps, it is possible to export maps in `R` to `WinBUGS` using the `sp2WB` function. We investigate this in the context of the homework assignment in Exercise 12.

6.4.4 Extending the CAR model

From the above, the reader may glean the idea that it might be easier to just work with the CAR model in (6.22), discarding the heterogeneity component. In a sense, this merely asserts that we are interested in a spatial model, we are interested in implementing spatial smoothing, and since we cannot separate the spatial structure from the heterogeneity, we adopt an appropriate spatial specification. In fact, we see the connection here with the discussion of the redundant pure error term in Section 6.2. With conditionally independent first stage Poisson counts, adding a heterogeneity model is redundant.

Instead, we might extend the CAR model to include two variance components to attempt to capture both heterogeneity and spatial dependence. A version was presented in the work of Leroux and Breslow (1999) and developed further in MacNab and Dean (2000). The key point is that the precision parameter in the intrinsic CAR model represents both overdispersion and spatial association. Letting $Q = D_W - W$, to parallel the architecture in a geostatistical model, we would like a covariance form, $\Sigma = \sigma^2 Q^{-1} + \tau^2 I$. Of course, Q^{-1} doesn't exist. So, instead, define $\Sigma = \omega^2 A^{-1}$ where $A = \lambda Q + (1-\lambda)I$ where $\lambda \in (0,1)$ so that A^{-1} exists and, in the likelihood, $\Sigma^{-1} = \frac{\lambda}{\omega^2} Q + \frac{1-\lambda}{\omega^2} I$, resembling the covariance matrix components associated with (6.22). The key difference is that, now, we introduce just a single set of random effects with covariance matrix Σ rather than two sets of random effects, separating the dependence structure.

For this new multivariate normal prior for, say, the set $\{xi\}$, we can immediately calculate the conditional normal distributions. In fact, $E(\xi_i|\{\xi_j, j \neq i\}) = \frac{\lambda}{1-\lambda+\lambda w_{i+}} \sum_{j\sim i} \xi_j$ and $\mathrm{Var}(\xi_i|\{\xi_j, j \neq i\}) = \frac{\omega^2}{1-\lambda+\lambda w_{i+}}$. When $\lambda = 1$, we have just an intrinsic CAr prior; when $\lambda = 0$, we have an independence model. And, as with the proper CAR, we see that the expected value of ξ_i is only a portion of the average of its neighbors if $\lambda \neq 1$.

6.5 General linear areal data modeling

By analogy with Section 6.2, the areal unit measurements Y_i that we model need not be restricted to counts, as in our disease mapping setting. They may also be binary events (say, presence or absense of a particular facility in region i), or continuous measurements (say, population density, i.e., a region's total population divided by its area).

Again formulating a hierarchical model, Y_i may be described using a suitable first-stage member of the exponential family. Now given $\boldsymbol{\beta}$ and ϕ_i, analogous to (6.18) the Y_i are conditionally independent with density,

$$f(y_i|\boldsymbol{\beta}, \phi_i, \gamma) = h(y_i, \gamma) \exp\{\gamma[y_i\eta_i - \psi(\eta_i)]\} \ , \tag{6.29}$$

where $g(\eta_i) = \mathbf{x}_i^T\boldsymbol{\beta} + \phi_i$ for some link function g with γ a dispersion parameter. The ϕ_i will be spatial random effects coming from a CAR model; the pairwise difference, intrinsic (IAR) form is most commonly used. As a result, we have a generalized linear mixed model with spatial structure in the random effects. Paralleling previous sections, it makes no sense to introduce independent heterogeneity effects, θ_i. Again, we are in a situation where the stochastic mechanism in f replaces these independent errors. Again, we recommend fitting models of the form (6.29) having only a spatial random effect. Computation is more stable and a "balanced" (or "fair") prior specification (as mentioned in connection with (6.26) above) is not an issue.

6.6 Comparison of point-referenced and areal data models

We conclude this chapter with a brief summary and comparison between point-referenced data models and areal unit data models. First, the former are defined with regard to an uncountable number of random variables. The process specification determines the n-dimensional joint distribution for the $Y(\mathbf{s}_i)$, $i = 1, \ldots, n$. For areal units, we envision only a single, finite, n-dimensional distribution for the Y_i, $i = 1, \ldots, n$, which we write down to begin with.

Next, with point-referenced data, we model association directly. For example, if $\mathbf{Y} = (Y(\mathbf{s}_1), \ldots, Y(\mathbf{s}_n))'$ we specify $\Sigma_{\mathbf{Y}}$ using isotropy (or anisotropy), stationarity (or nonstationarity), and so on. With areal data $\mathbf{Y} = (Y_1, \ldots, Y_n)'$ and CAR (or SAR) specifications, we instead model $\Sigma_{\mathbf{Y}}^{-1}$ directly. For instance, with CAR models, Brook's Lemma enables us to reconstruct $\Sigma_{\mathbf{Y}}^{-1}$ from the conditional specifications; $\Sigma_{\mathbf{Y}}^{-1}$ provides conditional association structure (as in Section 4.3) but says nothing about *unconditional* association structure. When $\Sigma_{\mathbf{Y}}^{-1}$ is full rank, the transformation to $\Sigma_{\mathbf{Y}}$ is very complicated, and very nonlinear. Positive conditional association can become negative unconditional association. If the CAR is defined through distance-based w_{ij}'s there need not be any corresponding distance-based order to the unconditional associations. See Besag and Kooperberg (1995), Conlon and Waller (1999), Hrafnkelsson and Cressie (2003), and Wall (2004) for further discussion.

With regard to formal specification, in the most commonly employed point-level Gaussian case, the process is specified through a valid covariance function. With CAR modeling, the specification is instead done through Markov random fields (Section 4.2) employing the Hammersley-Clifford Theorem to ensure a unique joint distribution.

Explanation is a common goal of point-referenced data modeling, but often an even more important goal is spatial prediction or interpolation (i.e., kriging). This may be done at new points, or for block averages (see Section 7.1). With areal units, again a goal is explanation, but now often is supplemented by smoothing. Here the interpolation problem is infrequent and, in any event, is to new areal units, the so-called modifiable areal unit problem (MAUP) as discussed in Sections 7.2 and 7.3.

Finally, with spatial processes, likelihood evaluation requires computation of a quadratic form involving $\Sigma_{\mathbf{Y}}^{-1}$ and the determinant of $\Sigma_{\mathbf{Y}}$. (With spatial random effects, this evaluation is deferred to the second stage of the model, but is still present.) With an increasing

number of locations, such computation becomes very expensive (computing time is greater than order n^2), and may also become unstable, due to the enormous number of floating point operations required. We refer to this situation as a "big N" problem (and devote Chapter 11 to it). On the other hand, with CAR modeling the likelihood (or the second-stage model for the random effects, as the case may be) is written down immediately, since this model parametrizes $\Sigma_{\mathbf{Y}}^{-1}$ (rather than $\Sigma_{\mathbf{Y}}$)). Full conditional distributions needed for MCMC sampling are immediate, and there is no big n problem. Also, for SAR models (though not of much interest in the context of hierarchical modeling because of the fact that the local conditional distributions are not available in simple, closed forms) the quadratic form is directly evaluated, while the determinant is usually evaluated efficiently (even for very large n) using sparse matrix methods; see, e.g., Pace and Barry (1997a,b).

6.7 Exercises

1. Derive the forms of the full conditionals for $\boldsymbol{\beta}, \sigma^2, \tau^2, \phi$, and $\mathbf{W}$ in the exponential kriging model (6.1) and (6.3).
2. Assuming the likelihood in (6.3), suppose that ϕ is fixed and we adopt the prior $p(\boldsymbol{\beta}, \sigma^2, \tau^2) \propto 1/(\sigma^2\tau^2)$, a rather standard noninformative prior often chosen in nonspatial analysis settings. Show that the resulting posterior $p(\boldsymbol{\beta}, \sigma^2, \tau^2|\mathbf{y})$ is *improper*.
3. Derive the form of $p(y_0|\mathbf{y}, \boldsymbol{\theta}, \mathbf{x}_0)$ in (6.5) via the usual conditional normal formulae; i.e., following Guttman (1982, pp. 69-72), if $\mathbf{Y} = (\mathbf{Y}_1^T, \mathbf{Y}_2^T)^T \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ where

$$\boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{pmatrix} \text{ and } \boldsymbol{\Sigma} = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix} ,$$

   then $\mathbf{Y}_2|\mathbf{Y}_1 \sim N(\boldsymbol{\mu}_{2.1}, \boldsymbol{\Sigma}_{2.1})$, where

$$\boldsymbol{\mu}_{2.1} = \boldsymbol{\mu}_2 + \Sigma_{21}\Sigma_{11}^{-1}(\mathbf{Y}_1 - \boldsymbol{\mu}_1) \text{ and } \boldsymbol{\Sigma}_{2.1} = \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12} .$$

4. In Expression (6.18), if $g(\theta) = \theta$ and the prior on $\boldsymbol{\beta}$ is a proper normal distribution,
   (a) Show that the full conditional distributions for the components of $\boldsymbol{\beta}$ are log-concave.
   (b) Show that the full conditional distributions for the $w(\mathbf{s}_i)$ are log-concave.
5. (a) Derive the variance and covariance relationships given in (3.11).
   (b) Derive the variance and covariance relationships given in (3.13).
6. The lithology data set (see `www.biostat.umn.edu/~brad/data2.html`) consists of measurements taken at 118 sample sites in the Radioactive Waste Management Complex region of the Idaho National Engineering and Environmental Laboratory. At each site, bore holes were drilled and measurements taken to determine the elevation and thickness of the various underground layers of soil and basalt. Understanding the spatial distribution of variables like these is critical to predicting fate and transport of groundwater and the (possibly harmful) constituents carried therein; see Leecaster (2002) for full details. For this problem, consider only the variables `Northing, Easting, Surf Elevation, Thickness`, and `A-B Elevation`, and only those records for which full information is available (i.e., extract only those data rows without an "NA" for any variable).
   (a) Produce image plots of the variables `Thickness, Surf Elevation`, and `A-B Elevation`. Add contour lines to each plot and comment on the descriptive topography of the region.
   (b) Taking `log(Thickness)` as the response and `Surf Elevation` and `A-B Elevation` as covariates, fit a univariate Gaussian spatial model with a nugget effect, using the exponential and Matérn covariance functions. You may start with flat priors for the covariate slopes, Inverse Gamma$(0.1, 0.1)$ priors for the spatial and nugget variances,

and a Gamma$(0.1, 0.1)$ prior for the spatial range parameter. Modify the priors and check for their sensitivity to the analysis. (*Hint:* You can use `WinBUGS` to fit the exponential model, but you must use `spBayes` for the Matérn.)

(c) Perform Bayesian kriging on a suitable grid of values and create image plots of the posterior mean residual surfaces for the spatial effects. Overlay the plots with contour lines and comment on the consistency with the plots from the raw data in part (a).

(d) Repeat the above for a purely spatial model (without a nugget) and compare this model with the spatial+nugget model using a model choice criterion (say, DIC).

7. The real estate data set (`www.biostat.umn.edu/~brad/data2.html`) consists of information regarding 70 sales of single-family homes in Baton Rouge, LA, during the month of June 1989. It is customary to model log-selling price.

(a) Obtain the empirical variogram of the raw log-selling prices.

(b) Fit an ordinary least squares regression to log-selling price using living area, age, other area, and number of bathrooms as explanatory variables. Such a model is usually referred to as a *hedonic* model.

(c) Obtain the empirical variogram of the residuals to the least squares fit.

(d) Using an exponential spatial correlation function, attempt to fit the model $Y(\mathbf{s}) = \mathbf{x}^T(\mathbf{s})\boldsymbol{\beta} + W(\mathbf{s}) + \epsilon(\mathbf{s})$ as in Equation (6.1) to the log-selling prices, obtaining estimates using `geoR`.

(e) Predict the actual selling price for a home at location (longitude, latitude) = (–91.1174, 30.506) that has characteristics LivingArea = 938 sqft, OtherArea = 332sqft, Age = 25yrs, Bedrooms = 3, Baths = 1, and HalfBaths = 0. (*Reasonability check:* The actual log-selling price for this location turned out to be 10.448.)

(f) Use `geoR` (grid-based integration routines) or `WinBUGS` (MCMC) to fit the above model in a Bayesian framework. Begin with the following prior specification: a flat prior for $\boldsymbol{\beta}$, IG(0.1,0.1) (`WinBUGS` parametrization) priors for $1/\sigma^2$ and $1/\tau^2$, and a Uniform(0,10) prior for ϕ. Also investigate prior robustness by experimenting with other choices.

(g) Obtain samples from the predictive distribution for log-selling price and selling price for the particular location mentioned above. Summarize this predictive distribution.

(h) Compare the classical and Bayesian inferences.

(i) *(advanced):* Hold out the first 20 observations in the data file, and fit the nonspatial (i.e., without the $W(\mathbf{s})$ term) and spatial models to the observations that remain. For both models, compute

- $\sum_{j=1}^{20}(Y(\mathbf{s}_{0j}) - \hat{Y}(\mathbf{s}_{0j}))^2$
- $\sum_{j=1}^{20} Var(Y(\mathbf{s}_{0j})|\mathbf{y})$
- the proportion of predictive intervals for $Y(\mathbf{s}_0)$ that are correct
- the proportion of predictions that are within 10% of the true value
- the proportion of predictions that are within 20% of the true value

Discuss the differences in predictive performance.

8. Suppose $Z_i = Y_i/n_i$ is the observed disease *rate* in each county, and we adopt the model $Z_i \overset{ind}{\sim} N(\eta_i, \sigma^2)$ and $\eta_i \overset{iid}{\sim} N(\mu, \tau^2)$, $i = 1, \ldots, I$. Find $E(\eta_i|y_i)$, and express it as a weighted average of Z_i and μ. Interpret your result as the weights vary.

9. In fitting model (6.22) with priors for the θ_i and ϕ_i given in (6.23) and (6.24), suppose we adopt the hyperpriors $\tau_h \sim G(a_h, b_h)$ and $\tau_c \sim G(a_c, b_c)$. Find closed form expressions for the full conditional distributions for these two parameters.

10. The full conditional (6.25) does *not* emerge in closed form, since the CAR (normal) prior is not conjugate with the Poisson likelihood. However, prove that this full conditional *is* log-concave, meaning that the necessary samples can be generated using the adaptive rejection sampling (ARS) algorithm of Gilks and Wild (1992).
11. Confirm algebraically that, taken together, the expressions

$$\phi_i|\boldsymbol{\phi}_{j\neq i} \sim N(\phi_i \mid \bar{\phi}_i \ , \ 1/(\tau_c m_i)) \ , \ i = 1, \ldots, I$$

are equivalent to the (improper) joint specification

$$p(\phi_1, \ldots, \phi_I) \propto \exp\left\{ -\frac{\tau_c}{2} \sum_{i \ adj \ j} (\phi_i - \phi_j)^2 \right\} \ ,$$

i.e., the version of (4.16) corresponding to the usual, adjacency-based CAR model (6.24).

12. The Minnesota Department of Health is charged with investigating the possibility of geographical clustering of the rates for the late detection of colorectal cancer in the state's 87 counties. For each county, the late detection rate is simply the number of regional or distant case detections divided by the total cases observed in that county. Information on several potentially helpful covariates is also available. The most helpful is the county-level estimated proportions of persons who have been screened for colorectal cancer, as estimated from telephone interviews available biannually between 1993 and 1999 as part of the nationwide Behavioral Risk Factor Surveillance System (BRFSS).

(a) Use the `poly.R` function in `R` available at `www.biostat.umn.edu/~brad/data2.html` to obtain a boundary file for the counties of Minnesota by executing

```
> source(``poly.R'')
> mkpoly(``minnesota'')
```

The result should be a file called `minnesota.txt`. Now open this file in `WinBUGS`, and pull down to `Import Splus` on the `Map` menu. Kill and restart `WinBUGS`, and then pull down to `Adjacency Tool` again from the `Map` menu; "minnesota" should now be one of the options! Click on `adj map` to see the adjacency map, `adj matrix` to print out the adjacency matrix, and `show region` to find any given region of interest.

(b) It is also possible to export a map from `R` to `WinBUGS` using the `sp2WB()` function in the `spdep` package. Create a spatial polygon object (e.g., `mn.poly` in Section 4.5.1) and pass it to `sp2WB()` as

```
> sp2WB(mn.poly, "mn_bugs.txt", Xscale = 1, plotorder = FALSE)
```

The file `"mn_bugs.txt"` can now be imported into `WinBUGS` as described in the part (a)

(c) Use `WinBUGS` to model the log-relative risk using (6.27), fitting the heterogeneity plus clustering (CAR) model to these data. You will find the observed late detections Y_i, expected late detections E_i and screening covariates x_i in `WinBUGS` format in the file `colorecbugs.dat` on the webpage `www.biostat.umn.edu/~brad/data2.html/`. Find and summarize the posterior distributions of α (the proportion of excess variability due to clustering) and β_1 (the screening effect). Does MCMC convergence appear adequate in this problem?

(d) To use `GeoBUGS` to map the raw and fitted SMR's, pull down to `Mapping Tool` on the `Map` menu. Customize your maps by playing with cuts and colors. (Remember you will have to have saved those Gibbs samples during your `WinBUGS` run in order to summarize them!)

(e) Since the screening covariate was estimated from the BRFSS survey, we should really account for survey measurement error, since this may be substantial for rural counties

having few respondents. Replace the observed covariate x_i in the log-relative risk model (6.27) by T_i, the true (unobserved) rate of colorectal screening in county i. Following Xia and Carlin (1998), we further augment our hierarchical model with

$$T_i \overset{iid}{\sim} N(\mu_0, 1/\lambda) \qquad \text{and} \qquad x_i|T_i \overset{ind}{\sim} N(T_i, 1/\delta) \ ,$$

That is, x_i is acknowledged as an imperfect (albeit unbiased) measure of the true screening rate T_i. Treat the measurement error precision λ and prior precision δ either as known "tuning parameters," or else assign them gamma hyperprior distributions, and recompute the posterior for β_1. Observe and interpret any changes.

(f) A more realistic errors-in-covariates model might assume that the precision of x_i given T_i should be proportional to the survey sample size r_i in each county. Write down (but do not fit) a hierarchical model that would address this problem.

Chapter 7

Spatial misalignment

In this chapter we tackle the problem of spatial misalignment. That is, with the explosion in spatial data collection, it is more and more the case that different spatial data layers are collected at different scales. For example, we may have one layer at point level, another at point level but at different locations, yet another for one set of areal units and a last over a different set of areal units. Standard GIS software can routinely create overlays and themes with such layers but this is primarily descriptive. Here we seek a formal inferential framework to deal with such misalignment. As a canonical example, consider an environmental justice setting where we seek to assess whether one group is adversely affected by, say, an environmental contaminant, compared with another. So, we might record exposure levels at monitoring stations, we might collect adverse health outcomes at the scale of zip or post codes, and we might learn about population groups at risk through census data at census tract scale. How might we assemble these layers to assess inequities?

As a result, two types of problems arise with such data settings. For the first, we seek an analysis of the spatial data at a different scale of spatial resolution than they were originally collected. For the second, we would be interested in a regression setting where we seek to use some set of spatially referenced variables to explain another where the variables have been collected at different spatial scales. For example, we might wish to obtain the spatial distribution of some variable at the county level, even though it was originally collected at the census tract level. We might have a very low-resolution global climate model for weather prediction, and seek a regional model or even more locally (i.e., at higher resolution). For areal unit data, our purpose might be simply to understand the distribution of a variable at a new level of spatial aggregation (the so-called *modifiable areal unit problem*, or MAUP). For data modeled at point level through a spatial process we would envision block averaging at different spatial scales (the so-called *change of support problem*, or COSP), again possibly for connection with another variable observed at a particular scale. For either type of data, our goal in the first case is typically one of spatial *interpolation*, while in the second it is one of spatial *regression*.

In addition to our presentation here, we also encourage the reader to look at the excellent review paper by Gotway and Young (2002). These authors give good discussions of (as well as both traditional and Bayesian approaches for) the MAUP and COSP, spatial regression, and the *ecological fallacy*. This last term refers to the fact that relationships observed between variables measured at the ecological (aggregate) level may not accurately reflect (and will often overstate) the relationship between these same variables measured at the individual level. Discussion of this problem dates at least to Robinson (1950); see Wakefield (2001, 2003, 2004) for more modern treatments of this difficult subject. As a simple graphical illustration, Figure 7.1 shows the incidence of low birth weight at three spatial scales, county, zipcode, and census tract, for the State of North Carolina. Visually, the nature of the spatial structure changes substantially according to scale.

As in previous sections, we group our discussion according to whether the data is suitably modeled using a spatial process as opposed to a CAR or SAR model. Here the former

Figure 7.1 *Incidence of low birth weight at three spatial scales, county, zipcode, and census tract, for the state of North Carolina.*

assumption leads to more general modeling, since point-level data may be naturally aggregated to block level, but the reverse procedure may or may not be possible; e.g., if the areal data are counts or proportions, what would the point-level variables be? However, since block-level summary data are quite frequent in practice (often due to confidentiality restrictions), methods associated with such data are also of great importance. We thus consider point-level and block-level modeling.

7.1 Point-level modeling

7.1.1 Gaussian process models

Consider a univariate variable that is modeled through a spatial process. In particular, assume that it is observed either at points in space, or over areal units (e.g., counties or zip codes), which we will refer to as *block* data. The *change of support problem* is concerned with inference about the values of the variable at points or blocks different from those at which it has been observed.

7.1.1.1 Motivating data set

A solution to the change of support problem is required in many health science applications, particularly spatial and environmental epidemiology. To illustrate, consider again the data set of ozone levels in the Atlanta, GA metropolitan area, originally reported by Tolbert et al. (2000). Ozone measures are available at between 8 and 10 fixed monitoring sites during the 92 summer days (June 1 through August 31) of 1995. Similar to Figure 1.3 (which shows 8-hour maximum ozone levels), Figure 7.2 shows the 1-hour daily maximum ozone measures at the 10 monitoring sites on July 15, 1995, along with the boundaries of the 162 zip codes in the Atlanta metropolitan area. Here we might be interested in predicting the ozone level at different points on the map (say, the two points marked **A** and **B**, which lie on opposite sides of a single city zip), or the average ozone level over a particular zip (say, one of the 36 zips falling within the city of Atlanta, the collection of which are encircled by

Figure 7.2 *Zip code boundaries in the Atlanta metropolitan area and 1-hour daily maximum ozone levels at the 10 monitoring sites for July 15, 1995.*

the dark boundary on the map). The latter problem is of special interest, since, in this case, relevant health outcome data are available only at the zip level. In particular, for each day and zip, we have the number of pediatric ER visits for asthma, as well as the total number of pediatric ER visits. Thus an investigation of the relationship between ozone exposure and pediatric asthma cannot be undertaken until the mismatch in the support of the two variables is resolved. Situations like this are relatively common in health outcome settings, since personal privacy concerns often limit statisticians' access to data other than at the areal or block level.

In many earth science and population ecology contexts, presence/absence is typically recorded at essentially point-referenced sites while relevant climate layers are often downscaled to grid cells at some resolution. A previous study of the Atlanta ozone data by Carlin et al. (1999) realigned the point-level ozone measures to the zip level by using an `ARC/INFO` universal kriging procedure to fit a smooth ozone exposure surface, and subsequently took the kriged value at each zip centroid as the ozone value for that zip. But this approach uses a single centroid value to represent the ozone level in the entire zip, and fails to properly capture variability and spatial association by treating these kriged estimates as observed values.

7.1.1.2 Model assumptions and analytic goals

Let $Y(\mathbf{s})$ denote the spatial process (e.g., ozone level) measured at location $\mathbf{s}$, for $\mathbf{s}$ in some region of interest D. In our applications $D \subset \Re^2$ but our development works in arbitrary dimensions. A realization of the process is a surface over D. For point-referenced data the realization is observed at a finite set of sites, say, $\mathbf{s}_i, i = 1, 2, \ldots, I$. For block data we assume the observations arise as block averages. That is, for a block $B \subset D$,

$$Y(B) = |B|^{-1} \int_B Y(\mathbf{s}) d\mathbf{s}\,, \tag{7.1}$$

where $|B|$ denotes the area of B (see, e.g., Cressie, 1993). The integration in (7.1) is not the usual integration of a function. Rather, it is an average of random variables, a realization of a stochastic process, hence a random or stochastic integral. Thus, the assumption of an underlying spatial process is only appropriate for block data that can be sensibly viewed as an averaging over point data; examples of this would include rainfall, pollutant level, temperature, and elevation. It would be inappropriate for, say, population, since there is no "population" at a particular point. It would also be inappropriate for most proportions. For instance, if $Y(B)$ is the proportion of college-educated persons in B, then $Y(B)$ is continuous but even were we to conceptualize an individual at every point, $Y(\mathbf{s})$ would be binary.

In general, we envision four possibilities. First, starting with point data $Y(\mathbf{s}_1), \ldots, Y(\mathbf{s}_I)$, we seek to predict at new locations, i.e., to infer about $Y(\mathbf{s}'_1), \ldots, Y(\mathbf{s}'_K)$ (points to points). Second, starting with point data, we seek to predict at blocks, i.e., to infer about $Y(B_1), \ldots, Y(B_K)$ (points to blocks). Third, starting with block data $Y(B_1), ..., Y(B_I)$, we seek to predict at a set of locations, i.e., to infer about $Y(\mathbf{s}'_1), \ldots, Y(\mathbf{s}'_K)$ (blocks to points). Finally, starting with block data, we seek to predict at new blocks, i.e., to infer about $Y(B'_1), \ldots, Y(B'_K)$ (blocks to blocks).

All of these predictions may be collected under the umbrella of kriging, as in Sections 2.4 and 6.1. Our kriging here will be implemented within the Bayesian framework enabling full inference (a posterior predictive distribution for every prediction of interest, joint distributions for all pairs of predictions, etc.) and avoiding asymptotics. We will however use rather noninformative priors, so that our results will roughly resemble those of a likelihood analysis.

Inference about blocks through averages as in (7.1) is not only formally attractive but demonstrably preferable to *ad hoc* approaches. One such approach would be to average over the observed $Y(\mathbf{s}_i)$ in B. But this presumes there is at least one observation in any B, and ignores the information about the spatial process in the observations outside of B. Another *ad hoc* approach would be to simply predict the value at some central point of B. But this value has larger variability than (and may be biased for) the block average.

In the next section, we develop the methodology for spatial data at a single time point; the general spatiotemporal case is similar and described in Section 11.2. Example 7.1 then applies our approaches to the Atlanta ozone data pictured in Figure 7.2.

7.1.2 Methodology for the point-level realignment

We start with a stationary Gaussian process specification for $Y(\mathbf{s})$ having mean function $\mu(\mathbf{s};\boldsymbol{\beta})$ and covariance function $C(\mathbf{s}-\mathbf{s}';\boldsymbol{\theta}) = \sigma^2\rho(\mathbf{s}-\mathbf{s}';\boldsymbol{\phi})$, so that $\boldsymbol{\theta} = (\sigma^2, \boldsymbol{\phi})^T$. Here μ is a trend surface with coefficient vector $\boldsymbol{\beta}$, while σ^2 is the process variance and $\boldsymbol{\phi}$ denotes the parameters associated with the stationary correlation function ρ. Beginning with point data observed at sites $\mathbf{s}_1, ..., \mathbf{s}_I$, let $\mathbf{Y}_s^T = (Y(\mathbf{s}_1), \ldots, Y(\mathbf{s}_I))$. Then

$$\mathbf{Y}_s \mid \boldsymbol{\beta}, \boldsymbol{\theta} \sim N(\boldsymbol{\mu}_s(\boldsymbol{\beta}), \sigma^2 H_s(\boldsymbol{\phi})) \ , \tag{7.2}$$

where $\boldsymbol{\mu}_s(\boldsymbol{\beta})_i = \mu(\mathbf{s}_i;\boldsymbol{\beta})$ and $(H_s(\boldsymbol{\phi}))_{ii'} = \rho(\mathbf{s}_i - \mathbf{s}_{i'};\boldsymbol{\phi})$.

Given a prior on $\boldsymbol{\beta}, \sigma^2$, and $\boldsymbol{\phi}$, models such as (7.2) are straightforwardly fit using simulation methods as described in Section 6.1, yielding posterior samples $(\boldsymbol{\beta}^*_g, \boldsymbol{\theta}^*_g)$, $g = 1, \ldots, G$ from $f(\boldsymbol{\beta}, \boldsymbol{\theta} \mid \mathbf{Y}_s)$.

Then prediction at a set of new locations $\mathbf{Y}_{s'}^T = (Y(\mathbf{s}'_1), ..., Y(\mathbf{s}'_K))$ is really just multiple kriging; we require only the predictive distribution,

$$f(\mathbf{Y}_{s'} \mid \mathbf{Y}_s) = \int f(\mathbf{Y}_{s'} \mid \mathbf{Y}_s, \boldsymbol{\beta}, \boldsymbol{\theta}) f(\boldsymbol{\beta}, \boldsymbol{\theta} \mid \mathbf{Y}_s) d\boldsymbol{\beta} d\boldsymbol{\theta} \ . \tag{7.3}$$

By drawing $\mathbf{Y}^*_{s',g} \sim f(\mathbf{Y}_{s'} \mid \mathbf{Y}_s, \boldsymbol{\beta}^*_g, \boldsymbol{\theta}^*_g)$ we obtain a sample from (7.3) via composition which provides any desired inference about $\mathbf{Y}_{s'}$ and its components.

Under a Gaussian process,

$$f\left(\left(\begin{array}{c}\mathbf{Y}_s\\ \mathbf{Y}_{s'}\end{array}\right)\middle|\boldsymbol{\beta},\boldsymbol{\theta}\right) = N\left(\left(\begin{array}{c}\boldsymbol{\mu}_s(\boldsymbol{\beta})\\ \boldsymbol{\mu}_{s'}(\boldsymbol{\beta})\end{array}\right),\ \sigma^2\left(\begin{array}{cc}H_s(\boldsymbol{\phi}) & H_{s,s'}(\boldsymbol{\phi})\\ H^T_{s,s'}(\boldsymbol{\phi}) & H_{s'}(\boldsymbol{\phi})\end{array}\right)\right), \tag{7.4}$$

with entries defined as in (7.2). Hence, $\mathbf{Y}_{s'}|\mathbf{Y}_s, \boldsymbol{\beta}, \boldsymbol{\theta}$ is distributed as

$$\begin{aligned} N\big(&\boldsymbol{\mu}_{s'}(\boldsymbol{\beta}) + H^T_{s,s'}(\boldsymbol{\phi})H^{-1}_s(\boldsymbol{\phi})(\mathbf{Y}_s - \mu_s(\boldsymbol{\beta})),\\ &\sigma^2[H_{s'}(\boldsymbol{\phi}) - H^T_{s,s'}(\boldsymbol{\phi})H^{-1}_s(\boldsymbol{\phi})H_{s,s'}(\boldsymbol{\phi})]\big). \end{aligned} \tag{7.5}$$

Sampling from (7.5) requires the inversion of $H_s(\boldsymbol{\phi}^*_g)$, which will already have been done in sampling $\boldsymbol{\phi}^*_g$, and then the square root of the $K \times K$ covariance matrix in (7.5).

For a single set B, it is straightforward to show (see Exercise 6.2) that $Y(B)\ simN(\mu_B(\boldsymbol{\beta}), \sigma^2_B(\boldsymbol{\theta}))$ where $\mu_B(\boldsymbol{\beta}) = \frac{1}{|B|}\int_B \mu(\mathbf{s}, \boldsymbol{\beta})d\mathbf{s}$ and $\sigma^2_B(\boldsymbol{\theta}) = \sigma^2 \frac{1}{|B|^2}\int_B\int_B \rho(\mathbf{s}-\mathbf{s}'; \boldsymbol{\phi})d\mathbf{s}d\mathbf{s}'$. Since $\rho < 1$, we immediately see that $\mathrm{var}(Y(B)) < \mathrm{var}(Y(\mathbf{s}))$ for any $\mathbf{s} \in B$ (and, in fact, this holds for the average of any finite set of points in B).

Now, turning to prediction for $\mathbf{Y}^T_B = (Y(B_1), ..., Y(B_K))$, the vector of averages over blocks $B_1, ..., B_K$, we again require the predictive distribution, which is now

$$f(\mathbf{Y}_B \mid \mathbf{Y}_s) = \int f(\mathbf{Y}_B \mid \mathbf{Y}_s; \boldsymbol{\beta}, \boldsymbol{\theta}) f(\boldsymbol{\beta}, \boldsymbol{\theta} \mid \mathbf{Y}_s) d\boldsymbol{\beta} d\boldsymbol{\theta}\ . \tag{7.6}$$

Under a Gaussian process, we now have

$$f\left(\left(\begin{array}{c}\mathbf{Y}_s\\ \mathbf{Y}_B\end{array}\right)\middle|\boldsymbol{\beta},\boldsymbol{\theta}\right) = N\left(\left(\begin{array}{c}\boldsymbol{\mu}_s(\boldsymbol{\beta})\\ \boldsymbol{\mu}_B(\boldsymbol{\beta})\end{array}\right),\ \sigma^2\left(\begin{array}{cc}H_s(\boldsymbol{\phi}) & H_{s,B}(\boldsymbol{\phi})\\ H^T_{s,B}(\boldsymbol{\phi}) & H_B(\boldsymbol{\phi})\end{array}\right)\right), \tag{7.7}$$

where

$$\begin{aligned} (\boldsymbol{\mu}_B(\boldsymbol{\beta}))_k &= E(Y(B_k) \mid \boldsymbol{\beta}) = |B_k|^{-1}\int_{B_k} \mu(\mathbf{s}; \boldsymbol{\beta})d\mathbf{s}\ ,\\ (H_B(\boldsymbol{\phi}))_{kk'} &= |B_k|^{-1}|B_{k'}|^{-1}\int_{B_k}\int_{B_{k'}} \rho(\mathbf{s}-\mathbf{s}'; \boldsymbol{\phi})d\mathbf{s}'d\mathbf{s}\ ,\\ \text{and}\quad (H_{s,B}(\boldsymbol{\phi}))_{ik} &= |B_k|^{-1}\int_{B_k} \rho(\mathbf{s}_i - \mathbf{s}'; \boldsymbol{\phi})d\mathbf{s}'\ . \end{aligned}$$

Analogously to (7.5), $\mathbf{Y}_B|\mathbf{Y}_s, \boldsymbol{\beta}, \boldsymbol{\theta}$ is distributed as

$$\begin{aligned} N\big(&\boldsymbol{\mu}_B(\boldsymbol{\beta}) + H^T_{s,B}(\boldsymbol{\phi})H^{-1}_s(\boldsymbol{\phi})(\mathbf{Y}_s - \boldsymbol{\mu}_s(\boldsymbol{\beta})),\\ &\sigma^2\left[H_B(\boldsymbol{\phi}) - H^T_{s,B}(\boldsymbol{\phi})H^{-1}_s(\boldsymbol{\phi})H_{s,B}(\boldsymbol{\phi})\right]\big)\ . \end{aligned} \tag{7.8}$$

The major difference between (7.5) and (7.8) is that in (7.5), given $(\boldsymbol{\beta}^*_g, \boldsymbol{\theta}^*_g)$, numerical values for all of the entries in $\boldsymbol{\mu}_{s'}(\boldsymbol{\beta})$, $H_{s'}(\boldsymbol{\phi})$, and $H_{s,s'}(\boldsymbol{\phi})$ are immediately obtained. In (7.8) every analogous entry requires an integration as above. Anticipating irregularly shaped B_k's, Riemann approximation to integrate over these regions may be awkward. Instead, noting that each such integration is an expectation with respect to a uniform distribution, we propose Monte Carlo integration. In particular, for each B_k we propose to draw a set of locations $\mathbf{s}_{k,\ell}$, $\ell = 1, 2, ..., L_k$, distributed independently and uniformly over B_k. Here L_k can vary with k to allow for very unequal $|B_k|$.

Hence, we replace $(\boldsymbol{\mu}_B(\boldsymbol{\beta}))_k$, $(H_B(\boldsymbol{\phi}))_{kk'}$, and $(H_{s,B}(\boldsymbol{\phi}))_{ik}$ with

$$(\widehat{\boldsymbol{\mu}}_B(\boldsymbol{\beta}))_k = L^{-1}_k \sum_\ell \mu(\mathbf{s}_{k,\ell}; \boldsymbol{\beta})\ ,$$

$$(\widehat{H}_B(\boldsymbol{\phi}))_{kk'} = L_k^{-1} L_{k'}^{-1} \sum_{\ell} \sum_{\ell'} \rho(\mathbf{s}_{k\ell} - \mathbf{s}_{k'\ell'}; \boldsymbol{\phi}) \ , \tag{7.9}$$

$$\text{and} \ \ (\widehat{H}_{s,B}(\boldsymbol{\phi}))_{ik} = L_k^{-1} \sum_{\ell} \rho(\mathbf{s}_i - \mathbf{s}_{k\ell}; \boldsymbol{\phi}) \ .$$

In our notation, the "hat" denotes a Monte Carlo integration that can be made arbitrarily accurate and has nothing to do with the data $\mathbf{Y}_s$. Note also that the same set of $\mathbf{s}_{k\ell}$'s can be used for each integration and with each $(\boldsymbol{\beta}_g^*, \boldsymbol{\theta}_g^*)$; we need only obtain this set once. In obvious notation we replace (7.7) with the $(I+K)$-dimensional multivariate normal distribution $\widehat{f}\left((\mathbf{Y}_s, \mathbf{Y}_B)^T \,\middle|\, \boldsymbol{\beta}, \boldsymbol{\theta}\right)$ with entries using (7.9).

It is important to note that we can *observe* $Y(B)$ but we cannot *sample* $Y(B)$. However, if we define $\widehat{Y}(B_k) = L_k^{-1} \sum_\ell Y(\mathbf{s}_{k\ell})$, then $\widehat{Y}(B_k)$ is a Monte Carlo integration for $Y(B_k)$ as given in (7.1). With an obvious definition for $\widehat{\mathbf{Y}}_B$, it is apparent that

$$\widehat{f}\left((\mathbf{Y}_s, \mathbf{Y}_B)^T \mid \boldsymbol{\beta}, \boldsymbol{\theta}\right) = f\left((\mathbf{Y}_s, \widehat{\mathbf{Y}}_B)^T \mid \boldsymbol{\beta}, \boldsymbol{\theta}\right) \tag{7.10}$$

where (7.10) is interpreted to mean that the approximate joint distribution of $(\mathbf{Y}_s, \mathbf{Y}_B)$ is the exact joint distribution of $\mathbf{Y}_s, \widehat{\mathbf{Y}}_B$. In practice, we will work with $\widehat{f}$, converting to $\widehat{f}(\mathbf{Y}_B \mid \mathbf{Y}_s, \boldsymbol{\beta}, \boldsymbol{\theta})$ to sample $\mathbf{Y}_B$ rather than sampling the $\widehat{Y}(B_k)$'s through the $Y(\mathbf{s}_{k\ell})$'s. But, evidently, we are sampling $\widehat{\mathbf{Y}}_B$ rather than $\mathbf{Y}_B$.

As a technical point, we might ask when $\widehat{\mathbf{Y}}_B \xrightarrow{P} \mathbf{Y}_B$. An obvious sufficient condition is that realizations of the $Y(\mathbf{s})$ process are almost surely continuous. In the stationary case, Kent (1989) provides sufficient conditions on $C(\mathbf{s}-\mathbf{t}; \boldsymbol{\theta})$ to ensure this. Alternatively, Stein (1999a) defines $Y(\mathbf{s})$ to be *mean square continuous* if $\lim_{\mathbf{h}\to 0} E(Y(\mathbf{s}+\mathbf{h}) - Y(\mathbf{s}))^2 = 0$ for all $\mathbf{s}$. But then $Y(\mathbf{s}+\mathbf{h}) \xrightarrow{P} Y(\mathbf{s})$ as $\mathbf{h} \to 0$, which is sufficient to guarantee that $\widehat{\mathbf{Y}}_B \xrightarrow{P} \mathbf{Y}_B$. Stein notes that if $Y(\mathbf{s})$ is stationary, we only require $c(\cdot; \boldsymbol{\theta})$ continuous at $\mathbf{0}$ for mean square continuity. (See Subsection 3.1.4 and Section 13.2 for further discussion of smoothness of process realizations.)

This leads us to an opportunity to make an important distinction between estimation and prediction in the spatial context. Suppose we assume a constant mean surface over the bounded set B, say, at height μ. Then we could consider $\hat{Y}(B)$ as an *estimator* of μ as well as a *predictor* of $Y(B)$. We note that $\hat{Y}(B) \to Y(B)$ in probability and, in fact, if the process covariance function is continuous at 0, we have mean square convergence, as we argued above. However, $\hat{Y}(B)$ need not converge in probability to μ. It suffices to look at $\text{var}(\hat{Y}(B))$ which, for a sample of size L, from (7.6), is $\sigma^2 \frac{1}{L^2} \sum_\ell \sum_{\ell'} \rho(\mathbf{s}_\ell - \mathbf{s}_{\ell'})$. In this double sum, the "diagonal" terms contribute σ^2/L and so will tend to 0 as $\ell \to \infty$. However, the remaining $L(L-1)$ terms in the double sum, when divided by L^2 need not tend to 0. For instance, with an isotropic covariance function that reaches 0 only when distance goes to ∞, we will have every entry in the remaining double sum bigger than $\rho(d_{max})$, where d_{max} is the largest pairwise distance between the locations. Therefore, the remaining sum is greater than $\frac{L(L-1)}{L^2}\rho(d_{max})$. And, as $L \to \infty$, this term will not tend to 0 since d_{max} is bounded because B is bounded. The result holds more generally, even with covariance functions having bounded support since, as $L \to \infty$, the proportion of points that will stay within a subset such that the maximum distance over the subset will be less than the upper support bound for the covariance function will also tend to ∞. In summary, we have consistent prediction but need not have consistency for estimation.

A related block averaging is associated with a binary process. Binary processes are routinely created from continuous processes using indicator functions. In any event, suppose $Z(\mathbf{s})$ is either 1 or 0. For instance, with presence/absence data, we observe an indicator of whether a particular species is present at location $\mathbf{s}$. Now, let $Z(B) = \frac{1}{|B|} \int_B Z(\mathbf{s}) d\mathbf{s}$. How

shall we interpret $Z(B)$? If we think of $Z(\mathbf{s})$ as a light bulb that is either on (1) or off (0) at each location, then $Z(B)$ is the proportion of light bulbs that are on in areal unit B. Evidently, while every $Z(\mathbf{s})$ is either 1 or 0, $Z(B)$ will be in $(0,1)$. This reminds us of the foregoing issue of scaling from blocks to points. However, suppose we define the binary variable $U(B)$ as the result of choosing a location at random in B and taking the value of the binary variable at that location. Then, we see that $P(U(B) = 1) = E(Z(B))$. These issues, in the context of modeling presence/absence data are considered in detail in Gelfand et al. (2005).

Returning to the prediction problem, starting with block data $\mathbf{Y}_B^T = (Y(B_1), \ldots, Y(B_I))$, analogous to (7.2) the likelihood is well defined as

$$f(\mathbf{Y}_B \mid \boldsymbol{\beta}, \boldsymbol{\theta}) = N(\boldsymbol{\mu}_B(\boldsymbol{\beta}) \, , \, \sigma^2 H_B(\boldsymbol{\phi})). \tag{7.11}$$

Hence, given a prior on $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$, the Bayesian model is completely specified. As above, evaluation of the likelihood requires integrations. So, we replace (7.11) with

$$\widehat{f}(\mathbf{Y}_B \mid \boldsymbol{\beta}, \boldsymbol{\theta}) = N(\widehat{\boldsymbol{\mu}}_B(\boldsymbol{\beta}) \, , \, \sigma^2 \widehat{H}_B(\boldsymbol{\phi})). \tag{7.12}$$

Simulation-based fitting is now straightforward, as below (7.2), albeit somewhat more time consuming due to the need to calculate $\widehat{\boldsymbol{\mu}}_B(\boldsymbol{\beta})$ and $\widehat{H}_B(\boldsymbol{\phi})$.

To predict for $\mathbf{Y}_{s'}$ we require $f(\mathbf{Y}_{s'} \mid \mathbf{Y}_B)$. As above, we only require $f(\mathbf{Y}_B, \mathbf{Y}_{s'} \mid \boldsymbol{\beta}, \boldsymbol{\theta})$, which has been given in (7.7). Using (7.10) we now obtain $\widehat{f}(\mathbf{Y}_{s'} \mid \mathbf{Y}_B, \boldsymbol{\beta}, \boldsymbol{\theta})$ to sample $\mathbf{Y}_{s'}$. Note that $\widehat{f}$ is used in (7.12) to obtain the posterior samples and again to obtain the predictive samples. Equivalently, the foregoing discussion shows that we can replace $\mathbf{Y}_B$ with $\widehat{\mathbf{Y}}_B$ throughout. To predict for new blocks $B'_1, ..., B'_K$, let $\mathbf{Y}_{B'}^T = (Y(B'_1), ..., Y(B'_K))$. Now we require $f(\mathbf{Y}_{B'} \mid \mathbf{Y}_B)$, which in turn requires $f\,(\mathbf{Y}_B, \mathbf{Y}_{B'} \mid \boldsymbol{\beta}, \boldsymbol{\theta})$. The approximate distribution $\widehat{f}\,(\mathbf{Y}_B, \mathbf{Y}_{B'} | \boldsymbol{\beta}, \boldsymbol{\theta})$ employs Monte Carlo integrations over the B'_k's as well as the B_i's, and yields $\widehat{f}(\mathbf{Y}_{B'} \mid \mathbf{Y}_B, \boldsymbol{\beta}, \boldsymbol{\theta})$ to sample $\mathbf{Y}_{B'}$. Again $\widehat{f}$ is used to obtain both the posterior and predictive samples.

Note that in all four prediction cases, we can confine ourselves to an $(I+K)$-dimensional multivariate normal. Moreover, we have only an $I \times I$ matrix to invert repeatedly in the model fitting, and a $K \times K$ matrix whose square root is required for the predictive sampling.

For the modifiable areal unit problem (i.e., prediction at new blocks using data for a given set of blocks), suppose we take as our point estimate for a generic new set B_0 the posterior mean,

$$E(Y(B_0) \mid \mathbf{Y}_B) = E\{\mu(B_0; \boldsymbol{\beta}) + \mathbf{H}_{B,B_0}^T(\boldsymbol{\phi}) H_B^{-1}(\boldsymbol{\phi})(\mathbf{Y}_B - \boldsymbol{\mu}_B(\boldsymbol{\beta})) \mid \mathbf{Y}_B\} \, ,$$

where $\mathbf{H}_{B,B_0}(\boldsymbol{\phi})$ is $I \times 1$ with i^{th} entry equal to $cov(Y(B_i), Y(B_0) \mid \boldsymbol{\theta})/\sigma^2$. If $\mu(\mathbf{s}; \boldsymbol{\beta}) \equiv \mu_i$ for $\mathbf{s} \in B_i$, then $\mu(B_0; \boldsymbol{\beta}) = |B_0|^{-1} \sum_i |B_i \cap B_0| \, \mu_i$. But $E(\mu_i \mid \mathbf{Y}_B) \approx Y(B_i)$ to a first-order approximation, so in this case $E(Y(B_0) \mid \mathbf{Y}_B) \approx |B_0|^{-1} \sum_i |B_i \cap B_0| \, Y(B_i)$, the areally weighted estimate.

Example 7.1 We now use the foregoing approach to perform point-point and point-block inference for the Atlanta ozone data pictured in Figure 7.2. Recall that the target points are those marked **A** and **B** on the map, while the target blocks are the 36 Atlanta city zips. The differing block sizes suggest use of a different L_k for each k in equation (7.9). Conveniently, our GIS (`ARC/INFO`) can generate random points over the whole study area, and then allocate them to each zip. Thus L_k is proportional to the area of the zip, $|B_k|$. Illustratively, our procedure produced 3743 randomly chosen locations distributed over the 36 city zips, an average L_k of nearly 104.

Suppose that log-ozone exposure $Y(\mathbf{s})$ follows a second-order stationary spatial Gaussian process, using the exponential covariance function $c(\mathbf{s}_i - \mathbf{s}_{i'}; \boldsymbol{\theta}) = \sigma^2 e^{-\phi \|\mathbf{s}_i - \mathbf{s}_{i'}\|}$. A

Figure 7.3 *Posterior point-point and point-block summaries, static spatial model, Atlanta ozone data for July 15, 1995.*

preliminary exploratory analysis of our data set suggested that a constant mean function $\mu(\mathbf{s}_i; \boldsymbol{\beta}) = \mu$ is adequate for our data set. We place the customary flat prior on μ, and assume that $\sigma^2 \sim IG(a, b)$ and $\phi \sim G(c, d)$. We chose $a = 3$, $b = 0.5$, $c = 0.03$, and $d = 100$, corresponding to fairly vague priors. We then fit this three-parameter model using an MCMC implementation, which ran 3 parallel sampling chains for 1000 iterations each, sampling μ and σ^2 via Gibbs steps and ϕ through Metropolis-Hastings steps with a $G(3, 1)$ candidate density. Convergence of the sampling chains was virtually immediate. We obtained the following posterior medians and 95% equal-tail credible intervals for the three parameters: for μ, 0.111 and (0.072, 0.167); for σ^2, 1.37 and (1.18, 2.11); and for ϕ, 1.62 and (0.28, 4.13).

Figure 7.3 maps summaries of the posterior samples for the 36 target blocks (city zips) and the 2 target points (A and B); specifically, the posterior medians, $q_{.50}$, upper and lower .025 points, $q_{.975}$ and $q_{.025}$, and the lengths of the 95% equal-tail credible intervals, $q_{.975} - q_{.025}$. The zip-level medians show a clear spatial pattern, with the highest predicted block averages occurring in the southeastern part of the city near the two high observed

readings (0.144 and 0.136), and the lower predictions in the north apparently the result of smoothing toward the low observed value in this direction (0.076). The interval lengths reflect spatial variability, with lower values occurring in larger areas (which require more averaging) or in areas nearer to observed monitoring stations (e.g., those near the southeastern, northeastern, and western city boundaries). Finally, note that our approach allows sensibly differing predicted medians for points A and B, with A being higher due to the slope of the fitted surface. Previous centroid-based analyses (like that of Carlin et al., 1999) would instead implausibly impute the same fitted value to both points, since both lie within the same zip.

7.2 Nested block-level modeling

We now turn to the case of variables available (and easily definable) only as block-level summaries. For example, it might be that disease data are known at the county level, but hypotheses of interest pertain to sociodemographically depressed census tracts. We refer to regions on which data are available as "source" zones and regions for which data are needed as "target" zones.

As mentioned earlier, the block-block interpolation problem has a rich literature and is often referred to as the *modifiable areal unit problem* (see, e.g., Cressie, 1996). In the case of an *extensive* variable (i.e., one whose value for a block can be viewed as a sum of sub-block values, as in the case of population, disease counts, productivity, or wealth), areal weighting offers a simple imputation strategy. While rather naive, such allocation proportional to area has a long history and is routinely available in GIS software.

The validity of simple areal interpolation obviously depends on the spatial variable in question being more or less evenly distributed across each region. For instance, Tobler (1979) introduced the so-called *pycnophylactic* approach. He assumed population density to be a continuous function of location, and proposed a simple "volume preserving" (with regard to the observed areal data) estimator of that function. This method is appropriate for continuous outcome variables but is harder to justify for count data, especially counts of human populations, since people do not generally spread out continuously over an areal unit; they tend to cluster.

Flowerdew and Green (1989) presented an approach wherein the variable of interest is count data and which uses information about the distribution of a binary covariate in the target zone to help estimate the counts. Their approach applies Poisson regression iteratively, using the EM algorithm, to estimate target zone characteristics. While subsequent work (Flowerdew and Green, 1992) extended this EM approach to continuous (typically normally distributed) outcome variables, neither of these papers reflects a fully inferential approach to the population interpolation problem.

In this section we follow Mugglin and Carlin (1998), and focus on the setting where the target zonation of the spatial domain D is a refinement of the source zonation, a situation we term *nested* misalignment. In the data setting we describe below, the source zones are U.S. census tracts, while the target zones (and the zones on which covariate data are available) are U.S. census block groups.

7.2.1 Methodology for nested block-level realignment

Consider the diagram in Figure 7.4. Assume that a particular rectangular tract of land is divided into two regions (I and II), and spatial variables (say, disease counts) y_1 and y_2 are known for these regions (the source zones). But suppose that the quantity of interest is Y_3, the unknown corresponding count in Region III (the target zone), which is comprised of subsections (IIIa and IIIb) of Regions I and II.

Figure 7.4 *Regional map for motivating example.*

As already mentioned, a crude way to approach the problem is to assume that disease counts are distributed evenly throughout Regions I and II, and so the number of affected individuals in Region III is just

$$y_1 \left[\frac{area(IIIa)}{area(I)} \right] + y_2 \left[\frac{area(IIIb)}{area(II)} \right] . \tag{7.13}$$

This simple areal interpolation approach is available within many GIS's. However, (7.13) is based on an assumption that is likely to be unviable, and also offers no associated estimate of uncertainty.

Let us now assume that the entire tract can be partitioned into smaller subsections, where on each subsection we can measure some other variable that is correlated with the disease count for that region. For instance, if we are looking at a particular tract of land, in each subsection we might record whether the land is predominantly rural or urban in character. We do this in the belief that this variable affects the likelihood of disease. Continuous covariates could also be used (say, the median household income in the subsection). Note that the subsections could arise simply as a refinement of the original scale of aggregation (e.g., if disease counts were available only by census tract, but covariate information arose at the census block group level), or as the result of overlaying a completely new set of boundaries (say, a zip code map) onto our original map. The statistical model is easier to formulate in the former case, but the latter case is of course more general, and is the one motivated by modern GIS technology (and to which we return in Section 7.3).

To facilitate our discussion in the former case, we consider a data set on the incidence of leukemia in Tompkins County, NY, that was originally presented and analyzed by Waller et al. (1994), and available on the web at `www.biostat.umn.edu/~brad/data/tompkins.dat`. As seen in Figure 7.5, Tompkins County, located in west-central New York state, is roughly centered around the city of Ithaca, NY. The county is divided into 23 census tracts, with each tract further subdivided into between 1 and 5 block groups, for a total of 51 such subregions. We have leukemia counts available at the tract level, and we wish to predict them at the block group level with the help of population counts and covariate information available on this more refined scale. In this illustration, the two covariates we consider are whether the block group is coded as "rural" or "urban," and whether or not the block group centroid is located within 2 kilometers of a hazardous chemical waste site. There are two waste sites in the county, one in the northeast corner and the other in downtown

Figure 7.5 *Map of Tompkins County, NY.*

Ithaca, near the county's center. (For this data set, we in fact have leukemia counts at the block group level, but we use only the tract totals in the model-fitting process, reserving the refined information to assess the accuracy of our results.) In this example, the unequal population totals in the block groups will play the weighting role that unequal areas would have played in (7.13).

Figure 7.6 shows a census tract-level disease map produced by the GIS `MapInfo`. The data record the block group-level population counts n_{ij} and covariate values u_{ij} and w_{ij}, where u_{ij} is 1 if block group j of census tract i is classified as urban, 0 if rural, and w_{ij} is 1 if the block group centroid is within 2 km of a waste site, 0 if not. Typical of GIS software, `MapInfo` permits allocation of the census tract totals to the various block groups proportional to block group area or population. We use our hierarchical Bayesian method to incorporate the covariate information, as well as to obtain variance estimates to accompany the block group-level point estimates.

As in our earlier disease mapping discussion (Subsection 6.4.1), we introduce a first-stage Poisson model for the disease counts,

$$Y_{ij}|m_{k(i,j)} \overset{ind}{\sim} Po(E_{ij}m_{k(i,j)}),\ i = 1, \ldots, I,\ j = 1, \ldots, J_i\ ,$$

where $I = 23$, J_i varies from 1 to 5, Y_{ij} is the disease count in block group j of census tract i, and E_{ij} is the corresponding "expected" disease count, computed as $E_{ij} = n_{ij}\lambda$ where n_{ij} is the population count in the cell and λ is the overall probability of contracting the disease. This "background" probability could be estimated from our data; here we take $\lambda = 5.597 \times 10^{-4}$, the crude leukemia rate for the eight-county region studied by Waller et al. (1994), an area that includes Tompkins County. Hence, $m_{k(i,j)}$ is the relative risk of contracting leukemia in block group (i, j), and $k = k(i, j) = 1, 2, 3$, or 4 depending on the covariate status of the block group. Specifically, we let

$$k(i,j) = \begin{cases} 1\ , & \text{if } (i,j) \text{ is rural, not near a waste site} \\ 2\ , & \text{if } (i,j) \text{ is urban, not near a waste site} \\ 3\ , & \text{if } (i,j) \text{ is rural, near a waste site} \\ 4\ , & \text{if } (i,j) \text{ is urban, near a waste site} \end{cases} .$$

Figure 7.6 *GIS map of disease counts by census tract, Tompkins County, NY.*

Defining $\mathbf{m} = (m_1, m_2, m_3, m_4)$ and again adopting independent and minimally informative gamma priors for these four parameters, we seek estimates of $p(m_k|\mathbf{y})$, where $\mathbf{y} = (y_{1.}, \ldots, y_{I.})$, and $y_{i.} = \sum_{j=1}^{J_i} y_{ij}$, the census tract disease count totals. We also wish to obtain block group-specific mean and variance estimates $E[Y_{ij}|\mathbf{y}]$ and $Var[Y_{ij}|\mathbf{y}]$, to be plotted in a disease map at the block group (rather than census tract) level. Finally, we may also wish to estimate the distribution of the total disease count in some conglomeration of block groups (say, corresponding to some village or city).

By the conditional independence of the block group counts we have $Y_{i.}|\mathbf{m} \overset{ind}{\sim} Po(\sum_{k=1}^{4} s_k m_k)$, $i = 1, \ldots, I$, where $s_k = \sum_{j:k(i,j)=k} E_{ij}$, the sum of the expected cases in block groups j of region i corresponding to covariate pattern k, $k = 1, \ldots, 4$. The likelihood $L(\mathbf{m}; \mathbf{y})$ is then the product of the resulting $I = 23$ Poisson kernels. After multiplying this by the prior distribution term $\prod_{k=1}^{4} p(m_k)$, we can obtain forms proportional to the four full conditional distributions $p(m_k|m_{l\neq k}, \mathbf{y})$, and sample these sequentially via univariate Metropolis steps.

Once again it is helpful to reparameterize to $\delta_k = \log(m_k)$, $k = 1, \ldots, 4$, and perform the Metropolis sampling on the log scale. We specify reasonably vague $Gamma(a, b)$ priors for the m_k by taking $a = 2$ and $b = 10$ (similar results were obtained with even less informative gamma priors unless a was quite close to 0, in which case convergence was unacceptably poor). For this "base prior," convergence obtains after 200 iterations, and the remaining 1800 iterations in 5 parallel MCMC chains are retained as posterior samples from $p(\mathbf{m}|\mathbf{y})$.

A second reparametrization aids in interpreting our results. Suppose we write

$$\delta_{k(i,j)} = \theta_0 + \theta_1 u_{ij} + \theta_2 w_{ij} + \theta_3 u_{ij} w_{ij} \ , \tag{7.14}$$

so that θ_0 is an intercept, θ_1 is the effect of living in an urban area, θ_2 is the effect of living near a waste site, and θ_3 is the urban/waste site interaction. This reparametrization expresses the log-relative risk of disease as a linear model, a common approach in spatial disease mapping (Besag et al., 1991; Waller et al., 1997). A simple 1-1 transformation converts our $(m_1^{(g)}, m_2^{(g)}, m_3^{(g)}, m_4^{(g)})$ samples to $(\theta_0^{(g)}, \theta_1^{(g)}, \theta_2^{(g)}, \theta_3^{(g)})$ samples on the new scale, which in turn allows direct investigation of the main effects of urban area and waste site proximity, as well as the effect of interaction between these two. Figure 7.7 shows the

Figure 7.7 *Posterior histograms of sampled log-relative risk parameters, Tompkins County, NY, data set.*

histograms of the posterior samples for $\theta_i,\ i = 0, 1, 2, 3$. We note that θ_0, θ_1, and θ_3 are not significantly different from 0 as judged by the 95% BCI, while θ_2 is "marginally significant" (in a Bayesian sense) at this level. This suggests a moderately harmful effect of residing within 2 km of a waste site, but no effect of merely residing in an urban area (in this case, the city of Ithaca). The preponderance of negative $\theta_3^{(g)}$ samples is somewhat surprising; we might have expected living near an urban waste site to be associated with an increased (rather than decreased) risk of leukemia. This is apparently the result of the high leukemia rate in a few rural block groups *not* near a waste site (block groups 1 and 2 of tract 7, and block group 2 of tract 20), forcing θ_3 to adjust for the relatively lower overall rate near the Ithaca waste site.

7.2.2 Individual block group estimation

To create the block group-level estimated disease map, for those census tracts having $J_i > 1$, we obtain a conditional binomial distribution for Y_{ij} given the parameters $\mathbf{m}$ and the census tract totals $\mathbf{y}$, so that

$$E(Y_{ij}|\mathbf{y}) = E[E(Y_{ij}|\mathbf{m}, \mathbf{y})] \approx \frac{y_{i.}}{G} \sum_{g=1}^{G} p_{ij}^{(g)} \ , \tag{7.15}$$

where p_{ij} is the appropriate binomial probability arising from conditioning a Poisson random variable on the sum of itself and a second, independent Poisson variable. For example, for $p_{11}^{(g)}$ we have

$$p_{11}^{(g)} = \frac{1617 m_1^{(g)}}{(1617 + 702) m_1^{(g)} + (1526 + 1368) m_3^{(g)}} \ ,$$

Figure 7.8 *GIS maps of interpolated (a) and actual (b) block group disease counts, Tompkins County, NY.*

as determined by the covariate patterns in the first four rows of the data set. Note that when $J_i = 1$ the block group total equals the known census tract total, hence no estimation is necessary.

The resulting collection of estimated block group means $E(Y_{ij}|\mathbf{y})$ are included in the data set on our webpage, along with the actual case counts y_{ij}. (The occasional noninteger values of y_{ij} in the data are not errors, but arise from a few cases in which the precise block group of occurrence is unknown, resulting in fractional counts being allocated to several block groups.) Note that, like other interpolation methods, the sum of the estimated cases in each census tract is the same as the corresponding sum for the actual case counts. The GIS maps of the $E(Y_{ij}|\mathbf{y})$ and the actual y_{ij} shown in Figure 7.8 reveal two pockets of elevated disease counts (in the villages of Cayuga Heights and Groton).

To get an idea of the variability inherent in the posterior surface, we might consider mapping the estimated posterior variances of our interpolated counts. Since the block

group-level variances do not involve aggregation across census tracts, these variances may be easily estimated as $Var(Y_{ij}|\mathbf{y}) = E(Y_{ij}^2|\mathbf{y}) - [E(Y_{ij}|\mathbf{y})]^2$, where the $E(Y_{ij}|\mathbf{y})$ are the estimated means (already calculated), and

$$\begin{aligned} E(Y_{ij}^2|\mathbf{y}) &= E[E(Y_{ij}^2|\mathbf{m},\mathbf{y})] = E[y_{i.}p_{ij}(1-p_{ij}) + y_{i.}^2(p_{ij})^2] \\ &\approx \frac{1}{G}\sum_{g=1}^{G}\left[y_{i.}p_{ij}^{(g)}(1-p_{ij}^{(g)}) + y_{i.}^2(p_{ij}^{(g)})^2\right] , \end{aligned} \tag{7.16}$$

where p_{ij} is again the appropriate binomial probability for block group (i, j); see Mugglin and Carlin (1998) for more details.

We remark that most of the census tracts are composed of homogeneous block groups (e.g., all rural with no waste site nearby); in these instances the resulting binomial probability for each block group is free of $\mathbf{m}$. In such cases, posterior means and variances are readily available without any need for mixing over the Metropolis samples, as in Equations (7.15) and (7.16).

7.2.3 Aggregate estimation: Block groups near the Ithaca, NY, waste site

In order to assess the number of leukemia cases we expect in those block groups within 2 km of the Ithaca waste site, we can sample the predictive distributions for these blocks, sum the results, and draw a histogram of these sums. Twelve block groups in five census tracts fall within these 2-km radii: all of the block groups in census tracts 11, 12, and 13, plus two of the three (block groups 2 and 3) in tract 6 and three of the four (block groups 2, 3, and 4) in tract 10. Since the totals in census tracts 11, 12, and 13 are known to our analysis, we need only sample from two binomial distributions, one each for the conglomerations of near-waste site block groups within tracts 6 and 10. Defining the sum over the 12 block groups as Z, we have

$$Z^{(g)} = Y_{6,(2,3)}^{(g)} + Y_{10,(2,3,4)}^{(g)} + y_{11,.} + y_{12,.} + y_{13,.} \ .$$

A histogram of these values is shown in Figure 7.9. The estimated median value of 10 happens to be exactly equal to the true value of 10 cases in this area. The sample mean, 9.43, is also an excellent estimate. Note that the minimum and maximum values in Figure 7.9, $Z = 7$ and $Z = 11$, are imposed by the data structure: there must be at least as many cases as the total known to have occurred in census tracts 11, 12, and 13 (which is 7), and there can be no more than the total number known to have occurred in tracts 6, 10, 11, 12, and 13 (which is 11).

Finally, we may again compare our results to those produced by a GIS under either area-based or population-based interpolation. The former produces a mean estimate of 9.28, while the latter gives 9.59. These are close to the Bayesian mean 9.43, but neither approach produces an associated confidence interval, much less a full graphical display of the sort given in Figure 7.9.

7.3 Nonnested block-level modeling

The approach of the previous section (see also Mugglin and Carlin, 1998, and Mugglin et al., 1999) offered a hierarchical Bayesian method for interpolation and smoothing of Poisson responses with covariates in the nested case. In the remainder of this section we develop a framework for hierarchical Bayesian interpolation, estimation, and spatial smoothing over *nonnested* misaligned data grids. In Subsection 7.3.1 we summarize a data set collected in response to possible contamination resulting from the former Feed Materials Production

Figure 7.9 *Histogram of sampled disease counts, total of all block groups having centroid within 2 km of the Ithaca, NY, waste site.*

Center (FMPC) in southwestern Ohio with the foregoing analytic goals. In Subsection 7.3.2 we develop the theory of our modeling approach in a general framework, as well as our MCMC approach and a particular challenge that arises in its implementation for the FMPC data. Finally in Example 7.2 we set forth the conclusions resulting from our analysis of the FMPC data.

7.3.1 Motivating data set

Risk-based decision making is often used for prioritizing cleanup efforts at U.S. Superfund sites. Often these decisions will be based on estimates of the past, present, and future potential health impacts. These impact assessments usually rely on estimation of the number of outcomes, and the accuracy of these estimates will depend heavily on the ability to estimate the number of individuals at risk. Our motivating data set is connected with just this sort of risk assessment.

In the years 1951–1988 near the town of Ross in southwestern Ohio, the former Feed Materials Production Center (FMPC) processed uranium for weapons production. Draft results of the Fernald Dosimetry Reconstruction Project, sponsored by the Centers for Disease Control and Prevention (CDC), indicated that during production years the FMPC released radioactive materials (primarily radon and its decay products and, to a lesser extent, uranium and thorium) from the site. Although radioactive liquid wastes were released, the primary exposure to residents of the surrounding community resulted from breathing radon decay products. The potential for increased risk of lung cancer is thus the focus of intense local public interest and ongoing public health studies (see Devine et al., 1998).

Estimating the number of adverse health outcomes in the population (or in subsets thereof) requires estimation of the number of individuals at risk. Population counts, broken down by age and sex, are available from the U.S. Census Bureau according to federal census block groups, while the areas of exposure interest are dictated by both direction and distance from the plant. Rogers and Killough (1997) construct an exposure "windrose," which consists of 10 concentric circular bands at 1-kilometer radial increments divided into 16

Figure 7.10 *Census block groups and 10-km windrose near the FMPC site, with 1990 population density by block group and 1980 structure density by cell (both in counts per km^2).*

compass sectors (N, NNW, NW, WNW, W, etc.). Through the overlay of such a windrose onto U.S. Geological Survey (USGS) maps, they provide counts of the numbers of "structures" (residential buildings, office buildings, industrial building complexes, warehouses, barns, and garages) within each subdivision (*cell*) of the windrose.

Figure 7.10 shows the windrose centered at the FMPC. We assign numbers to the windrose cells, with 1 to 10 indexing the cells starting at the plant and running due north, then 11 to 20 running from the plant to the north-northwest, and so on. Structure counts are known for each cell; the hatching pattern in the figure indicates the areal density (structures per square kilometer) in each cell.

Also shown in Figure 7.10 are the boundaries of 39 Census Bureau block groups, for which 1990 population counts are known. These are the source zones for our interpolation problem. Shading intensity indicates the population density (persons per square kilometer) for each block group. The intersection of the two (nonnested) zonation systems results in 389 regions we call *atoms,* which can be aggregated appropriately to form either cells or block groups.

The plant was in operation for 38 years, raising concern about the potential health risks it has caused, a question that has been under active investigation by the CDC for some time. Present efforts to assess the impact of the FMPC on cancer morbidity and mortality require the analysis of this misaligned data set; in particular, it is necessary to interpolate gender- and age group-specific population counts to the windrose exposure cells. These numbers of persons at risk could then be combined with cell-specific dose estimates obtained by Killough et al. (1996) and estimates of the cancer risk per unit dose to obtain expected numbers of excess cancer cases by cell.

In fact, such an expected death calculation was made by Devine et al. (1998), using traditional life table methods operating on the Rogers and Killough (1997) cell-level population estimates (which were in turn derived simply as proportional to the structure counts).

Figure 7.11 *Illustrative representation of areal data misalignment.*

However, these estimates were only for the total population in each cell; sex and age group-specific counts were obtained by "breaking out" the totals into subcategories using a standard table (i.e., the *same* table in each cell, regardless of its true demographic makeup). In addition, the uncertainty associated with the cell-specific population estimates was quantified in a rather ad hoc way.

7.3.2 Methodology for nonnested block-level realignment

We confine our model development to the case of two misaligned spatial grids. Given this development, the extension to more than two grids will be conceptually apparent. The additional computational complexity and bookkeeping detail will also be evident.

Let the first grid have regions indexed by $i = 1, ..., I$, denoted by B_i, and let $S_B = \bigcup_i B_i$. Similarly, for the second grid we have regions C_j, $j = 1, ..., J$ with $S_C = \bigcup_j C_j$. In some applications $S_B = S_C$, i.e., the B-cells and the C-cells offer different partitions of a common region. Nested misalignment (e.g., where each C_j is contained entirely in one and only one B_i) is evidently a special case. Another possibility is that one data grid contains the other; say, $S_B \subset S_C$. In this case, there will exist some C cells for which a portion lies outside of S_B. In the most general case, there is no containment and there will exist B-cells for which a portion lies outside of S_C and C-cells for which a portion lies outside of S_B. Figure 7.11 illustrates this most general situation.

Atoms are created by intersecting the two grids. For a given B_i, each C-cell which intersects B_i creates an atom (which possibly could be a union of disjoint regions). There may also be a portion of B_i which does not intersect with any C_j. We refer to this portion as the *edge* atom associated with B_i, i.e., a B-edge atom. In Figure 7.11, atoms B_{11} and B_{21} are B-edge atoms. Similarly, for a given C_j, each B-cell which intersects with C_j creates an atom, and we analogously determine C-edge atoms (atoms C_{11} and C_{22} in Figure 7.11). It is crucial to note that each nonedge atom can be referenced relative to an appropriate B-cell, say, B_i, and denoted as B_{ik}. It also can be referenced relative to an appropriate C cell, say C_j, and denoted by $C_{j\ell}$. Hence, there is a one-to-one mapping within $S_B \bigcap S_C$ between the set of ik's and the set of $j\ell$'s, as shown in Figure 7.11 (which also illustrates our convention of indexing atoms by area, in descending order). Formally we can define the function c on nonedge B-atoms such that $c(B_{ik}) = C_{j\ell}$, and the *inverse* function b on C-atoms such that $b(C_{j\ell}) = B_{ik}$. For computational purposes we suggest creation of "look-up" tables to specify these functions. (Note that the possible presence of both types of edge cell precludes a single

"ij" atom numbering system, since such a system could index cells on either S_B or S_C, but not their union.)

Without loss of generality we refer to the first grid as the *response* grid, that is, at each B_i we observe a response Y_i. We seek to explain Y_i using a variety of covariates. Some of these covariates may, in fact, be observed on the response grid; we denote the value of this vector for B_i by $\mathbf{W}_i$. But also, some covariates are observed on the second or *explanatory* grid. We denote the value of this vector for C_j by $\mathbf{X}_j$.

We seek to explain the observed Y's through both $\mathbf{X}$ and $\mathbf{W}$. The misalignment between the $\mathbf{X}$'s and Y's is the obstacle to standard regression methods. What levels of $\mathbf{X}$ should be assigned to Y_i? We propose a fully model-based approach in the case where the Y's and X's are aggregated measurements. The advantage of a model-based approach implemented within a Bayesian framework is full inference both with regard to estimation of model parameters and prediction using the model.

The assumption that the Y's are aggregated measurements means Y_i can be envisioned as $\sum_k Y_{ik}$, where the Y_{ik} are unobserved or latent and the summation is over all atoms (including perhaps an edge atom) associated with B_i. To simplify, we assume that the X's are also scalar aggregated measurements, i.e., $X_j = \sum_\ell X_{j\ell}$ where the summation is over all atoms associated with C_j. As for the $\mathbf{W}$'s, we assume that each component is either an aggregated measurement or an *inheritable* measurement. For component r, in the former case $W_i^{(r)} = \sum_k W_{ik}^{(r)}$ as with Y_i; in the latter case $W_{ik}^{(r)} = W_i^{(r)}$.

In addition to (or perhaps in place of) the $\mathbf{W}_i$ we will introduce B-cell random effects μ_i, $i = 1, ..., I$. These effects are employed to capture spatial association among the Y_i's. The μ_i can be given a spatial prior specification. A Markov random field form (Besag, 1974; Bernardinelli and Montomoli, 1992), as described below, is convenient. Similarly we will introduce C-cell random effects ω_j, $j = 1, ..., J$ to capture spatial association among the X_j's. It is assumed that the latent Y_{ik} inherit the effect μ_i and that the latent $X_{j\ell}$ inherit the effect ω_j.

For aggregated measurements that are counts, we assume the latent variables are conditionally independent Poissons. As a result, the observed measurements are Poissons as well and the conditional distribution of the latent variables given the observed is a product multinomial. We note that it is not required that the Y's be count data. For instance, with aggregated measurements that are continuous, a convenient distributional assumption is conditionally independent gammas, in which case the latent variables would be rescaled to product Dirichlet. An alternative choice is the normal, whereupon the latent variables would have a distribution that is a product of conditional multivariate normals. In this section we detail the Poisson case.

As mentioned above, area naturally plays an important role in allocation of spatial measurements. Letting $|A|$ denote the area of region A, if we apply the standard assumption of allocation proportional to area to the $X_{j\ell}$ in a stochastic fashion, we would obtain

$$X_{j\ell} \mid \omega_j \sim Po(e^{\omega_j}|C_{j\ell}|) \ , \tag{7.17}$$

assumed independent for $\ell = 1, 2, ..., L_j$. Then $X_j \mid \omega_j \sim Po(e^{\omega_j}|C_j|)$ and $(X_{j1}, X_{j2}, ..., X_{j,L_j} \mid X_j, \omega_j) \sim Mult(X_j; q_{j1}, ..., q_{j,L_j})$ where $q_{j\ell} = |C_{j\ell}|/|C_j|$.

Such strictly area-based modeling cannot be applied to the Y_{ik}'s since it fails to connect the Y's with the X's (as well as the $\mathbf{W}$'s). To do so we again begin at the atom level. For nonedge atoms we use the previously mentioned look-up table to find the $X_{j\ell}$ to associate with a given Y_{ik}. It is convenient to denote this $X_{j\ell}$ as X'_{ik}. Ignoring the $\mathbf{W}_i$ for the moment, we assume

$$Y_{ik} \mid \mu_i, \theta_{ik} \sim Po\left(e^{\mu_i}|B_{ik}|\, h(X'_{ik}/|B_{ik}|\,;\, \theta_{ik})\right) \ , \tag{7.18}$$

independent for $k = 1, \ldots, K_i$. Here h is a preselected parametric function, the part of the model specification that adjusts an expected proportional-to-area allocation according to

X'_{ik}. Since (7.17) models expectation for $X_{j\ell}$ proportional to $|C_{j\ell}|$, it is natural to use the *standardized* form $X'_{ik}/|B_{ik}|$ in (7.18). Particular choices of h include $h(z\,;\,\theta_{ik}) = z$ yielding $Y_{ik} \mid \mu_i \sim Po(e^{\mu_i} X'_{ik})$, which would be appropriate if we choose not to use $|B_{ik}|$ explicitly in modeling $E(Y_{ik})$. In our FMPC implementation, we actually select $h(z;\theta_{ik}) = z+\theta_{ik}$ where $\theta_{ik} = \theta/(K_i|B_{ik}|)$ and $\theta > 0$; see Equation (7.23) below and the associated discussion.

If B_i has no associated edge atom, then

$$Y_i \mid \mu_i, \boldsymbol{\theta}, \{X_{j\ell}\} \sim Po\left(e^{\mu_i} \sum_k |B_{ik}|\, h(X'_{ik}/|B_{ik}|\,;\,\theta_{ik})\right) . \tag{7.19}$$

If B_i has an edge atom, say B_{iE}, since there is no corresponding $C_{j\ell}$, there is no corresponding X'_{iE}. Hence, we introduce a latent X'_{iE} whose distribution is determined by the nonedge atoms that are neighbors of B_{iE}. Paralleling Equation (7.17), we model X'_{iE} as

$$X'_{iE} \mid \omega_i^* \sim Po(e^{\omega_i^*}|B_{iE}|) , \tag{7.20}$$

thus adding a new set of random effects $\{\omega_i^*\}$ to the existing set $\{\omega_j\}$. These two sets together are assumed to have a single CAR specification. An alternative is to model $X'_{iE} \sim Po\left(|B_{iE}|\left(\sum_{N(B_{iE})} X'_t / \sum_{N(B_{iE})} |B_t|\right)\right)$, where $N(B_{iE})$ is the set of neighbors of B_{iE} and t indexes this set. Effectively, we multiply $|B_{iE}|$ by the overall count per unit area in the neighboring nonedge atoms. While this model is somewhat more data-dependent than the (more model-dependent) one given in (7.20), we remark that it can actually lead to better MCMC convergence due to the improved identifiability in its parameter space: the spatial similarity of the structures in the edge zones is being modeled directly, rather than indirectly via the similarity of the ω_i^* and the ω_j.

Now, with an X'_{ik} for all ik, (7.18) is extended to all B-atoms and the conditional distribution of Y_i is determined for all i as in (7.19). But also $Y_{i1}, ..., Y_{ik_i}|Y_i, \mu_i, \theta_{ik}$ is distributed Multinomial$(Y_i; q_{i1}, ..., q_{ik_i})$, where $q_{ik} = |B_{ik}|\, h(X'_{ik}/\,|B_{ik}|\,;\theta_{ik})/\sum_k |B_{ik}|\, h(X'_{ik}/\,|B_{ik}|\,;\theta_{ik})$.

To capture the spatial nature of the B_i we may adopt an IAR model for the μ_i, i.e.,

$$p(\mu_i \mid \mu_{i',i'\neq i}) = N\left(\sum_{i'} w_{ii'}\mu_{i'}/w_{i.}\ ,\ 1/(\lambda_\mu w_{i.})\right) \tag{7.21}$$

where $w_{ii} = 0$, $w_{ii'} = w_{i'i}$ and $w_{i.} = \sum_{i'} w_{ii'}$. Below, we set $w_{ii'} = 1$ for $B_{i'}$ a neighbor of B_i and $w_{ii'} = 0$ otherwise, the standard "0-1 adjacency" form.

Similarly we assume that

$$f(\omega_j \mid \omega_{j',j'\neq j}) = N\left(\sum_{j'} v_{jj'}\omega_{j'}/v_{j.}\ ,\ 1/(\lambda_\omega v_{j.})\right) .$$

We adopt a proper Gamma prior for λ_μ and also for λ_ω. When $\boldsymbol{\theta}$ is present we require a prior that we denote by $f(\boldsymbol{\theta})$. The choice of $f(\boldsymbol{\theta})$ will likely be vague but its form depends upon the adopted parametric form of h.

The entire specification can be given a representation as a graphical model, as in Figure 7.12. In this model the arrow from $\{X_{j\ell}\} \to \{X'_{ik}\}$ indicates the inversion of the $\{X_{jl}\}$ to $\{X'_{ik}\}$, augmented by any required edge atom values X'_{iE}. The $\{\omega_i^*\}$ would be generated if the X'_{iE} are modeled using (7.20). Since the $\{Y_{ik}\}$ are not observed, but are distributed as multinomial given the fixed block group totals $\{Y_i\}$, this is a predictive step in our model, as indicated by the arrow from $\{Y_i\}$ to $\{Y_{ik}\}$ in the figure. In fact, as mentioned above the further predictive step to impute Y'_j, the Y total associated with X_j in the jth target zone, is of key interest. If there are edge atoms C_{jE}, this will require a model for the associated

Figure 7.12 *Graphical version of the model, with variables as described in the text. Boxes indicate data nodes, while circles indicate unknowns.*

Y'_{jE}. Since there is no corresponding B-atom for C_{jE} a specification such as (7.18) is not appropriate. Rather, we can imitate the above modeling for X'_{iE} using (7.20) by introducing $\{\mu_j^*\}$, which along with the μ_i follow the prior in (7.21). The $\{\mu_j^*\}$ and $\{Y'_{jE}\}$ would add two consecutive nodes to the right side of Figure 7.12, connecting from λ_μ to $\{Y'_j\}$.

The entire distributional specification overlaid on this graphical model has been supplied in the foregoing discussion and (in the absence of C_{jE} edge atoms, as in Figure 7.10) takes the form

$$\begin{aligned}&\textstyle\prod_i f(Y_{i1},...,Y_{ik_i} \mid Y_i, \boldsymbol{\theta}) \prod_i f(Y_i \mid \mu_i, \boldsymbol{\theta}, \{X'_{ik}\})\, f(\{X'_{ik}\} \mid \omega_i^*, \{X_{j\ell}\})\\ &\quad \times \textstyle\prod_j f(X_{j1},...,X_{jL_j} \mid X_j) \prod_j f(X_j \mid \omega_j)\\ &\quad \times f(\{\mu_i\} \mid \lambda_\mu)\, f(\lambda_\mu)\, f(\{\omega_j\}, \{\omega_i^*\} \mid \lambda_\omega)\, f(\lambda_\omega)\, f(\boldsymbol{\theta})\,.\end{aligned} \tag{7.22}$$

Bringing in the $\mathbf{W}_i$ merely revises the exponential term in (7.18) from $\exp(\mu_i)$ to $\exp(\mu_i + \mathbf{W}_{ik}^T\boldsymbol{\beta})$. Again, for an inherited component of $\mathbf{W}_i$, say, $W_i^{(r)}$, the resulting $W_{ik}^{(r)} = W_i^{(r)}$. For an aggregated component of $\mathbf{W}_i$, again, say, $W_i^{(r)}$, we imitate (7.17) assuming $W_{ik}^{(r)} \mid \mu_i^{(r)} \sim Po(e^{\mu_i^{(r)}}|B_{ik}|)$, independent for $k = 1, ..., K_i$. A spatial prior on the $\mu_i^{(r)}$ and a Gaussian (or perhaps flat) prior on $\boldsymbol{\beta}$ completes the model specification.

Finally, on the response grid, for each B_i rather than observing a single Y_i we may observe Y_{im}, where $m = 1, 2, ..., M$ indexes levels of factors such as sex, race, or age group. Here we seek to use these factors, in an ANOVA fashion, along with the X_j (and $\mathbf{W}_i$) to explain the Y_{im}. Ignoring $\mathbf{W}_i$, the resultant change in (7.18) is that Y_{ikm} will be Poisson with μ_i replaced by μ_{im}, where μ_{im} has an appropriate ANOVA form. For example, in the case of sex and age classes, we might have a sex main effect, an age main effect, and a sex-age interaction effect. In our application these effects are not nested within i; we include only a spatial overall mean effect indexed by i.

Regarding the MCMC implementation of our model, besides the usual concerns about appropriate choice of Metropolis-Hastings candidate densities and acceptability of the resulting convergence rate, one issue deserves special attention. Adopting the identity function for h in (7.18) produces the model $Y_{ik} \sim Po\,(e^{\mu_i}\,(X'_{ik}))$, which in turn implies $Y_{i.} \sim Po\,(e^{\mu_i}(X'_{i.}))$. Suppose however that $Y_{i.} > 0$ for a particular block group i, but in some MCMC iteration no structures are allocated to any of the atoms of the block group. The result is a flawed probabilistic specification. To ensure $h > 0$ even when $z = 0$, we revised our model to $h(z\,;\,\theta_{ik}) = z + \theta_{ik}$ where $\theta_{ik} = \theta/(K_i|B_{ik}|)$ with $\theta > 0$, resulting in

$$Y_{ik} \sim Po\left(e^{\mu_i}\left(X'_{ik} + \frac{\theta}{K_i}\right)\right)\,. \tag{7.23}$$

This adjustment eliminates the possibility of a zero-valued Poisson parameter, but does allow for the possibility of a nonzero population count in a region where there are no structures observed. When conditioned on $Y_{i\cdot}$, we find $(Y_{i1}, \ldots, Y_{iK_i} \mid Y_{i\cdot}) \sim \text{Mult}(Y_{i\cdot}\,;\, p_{i1}, \ldots, p_{iK_i})$, where

$$p_{ik} = \frac{X'_{ik} + \theta/K_i}{X'_{i\cdot} + \theta} \text{ and } Y_{i\cdot} \sim Po\left(e^{\mu_i}(X'_{i\cdot} + \theta)\right) . \tag{7.24}$$

Our basic model then consists of (7.23) to (7.24) together with

$$\begin{gathered} \mu_i \overset{iid}{\sim} N\left(\eta_\mu, 1/\tau_\mu\right), \;\; X_{jl} \sim Po\left(e^{\omega_j}|C_{jl}|\right) \Rightarrow X_{j\cdot} \sim Po\left(e^{\omega_j}|C_j|\right) , \\ (X_{j1}, \ldots, X_{jL_j} \mid X_{j\cdot}) \sim \text{Mult}(X_{j\cdot}\,;\, q_{j1}, \ldots, q_{jL_j}), \text{ where } q_{jl} = |C_{jl}|/|C_j|, \\ X'_{iE} \sim Po\left(e^{\omega_i^*}|B_{iE}|\right) , \text{ and } (\omega_j, \omega_i^*) \sim \text{CAR}(\lambda_\omega) , \end{gathered} \tag{7.25}$$

where X'_{iE} and ω_i^* refer to edge atom structure counts and log relative risk parameters, respectively. While θ could be estimated from the data, in our implementation we simply set $\theta = 1$; Mugglin et al. (2000, Sec. 6) discuss the impact of alternate selections.

Example 7.2 *(FMPC data analysis).* We turn now to the particulars of the FMPC data analysis, examining two different models in the context of the misaligned data as described in Section 7.3.1. In the first case we take up the problem of total population interpolation, while in the second we consider age- and sex-specific population interpolation.

7.3.2.1 Total population interpolation model

We begin by taking $\eta_\mu = 1.1$ and $\tau_\mu = 0.5$ in (7.25). The choice of mean value reflects the work of Rogers and Killough (1997), who found population per household (PPH) estimates for four of the seven townships in which the windrose lies. Their estimates ranged in value from 2.9 to 3.2, hence our choice of $\eta_\mu = 1.1 \approx \log(3)$. The value $\tau_\mu = 0.5$ is sufficiently small to make the prior for μ_i large enough to support all feasible values of μ_i (two prior standard deviations in either direction would enable PPH values of 0.18 to 50.8).

For $\boldsymbol{\omega} = \{\omega_j, \omega_i^*\}$ we adopted a CAR prior and fixed $\lambda_\omega = 10$. We did not impose any centering of the elements of $\boldsymbol{\omega}$ around 0, allowing them to determine their own mean level in the MCMC algorithm. Since most cells have four neighbors, the value $\lambda_\omega = 10$ translates into a conditional prior standard deviation for the ω's of $\sqrt{1/(10 \cdot 4)} = .158$, hence a marginal prior standard deviation of roughly $.158/.7 \approx .23$ (Bernardinelli et al., 1995). In any case, we found $\lambda_\omega < 10$ too vague to allow MCMC convergence. Typical posterior medians for the ω's ranged from 2.2 to 3.3 for the windrose ω_j's and from 3.3 to 4.5 for the edge ω_i^*s.

Running 5 parallel sampling chains, acceptable convergence obtains for all parameters within 1,500 iterations. We discarded this initial sample and then continued the chains for an additional 5,000 iterations each, obtaining a final posterior sample of size 25,000. From the resulting samples, we can examine the posterior distributions of any parameters we wish. It is instructive first to examine the distributions of the imputed structure counts X_{jl}. For example, consider Figure 7.13, which shows the posterior distributions of the structure counts in cell 106 (the sixth one from the windrose center in the SE direction), for which $L_j = 4$. The known cell total $X_{106,\cdot}$ is 55. Note that the structure values indicated in the histograms are integers. The vertical bars in each histogram indicate how the 55 structures would be allocated if imputed proportionally to area. In this cell we observe good general agreement between these naively imputed values and our histograms, but the advantage of assessing variability from the full distributional estimates is immediately apparent.

Population estimates per cell for cells 105 through 110 (again in the SE direction, from the middle to outer edge of the windrose) are indicated in Figure 7.14. Vertical bars here represent estimates calculated by multiplying the number of structures in the cell by a fixed (map-wide) constant representing population per household (PPH), a method roughly

Figure 7.13 *Posterior distributions of structure estimates for the four atoms of cell 106 (SE6). Vertical bars represent structure values if imputed proportionally to area. Here and in the next figure, "median" denotes posterior median, and "95% BCI" denotes the equal-tail Bayesian confidence interval.*

equivalent to that employed by Rogers and Killough (1997), who as mentioned above actually used four different PPH values. Our reference lines use a constant value of 3 (the analogue of our prior mean). While cells 105 and 106 indicate good general agreement in these estimates, cells 107 through 110 display markedly different population estimates, where our estimates are substantially higher than the constant-PPH estimates. This is typical of cells toward the outer edge of the southeast portion of the windrose, since the suburbs of Cincinnati encroach on this region. We have population data only (no structures) in the southeastern edge atoms, so our model must estimate both the structures and the population in these regions. The resulting PPH is higher than a mapwide value of 3 (one would expect suburban PPH to be greater than rural PPH) and so the CAR model placed on the $\{\omega_j, \omega_i^*\}$ parameters induces a spatial similarity that can be observed in Figure 7.14.

We next implement the $\{Y_i.\} \to \{Y_{ik}\}$ step. From the resulting $\{Y_{ik}\}$ come the $\{Y'_{j.}\}$ cell totals by appropriate reaggregation. Figure 7.15 shows the population densities by atom $(Y_{ik}/|B_{ik}|)$, calculated by taking the posterior medians of the population distributions for each atom and dividing by atom area in square kilometers. This figure clearly shows the encroachment by suburban Cincinnati on the southeast side of our map, with some spatial smoothing between the edge cells and the outer windrose cells. Finally, Figure 7.16 shows population densities by cell $(Y'_{j.}/|C_j|)$, where the atom-level populations have been aggregated to cells before calculating densities. Posterior standard deviations, though not shown, are also available for each cell. While this figure, by definition, provides less detail than Figure 7.15, it provides information at the scale appropriate for combination with the exposure values of Killough et al. (1996). Moreover, the scale of aggregation is still fine enough to permit identification of the locations of Cincinnati suburban sprawl, as well as the communities of Ross (contained in cells ENE 4-5 and NE 4), Shandon (NW 4-5), New Haven (WSW 5-6), and New Baltimore (SSE 4-5).

Figure 7.14 *Posterior distributions of populations in cells 105 to 110. Vertical bars represent estimates formed by multiplying structures per cell by a constant population per household (PPH) of 3.0.*

7.3.2.2 Age and sex effects

Recall from Section 7.3.1 that we seek population counts not only by cell but also by sex and age group. This is because the dose resulting from a given exposure will likely differ depending on gender and age, and because the risk resulting from that dose can also be affected by these factors. Again we provide results only for the year 1990; the extension to other timepoints would of course be similar. Population counts at the block group level by sex and age group are provided by the U.S. Census Bureau. Specifically, age is recorded as counts in 18 quinquennial (5-year) intervals: 0–4, 5–9, . . . , 80–84, and 85+. We consider an additive extension of our basic model (7.23)–(7.25) to the sex and age group-specific case; see Mugglin et al. (2000) for results from a slightly more complex additive-plus-interaction model.

We start with the assumption that the population counts in atom k of block group i for gender g at age group a is Poisson-distributed as

$$Y_{ikga} \sim Po\left(e^{\delta_{iga}}\left(X'_{ik} + \frac{\theta}{K_i}\right)\right), \quad \text{where} \quad \delta_{iga} = \mu_i + g\alpha + \sum_{a=1}^{17} \beta_a I_a \ ,$$

g=0 for males and 1 for females, and I_a is a 0–1 indicator for age group a ($a = 1$ for ages 5-9, $a = 2$ for 10-14, etc.). The μ_i are block group-specific baselines (in our parametrization, they are the logs of the fitted numbers of males in the 0–4 age bracket), and α and the $\{\beta_a\}$ function as main effects for sex and age group, respectively. Note the α and $\{\beta_a\}$ parameters are not specific to any one block group, but rather apply to all 39 block groups in the map.

With each μ_i now corresponding only to the number of baby boys (not the whole population) in block group i, we expect its value to be decreased accordingly. Because there are 36 age-sex divisions, we modified the prior mean η_μ to $-2.5 \approx \log(3/36)$. We placed vague independent $N(0, 10^2)$ priors on α and the βs, and kept all other prior values the same as in Section 7.3.2.1. Convergence of the MCMC algorithm obtains in about 1,500 iterations.

Figure 7.15 *Imputed population densities* (persons/km^2) *by atom for the FMPC region.*

(The slowest parameters to converge are those pertaining to the edge atoms, where we have no structure data. Some parameters converge much faster: the α and β_a parameters, for example, converge by about 500 iterations.) We then ran 5,000 iterations for each of 5 chains, resulting in a final sample of 25,000.

Population interpolation results are quite similar to those outlined in Section 7.3.2.1, except that population distributions are available for each cell at any combination of age and sex. While we do not show these results here, we do include a summary of the main effects for age and sex. Table 7.1 shows the posterior medians and 2.5% and 97.5% quantiles for the α and β_a parameters. Among the β_a parameters, we see a significant negative value of β_4 (ages 20–24), reflecting a relatively small group of college-aged residents in this area. After a slight increase in the age distribution for ages 30–44, we observe increasingly negative values as a increases, indicating the expected decrease in population with advancing age.

7.4 A data assimilation example

Here, we offer a brief discussion of an example from Wikle and Berliner (2005) regarding wind vector data. The data consists of monitoring observations at one areal scale along with computer model output at a different areal scale. The objective is to fuse the data to provide inference at a third spatial scale. So, this is a misalignment or change of support problem. However, this problem also falls under the classification of data assimilation which is the topic of Section 15.1. In particular, the two sources of data are daily wind satellite data and computer model output from a weather center over the period 15 September 1996-29 June 1997. There are satellite-based wind estimates from a NASA Scatterometer (NSCAT) at 0.5 degree spatial resolution, not on a regular grid, along with National Center for Environmental Prediction (NCEP) analysis of wind direction at 2.5 degree resolution on a regular grid. The goal is to predict surface streamfunction at a resolution of 1.0 degree. Adopting our usual paradigm, [Data | Process, Parameters][Process | Parameters][Parameters] we

Figure 7.16 *Imputed population densities* (persons/km^2) *by cell for the FMPC windrose.*

have measurement data, Z, from the two sources and we let Y denote the true underlying process which operates at point level.

Let $A_i, i = 1, 2, .., n_a$ denote the .5 degree resolution grid, $B_j, j = 1, 2, ..., n_b$ the grid at 1.0 degree resolution and $C_k, k = 1, 2, ..., n_c$, the sets at 2.5 degree resolution. Though the scales are nested the sets need not be. We let $\mathbf{Z}_A = (Z(A_1), ..., Z(A_{n_a}))^T$ denote the,observations on the subgrid and $\mathbf{Z}_C = (Z(C_1), ..., Z(C_{n_c}))^T$ denote the observations on the supergrid.

Next, denote by $Y_D = \{Y(\mathbf{s}) : \mathbf{s} \in D\}$ the true spatial process and employ block averaging to obtain the true process values at the three spatial scales. That is, in general $Y(S) = \int_S Y(\mathbf{s})d\mathbf{s}$. Hence, $\mathbf{Y}_A = (Y(A_1), ..., Y(A_{n_a}))^T$ is the true subgrid wind vector process, $\mathbf{Y}_C = (Y(C_1), ..., Y(C_{n_c}))^T$ is the true supergrid process and $\mathbf{Y}_B = (Y(B_1), ..., Y(B_{n_b}))^T$ is the true process on the desired prediction grid. A simple measurement error model is introduced for the data, i.e., $\mathbf{Z}_A = \mathbf{Y}_A + \mathbf{e}_A$, where $\mathbf{e}_A \sim N(0, \sigma_a^2 I_{n_a})$ and $\mathbf{Z}_C = \mathbf{Y}_C + \mathbf{e}_C$, where $\mathbf{e}_C \sim N(0, \sigma_c^2 I_{n_c})$. Wikle and Berliner (2005) treat the two measurement error models as independent.

The remaining challenge is to align the $Y(A_i)$'s and, $Y(C_k)$ with the $Y(B_j)$'s. This requires creating a highest resolution partition (in the spirit of the previous section) by intersecting all of the sets in A, B, and C. We omit the details and encourage the reader to consult the Wikle and Berliner paper for full details.

7.5 Misaligned regression modeling

The methods of the preceding sections allow us to realign spatially misaligned data. The results of such methods may be interesting in and of themselves, but in many cases our real interest in data realignment will be as a precursor to fitting *regression* models relating the (newly realigned) variables.

For instance, Agarwal, Gelfand, and Silander (2002) apply the ideas of Section 7.2 in

Effect	Parameter	Median	2.5%	97.5%
Gender	α	0.005	–0.012	0.021
Ages 5–9	β_1	0.073	0.033	0.116
Ages 10–14	β_2	0.062	0.021	0.106
Ages 15–19	β_3	–0.003	–0.043	0.041
Ages 20–24	β_4	–0.223	–0.268	–0.177
Ages 25–29	β_5	–0.021	–0.063	0.024
Ages 30–34	β_6	0.137	0.095	0.178
Ages 35–39	β_7	0.118	0.077	0.160
Ages 40–44	β_8	0.044	0.001	0.088
Ages 45–49	β_9	–0.224	–0.270	–0.180
Ages 50–54	β_{10}	–0.404	–0.448	–0.357
Ages 55–59	β_{11}	–0.558	–0.609	–0.507
Ages 60–64	β_{12}	–0.627	–0.677	–0.576
Ages 65–69	β_{13}	–0.896	–0.951	–0.839
Ages 70–74	β_{14}	–1.320	–1.386	–1.255
Ages 75–79	β_{15}	–1.720	–1.797	–1.643
Ages 80–84	β_{16}	–2.320	–2.424	–2.224
Ages 85+	β_{17}	–2.836	–2.969	–2.714

Table 7.1 *Quantiles and significance of gender and age effects for the age-sex additive model.*

a rasterized data setting. Such data are common in remote sensing, where satellites can collect data (say, land use) over a pixelized surface, which is often fine enough so that town or other geopolitical boundaries can be (approximately) taken as the union of a collection of pixels.

The focal area for the Agarwal et al. (2002) study is the tropical rainforest biome within Toamasina (or Tamatave) Province of Madagascar. This province is located along the east coast of Madagascar, and includes the greatest extent of tropical rainforest in the island nation. The aerial extent of Toamasina Province is roughly 75,000 square km. Four georeferenced GIS coverages were constructed for the province: town boundaries with associated 1993 population census data, elevation, slope, and land cover. Ultimately, the total number of towns was 159, and the total number of pixels was 74,607. For analysis at a lower resolution, the above 1-km raster layers are aggregated into 4-km pixels.

Figure 7.17 shows the town-level map for the 159 towns in the Madagascar study region. In fact, there is an escarpment in the western portion where the climate differs from the rest of the region. It is a seasonally dry grassland/savanna mosaic. Also, the northern part is expected to differ from the southern part, since the north has fewer population areas with large forest patches, while the south has more villages with many smaller forest patches and more extensive road development, including commercial routes to the national capital west of the study region. The north and south regions with a transition zone were created as shown in Figure 7.17.

The joint distribution of land use and population count is modeled at the pixel level. Let L_{ij} denote the land use value for the jth pixel in the ith town and let P_{ij} denote the population count for the jth pixel in the ith town. Again, the L_{ij} are observed but only $P_{i.} = \sum_j P_{ij}$ are observed at the town level. Collect the L_{ij} and P_{ij} into town-level vectors $\mathbf{L}_i$ and $\mathbf{P}_i$, and overall vectors $\mathbf{L}$ and $\mathbf{P}$.

Covariates observed at each pixel include an elevation, E_{ij}, and a slope, S_{ij}. To capture spatial association between the L_{ij}, pixel-level spatial effects φ_{ij} are introduced; to capture spatial association between the $P_{i.}$, town-level spatial effects δ_i are introduced. That is, the spatial process governing land use may differ from that for population.

Figure 7.17 *Northern and southern regions within the Madagascar study region, with population overlaid.*

The joint distribution, $p(\mathbf{L}, \mathbf{P} | \{E_{ij}\}, \{S_{ij}\}, \{\varphi_{ij}\}, \{\delta_i\})$ is specified by factoring it as

$$p(\mathbf{P} \mid \{E_{ij}\}, \{S_{ij}\}, \{\delta_i\})\, p(\mathbf{L} \mid \mathbf{P}, \{E_{ij}\}, \{S_{ij}\}, \{\varphi_{ij}\}) \,. \tag{7.26}$$

Conditioning is done in this fashion in order to explain the effect of population on land use. Causality is *not* asserted; the conditioning could be reversed. (Also, implicit in (7.26) is a marginal specification for $\mathbf{L}$ and a conditional specification for $\mathbf{P} \mid \mathbf{L}$.)

Turning to the first term in (7.26), the P_{ij} are assumed conditionally independent given the E's, S's, and δ's. In fact, we assume $P_{ij} \sim$Poisson(λ_{ij}), where

$$\log \lambda_{ij} = \beta_0 + \beta_1 E_{ij} + \beta_2 S_{ij} + \delta_i \,. \tag{7.27}$$

Thus $\mathrm{P}_{i.} \sim$ Poisson$(\lambda_{i.})$, where $\log \lambda_{i.} = \log \sum_j \lambda_{ij} = \log \sum_j \exp(\beta_0 + \beta_1 E_{ij} + \beta_2 S_{ij} + \delta_i)$. In other words, the P_{ij} inherit the spatial effect associated with $P_{i.}$. Also, $\{P_{ij}\} | P_{i.} \sim$ Multinomial$(P_{i.}; \{\gamma_{ij}\})$, where $\gamma_{ij} = \lambda_{ij}/\lambda_{i.}$.

In the second term in (7.26), conditional independence of the L_{ij} given the P's, E's, S's, and φ's is assumed. To facilitate computation, we aggregate to 4 km $\times$ 4 km resolution. (The discussion regarding Figure 4.2 in Subsection 4.1.1 supports this.) Since L_{ij} lies between 0 and 16, it is assumed that $L_{ij} \sim$Binomial$(16, q_{ij})$, i.e., that the sixteen 1 km $\times$ 1 km pixels that comprise a given 4 km $\times$ 4 km pixel are i.i.d. Bernoulli random variables with q_{ij} such that

$$\log\left(\frac{q_{ij}}{1 - q_{ij}}\right) = \alpha_0 + \alpha_1 E_{ij} + \alpha_2 S_{ij} + \alpha_3 P_{ij} + \varphi_{ij} \,. \tag{7.28}$$

For the town-level spatial effects, a conditionally autoregressive (CAR) prior is assumed using only the adjacent towns for the mean structure, with variance τ_δ^2, and similarly for the pixel effects using only adjacent pixels, with variance τ_φ^2.

Figure 7.18 *Graphical representation of the land use-population model.*

To complete the hierarchical model specification, priors for $\boldsymbol{\alpha}, \boldsymbol{\beta}, \tau_\delta^2$, and τ_φ^2 (when the φ_{ij} are included) are required. Under a binomial, with proper priors for τ_δ^2 and τ_φ^2, a flat prior for $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ will yield a proper posterior. For τ_δ^2 and τ_φ^2, inverse Gamma priors may be adopted. Figure 7.18 offers a graphical representation of the full model.

We now present a brief summary of the data analysis. At the 4 km x 4 km pixel scale, two versions of the model in (7.28) were fit, one with the φ_{ij} (Model 2) and one without them (Model 1). Models 1 and 2 were fitted separately for the northern and southern regions. The results are summarized in Table 7.2, point (posterior median) and interval (95% equal tail) estimate. The population-count model results are little affected by the inclusion of the φ_{ij}. For the land-use model this is not the case. Interval estimates for the fixed effects coefficients are much wider when the φ_{ij} are included. This is not surprising from the form in (7.28). Though the P_{ij} are modeled and are constrained by summation over j and though the ϕ_{ij} are modeled dependently through the CAR specification, since neither is observed, strong collinearity between the P_{ij} and ϕ_{ij} is expected, inflating the variability of the α's.

Specifically, for the population count model in (7.27), in all cases the elevation coefficient is significantly negative; higher elevation yields smaller expected population. Interestingly, the elevation coefficient is more negative in the north. The slope variable is intended to provide a measure of the differential in elevation between a pixel and its neighbors. However, a crude algorithm is used within the ARC/INFO software for its calculation, diminishing its value as a covariate. Indeed, higher slope would typically encourage lower expected population. While this is roughly true for the south under either model, the opposite emerges for the north. The inference for the town-level spatial variance component τ_δ^2 is consistent across all models. Homogeneity of spatial variance for the population model is acceptable.

Turning to (7.28), in all cases the coefficient for population is significantly negative. There is a strong relationship between land use and population size; increased population increases the chance of deforestation, in support of the primary hypothesis for this analysis. The elevation coefficients are mixed with regard to significance. However, for both Models 1 and 2, the coefficient is always at least .46 larger in the north. Elevation more strongly encourages forest cover in the north than in the south. This is consistent with the discussion

Model:	M_1		M_2	
Region:	North	South	North	South
Population model parameters:				
β_1 (elev)	–.577 (–.663,–.498)	–.245 (–.419,–.061)	–.592 (–.679,–.500)	–.176 (–.341,.019)
β_2 (slope)	.125 (.027,.209)	–.061 (–.212,.095)	.127 (.014,.220)	–.096 (–.270,.050)
τ_{δ^2}	1.32 (.910,2.04)	1.67 (1.23,2.36)	1.33 (.906,1.94)	1.71 (1.22,2.41)
Land use model parameters:				
α_1 (elev)	.406 (.373,.440)	–.081 (–.109,–.053)	.490 (.160,.857)	.130 (–.327,.610)
α_2 (slope)	.015 (–.013,.047)	.157 (.129,.187)	.040 (–.085,.178)	–.011 (–.152,.117)
α_3 ($\times 10^{-4}$)	–5.10 (–5.76,–4.43)	–3.60 (–4.27,–2.80)	–4.12 (–7.90,–.329)	–8.11 (–14.2,–3.69)
τ_{φ_2}	—	—	6.84 (6.15,7.65)	5.85 (5.23,6.54)

Table 7.2 *Parameter estimation (point and interval estimates) for Models 1 and 2 for the northern and southern regions.*

of the preceding paragraph but, apparently, the effect is weaker in the presence of the population effect. Again, the slope covariate provides inconsistent results; but is insignificant in the presence of spatial effects. Inference for the pixel-level spatial variance component does not criticize homogeneity across regions. Note that τ_φ^2 is significantly larger than τ_δ^2. Again, this is expected. With a model having four population parameters to explain 3186 $q'_{ij}s$ as opposed to a model having three population parameters to explain 115 $\lambda'_i s$, we would expect much more variability in the $\varphi'_{ij}s$ than in the $\delta'_i s$. Finally, Figure 7.19 shows the imputed population at the 4 km $\times$ 4 km pixel level.

The approach of Section 7.3 will be difficult to implement with more than two mutually misaligned areal data layers, due mostly to the multiple labeling of atoms and the needed higher-way look-up table. However, the approach of this section suggests a simpler strategy for handling this situation. First, rasterize all data layers to a common scale of resolution. Then, build a suitable latent regression model at that scale, with conditional distributions for the response and explanatory variables constrained by the observed aggregated measurements for the respective layers.

Zhu, Carlin, and Gelfand (2003) consider regression in the point-block misalignment setting, illustrating with the Atlanta ozone data pictured in Figure 7.2. Recall that in this setting the problem is to relate several air quality indicators (ozone, particulate matter, nitrogen oxides, etc.) and a range of sociodemographic variables (age, gender, race, and a socioeconomic status surrogate) to the response, pediatric emergency room (ER) visit counts for asthma in Atlanta, GA. Here the air quality data is collected at fixed monitoring stations (point locations) while the sociodemographic covariates and response variable is collected by zip code (areal summaries). In fact, the air quality data is available as daily averages at each monitoring station, and the response is available as daily counts of visits

Figure 7.19 *Imputed population (on the square root scale) at the pixel level for north and south regions.*

in each zip code. Zhu et al. (2003) use the methods of Section 7.1 to realign the data, and then fit a Poisson regression model on this scale. Since the data also involves a temporal component, we defer further details until Subsection 11.7.4.

7.6 Exercises

1. Suppose we estimate the average value of some areal variable $Y(B)$ over a block B by the predicted value $Y(\mathbf{s}^*)$, where $\mathbf{s}^*$ is some central point of B (say, the population-weighted centroid). Prove that $Var(Y(\mathbf{s}^*)) \geq Var(Y(B))$ for any $\mathbf{s}^*$ in B. Is this result still true if $Y(\mathbf{s})$ is nonstationary?
2. Derive the form for $H_B(\boldsymbol{\phi})$ given below (7.7). (*Hint:* This may be easiest to do by gridding the B_k's, or through a limiting Monte Carlo integration argument.)
3. Suppose g is a differentiable function on $\Re^+$, and suppose $Y(\mathbf{s})$ is a mean-zero stationary process. Let $Z(\mathbf{s}) = g(Y(\mathbf{s}))$ and $Z(B) = \frac{1}{|B|}\int_B Z(\mathbf{s})d\mathbf{s}$. Approximate $Var(Z(B))$ and $Cov(Z(B), Z(B'))$. (*Hint:* Try the delta method here.)
4. Define a process (for convenience, on $\Re^1$) such that $\widehat{Y}(B)$ defined as above (7.10) does *not* converge almost surely to $Y(B)$.
5. Consider the scallop data sites formed by the rectangle having opposite vertices (73.0W, 39.5N) and (72.5W, 40.0N) (refer to Figure 7.20). This rectangle includes 20 locations; the full scallop data are provided in `www.biostat.umn.edu/~brad/data/myscallops.dat`, which includes our transformed variable, `log(tcatch+1)`.

Figure 7.20 *1993 scallop data, with rectangle over which a block average is desired.*

(a) Krige the block average of `log(tcatch+1)` for this region by simulating from the posterior predictive distribution given all of the 1993 data. Adopt the model and prior structure in Example 7.1, and use Equation (7.6) implemented through (7.12) to carry out the generation.

(b) Noting the caveats regarding vague priors mentioned just below Equation (6.3), change to a more informative prior specification on the spatial variance components. Are your findings robust to this change?

6. Suppose that Figure 7.21 gives a (nested) subdivision of the region in Figure 7.4, where we assume the disease count in each subsection is Poisson-distributed with parameter m_1 or m_2, depending on which value (1 or 2) a subregional binary measurement assumes. Suppose further that these Poisson variables are independent given the covariate. Let the observed disease counts in Region I and Region II be $y_1 = 632$ and $y_2 = 311$, respectively, and adopt independent $Gamma(a, b)$ priors for m_1 and m_2 with $a = 0.5$ and $b = 100$, so that the priors have mean 50 (roughly the average observed count per subregion) and variance 5000.

(a) Derive the full conditional distributions for m_1 and m_2, and obtain estimates of their marginal posterior densities using MCMC or some other approach. (*Hint:* To improve the numerical stability of your algorithm, you may wish to transform to the log scale. That is, reparametrize to $\delta_1 = \log(m_1)$ and $\delta_2 = \log(m_2)$, remembering to multiply by the Jacobian $(\exp(\delta_i), i = 1, 2)$ for each transformation.)

(b) Find an estimate of $E(Y_3|\mathbf{y})$, the predictive mean of Y_3, the total disease count in the shaded region. (*Hint:* First estimate $E(Y_{3a}|\mathbf{y})$ and $E(Y_{3b}|\mathbf{y})$, where Y_{3a} and Y_{3b} are the subtotals in the left (Region I) and right (Region II) portions of Region III.)

Figure 7.21 *Subregional map for motivating example.*

(c) Obtain a sample from the posterior predictive distribution of Y_3, $p(y_3|\mathbf{y})$. Is your answer consistent with the naive one obtained from equation (7.13)?

7. For the Tompkins County data, available on our website at address `www.biostat.umn.edu/~brad/data/tompkins.dat` and with supporting information on `StatLib` at `lib.stat.cmu.edu/datasets/csb/`, obtain smoothed estimates of the underlying block group-level relative risks of disease by modifying the log-relative risk model (7.14) to

$$\delta_{k(i,j)} = \theta_0 + \theta_1 u_{ij} + \theta_2 w_{ij} + \theta_3 u_{ij} w_{ij} + \phi_k \ ,$$

where we assume

(a) $\phi_k \overset{iid}{\sim} N(0, 1/\tau)$ (*global* smoothing), and

(b) $\boldsymbol{\phi} \sim CAR(\lambda)$, i.e., $\phi_k \mid \phi_{k' \neq k} \sim N\left(\bar{\phi}_k \ , \ \frac{1}{\lambda n_k}\right)$ (*local* smoothing).

Do your estimates significantly differ? How do they change as you change λ?

8. For the FMPC data and model in Section 7.3,

(a) Write an explicit expression for the full Bayesian model, given in shorthand notation in Equation (7.22).

(b) For the full conditionals for μ_i, X_{ji}, and X'_{iE}, show that the Gaussian, multinomial, and Poisson (respectively) are sensible choices as Metropolis-Hastings proposal densities, and give the rejection ratio (5.21) in each case.

Chapter 8

Modeling and Analysis for Point Patterns

8.1 Introduction

As we noted in Chapter 1, point patterns form the third type of spatial data that we collect. Of the three data types, in our view, spatial and space-time point patterns are the least developed in terms of Bayesian development and application. There is a consequential formal theoretical literature and there is by now a substantial body of exploratory tools. We shall explore both of these parts in Sections 8.2 and 8.3. In Section 8.4 we look at basic modeling specifications. Here, it is evident that in the modeling side, in particular, the hierarchical approach through fully Bayesian modeling has received much less attention. In Section 8.5 we take up the problem of generating point patterns, potentially useful for simulation-based model fitting. In Section 8.6 we extend the class of models to Neyman-Scott and Gibbs processes, again, with an eye toward Bayesian model fitting. In Section 8.7 we consider marked point processes. Section 8.8 will look at space-time point patterns, i.e., how can we learn about the evolution of point patterns in time? We conclude in Section 8.9 with a few special topics, arguably, areas which need more development for the practitioner. Several exercises are presented at the end. Also, we believe that the inferential aspects of this chapter will be best appreciated after absorbing Chapters 5 and 6.

Examples of spatial point patterns arise in various contexts. For instance, in looking at ecological processes, we may be interested in the pattern of occurrences of species, e.g., the pattern of trees in a forest, say junipers and pinions. What sort of pattern do the species reveal? Do the two species present *different* patterns? Are there environmental/habitat features which explain the observed patterns? Do the species respond differently to the available environment? In epidemiology, in particular so-called spatial epidemiology, we seek to find pattern in disease cases, perhaps different pattern for cases vs. controls. With breast cancer cases where a woman may elect one of say, two treatment options – mastectomy or radiation – do point patterns differ according to option? In *syndromic surveillance* we seek to identify disease outbreaks. Here, we would be looking for clustering of cases. Additionally, we might be looking at the evolution of a pattern over time. Examples include recent outbreaks involving e-coli, H1N1, swine flu, bird flu, and bovine tuberculosis. Another variant is to investigate the point pattern of invasive species. Here, we see a picture of locations that is not yet in equilibrium. We might be able to gather subsequent time-slices of the pattern but our primary inferential objective might be to anticipate where the species will eventually appear given where it is now. Again, environmental/habitat features that explain where it currently resides can help to address this question. Finally, we might look at the evolution/growth of a city, i.e., urban development. We could look at the pattern of development of single family homes or of commercial property over time.

Again, point patterns consider the setting where the randomness is associated with the locations of the points themselves. We are immediately led to attempt to clarify what we mean by "no spatial pattern." What do we mean by a *uniform* distribution of points? This is referred to as complete spatial randomness and will be developed in Section 8.2. Moreover, it is at the heart of several of the tools in Section 8.3. We can further imagine a collection

of point patterns, each indexed by a level of a variable, a so-called "mark," resulting in a *marked* point pattern. Conceptually, marks may be discrete or continuous. With discrete marks, natural interest would be in comparing point patterns across marks. With continuous marks, we may be more interested in the joint distribution of marks and locations. In a sense, this is the reverse of our earlier specifications where we have a measurement variable (e.g., a mark) at a given location. That is, geostatistical analysis typically seeks to infer about spatially continuous phenomena given observation at a fixed finite set of locations. In marked point pattern analysis, the randomness in locations, along with associated marks, is analyzed. We shall illuminate this point further in Section 8.7.

For a specified, bounded region D, we will denote the realization as $\mathbf{s}_i, i = 1, 2..., n$ where, again, both n and the $\mathbf{s}_i$ are random. Below, we will consider the role of D more carefully. Are we seeing a finite realization of an infinite point pattern as a result of imposing D (in which case, we might need to worry about edge effects and the shape of D might matter). Or, are we seeing a finite point pattern associated with a specified D (conceptually, perhaps an island or a forest or the limits of a city)? The modeling for these two settings will not be the same and, it may be argued that, in practice, the choice between these two settings is arbitrary. In fact, we will focus more on the second case since, arguably, it is better suited to application as it allows more flexible modeling with easier model fitting and analysis.

Again, we need not have variables at locations, just the pattern of points provided by the locations. We seek to extract relatively crude features of the patterns. Evidently, complete randomness or spatial homogeneity is a place to start, an assumption which we hope to criticize on several accounts. The first is because, realistically, it cannever be operating in practice. A second is because we seek to shed light on where there is departure from randomness and what its nature might be. Third, such departure can result from environmental features in which case we would like to develop regression models to explain why we observe the pattern that we do. Fourth, alternatively, the pattern may reveal a form of clustering or attraction, possibly of inhibition or repulsion, perhaps regular or systematic behavior which, again, we would seek to explain.

Figure 8.1 shows displays of *spatial homogeneity* for six samples each of 30 points. The plots reveal that the eye cannot easily assess complete randomness; it tends to "see" structure. Here, the analogy is with seeing functional relationship in a scatterplot. The eye will tend to respond to artifacts in the randomness. By contrast Figure 8.2 shows clustering and systematic pattern. Now, still with a small number of points, we can see real structure, real departure from homogeneity.

We can take this a bit further in Figure 8.3 where we show an *intensity* surface which is used to generate point patterns. The surface is supplied in both a perspective plot and a contour plot. We formalize the notion of an intensity in the next section but, for now, we merely conceptualize it as a surface which is such that in regions where it is high we expect to see more points and in regions where it is low we expect to see fewer points.

Realizations of point patterns from this intensity surface are shown in Figure 8.4, with some contours overlaid (dashed lines) to reveal the nature of resultant point patterns and their inherent variability.

A noteworthy remark with regard to point patterns is that often we are prevented from seeing them. Particularly with regard to public health data, points may be aggregated to geographic units, e.g., census units, zip/post codes, counties, for confidentiality/privacy reasons. That is, we may be denied the opportunity to analyze the data at point level resolution. Such aggregation returns us to counts associated with areal units which may be analyzed using the disease mapping methodology presented in Chapter 4. In this regard, we will again confront the ecological fallacy, offered in Section 8.4.3.1. Aggregating points with covariate marks to counts over units, with an associated explanatory variable for the units, typically does not arise through simple scaling as we elaborate in Section 8.4.

Figure 8.1 *The panels depict* spatial homogeneity *for six samples each of 30 points. The plots reveal that the eye cannot easily assess complete randomness and tends to look for structure.*

Figure 8.2 *Clustering and systematic (regular) pattern.*

There are various books that devote all or a portion of their space to point patterns. In particular, Cressie (1993) presents a fairly detailed, formal discussion at a fairly high technical level. Peter Diggle has made substantial contributions through his website and his book (2003), which is accessible and now a classic but is not broad. Møller and Waagepetersen (2004) offer a model driven perspective that is likelihood based and fairly technical. Waller and Gotway (2004) provide an easy read, focusing primarily on what we view as exploratory tools, little on modeling.

Figure 8.3 *Intensity surface used to generate point patterns.*

Figure 8.4 *Realizations from the intensity surface in Figure 8.3 with overlaid contours shown as dashed lines.*

We recommend the recent book by Illian, Penttinen, Stoyan, and Stoyan (2008). It offers a fair bit of technical development along with a data analytic perspective, paying attention to *modeling*. It has a fresh and engaging feel but, by its own admission, emphasizes descriptive (summary) statistics for features of the process model in order to reveal the nature of departures from complete spatial randomness. In fairness, they note that many processes for point patterns are described constructively (in terms of such features), avoiding likelihood specification or making it infeasible. They also indicate that they would prefer to go beyond the exploratory stage when possible. Evidently, consistent with the fully model-based character of this book, leading to posterior inference, we will focus on specifications where this is possible. Finally, we note the recent Handbook of Spatial Statistics (Gelfand et al., 2010) which presents a very current review of the field and devotes 161 pages spanning 7 chapters to theory and inference for point patterns.

8.2 Theoretical development

We focus on point patterns over $D \subset R^2$. Much theoretical work is over a subregion of R^d since many of the features of interest are geometric in nature, e.g., distance between points or number of points in a sphere, and these can be formulated for general d dimensional locations. Of course, there is a substantial literature for point patterns over an interval on R^1 (usually specified as $(0, T]$). These are sometimes referred to as *counting processes*, emphasizing the number of events as well as their locations (see, e.g., Andersen et al., 1995; Fleming and Harrington, 2005). Working in R^1 offers the advantage of a well-defined "history." We can build models to capture what we expect in the future, given what we have seen thus far. That is, we can view the points as arriving sequentially and model them accordingly. Alternatively, we might view the entire pattern retrospectively, once we have reached time T. If so, we might consider the event times as conditionally independent given some distribution over $(0, T]$. This latter approach is widely adopted for point patterns over R^2 since we have no ordering of the points. It is the approach below through nonhomogeneous Poisson processes, more generally, Cox processes. However, in Section 8.6 we clarify that, in order to achieve inhibition or clustering behavior, we might work with pairwise interaction specifications which remove this assumption.

There are settings in R^1 where the scale is not time but, rather, some other continuous classification such as size. For instance, in population demography with regard to trees, we can consider a point pattern of tree diameters (usually referred to as diameter at breast height – DBH) or of basal areas, observed in some region of interest. In fact, over time, we may see several censuses of such diameters. This could enable assessment of forest dynamics, for example, whether the number of individuals is changing as well as whether the distribution of size is changing. See, the recent work of Ghosh, Gelfand, and Clark (2012) in this regard.

Returning to R^2, we consider a bounded, connected subset D. We denote a random realization of a point pattern by $\mathbf{S}$ with elements $\mathbf{s}_1, \ldots, \mathbf{s}_n$. Here, $\mathbf{S}$ is random and so are any features we calculate from it. A probabilistic model for $\mathbf{S} \in D$ must place a distribution over all possible realizations in D. Evidently, this is where the modeling challenge emerges. In practice, it will be easier to specify features/functionals of this distribution than it will be to specify the distribution. Perhaps, this is not surprising since the required distribution must be over a countable set in order to provide the number of points and then, jointly over the continuous domain, D, in order to locate the set of points.

However, if we continue, in a formal sense, then, in order to specify a probabilistic model for $\mathbf{S}$, we can identify two ingredients. One is the distribution for $N(D)$, the number of points in D. Evidently, this is a distribution over the set $n \in \{0, 1, .., \infty\}$. The second is, for any n, a multivariate density over D^n, for any n, say, $f(\mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_n)$. We will call f a *location density* and, since points are unordered/unlabeled, f must be symmetric in its arguments. The implication is that, with $\partial\mathbf{s}$ denoting an arbitrarily small circular neighborhood around $\mathbf{s}$,

$$\begin{aligned} P(N(\partial\mathbf{s}_1) = 1, N(\partial\mathbf{s}_2) = 1, \ldots, N(\partial\mathbf{s}_n) = 1 \mid N(D) = n) \\ \approx f(\mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_n) \prod_i |\partial\mathbf{s}_i| \,, \end{aligned} \tag{8.1}$$

where $|\partial\mathbf{s}|$ denotes the area of $\partial\mathbf{s}$. An additional implication is that the likelihood will take the form:

$$L(\mathbf{S}) = P(N(D) = n) n! f(\mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_n). \tag{8.2}$$

The first term on the right side of (8.2) appears because the number of points is random. The $n!$ appears in (8.2) because the unordered points can be assigned to the n locations in $n!$ ways. A further thought at this point is that, according to the above, we need to specify

f consistently over all $\mathbf{S}$. This sort of consistency condition reminds us of the requirements for providing a model for point-referenced data back in Chapter 2. If we think of $N(D)$ fixed at n, then we only need to provide a finite n-dimensional distribution for f, as with the discrete spatial data models in Chapter 4. In fact, both paths will be discussed below in attempting to supply stochastic models for $\mathbf{S}$.

We can now define a stationary point pattern model. If $f(\mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_n) = f(\mathbf{s}_1 + \mathbf{h}, \mathbf{s}_2 + \mathbf{h}, \ldots, \mathbf{s}_n + \mathbf{h})$ for all n, $\mathbf{s}_i$, and $\mathbf{h}$, we say that the point process is stationary. This condition would naturally be proposed over R^2, in which case it would hold, suitably, over D. Note again, that stationarity is a model property, not a model specification.

8.2.1 Counting measure

Given a point pattern, arguably the simplest feature to think about is the number of points in a specified set. To this end, analogous to the definition of $N(D)$, we introduce count variables, $N(B)$ where, for set B, $N(B)$ is number of points in set B. That is, $N(B) = \sum_{\mathbf{s}_i \in \mathbf{S}} 1(\mathbf{s}_i \in B)$. Note that $N(B)$ is computed by looking at the points in $\mathbf{S}$ individually, which we will call a first order property. This is opposed to looking at objects based upon say, pairs of points, which we will call a second order property below. Since the number of sets B is uncountable, formally, we will need to specify a counting measure over a $\sigma-$algebra of measurable sets. Moreover, since the point pattern is random, so are the $N(B)$. We need a random counting measure in order to specify a joint model for an uncountable number of random counts. We formalize this through finite dimensional distributions, i.e., the joint distribution for a finite set of count variables. Initially, we do this through Poisson processes, as is customarily done due to the convenient distribution theory but in Section 8.6 we consider more general cluster/mixture and Gibbs process models. We also briefly digress to consider first and second order moment measures.

We recall the definition of a Poisson process over a set D, driven by the intensity function $\lambda(\mathbf{s})$. For $B \subseteq D$, $N(B) \sim \text{Po}(\lambda(B))$ where $\lambda(B) = \int_B \lambda(\mathbf{s}) d\mathbf{s}$.[1] In addition, if B_1 and B_2 are disjoint, then $N(B_1)$ and $N(B_2)$ are independent (see Cressie, 1993, p. 620 or Illian et al., 2008 p. 118). This definition clarifies the need for a bounded set; otherwise, for some B's the integral could be ∞. We also note that we can define the random Poisson measure induced by $\lambda(\mathbf{s})$. We may view this as $\lim_{\partial \mathbf{s} \to 0} \frac{N(\partial \mathbf{s})}{|\partial \mathbf{s}|} = N(\mathbf{s})$ or equivalently, $N(B) = \int_B N(\mathbf{s}) d\mathbf{s}$

Evidently, $E(N(B)) = \text{var}(N(B)) = \lambda(B)$. The independence of disjoint sets immediately implies that $f(\mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_n) = \Pi_i f(\mathbf{s}_i) = \Pi_i \lambda(\mathbf{s}_i)/\lambda(D)$ where $\lambda(D) = \int_D \lambda(\mathbf{s}) d\mathbf{s}$. In fact, $P(N(\partial \mathbf{s}) = 1) \approx E(N(\partial \mathbf{s})) = \lambda(\partial \mathbf{s}) \approx \lambda(\mathbf{s})|\partial \mathbf{s}|$ which is usually written as $\lambda(\mathbf{s}) d\mathbf{s}$.

This formalization allows us to specify the notion of complete spatial randomness, equivalently, spatial homogeneity. This arises when $\lambda(\mathbf{s}) = \lambda$, i.e., we have a constant surface over D and we refer to this as a homogeneous Poisson process (HPP). Evidently, $\lambda(B) = \lambda|B|$ where $|B|$ is the area of B, that is, the expected number of points in set B is proportional to the area of B. The total number of points expected over D is $\lambda|D|$.

From a different perspective, we note that stationarity of the process implies that $\lambda(\mathbf{s}) = \lambda$ for all $\mathbf{s}$ and thus, $\lambda(B) = \lambda|B|$ for all $B \subseteq D$. This is evident since stationarity implies that $f(\mathbf{s}) = f(\mathbf{s} + \mathbf{h})$ for all $\mathbf{s}$ and $\mathbf{h}$. In different terms, if $\lambda(B) = \lambda(B + \mathbf{h})$ for all $B \subseteq D$ and $\mathbf{h}$, then standard real analysis shows that $\lambda(\mathbf{s})$ is constant, i.e., the unique measure satisfying this condition is proportional to Lebesgue measure. It is also clear that $f(\mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_n) = 1/|D|^n$.

It is important to emphasize that the HPP is only one (arguably, the simplest, nontrivial) stationary process specification. It is the one that specifies a constant intensity with conditionally independent locations. More general models include the stationary Gibbs processes

[1] It is worth noting that this is a double integral over two-dimensional locations and, for circular regions it may be more easily calculated using polar coordinates.

which we consider in Section 8.6.3. Also, as we shall see below, with stationary processes, one can ask questions regarding so-called *typical* points. A typical point is one of the locations in a point pattern realization, $\mathbf{S}$. Under stationarity, the questions address issues such as the probability of no other points in $\mathbf{S}$ within a specified distance of a typical point or the expected number of points in $\mathbf{S}$ within a specified distance from a typical point. We discuss these ideas further in Section 8.3. Of course, stationarity, hence, a constant intensity would be appropriate only for certain types of data collection. In particular, customary applications would be to physical processes in a homogeneous environment, for example, interacting particle models.

In most environmental data collection settings, it is anticipated that the environment will not be homogeneous and that the heterogeneity of the environment affects the first order intensity of the process. Within the Poisson process setting, this suggests the more general case when $\lambda(\mathbf{s})$ is not constant which we refer to the model as a nonhomogeneous Poisson process (NHPP), in some literature, an *inhomogeneous* Poisson process. So, the NHPP is not stationary. Again, there are nonstationary process beyond the NHPP; the NHPP has conditionally independent locations with local density, $f(\mathbf{s}) = \lambda(\mathbf{s})/\lambda(D)$.

When $\lambda(\mathbf{s})$ is random we refer to the NHPP as a Cox process. For us, we should take care with regard to this distinction since, from a Bayesian inference perspective, whenever $\lambda(\mathbf{s})$ is unknown, it will be assumed to be random. The Cox process notion thinks of $\lambda(\mathbf{s})$ as a realization of a stochastic process. For example, $\log\lambda(\mathbf{s})$ might be a realization from a Gaussian process, say, restricted to D. Then, the Cox process is, in fact, the marginal process obtained by integrating over the randomness in $\lambda(\mathbf{s})$. But, clearly it is a hierarchical (two-stage) specification, i.e., draw $\lambda(\mathbf{s})$ and then draw $\mathbf{S}$ given $\lambda(\mathbf{s})$; so is naturally amenable to Bayesian inference.

We now formalize the notion of counting measure. For any set $B \subset D$, let $N(B)$ count the number of points in B. Then, $N(B) \in \{0, 1, 2, ...\}$. We say that $N(B)$ is a counting measure if $N(B) < \infty, \forall B \in \mathcal{B}$ where $\mathcal{B}$ is a $\sigma-$algebra of sets intersected with D. Counting measure satisfies the usual countable additivity property thus enabling $N(B_1 \cup B_2)$ and $N(B_1 \cap B_2)$. In particular, we can argue that a realization of a point pattern is equivalent to a realization of a counting measure. That is, given the point pattern, we can immediately assign $N(B)$ to all of the measurable sets. But, conversely (and informally), given $N(B)$ for a $\sigma-$algebra of sets, by selecting a suitable increasing or decreasing sequence of sets we can identify the location of a point through the change in counting measure. Again, $\{N(\cdot)\}$ over $\mathcal{B}$ is random with a distribution induced by that for $\mathbf{S}$ over $\mathcal{S}$ (where $\mathcal{S}$ denotes the set of all possible point pattern realizations over D); conversely, a distribution specified over the $\{N(\cdot)\}$ over the $\mathcal{B}$ would induce a distribution for $\mathbf{S}$ over $\mathcal{S}$. In any event, we refer to either collection of variables as a *spatial point process*. Evidently, a NHPP is an example of a spatial point process. In fact, it is an example where the distribution over the space of point patterns and the distribution over the σ-algebra of sets can both be written down explicitly. In more general cases, we specify the distribution in (8.2), through a distribution for $N(D)$ and a location density, $f(\mathbf{s}_1, \ldots, \mathbf{s}_n)$. (In some cases, such as in Section 8.6.1, we specify the process constructively, without an explicit location density.)

8.2.2 Moment measures

Next, we turn to the notion of moment measures, also product densities, as in Illian et al. (2008). We start with first order properties, i.e., the first moment measure, $\{E(N(B)): B \in \mathcal{B}\}$. With regard to this collection of sets, given $\lambda(\mathbf{s})$, we can compute $E(N(B))$ in the usual way, i.e., $E(N(B)) = \int_B \lambda(\mathbf{s})d\mathbf{s}$. However, given that the collection, $\{E(N(B)): B \in \mathcal{B}\}$, is a measure, can we extract the *first-order* intensity? The approach is to take limits in the form, $\lambda(\mathbf{s}) = \lim_{|\partial\mathbf{s}|\to 0} \frac{E(N(\partial\mathbf{s}))}{|\partial\mathbf{s}|}$, where, as above, $\partial\mathbf{s}$ is a neighborhood of $\mathbf{s}$.

The intuition is that $E(N(\partial\mathbf{s})) = \int_{\partial\mathbf{s}} \lambda(\mathbf{s}')d\mathbf{s}' \approx \lambda(\mathbf{s})|\partial\mathbf{s}|$. An analogy may be made with obtaining the probability density function given the cumulative distribution function. That is, we "build up" from $\lambda(\mathbf{s})$, "scale down" from $\{E(N(B)) : B \in \mathcal{B}\}$. Also, we note that, if $f(\mathbf{s}_1, \ldots, \mathbf{s}_n) = \Pi_i f(\mathbf{s}_i)$, then $\lambda(\mathbf{s}) = f(\mathbf{s})\lambda(D)$ and this conclusion does not depend upon a NHPP assumption. In fact, by the conditional independence of the locations, we see that, given $N(D) = n$, $N(B) \sim Bi(n, P(B))$ where $P(B) = \int_B f(\mathbf{s})d\mathbf{s}$. So, $E(N(B)) = E(E(N(B)|N(D) = n)) = E(nP(B)) = E(n\int_B f(\mathbf{s})d\mathbf{s}) = \int_B E(n)f(\mathbf{s})d\mathbf{s} = \int_B \lambda(D)f(\mathbf{s})d\mathbf{s}$ and thus, $\lambda(\mathbf{s}) = f(\mathbf{s})\lambda(D)$.

A remarkably elegant result for computing certain expectations with regard to point patterns is Campbell's Theorem. We start with a first order version whose proof is left as an exercise. It will often be of interest to compute features associated with individual points, e.g., as above, whether or not they are in a given set or whether or not they are within a specified distance from a given point. If the feature is denoted by $g(\mathbf{s})$, then we might be interested in the value of this feature summed over the points in $\mathbf{S}$, i.e., $\sum_{\mathbf{s}_i \in \mathbf{S}} g(\mathbf{s}_i)$. Campbell's Theorem provides the expectation of this variable, i.e.,

$$E_{\mathbf{S}}\left(\sum_{\mathbf{s}_i \in \mathbf{S}} g(\mathbf{s}_i)\right) = \int g(\mathbf{s})\lambda(\mathbf{s})d\mathbf{s}. \tag{8.3}$$

Evidently, applied to the indicator of whether $\mathbf{s}_i \in B$, we obtain $E(N(B)) = \int_B \lambda(\mathbf{s})d\mathbf{s}$ which, as a result, enables the proof of the theorem. Moreover, whenever $g(\mathbf{s})$ is applied only over D, from above, the right side of (8.3) becomes $|D|E(g(\mathbf{s}))$.

For second-order properties, we need to consider the set $\{E(N(B_1)N(B_2)) : B_1, B_2 \in \mathcal{B}\}$, again as a measure. Now, the object that we seek is the second-order (factorial) intensity, $\gamma(\mathbf{s}, \mathbf{s}')$, such that $E(N(B_1)N(B_2)) = \int_{B_1}\int_{B_2} \gamma(\mathbf{s}, \mathbf{s}')d\mathbf{s}'d\mathbf{s}$. Again, given $\gamma(\cdot, \cdot)$, we can integrate to obtain the set of product expectations. Now, the converse finds us seeking to extract $\gamma(\mathbf{s}, \mathbf{s}')$ from the measure defined by set of product expectations. Again, we use a limiting form, i.e., $\gamma(\mathbf{s}, \mathbf{s}') = \lim_{|\partial\mathbf{s}|\to 0, |\partial\mathbf{s}'|\to 0} \frac{E(N(\partial\mathbf{s})N(\partial\mathbf{s}'))}{|\partial\mathbf{s}||\partial\mathbf{s}'|}$. Again, we build up from $\gamma(\mathbf{s}, \mathbf{s}')$, scale down from $\{E(N(B_1)N(B_2)) : B_1, B_2 \in \mathcal{B}\}$. Analogous to the first order case, if $f(\mathbf{s}_1, \ldots, \mathbf{s}_n) = \Pi_i f(\mathbf{s}_i)$, then $\gamma(\mathbf{s}, \mathbf{s}') = \lambda(\mathbf{s})\lambda(\mathbf{s}') = \lambda^2(D)f(\mathbf{s})f(\mathbf{s}')$. We leave this proof as an exercise. We sometimes talk of the *reweighted* second-order intensity as $\gamma(\mathbf{s}, \mathbf{s}')/\lambda(\mathbf{s})\lambda(\mathbf{s}')$, also known as the *general pair correlation function*. Under stationarity this correlation function simplifies to $\gamma(\mathbf{s}, \mathbf{s}')/\lambda^2$ and, in fact, equals 1 under CSR.

More generally, we have that

$$\mathrm{E}[N(B_1)N(B_2)] = \mathrm{E}[N(B_1)]\mathrm{E}[N(B_2)] + \mathrm{var}[N(B_1 \cap B_2)]$$

since $\mathrm{cov}(N(B_1), N(B_2)) = \mathrm{var}(N(B_1 \cap B_2))$. This result simplifies to $\mathrm{E}(N(B_1)N(B_2)) = \mathrm{E}(N(B_1))\mathrm{E}(N(B_2))$ when B_1 and B_2 are disjoint.

There is a bivariate version of Campbell's Theorem as follows. Suppose g is a feature which is a function of two arguments, $g(\mathbf{s}, \mathbf{s}')$, e.g., the distance between $\mathbf{s}$ and $\mathbf{s}'$. Suppose we are interested in the value of this feature summed over pairs of point in $\mathbf{S}$, $\sum_{\mathbf{s}_i, \mathbf{s}_j \in \mathbf{S}, i \neq j} g(\mathbf{s}_i, \mathbf{s}_j)$. Then, Campbell's theorem provides the expectation of this variable over $\mathbf{S}$, i.e.,

$$E_{\mathbf{S}}\left(\sum_{\mathbf{s}_i, \mathbf{s}_j \in \mathbf{S}, i \neq j} g(\mathbf{s}_i, \mathbf{s}_j)\right) = \int\int g(\mathbf{s}, \mathbf{s}')\gamma(\mathbf{s}, \mathbf{s}')d\mathbf{s}d\mathbf{s}'. \tag{8.4}$$

This result can again be built up from indicator functions and, in fact, when $g(\mathbf{s}, \mathbf{s}') = 1(\mathbf{s} \in B_1, \mathbf{s}' \in B_2)$, we obtain

$$E(N(B_1)N(B_2)) = \int_{B_1}\int_{B_2} \gamma(\mathbf{s}, \mathbf{s}')d\mathbf{s}'d\mathbf{s} + \int_{B_1 \cap B_2} \lambda(\mathbf{s})d\mathbf{s}\ .$$

If the spatial point process is stationary, then $\gamma(\mathbf{s},\mathbf{s}') = \gamma(\mathbf{s}-\mathbf{s}')$. If $\gamma(\mathbf{s},\mathbf{s}') = \gamma(||\mathbf{s}-\mathbf{s}'||)$, we say that the spatial point process is isotropic. We link an isotropic γ to the K-function in Section 8.3.3 below.

Further insight can be obtained by noticing that, if $\partial\mathbf{s}$ is sufficiently small, $P(N(\partial\mathbf{s}) > 1)$ will be negligible so $E(N(\partial\mathbf{s})) \approx P(N(\partial\mathbf{s}) = 1) \approx P(N(\partial\mathbf{s}) > 0)$. Similarly, $E(N(\partial\mathbf{s})N(\partial\mathbf{s}')) \approx P(N(\partial\mathbf{s}) > 0, N(\partial\mathbf{s}') > 0)$. But, since $E(N(\partial\mathbf{s})N(\partial\mathbf{s}')) \approx \gamma(\mathbf{s},\mathbf{s}')|\partial\mathbf{s}||\partial\mathbf{s}'|$, we find that $\gamma(\mathbf{s},\mathbf{s}') \approx P(N(\partial\mathbf{s}) > 0, N(\partial\mathbf{s}') > 0)/|\partial\mathbf{s}||\partial\mathbf{s}'|$. That is, $\gamma(\mathbf{s},\mathbf{s}')|\partial\mathbf{s}||\partial\mathbf{s}'|$ is the probability of a point of $\mathbf{S}$ in $\partial\mathbf{s}$ and a point of $\mathbf{S}$ in $\partial\mathbf{s}'$, with an evident limiting interpretation as an intensity. For instance, if γ is isotropic, $\gamma(||\mathbf{s}-\mathbf{s}'||)$ can be interpreted loosely as the density for inter-point distances.

We conclude this subsection with a brief discussion of the conditional or Papangelou intensity. Consider the notation, $\lambda(\mathbf{s}|\mathbf{S})$ where $\mathbf{s}$ is a fixed location and $\mathbf{S}$ is a realization of the point process. How might we interpret this function for a given location $\mathbf{s}$ and a given realization $\mathbf{S}$? As above, $\lambda(\partial\mathbf{s}|\mathbf{S}) \approx \lambda(\mathbf{s}|\mathbf{S})|\partial\mathbf{s}|$ But also, $\lambda(\partial\mathbf{s}|\mathbf{S}) \approx P(N(\partial\mathbf{s}) = 1 \mid \mathbf{S})$. That is, $\lambda(\partial\mathbf{s}|\mathbf{S})$ is roughly the probability that there is a point of $\mathbf{S}$ in $\partial\mathbf{s}$ and the rest of the $\mathbf{s}_i \in \mathbf{S}$ lie outside of $\partial\mathbf{s}$. Formally, this suggests that, with n random, we view

$$\lambda(\partial\mathbf{s}|\mathbf{S}) = \int_{\partial\mathbf{s}} \frac{f(\mathbf{u},\mathbf{s}_1,\ldots,\mathbf{s}_n)}{f(\mathbf{s}_1,\ldots,\mathbf{s}_n)} d\mathbf{u} \approx \frac{f(\mathbf{s},\mathbf{s}_1,\ldots,\mathbf{s}_n)}{f(\mathbf{s}_1,\ldots,\mathbf{s}_n)}|\partial\mathbf{s}|. \tag{8.5}$$

So, in the limit $\lambda(\mathbf{s}|\mathbf{S}) = \dfrac{f(\mathbf{s},\mathbf{s}_1,\ldots,\mathbf{s}_n)}{f(\mathbf{s}_1,\ldots,\mathbf{s}_n)}$ presuming the denominator is not zero. It may be shown that $\lambda(\mathbf{s}) = E_{\mathbf{S}}(\lambda(\mathbf{s}|\mathbf{S}))$ (see Exercise 5 at the end of this chapter or refer to Møller and Waagepetersen, 2007). Evidently, for conditionally independent locations $\lambda(\mathbf{s}|\mathbf{S}) = f(\mathbf{s}) = \lambda(\mathbf{s})/\lambda(D)$. The conditional intensity will also take a convenient explicit form for Gibbs processes (pairwise interaction processes) as we show in Section 8.6.3 which also facilitates simulation of these processes (and will connect with the Markov random field theory presented in Chapter 4). In general, $\lambda(\mathbf{s}|\mathbf{S})$ is random since $\mathbf{S}$ is random. We leave for an exercise the proof that $E_{\mathbf{S}}(\lambda(\mathbf{s}|\mathbf{S})) = \lambda(\mathbf{s})$.

Returning to the second moment measure or product intensity, using the foregoing infinitesimal argument, we can view $\gamma(\mathbf{s},\mathbf{s}')/\lambda(\mathbf{s}')$ as the intensity at $\mathbf{s}$ given a point at $\mathbf{s}'$. In particular, under stationarity, we can write $\gamma(\mathbf{s}-\mathbf{s}')/\lambda = \pi(\mathbf{s}-\mathbf{s}')$ where $\pi(\mathbf{s}-\mathbf{s}')|\partial\mathbf{s}| \approx P(N(\partial\mathbf{s}) > 0)|N(\partial\mathbf{s}') > 0)$, i.e., the probability that there is a point in $\partial\mathbf{s}'$ given there is a point in $\partial\mathbf{s}$.

8.3 Diagnostic tools

Here, we take up what has, historically, been the "bread and butter" of point pattern data analysis and is incorporated into most of the software for analyzing point pattern data. Arguably the best at present, in this regard, is "spatstat, An R package for spatial statistics," A. Baddeley and R. Turner (2005), which we use in conjunction with several of the examples below. We look at approaches for studying departure from spatial homogeneity. In particular, we look at the traditional distance based approaches, G-functions, F-functions, and K-functions. We also look at empirical intensity estimates. Again, we view this work as exploratory. Hence, our review is brief; there are many other texts that develop this material in full detail. See, e.g., Illian et al. (2008). A useful perspective here is to recognize that all of these exploratory tools (and many further ones which we do not present here) are descriptive or summary measures from a sample to investigate a process feature. They are nonparametric in nature; in the above cases, they are based upon first or second order properties, taking forms that are analogues of empirical c.d.f.'s, typically offering no associated uncertainty.

In other words, if you assume less, say, only the existence of second order measures, then you return less with regard to inference. As noted in Section 8.1, from the Bayesian point of view, we will require the full probabilistic specification for the model. Then, if we can fit the model, we will enjoy the full Bayesian benefit – posterior inference for any model features of interest. These features are merely functionals of the specification. Hence, we experience the usual pluses and minuses of working within the Bayesian paradigm. On the plus side, fitting models with sampling-based methods yields posterior samples for model unknowns which can be converted to Monte Carlo approximations for these features. On the minus side, MCMC model fitting can be very challenging for many of these models. Some of the process specifications include intractable normalizing constants which are functions of the model parameters; for others, specifications are provided constructively, not yielding an explicit likelihood. We discuss these issues in detail in Section 8.6 below.

8.3.1 Exploratory data analysis; investigating complete spatial randomness

The most elementary way to investigate complete spatial randomness (CSR) proceeds from the definition of an HPP. In particular, if we look at a collection of cells in D, all of equal area, mutually disjoint, not necessarily exhaustive, then, given λ, the observed counts for these cells are i.i.d. Hence, we can compute $\bar{N}$, the mean of the cell counts. As well, we can compute S_N^2, the sample variance of the counts. Then, if we look at $S_N^2/\bar{N}$, under CSR we would expect this to be near one; substantial departure would criticize CSR. Other functions of $\bar{N}$ and S_N^2 have been considered in the literature (see, e.g., Cressie, 1993, p. 590) with hypothesis tests proposed. However, confining our diagnostics to these two moment estimates seems to drastically discard information in the point pattern.

In the same spirit, given a collection of counts, we can examine the i.i.d. Poisson assumption through a chi-square test, i.e., comparing, respectively, across the collection of cells, the observed number of 0's, 1's, etc., with the expected number of 0's, 1's, etc. We would do some sort of aggregation for the large counts, develop an empirical estimate of λ (e.g., total count divided by total area), and employ a χ^2 distribution with degrees of freedom one less than the total number of cells for the χ^2 statistic (again, see Cressie, 1993, pp. 688-689).

8.3.2 G and F functions

As suggested in Section 8.2.1 (and below), if we confine ourselves to stationary processes, we can think in terms of typical points and notions like the probability of being within distance d of a typical point or the expected number of points within distance d of a typical point. In what follows, we formally define these process features and then provide the customary sample estimates for them. The recurring theme here is that, under CSR, these notions have simple explicit forms. The goal of the exploratory data analysis is to see if the sample estimate supports CSR or not. It is important to appreciate that these conceptual quantities are only sensible for stationary processes. For nonstationary processes, we will be focused on estimating the associated first and, possibly, second order characteristics (Section 8.3.4 below).

Let us view the process over all of $\Re^2$, i.e., a countably infinite point pattern. Consider the random variable $N(\mathbf{s}, d; \mathbf{S})$, where $\mathbf{s} \in \mathbf{S}$, $\partial_d \mathbf{s}$ is a circle of radius d centered at $\mathbf{s}$, and N counts the number of points in the circle from $\mathbf{S}$, excluding $\mathbf{s}$. By stationarity, $N(\mathbf{s}, d; \mathbf{S}) \sim N(\mathbf{0}, d; \mathbf{S} - \mathbf{s})$, where $\mathbf{S} - \mathbf{s}$ is the translation of $\mathbf{S}$ by $\mathbf{s}$. This distributional result is equivalent to saying that every point in $\mathbf{S}$ is a *typical* point, in the sense that each one can be viewed as equivalent to $\mathbf{0}$ under translation.

Under restriction to a bounded set D, consider

$$E_{\mathbf{S}}\left(\sum_{\mathbf{s}_i \in \mathbf{S}, \mathbf{S} \in D} 1(N(\mathbf{s}_i, d; \mathbf{S}) > 0)\right) = \lambda |D| P(N_D(\mathbf{s}, d; \mathbf{S}) > 0). \tag{8.6}$$

The right side of (8.6) follows by noting that the expectation over $\mathbf{S}$ can be calculated in two stages, over $\mathbf{S}$ given $N(D)$ and then over $N(D)$. Here, $N_D(\mathbf{s}, d; \mathbf{S})$ is the count under restriction of the random $\mathbf{S}$ to D and, using the left side of (8.6), we see an obvious Monte Carlo integration for it. Moreover, it is clear that this integration arises in two stages. First, $\mathbf{S}$ is sampled, then N is calculated, given $\mathbf{S}$. Again, note that to obtain (8.6) requires restriction to a bounded set D and enables a Monte Carlo integration for $P(N_D(\mathbf{s}, d; \mathbf{S}) > 0)$, not for $P(N(\mathbf{s}, d; \mathbf{S}) > 0) \geq P(N_D(\mathbf{s}, d; \mathbf{S}) > 0)$. Empirical estimation of the latter requires an *edge correction* (see below).

In the literature, $P(N(\mathbf{s}, d; \mathbf{S}) > 0)$ is denoted by $G(d)$. That is, this probability does not depend upon $\mathbf{s}$, consistent with the notion of a typical point. We see that $G(d)$ increases in d and, in fact, can be viewed as a c.d.f. in distance d. An alternative to $G(d)$ in the literature is $F(d)$ where now $N(\mathbf{s}, d; \mathbf{S})$ would assume $\mathbf{s}$ is not in $\mathbf{S}$. We might distinguish these two definitions of event N by subscript, say N_G and N_F, (sometimes, N^- and N). More importantly, $G(d)$ need not equal $F(d)$. For instance, inhibition might preclude two points in $\mathbf{S}$ from being within distance d of each other. Comparison of estimates of G and F may be informative.

More informally, think of $G(d)$ as the "nearest neighbor" distribution, i.e., the c.d.f. of the nearest neighbor distance, event to event, i.e., at an observed event, $G(d) = Pr(\texttt{nearest event} \leq d)$. Similarly, think of $F(d)$ as the "empty space" distribution, i.e., for an arbitrary location, the c.d.f. of the nearest neighbor distance, point to event, $F(d) = Pr(\texttt{nearest event} \leq d)$. Under CSR, $G(d) = F(d) = 1 - \exp(-\lambda\pi d^2)$. That is, given d, the outcome $\texttt{nearest event} \leq d$ occurs if there is at least one event within a circle of radius d. With constant intensity λ, we know that the number of events in this circle follows a $\texttt{Po}(\lambda\pi d^2)$ distribution from which the expression for $G(d) = F(d)$ follows. The above reveals that, if $X \sim G$, then $Z = \pi X^2 \sim \texttt{Exp}(\lambda)$ from which it is easy to obtain the mean and variance of the distribution G. We leave this as an exercise. Moreover, since $2\pi\lambda X^2 \sim \chi^2_2$, we note that G places a lot of mass on small distances. We expect to see some clustering under CSR.

The empirical c.d.f. for G, $\hat{G}(d)$, arises from the n nearest neighbor distances (nearest neighbor distance for $\mathbf{s}_1$, for $\mathbf{s}_2$, etc.). Denote this set by $\{d_1, d_2, ..., d_n\}$. The empirical c.d.f. for F is different since the number of "points" is arbitrary. That is, $\hat{F}(d)$ is the empirical c.d.f. arising from the m nearest neighbor distances associated with a randomly selected set of m points in D. Evidently, m is arbitrary and $\hat{G} \neq \hat{F}$ though, as above, we will be interested in looking at the difference.

With restriction to D, it is clear that we will need an edge correction if, for $\mathbf{s}_i$, $d > b_i$, where b_i is the *distance* from $\mathbf{s}_i$ to edge of D. Introducing this into the empirical c.d.f., we take

$$\hat{G}(d) = \frac{\sum_i I(d_i \leq d < b_i)}{\sum_i I(d < b_i)} \tag{8.7}$$

with a similar form for $\hat{F}$. The rationale for (8.7) is that, if $d > b_i$, then the event $\{d_i < d\}$ is not observed. Evidently, (8.7) is only sensible when d is not large.

Comparison of $\hat{G}$ and $\hat{F}$ with $G = F$ under CSR, is usually through a customary theoretical Q-Q plot.[2] Shorter tails suggest clustering/attraction, i.e., nearest neighbor distances are shorter than expected. Longer tails suggest inhibition/repulsion, i.e., nearest neighbor distances are longer than expected. Another potentially useful function is the J function, $J(d) = \frac{1-G(d)}{1-F(d)}$. This function avoids comparison with CSR (though it equals 1 in that case), rather bringing, upon reflection, the interpretation of clustering for $J(d) < 1$ and inhibition for $J(d) > 1$. $\hat{J}(d) = \frac{1-\hat{G}(d)}{1-\hat{F}(d)}$ is the customary estimate of $J(d)$. Clearly, diagnostic tools, the

[2]Note that $\hat{G}$ and $\hat{F}$ are free of λ while the expression for $G = F$ requires it. However, it is simple to obtain quantiles of G with say $\lambda = 1$ to compare with the ordered d_i.

F and G functions provide more information than the foregoing χ^2 test. We do note that, technically, $\hat{G}(d)$ is not exactly an empirical c.d.f. since the d_i's are not independent. The same is true for $\hat{F}$. For formal testing of CSR, using say a Cramer-Von Mises test statistic, this would be problematic and perhaps a version of a Monte Carlo test would be preferred. However, with EDA intentions, perhaps this issue can be ignored.

8.3.3 The K function

Again, for a stationary process, continuing with the notation from Section 8.3.2, now consider $E(N(\mathbf{s}, d; \mathbf{S})$ which is the expected number of points in $\partial_d \mathbf{s}$, a circle of radius d, centered at $\mathbf{s}$, when $\mathbf{s} \in \mathbf{S}$ but not including $\mathbf{s}$. Using the foregoing notation and a similar calculation, we have

$$E_{\mathbf{S}} \sum_{\mathbf{s}_i \in \mathbf{S}, \mathbf{S} \in D} N(\mathbf{s}_i, d; \mathbf{S}) = \lambda |D| E(N_D(\mathbf{s}, d; \mathbf{S})). \tag{8.8}$$

We have formalized an expectation of interest. That is, with respect to the set D, the right side is the expected number of points from a random $\mathbf{S}$ within distance d of a point in $\mathbf{S}$. Again, the left side of the equality motivates a natural Monte Carlo integration. Again, empirical estimation of $E(N(\mathbf{s}, d; \mathbf{S})) \geq E(N_D(\mathbf{s}, d; \mathbf{S}))$ requires edge correction (see below). $E(N(\mathbf{s}, d; \mathbf{S})$ does not depend on $\mathbf{s}$. E and P with respect to $N(\mathbf{s}, d; \mathbf{S})$ are referred to as Palm characteristics, features of so-called Palm distributions (see Daley and Vere-Jones, 2003, for development). In the literature, we find $E(N(\mathbf{s}, d; \mathbf{S}) \equiv \lambda K(d)$.

The K function was introduced by Ripley (1977) and its estimator (there are several in the literature by now; see Illian et al., 2008, p. 231) is a widely used descriptive statistic. Informally, rather than nearest neighbor distance, the K function considers the *expected number* of points within distance d of an arbitrary point. Under stationarity, this expectation is the same for any point. Under CSR we can calculate it explicitly.

$$K(d) = (\lambda)^{-1} \mathrm{E}(\text{number of points within } d \text{ of an arbitrary point})$$

The scaling by $1/\lambda$, along with stationarity, enables us to have $K(d)$ free of λ. Under CSR, $K(d) = \lambda \pi d^2 / \lambda = \pi d^2$, i.e., the area of a circle of radius d. A customary estimate of $K(d)$ is

$$\hat{K}(d) = (\hat{\lambda})^{-2} \sum_i \sum_{j \neq i} 1(d_{ij} \equiv ||\mathbf{s}_i - \mathbf{s}_j|| \leq d)/n \tag{8.9}$$

which may be written as $(n\hat{\lambda})^{-1} \sum_i r_i$ where r_i is the number of $\mathbf{s}_j$ within d of $\mathbf{s}_i$. Here, $\hat{\lambda} = n/|D|$ where $|D|$ is the area of D.

Once again, an edge correction is needed when $\mathbf{s}_i$ is too near the boundary of D. More precisely, in (8.9), we place $\frac{1}{w_{ij}}$ inside the double sum. w_{ij} is the conditional probability that an event is in D given that it is exactly distance d_{ij} from $\mathbf{s}_i$. This probability is approximated as the proportion of the circumference of the circle centered at $\mathbf{s}_i$ with radius $\|\mathbf{s}_i - \mathbf{s}_j\|$ that lies within D.

As with G and F, we compare $\hat{K}(d)$ with $K(d) = \pi d^2$. Regularity/inhibition implies $\hat{K}(d) < \pi d^2$; clustering implies $\hat{K}(d) > \pi d^2$. A plot which has been proposed in the literature (see, for example, Cressie, 1993)) is $L(d)$ vs d where $L(d) = \sqrt{\frac{\hat{K}(d)}{\pi}} - d$. Evidently, $L(d) = 0$ under CSR, suggesting that we look for peaks and valleys in the plot. For instance, a peak at distance d would suggest clustering at that distance.

Under isotropy, we can connect $K(d)$ to $\gamma(d)$, where $\gamma(d)$ is the second moment measure, presented in the previous section. The relationship is $\gamma(d) = \frac{\lambda^2 K'(d)}{2\pi d}$. We offer a simple proof. Recall that $\gamma(\mathbf{s}, \mathbf{s}')|\partial \mathbf{s}||\partial \mathbf{s}'|$ is approximately the probability that there is a point

of $\mathbf{S}$ in each of $\partial\mathbf{s}$ and $\partial\mathbf{s}'$. Hence, given there is a point in $\mathbf{S}$ at $\mathbf{s}$, in the isotropic case, $\gamma(r)/\lambda$ is the conditional probability of a point in $\mathbf{S}$ at distance r from $\mathbf{s}$ given there is a point at $\mathbf{s}$ (from the end of Section 8.2.2). If we sweep out the area of the circle of radius d around $\mathbf{s}$ using concentric circles of increasing radius r centered at $\mathbf{s}$, we have that $\int_0^d 2\pi r\gamma(r)dr/\lambda$ is the expected number of points within distance d of $\mathbf{s}$. That is, $\frac{2\pi}{\lambda}\int_0^d r\gamma(r)dr = E(N(\mathbf{s}, d; \mathbf{S}) = \lambda K(d)$. So, $\frac{\lambda^2 K(d)}{2\pi} = \int_0^d r\gamma(r)dr$. Differentiating both sides with respect to d yields $\frac{\lambda^2 K'(d)}{2\pi} = d\gamma(d)$ from which the result follows.

As an aside, in the literature we find the inhomogeneous K function associated with a general $\lambda(\mathbf{s})$ rather than a constant λ (Baddeley et al., 2001). It is defined through the pair correlation function, $g(\mathbf{s}, \mathbf{s}')\frac{\gamma(\mathbf{s},\mathbf{s}')}{\lambda(\mathbf{s})\lambda(\mathbf{s}')}$. In particular, if $g(\mathbf{s}, \mathbf{s}') = g_o(||\mathbf{s} - \mathbf{s}'||)$, the associated inhomogeneous K function, $K_o(d)$, satisfies $K_o(d) = 2\pi\int_0^d rg_o(r)dr$, analogous to the calculation of the preceding paragraph. In the case of a NHPP, it may be shown that $K_o(d)$ is still equal to πd^2. The modified version of (8.9) would replace $\hat{\lambda}^2$ with $\hat{\lambda}(\mathbf{s})\hat{\lambda}(\mathbf{s}')$, moved to the denominator in the respective terms of the summation.

Returning to the homogeneous case, the awkwardness in estimating $\gamma(d)$ compared with the ease of estimation and interpretation of $K(d)$ have led to the dominant usage of the latter. However, Illian et al. (2008) prefer the use of an estimator of $\gamma(d)$, in fact, of the pair correlation coefficient, $\gamma(d)/\lambda^2$. The analogue of (8.9) for γ is given by

$$\frac{1}{2\pi d}\sum_{\mathbf{s}_i,\mathbf{s}_j\in\mathbf{S}_{obs}} h(d - d_{ij})w_{ij} \tag{8.10}$$

where $d_{ij} = ||\mathbf{s}_i - \mathbf{s}_j||$, h is a kernel function, and again, w_{ij} is as above. We see that (8.10) is a customary kernel estimator, based upon the d_{ij}'s, adjusted for edge effects. Of course, the d_{ij} are not independent.

We illustrate the EDA using the foregoing diagnostic tools. In particular, we work with a set of data from Duke Forest which considers three tree species prevalent in Duke Forest in North Carolina. The three species are ACRU: *Acer rubrum* (red maple), COFL: *Cornus florida* (flowering dogwood), and OXAR: *Oxydendrum arboretum* (sourwood) and the point patterns of their locations are from the Blackwood region in Duke Forest. Indeed, in Figure 8.5 we show the observed point pattern for each species. The region is rectangular and the coordinates shown are local in meters. Figure 8.6 provides the theoretical Q-Q plots for the G function and the F function (adding $m = 100$ points) for the three species in the Duke Forest data. There does not seem to be strong evidence for departure from CSR in these plots (perhaps a bit for COFL). Figure 8.7 presents the theoretical Q-Q plots for the J function (top) and the L function (bottom) for the three species in the Duke Forest dataset. Compared to the F and G plots, the J plots here are a bit more revealing. For COFI, there

Figure 8.5 *The observed point patterns for the three species ACRU, COFL and OXAR in the Duke Forest dataset.*

Figure 8.6 *The theoretical Q-Q plots for the G function (top) and the F function (bottom) for the three species in the Duke Forest dataset.*

seems to be considerable evidence of clustering and, to a lesser extent for ACRU. Finally, the lower panel in Figure 8.7 provides the L plots, taken from the K functions. Here, for all species we see evidence of clustering with peaks in the vicinity of $d = 6$ to 8 meters.

Figure 8.7 *The theoretical Q-Q plots for the J function (top) and the L function (bottom) for the three species in the Duke Forest dataset.*

8.3.4 Empirical estimates of the intensity

For nonstationary processes, moving away from CSR leads to interest in estimating the first order intensity, $\lambda(\mathbf{s})$. In the absence of a model for $\lambda(\mathbf{s})$, we briefly discuss empirical estimates. The first is the analogue of a histogram, the second of a kernel density estimate.

Imagine a refined grid over D. Then, as above, $\lambda(\partial\mathbf{s}) = \int_{\partial\mathbf{s}} \lambda(\mathbf{s})d\mathbf{s} \approx \lambda(\mathbf{s})|\partial\mathbf{s}|$. So, for grid cell A_l, assume $\lambda(\mathbf{s})$ is constant over A_l. Then, the natural estimate is $N(A_l)/|A_l|$. Evidently, a picture of this estimate will reveal a two-dimensional step surface which we might call a tile surface. It's appearance will resemble a two-dimensional histogram but the area under the surface will be the number of points in the pattern.

Kernel density estimates are widely used, providing a smoothing of a histogram. In the same spirit, a kernel intensity estimate takes the form

$$\hat{\lambda}_\tau(\mathbf{s}) = \sum_i h(||\mathbf{s} - \mathbf{s}_i||/\tau)/\tau^2, \mathbf{s} \in D. \tag{8.11}$$

In (8.11), h is a radially symmetric bivariate pdf (usually a bivariate normal) while τ is "bandwidth" which controls the smoothness of $\hat{\lambda}_\tau(\mathbf{s})$. The power τ^2 reflects the fact that the scaling is done in R^2. Finally, note that we don't divide by n, as with kernel density estimates. The reason is that we *cumulate* intensity rather than *normalizing* density.

The second order intensity, γ, is defined in general but would be difficult to estimate in the absence of replications. Furthermore, the K function is not defined for a nonstationary situation except in a special case. That special case was noted in Section 8.2.2, a so-called *intensity reweighted stationary process*. This process has the feature that the pair correlation function, $\gamma(\mathbf{s}_1, \mathbf{s}_2)/\lambda(\mathbf{s}_1)\lambda(\mathbf{s}_2)$ (where $\lambda(\mathbf{s})$ is strictly positive) is stationary, in fact, isotropic, say, $g(d)$ where $d = \|\mathbf{s}_1 - \mathbf{s}_2\|$. The associated intensity reweighted K function would become $K_{rew}(d) = 2\pi \int_0^d rg(r)dr$ (following the argument of the previous subsection).

As a result the natural extension of the estimate $K(d)$ becomes

$$\hat{K}_{rew}(d) = |D|^{-1} \sum_i \sum_{j \neq i} I(d_{ij} \equiv ||\mathbf{s}_i - \mathbf{s}_j|| \leq d)/\hat{\lambda}(\mathbf{s}_i)\hat{\lambda}(\mathbf{s}_j)$$

with say, a kernel estimate of the intensity. Similarly, the natural estimate of $\gamma(\mathbf{s}_1, \mathbf{s}_2)$ would be $\hat{g}(d)\hat{\lambda}(\mathbf{s}_1)\hat{\lambda}(\mathbf{s}_2)$ where

$$\hat{g}(d) = \frac{1}{2\pi d|D|} \sum_{\mathbf{s}_i, \mathbf{s}_j \in \mathbf{S}_{obs}} h(d - d_{ij})w_{ij}/\hat{\lambda}(\mathbf{s}_i)\hat{\lambda}(\mathbf{s}_j).$$

It is evident (and noted in Baddeley et al., 2000), that these estimates will typically be unstable and badly biased due to the intensity estimates in the denominators. A further concern is that stationarity of the second order reweighted intensity is restrictive. See Section 8.6 in this regard.

To illustrate, we consider a tropical rainforest tree dataset consisting of the locations of 3605 trees (left panel, Figure 8.8(a)). Provided covariates include elevation and slope (center and right panels, Figure 8.8(a)) and suggest that a constant intensity is not operating. This data has been analyzed in Møller and Waagepetersen (2007) as well as in the `spatstat` `R` package (Baddeley and Turner, 2005). Here, in Figure 8.8(b), we present the associated kernel (without and with edge correction) and tiled intensity estimates. We see the very high peak associated with the very high concentration of points in the northeastern part of the region.

8.4 Modeling point patterns; NHPP's and Cox processes

Here, we confine ourselves to modeling of point patterns using NHPP's, deferring alternative specifications to Section 8.6. Arguably, NHPP's are the most widely used models for point

Figure 8.8 *Tropical rainfall tree data. Panel (a) shows the number of locations (left), elevation (center) and slope (right). Panel (b) presents the associated kernel and tiled intensity estimates.*

patterns. We begin by providing the NHPP likelihood, developed from the joint density in two distinct ways. In fact, the likelihood has already been alluded to in Section 8.2 through Expression (8.2) and the discussion of the conditionally independent form of the location density in Section 8.2.1.

First, from Section 8.2.1, given $N(D) = n$, the location density,

$$f(\mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_n | N(D) = n) = \prod_i \frac{\lambda(\mathbf{s}_i)}{(\lambda(D))^n} ,$$

where, again $\lambda(D) = \int_D \lambda(\mathbf{s}) d\mathbf{s}$. So, the "joint density",

$$f(\mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_n, N(D) = n) = \prod_i \frac{\lambda(s_i)}{(\lambda(D))^n} \times (\lambda(D))^n \frac{\exp(-\lambda(D))}{n!} .$$

Thus, the likelihood becomes

$$L(\lambda(\mathbf{s}), \mathbf{s} \in D; \mathbf{s}_1, \ldots, \mathbf{s}_n) = \prod_i \lambda(\mathbf{s}_i) \exp(-\lambda(D)) \tag{8.12}$$

Alternatively, partition D into a fine grid. The Poisson assumption implies that the likelihood will be a product over the grid cells, that is, $\prod_l \exp(-\lambda(A_l))(\lambda(A_l))^{N(A_l)}$. Regardless of the grid, the product of the exponential terms is $\exp(-\lambda(D))$. Moreover, as the grid becomes finer, $N(A_l) = 1$ or 0 according to whether there is a $\mathbf{s}_i$ in A_l or not. In the limit, we obtain (8.12).

A key point to note in (8.12) is that the likelihood is a function of the entire intensity surface. We will have an uncountable dimensional model unless we provide a parametric form for $\lambda(\mathbf{s})$. In the latter case, we can replace $\lambda(\mathbf{s})$ with its associated parameters and L becomes a function of those parameters. In the former case we have *nonparametric* specification; for example, $\lambda(\mathbf{s})$ could be a realization of a log Gaussian process, i.e., $\lambda(\mathbf{s}) = \exp(Z(\mathbf{s}))$ where $Z(\mathbf{s})$ is a Gaussian process. In the latter case, the integral in (8.12) is an ordinary integral, i.e., an integral with respect to a function. In the nonparametric case, with a GP, the integral is with regard to a random process realization; we have a stochastic integral. Such integrals have been discussed in Chapter 7 but may be more challenging to handle here because,

for example, with a log Gaussian process, $\lambda(D)$ is not linear in the process. Strategy for *approximation* of $\lambda(D)$ has to be considered with care. We provide further discussion on this issue below.

8.4.1 Parametric specifications

Armed with the likelihood, we turn to modeling $\lambda(\mathbf{s})$. One form that has appeared in the literature is $\lambda(\mathbf{s}) = \sigma\lambda_0(\mathbf{s})$ with $\lambda_0(\mathbf{s})$ known, σ unknown. This seems, patently, silly; where would λ_0 come from? how would we know the intensity up to scale factor?

A more common choice is to assume that $\lambda(s)$ is a tiled surface over a grid. This requires specifying λ_l for A_l. Such tiling naturally occurs when covariate information is only available at, say, grid cell scale because then it will be impossible to learn about the intensity at higher resolution. With such tiling, to incorporate spatial structure into λ, rather than the Gaussian process specification that we mentioned above for $\lambda(\mathbf{s})$, we might now employ a Markov random field model for the λ_l's. In fact, this will take us to the realm of disease mapping models, the subject of Chapters 4 and 6, with elaboration below.

A convenient choice for λ would be a parametric function, $\lambda(\mathbf{s};\theta)$. The challenge here is to specify a rich enough class. Simple polynomial forms will not be flexible enough; they will suffer the phenomenon of the "tail wagging the dog." That is, if they are high in some places they will be forced to be low in others. Moreover, they must be nonnegative over D. More flexible choices will be through the use of basis functions, e.g., a collection of two-dimensional basis functions (usually orthonormal), yielding $\lambda(\mathbf{s};\boldsymbol{\theta})$ of the form, $\sum_{k=1}^{K} a_k g_k(\mathbf{s})$, where the g_k are a set of K basis functions (see, e.g., Fahrmeir, Kneib and Lang, 2007). Again, the nonnegativity requirement may be awkward.

In this spirit, we might write $\lambda(\mathbf{s};\theta) = \lambda f(\mathbf{s};\theta)$ where f is a bivariate density function truncated to D. Such a form has the implicit and appealing interpretation that $\lambda = \lambda(D)$ along with the assurance that $f \geq 0$. However, to create sufficiently rich choices for f, we would turn to mixture models, e.g., $f(\mathbf{s}) = \sum_{k=1}^{K} p_k f_k(\mathbf{s})$. Fitting such models introduces challenges. First, identifying the components is a well-known difficulty (see, e.g., Celeux, Hurn and Robert, 2000). Second, there is the potentially awkward restriction of the component densities to D. Normalizing these densities to D when D is a general geographic region will require numerical integration. Additionally, the componentwise normalizing constants will be functions of the respective component parameters.

The foregoing treats the locations as explanatory variables in the specification for $\lambda(\mathbf{s})$. So, as an alternative, $\lambda(\mathbf{s})$ may be viewed as a trend surface form. We could write $\log\lambda(\mathbf{s})$ as a trend surface in latitude and longitude or in eastings and northings. Of course, this will still result in a limiting polynomial form but working on the log scale mitigates the nonnegativity concern.

In practice, it is most typical to specify $\log\lambda(\mathbf{s}) = \mathbf{X}^T(\mathbf{s})\boldsymbol{\gamma}$. That is, we envision spatial covariates to drive the point pattern and we express this through a regression model on the log scale. For example, with regard to the distribution of a species, we might expect more points where elevation is lower or where temperature is higher. If so, we will need to calculate $\int_D e^{\mathbf{X}^T(\mathbf{s})\boldsymbol{\gamma}} d\mathbf{s}$ to obtain (8.12). This integral can be problematic since rarely will $\mathbf{X}(\mathbf{s})$ be given in functional form – consider, for example elevation as a covariate. A numerical integration will be needed.

In this regard, as noted above, often tiling is imposed with such covariates in the sense that the covariate surface may only be resolved to subregions, e.g., population density at census scales or, with regard to climate variables, mean annual temperature or a drought index at some grid box scale. Now, the issue of the ecological fallacy may emerge (See Chapter 7). More precisely, for any subregion B, we would need $\int_B e^{\mathbf{X}^T(\mathbf{s})\boldsymbol{\gamma}} d\mathbf{s}$. What we have is $e^{\mathbf{X}^T(B)\boldsymbol{\gamma}}$ which can be quite different unless we are comfortable with $\mathbf{X}^T(\mathbf{s})$ essentially

constant for $\mathbf{s} \in B$, say $\mathbf{X}^T(B)$. If so, then we only require a simple rescaling, $|B|e^{\mathbf{X}^T(\mathbf{s})\boldsymbol{\gamma}} = \int_B e^{\mathbf{X}^T(\mathbf{s})\boldsymbol{\gamma}} d\mathbf{s}$. In the absence of finer covariate resolution, we cannot do better with regard to the ecological fallacy.

8.4.2 Nonparametric specifications

Returning to the above, suppose we take $\lambda(\mathbf{s})$ to be a process realization. Writing $\lambda(\mathbf{s}) = g(\mathbf{X}(\mathbf{s})^T\boldsymbol{\gamma})\lambda_0(\mathbf{s})$, we insist that $g(\cdot) \geq 0$ and can think of $\lambda_0(\mathbf{s})$ as the *error* process, a realization of a positive stochastic process which, in the interest of centering, might naturally have mean 1. We recall, from Section 8.2.1, that this specification is a Cox process. In fact, conditional on $\{\lambda_0(\mathbf{s}), \mathbf{s} \in D\}$ (and $\boldsymbol{\gamma}$), we have a NHPP. Suppose $\lambda(\mathbf{s}) = \exp(Z(\mathbf{s}))$ where $Z(\mathbf{s})$ is a realization from a spatial Gaussian process with mean, say, $\mathbf{X}^T(\mathbf{s})\boldsymbol{\gamma}$ and covariance function $\sigma^2\rho(\cdot)$. Then, we refer to this specification as a log Gaussian Cox process (LGCP), noting that it provides a prior for $\lambda(\mathbf{s})$, analogous to the parametric case, $\lambda(\mathbf{s};\boldsymbol{\theta})$, where we have a prior on $\boldsymbol{\theta}$. As earlier, we might write this as $\mathbf{X}^T(\mathbf{s})\boldsymbol{\gamma} + w(\mathbf{s})$ so that $\log\lambda_0(\mathbf{s}) = w(\mathbf{s})$[3]. Conditional on $w(\mathbf{s})$, we immediately have the first and second moments for this process. Calculation of the marginal product densities is left as an exercise. In fact, if $g(\mathbf{X}^T(\mathbf{s})\boldsymbol{\gamma}) = \lambda$, marginally, the process is stationary.

With a Cox process, the further challenge is the evaluation of the likelihood which requires stochastic integration, $\int_D e^{\mathbf{X}^T(\mathbf{s})\boldsymbol{\gamma}+w(\mathbf{s})}d\mathbf{s}$. Evidently, such an integral does not have a closed form expression. Furthermore, approximation is tricky. Suppose we can dispense with the ecological fallacy concerns as in the previous subsection. In particular, suppose $\mathbf{X}(\mathbf{s})$ is tiled to M subregions, $B_i, i = 1, 2, \ldots, M$. Then,

$$\int_D e^{\mathbf{X}^T(\mathbf{s})\boldsymbol{\gamma}+w(\mathbf{s})}d\mathbf{s} = \sum_{m=1}^{M}\int_{B_m} e^{\mathbf{X}^T(\mathbf{s})\boldsymbol{\gamma}+w(\mathbf{s})}d\mathbf{s} = \sum_{m=1}^{M} e^{\mathbf{X}^T(B_m)\boldsymbol{\gamma}}\int_{B_m} e^{w(\mathbf{s})}d\mathbf{s}.$$

The required stochastic integrations are clear, one for each B_m. For a given B_m, if we divide the integral by $|B_m|$, we can view the integral as the expectation of $\exp(w(\mathbf{s}))$ with respect to a uniform distribution over B_m. Here, we find ourselves in similar territory to the block averaging discussed in Section 7.1. Hence, it is natural to attempt a Monte Carlo approximation to the integral, drawing random $\mathbf{s}_j, j = 1, 2, \ldots, J$ over B_m and taking the approximation, $|B_m|\sum_j \exp(w(\mathbf{s}_j))/J$. A critical difference between the Monte Carlo approximation proposed here and that of Section 7.1 is that there we integrate the process directly while here we integrate a nonlinear function of the process. For the former, we can argue that the behavior of the so-called *quadratic variation* is such that the difference between the stochastic integral and the Monte Carlo approximation tends to 0 in probability as $J \to \infty$. With a nonlinear function, this need not be the case.

The customary fix is to work with a finite dimensional process. One version would simply replace $\int_{B_m} e^{w(\mathbf{s})}d\mathbf{s}$ with e^{ϕ_m} where $\{\phi_i\}$ follow a CAR model (Chapter 3 but see below). Then, $\lambda(D)$ is approximated by $\sum_{m=1}^{M} e^{\mathbf{X}^T(B_m)\boldsymbol{\gamma}+\phi_m}$. An alternative is to specify the ϕ_m as the value of a realization of a mean 0 GP at a set of *representative points*, i.e., at a suitable point within each of the B_m's. Now $\{\phi_m\}$ will follow a familiar multivariate normal distribution. There is substantial literature on likelihood approximation in the Cox process case. See, for example, Wolpert and Ickstadt (1998), Beneš et al. (2005), and the book of Møller and Waggepetersen (2004).

[3]It is customary in the literature to make $w(\mathbf{s})$ have mean 0 but, to provide $\lambda_0(\mathbf{s})$ with mean 1 implies making $E(w(\mathbf{s})) = -\sigma^2/2$.

8.4.3 Bayesian modeling details

Bayesian modeling is straightforward for the setting we have described thus far. For parametric cases, we write the likelihood as $L(\boldsymbol{\theta}; \mathbf{s}_1, \ldots, \mathbf{s}_n)$ with a prior on a finite dimensional $\boldsymbol{\theta}$ as usual. The parametric model may be a trend surface so that the $\boldsymbol{\theta}$ are associated coefficients. However, as noted above, it would be customary to include covariates in $\lambda(\mathbf{s}; \boldsymbol{\theta})$, so that some of the $\boldsymbol{\theta}$ become regression coefficients. In any event, with a prior on $\boldsymbol{\theta}$, the specification is complete. The result, with regard to model fitting is to replace conditioning on λ_D with conditioning on $\boldsymbol{\theta}$. Again, we will need to calculate $\int_D \lambda(s; \theta) ds$. According to the specification for $\lambda(\mathbf{s})$, perhaps this can be done explicitly. If not, we would likely resort to numerical integration.

For the nonparametric case, again, $\lambda(\mathbf{s}) = \exp(Z(\mathbf{s}))$ where $Z(\mathbf{s}) = \mathbf{X}^T(\mathbf{s})\boldsymbol{\gamma} + w(\mathbf{s})$ is a realization of a GP, i.e., a GP prior on $\log\lambda_D$. Following the remark at the end of the previous subsection, we can only work with a finite dimensional distribution. We replace λ_D with $\lambda(\mathbf{s}_l^*)$ where the $\mathbf{s}_l^*$ are centroids associated with a suitably fine grid over D and yield a tiled or step surface. This introduces a multivariate normal prior for λ on the Z scale. Still, we need $\lambda(\mathbf{s})$ at every $\mathbf{s} \in D$ and we achieve this by using the $Z(\mathbf{s}_l^*)$ to create a tiled surface over D. This converts the likelihood from $L(\lambda_D; s_1, ...s_n)$ to $L(\boldsymbol{\theta}, \{w(\mathbf{s}_l^*)\}, \{w(\mathbf{s}_i), i = 1, 2, .., n\}; \mathbf{s}_1, \ldots, \mathbf{s}_n)$ where $\boldsymbol{\theta}$ denotes the remaining model parameters. Evidently, this is an approximation. However, the finite dimensional set of $Z(\mathbf{s}_l^*)$'s allows a convenient numerical integration.

Prior specification for the covariance function of $Z(\mathbf{s})$, say, σ^2 and ϕ with an exponential covariance choice, is not apparent nor is the potential to learn much about them. Recall the challenges articulated in Chapter 3 in the geostatistical case. Here, we are in a much more difficult setting. In the geostatistical case we observe values on the process surface, perhaps up to nugget error, to learn about the dependence structure associated with the surface. Here, we never see any realizations associated with the intensity surface; we only see a point pattern resulting from it. How much information is there in the point pattern regarding σ^2 and ϕ? Evidently, this is not a Bayesian issue; any fitting package will be challenged in this regard. Practically, we will have to depend upon the covariates, $\mathbf{X}(\mathbf{s})$, to inform about the intensity and hope that the errors are small. In the absence of useful covariates, e.g., say, a constant λ as above, the $\lambda_0(\mathbf{s})$ will reflect any random clustering and gaps in the observed point pattern and cannot possibly produce an estimated intensity which is roughly constant. (The same would be true for a kernel intensity estimator.) So, in our experience, we need very informative priors for σ^2 and ϕ. In fact, we typically fix ϕ to be relatively small compared with the size of D. Then, at least we have σ^2 identified (recall the discussion in Chapter 6 following the work of Zhang, 2004). Even so, Bayesian learning may be limited and it will still be helpful to adopt an informative appropriately centered inverse gamma or log normal prior for σ^2. A convenient way to do this is through EDA based upon minimum contrast estimation using the K function or the paired correlation function. See, e.g., Diggle (2003) or Waagepetersen and Guan (2009).

Model fitting requires updating $\boldsymbol{\gamma}, \sigma^2$ (and perhaps ϕ) along with the the w's associated with the $\mathbf{s}_l^*$ as well as the w's associated with the $\mathbf{s}_i$. We use standard Gaussian proposals for the γ's and, depending upon the prior, an inverse gamma or log normal proposal for σ^2. With m representative points, the latter requires updating an $m+n$ dimensional latent multivariate normal random variable. There are many MCMC algorithms for implementing this sort of updating. A particularly efficient choice is elliptical slice sampling as developed in the work of Murray et al. (2010) and Murray and Adams (2010). In the spirit of our discussion of slice sampling in Appendix A.2, this approach introduces an auxiliary variable under a novel parametrization.

8.4.3.1 The "poor man's" version; revisiting the ecological fallacy

Due to the computational demands of fitting a fully Bayesian LGCP model, it is sometimes the case that these models are fitted by aggregating the points to cells in a grid. We refer to this as the "poor man's" version of model fitting. In particular, we overlay a regular grid on D after which we work with the likelihood arising from the Poisson counts associated with the grid. Effectively, we reduce the problem to the disease mapping setting of Chapters 4 and 6. In fact, we now model the λ_i's associated with the grid cells. As in the disease mapping case, we can introduce covariates, now at the scale of the grid. Moreover, we could now introduce a CAR model for the spatial random effects, replacing the Gaussian process. The resultant specification for the intensity for grid cell i would become

$$\log\lambda_i = \mathbf{X}_i^T\boldsymbol{\beta} + \phi_i. \tag{8.13}$$

Such a model is routine to fit, in fact, easily using `WinBUGS` as discussed in Section 6.4. An obvious difference between this setting and that for disease mapping is the fact that, for the latter, the areal units are specified. For the poor man's version we have no such provision. We can select any number of grid cells, any sizes, orientations, etc. Evidently, there will be concern with regard to sensitivity to the choice of grid. We know that the appearance of the pattern with regard to the grid cells will be scale dependent. So, we recommend taking the effort to fit at the highest resolution, the scale of the point pattern itself.

In fact, such grid cell approximation reminds us of the foregoing concern, the ecological fallacy. With no spatial random effects and intensity over D, $\lambda(\mathbf{s}) = \exp(\mathbf{X}(\mathbf{s})^T\boldsymbol{\beta})$, we would have $\lambda_i = \int_{A_i}\lambda(\mathbf{s})d\mathbf{s} = \int_{A_i}\exp(\mathbf{X}(\mathbf{s})^T\boldsymbol{\beta}d\mathbf{s})$. Clearly, this need not at all be close to $\exp(\int_{A_i}\mathbf{X}(\mathbf{s})^T\boldsymbol{\beta})d\mathbf{s} = \exp(\mathbf{X}_i^T\boldsymbol{\beta})$ which would arise using (8.13). Bringing in spatial random effects only compounds the problem. Now, ignoring the covariate term, we seek $\int_{A_i}\lambda(\mathbf{s})d\mathbf{s} = \int_{A_i}\exp(Z(\mathbf{s})d\mathbf{s}$ which evidently, is not $\exp(\int_{A_i}Z(\mathbf{s})d\mathbf{s}) = \exp(Z(A_i))$. Moreover, starting with a Gaussian process, the collection $\{Z(A_i)\}$ is a set of block averages, with a joint normal random distribution that has no connection with a CAR model for the set $\{\phi_i\}$.

8.4.4 Examples

Returning to the tropical rainforest tree data, we consider four models here. In all cases we employ the two covariates, elevation and slope, resulting in three regression coefficients. We fit a NHPP model, we fit a LGCP model, and we fit two versions of the poor man's model, one using a 20×10 grid, the other using a 100×50 grid. In Figure 8.9(a) we show the results of the NHPP fit. All three coefficients are significant but the estimated intensity surface (Figure 8.9(b)) is poor; the regression form cannot capture the extreme peaks (compare with the left panel in Figure 8.8(a) or Figure 8.11 below). For the LGCP, in Figure 8.10 we see that, again all three regression coefficients are significant. We also provide a comparison of the prior to posterior for σ^2 to show the Bayesian learning. Figure 8.11 shows the point pattern and the intensity surface estimates. Three posterior mean intensity surface estimates are provided, $\lambda(\mathbf{s})$, $\lambda_0(\mathbf{s})$, and $z_0(\mathbf{s}) = \mathtt{log}\lambda_0(\mathbf{s})$. Now, we see the substantial local GP adjustment and we do a much better job with the estimated $\lambda(\mathbf{s})$. Finally, we show the results of the poor man's fitting for the 20×10 and 100×50 grids in Figure 8.12. Panels (a) and (b) correspond to the 20×10 grid and panels (c) and (d) correspond to the 100×50 grid. Again, we see general agreement regarding the coefficients while the higher resolution poor man's estimated intensity surfaces are very similar to those of the LGCP.

(a) (b)

Figure 8.9 *Estimates from the NHPP model. Left panel shows the estimated posterior distributions for the three coefficients along with their MCMC chains. Right panel depicts the estimated intensity surface.*

Figure 8.10 *Posterior estimates for the three regression coefficients and σ^2 from the LGCP model.*

Figure 8.11 *Estimates from the LGCP model: Point pattern and the intensity estimates.*

Figure 8.12 *Results for the poor man's fitting for the* 20×10 *((a) and (b)) and* 100×50 *((c) and (d)) grids.*

8.5 Generating point patterns

We briefly digress to remind the reader of how we customarily generate realizations under a NHPP. First, consider generating a realization from an HPP with intensity λ. As the foregoing discussion suggests, under CSR, we draw $n \sim \texttt{Po}(\lambda|D|)$. Given n, we sample n locations uniformly over D.

Next, consider generating a realization from a NHPP given $\lambda(\mathbf{s})$, where $\lambda(\mathbf{s})$ is a function over D. An initial thought, following the derivation of the likelihood, might be to first draw n, now from $\texttt{Po}(\lambda(D)$ and then, given n, place the locations according to the distribution, $\lambda(\mathbf{s})/\lambda(D)$. However, in practice, a more convenient approach is to employ location-dependent thinning. In particular, compute $\lambda_{max} = \max_{\mathbf{s} \in D} \lambda(\mathbf{s})$. Next, sample $n \sim \texttt{Po}(\lambda_{max}|D|)$. Then, given n, sample n locations uniformly over D. Finally, "thin" these samples using a *rejection method* by retaining $\mathbf{s}_i$ with probability $\lambda(\mathbf{s}_i)/\lambda_{max}$. We leave it to an exercise to show that such thinning does, indeed, provide samples from the desired NHPP.

Next, what happens if $\lambda(\mathbf{s})$ is a realization of a log Gaussian process, as above. Perhaps, a first question to ask is why we would wish to sample a point pattern from such a λ? That is, with $\log \lambda(\mathbf{s}) = \mathbf{X}^T(\mathbf{s})\boldsymbol{\gamma} + w(\mathbf{s})$, why wouldn't we generate, ignoring $w(\mathbf{s})$, following the recipe above, to create realizations driven by the covariates? One reason might be to see the effect of the additional randomness induced by the inclusion of $w(\mathbf{s})$. A second reason is that we may need to generate posterior point patterns in order to use Campbell's theorem to provide Monte Carlo integrations for posterior expectations of interest. Additionally, we may need to generate such realizations in order to implement MCMC for certain point pattern models (see Section 8.6.3). In any event, the generation would proceed in two steps, first generation of $w(\mathbf{s})$ and then, generation of the sample of locations given $w(\mathbf{s})$ along with $\mathbf{X}^T(\mathbf{s})\boldsymbol{\gamma}$. The latter is achieved as above. The former is, again, an uncountable dimensional random variable and can only be implemented finitely, generating a multivariate normal vector over a selected lattice in D and then treating the realization as a tiled surface

over D with the lattice locations as representative points, e.g., centroids. Evidently, such a construction yields a Cox process.

The above are examples of a general strategy called thinning which revises samples under one point pattern intensity to sample under another. In its simplest form, we might implement p-thinning. We employ a constant probability p of thinning, regardless of location. That is, each point $\mathbf{s}_i \in \mathbf{S}$ is retained, independently, with probability p. We note, as an exercise, that, applied to a realization from an HPP with intensity λ, this yields a realization from an HPP with intensity $p\lambda$. Note that this is **not** equivalent to randomly sampling a proportion p of the observations from the HPP with intensity λ. (Why? Under thinning, the $\mathbf{s}_i \in \mathbf{S}$ are retained independently; we obtain a random number of points, as it should be for a random point pattern.)

Next, we might consider $p(\mathbf{s})$ thinning. Here, we imagine a thinning surface, $p_D = \{p(\mathbf{s}), \mathbf{s} \in D, 0 \leq p(\mathbf{s}) \leq 1\}$. Now, each point $\mathbf{s}_i \in \mathbf{S}$ is retained independently with a local probability $p(\mathbf{s}_i)$. The foregoing exercise concludes that thinning using $p(\mathbf{s}) = \lambda(\mathbf{s})/\max_{\mathbf{s}\in D} \lambda(\mathbf{s})$ thins a realization of an HPP with intensity $\max_{\mathbf{s}\in D} \lambda(\mathbf{s})$ to a realization from a NHPP with intensity $\lambda(\mathbf{s})$. More generally, all nonstationary point processes that arise from $p(\mathbf{s})$ thinning of a stationary point process are such that the second order reweighted intensity function or pair correlation functions for the original stationary and the reweighted nonstationary process are identical. We leave this as an exercise.

A further version of thinning would view the set $p_D = \{p(\mathbf{s}), \mathbf{s} \in D, 0 \leq p(\mathbf{s}) \leq 1\}$ as a random realization of a process, i.e., of a random field over D. In principle, we could imitate the approach of the previous paragraph, but now we would create $p_{D^*} = \{p(\mathbf{s}) \equiv \lambda(\mathbf{s})/\max_{\mathbf{s}\in D^*} \lambda(\mathbf{s})\}$; now D^* would be the set of representative points in D. Alternatively, p_D could be achieved by logit or c.d.f. transformation of a GP realization. The GP specification could include suitable covariates for the mean surface which would be reflected in the corresponding p_D surface. In principle, the mean surface could have unknown regression parameters. To our knowledge, such models have not been investigated in the literature.

Mechanisms besides thinning exist in the literature for transforming point patterns from one process model to those from another. For instance, we could imagine more general dependent thinning. Additional possibilities include displacement, censoring, and superposition. See, e.g, Baddeley and van Lieshout (1993, Section 5), Lund, Penttinen, Rudemo (1999), Lund and Rudemo (2000) and also Section 8.9.1.

8.6 More general point pattern models

8.6.1 Cluster processes

Many observed point process realizations exhibit patterns of clustering. The notion of *clustering* or even a *cluster* is not well defined. We would recognize it as variation in point density across the region D. Informally, we would think of it as groups of points with interpoint distances that are shorter than the average distance across the pattern. In a sense it is easier to formulate models for spatial clustering than it may be to assert it or explain it for observed point patterns. We offer the family of Neyman Scott models in Section 8.6.1 and, more generally, shot noise processes in Section 8.6.2. However, recall that, even HPP's will exhibit clustering (as we noted from the behavior of the G function in Section 8.3.2); $G(d)$ places a lot of mass on small distances. Moreover, as remarked in Diggle et al. (2007), there is a "fundamental ambiguity between heterogeneity and clustering," the first due to the spatial variation in the first order intensity, $\lambda(\mathbf{s})$, the second due to potential stochastic dependence between the point locations, arising from the (stationary) second order intensity $\gamma(||\mathbf{s}-\mathbf{s}'||)$. These two phenomena are "difficult to disentangle." Indeed, detection of clusters in an observed point pattern is a very challenging problem. There are various algorithmic

cluster detection procedures (see, e.g., Illian et al., 2008, p. 373 for some discussion) but it will not be possible to assess how well they work since there is no notion of "truth."

The other side of the coin is inhibition or repulsion. Many physical processes are assumed theoretically to discourage points from being too close to each other. In ecological processes, competition for resources may discourage plants from being too proximate with each other. In a sense, such behavior is more well-defined than clustering and, as a result, it is simpler to specify models to capture this. We offer the family of Gibbs process models in Section 8.6.3.

In this regard, an important take home message emerges below. If you want to model clustering, use the Neyman-Scott or shot noise processes of Sections 8.6.1.1 and 8.6.2. If you want to model inhibition or repulsion, use the Gibbs processes of Section 8.6.3.

8.6.1.1 Neyman-Scott processes

Neyman-Scott processes offer one clustering strategy for enriching NHPP specifications. At the first stage, "parents" are generated using an HPP (more generally, a NHPP). At the second stage, "children" are generated, associated with respective parents. In fact this typically requires two stochastic stages: first generate the number of children associated with each parent, then generate locations for each child relative to its parent. Typically, the parents are removed and the point pattern realization is that of the entire set of children.

In particular, suppose we generate *parent* events from a NHPP with $\lambda(\mathbf{s})$, say, K, and their locations, say, $\boldsymbol{\mu}_k, k = 1, 2, \ldots, K$. Next, suppose each parent produces a random (but i.i.d.) number of offspring, N_k, where the N_k are i.i.d. according to a distribution on the integers, say, g. Typically, $g = \texttt{Po}(\delta)$ but, in some cases, it might make sense to have g be a mixture of, say, two Poisson distributions to allow for *small* and *large* numbers of offspring. Next we need to locate the offspring relative to the parent. For the kth parent, suppose this is done by assigning positions according to i.i.d. draws from a bivariate density, $f(\mathbf{s}; \boldsymbol{\mu}_k)$, i.e., a density centered at $\boldsymbol{\mu}_k$ (usually with radial symmetry, usually a Gaussian). The case where the bivariate density in $N(\boldsymbol{\mu}_k, \sigma^2 I)$ is referred to as the (modified) Thomas process. Again, only the offspring are retained to yield the point pattern.

Note that, in the above specification for the Neyman-Scott process, conditional on the number of parents, K, and their locations $\boldsymbol{\mu}_k, k = 1, 2, \ldots, K$, we can combine the steps of generating the number of children and their locations. That is, generate N i.i.d $\sim g_K$ and generate $\mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_N$ i.i.d $\sim \sum_{k=1}^{K} \frac{1}{K} f(\mathbf{s}; \boldsymbol{\mu}_k, \Sigma)$ (yielding conditionally independent locations). Here, we envision g_K as the distribution of the sum of K i.i.d. variates from g (this distribution is easy to obtain when g is a member of the exponential family). This allows us to connect Neyman-Scott processes to mixture models for the intensity for the locations, as in Section 8.4.1, i.e., the earlier λf formulation for the intensity surface, with λ being the expected value parameter for g_K and f being the mixture of f_k's, here with uniform weights across components rather than random weights. With Gaussian choices for the f_k, we might work with $\sigma^2 I_{2\times 2}$ as the covariance matrix. In fact, we could introduce a more general Σ but would likely work with a common choice across the k's. So, with regard to generalization and application, it may be useful to interpret a Neyman-Scott process through mixtures.

As a simple special case, we note the compound Poisson process. This process arises by setting the variance of the bivariate offspring density to a point mass at 0. Then, all of the children cluster at the $\boldsymbol{\mu}_k$. We obtain a count at each $\boldsymbol{\mu}_k$ which can be interpreted as a "mark" at that location.

Again, the case above where $\Sigma = \sigma^2 I_{2\times 2}$ is referred to as the (modified) Thomas process in the literature (Illian et al., 2008, p. 377). Here the density of distance, r, of offspring from the parent is evidently a Rayleigh, i.e., $\frac{r}{\sigma^2} e^{-\frac{r^2}{2\sigma^2}}, r \geq 0$. We can obtain the distribution of the distance, d between two random points in the same cluster. It is easy to see that $d \sim 2r$.

As another example, consider the Matérn process. Here, we restrict the N_k offspring of

the parent at $\boldsymbol{\mu}_k$ to be uniformly distributed in a circle of radius R around $\boldsymbol{\mu}_k$. R is a model parameter and the density of distance from the parent is $\frac{2r}{R^2}, 0 \leq r \leq R$. Given R, we can obtain the distribution of the distance between two random points in the same cluster. It is $f_{interpoint}(d) = \frac{4d}{\pi R^2}\left(\cosh\frac{d}{2R} - \frac{d}{2R}\sqrt{1-\frac{d^2}{4R^2}}\right), 0 \leq d \leq 2R$. We leave this as an exercise.

It is known that the set of Neyman-Scott processes is equivalent to the set of Cox processes (see, e.g., Cressie, 1993, p. 663). This should not be surprising since both begin with a first stage NHPP and then introduce process realizations to provide the second stage.

In extending this modeling idea a bit, we can imagine that there are parents outside of D who have children inside D and parents inside D who have some children outside of D. Careful investigation of this situation is taken up in Chakraborty and Gelfand (2009) which we review in Section 8.9. The goal there is to consider measurement error in point patterns so that there is an observed point pattern along with a true unobserved point pattern (i.e., with the true locations). With a simple measurement error model for observed varying around true, say a bivariate normal distribution, the challenge falls to modeling the latent true point pattern. Now the above setting emerges. Due to measurement error, the true value can be in D and the observed value outside or vice versa.

8.6.2 Shot noise processes

We offer a few words regarding the so-called shot noise process. The basic idea is, again, to define a process in two stages. For a region D, draw a point pattern $\mathbf{S}$ over D, say, from a homogeneous or nonhomogeneous Poisson process. Then, assign a random mass to each sampled location. The process realization at $\mathbf{s}$ is $Y(\mathbf{s}) = \sum_{\mathbf{s}_i \in \mathbf{S}} h(\mathbf{s}-\mathbf{s}_i; m(\mathbf{s}_i))$. Forms for h include $h(\mathbf{s}-\mathbf{s}_i; m(\mathbf{s}_i)) = f(\mathbf{s}-\mathbf{s}_i)m(\mathbf{s}_i)$ where f is a density over D and $m(\mathbf{s}_i$ is a positive random variable. m's might be i.i.d., free of $\mathbf{s}$ or a regression on some covariate, say, $X(\mathbf{s})$ over D or a process realization over D, like a log Gaussian process or a gamma process. So, $m(\mathbf{s}_i)$ denotes the contribution to $Y(\mathbf{s})$ from the point at $\mathbf{s}_i$ and $Y(\mathbf{s})$ accumulates the "shots" arising from the realization, $\mathbf{S}$. In essence, if we have a realization of a marked point process (Section 8.7), $\mathbf{S}_M = \{\mathbf{s}_i, m(\mathbf{s}_i\}$, the *impulse* function h provides what is referred to as a shot noise random field over all of D.

We also can make connections to implementation of constructions such as kernel convolution (see Section 3.8.2) and predictive processes (see Section 12.4). In both cases the marks are Gaussian, independent in the former, from a GP in the latter. For kernel convolution, f is a suitable kernel function; for predictive processes it arises from the correlation function for the process of interest (see Section 12.4). The wrinkle added by the shot noise process is that the $\mathbf{s}_i$'s are a realization from a point process.

Generically, we can compute $E_{\mathbf{S}}Y(\mathbf{s})$ as a single integral over $\mathbf{S}$, given $m(\cdot)$, i.e., $\int h(\mathbf{s}-\mathbf{s}'; m(\mathbf{s}'))\lambda(\mathbf{s}')d\mathbf{s}'$ or as a double integral over the randomness in m using Campbell's theorem.

We introduce shot noise processes here as models for intensities. That is, we view $Y(\mathbf{s})$ as $\lambda(\mathbf{s})$, creating a random intensity. We supply a Cox process that is an alternative to a log Gaussian Cox process. More specifically, suppose, as in Section 8.4.2, we write $\lambda(\mathbf{s}) = e^{X^T(\mathbf{s})\beta}\lambda_0(\mathbf{s})$ where the exponential term is the usual deterministic specification we have used before and now, $\lambda_0(\mathbf{s})$ is a mean 1 shot noise process so that $\lambda(\mathbf{s})$ is *centered* around the deterministic component. Suppose we adopt the form $\lambda_0(\mathbf{s}) = \sum_{\mathbf{s}_i \in \mathbf{S}} f(\mathbf{s}-\mathbf{s}_i)m(\mathbf{s}_i)$, with $\mathbf{S}$ drawn from an HPP(λ) and $m(\mathbf{s}_i)$ a constant, m. From Campbell's theorem, we have $E(\lambda_0(\mathbf{s})) = m\lambda = 1$ so $m = 1/\lambda$ (and, then, of course, $E(\lambda_0(D)) = |D|$).

The likelihood arises in two stages, following the process definition. That is, for $\mathbf{S}_{obs}$, we have

$$L(\beta, \lambda_0(\mathbf{s}), \mathbf{s} \in D; \mathbf{S}_{obs}) = e^{-\int_D e^{X^T(\mathbf{s})\beta}\lambda_0(\mathbf{s})d\mathbf{s}} \Pi_i e^{X^T(\mathbf{s}_i)\beta}\lambda_0(\mathbf{s}_i). \tag{8.14}$$

Approximation to the stochastic integral would be done using representative points as described in Section 8.4. However, note that the likelihood in (8.14) is conditional on $\mathbf{S}$, a realization from an HPP(λ), i.e., $\lambda_0(\mathbf{s})$ is a function of $\mathbf{S}$. That is, we first draw $\mathbf{S}$ given λ and then we draw $\mathbf{S}_{obs}$ given β and $\mathbf{S}$. In the context of this book, such hierarchical specification is routine and causes no problems.

A version considered in, e.g., Møller and Waggepetersen (2007) extends the point pattern realization to a Poisson process over $D \times R^+$, yielding a point pattern which we denote as $\{(\mathbf{s}_i, \lambda_i)\}$. Then, $\lambda(\mathbf{s}) = \sum_{\{(\mathbf{s}_i,\lambda_i)\}} \lambda_i f(\mathbf{s} - \mathbf{s}_i)$. We see that $\lambda(\mathbf{s}) = \sum_i \lambda_i(\mathbf{s})$ so that realizations from $\lambda(\mathbf{s})$ arise as a superposition or union of independent Poisson processes with intensities $\tilde{\lambda}_i(\mathbf{s}) = \lambda_i f(\mathbf{s} - \mathbf{s}_i)$ and so provide a natural clustering process.

In fact, we can see that the Neyman-Scott process can be viewed as a special case of the shot noise process. Simply, let the $\mathbf{s}_i$ play the role of the $\boldsymbol{\mu}_i$, the cluster centers, drawn from an HPP or a NHPP over D and set the λ_i all to a constant, say, δ. Then locally, at $\boldsymbol{\mu}_i$, we expect δ offspring.

8.6.3 Gibbs processes

Having invested a large part of a chapter (Chapter 4) on the development of Markov random fields and, in particular, on the formalization of Gibbs distributions, it is attractive that here, in the context of developing point process models that are not Poisson, we can revisit these specifications. In fact, we can define the probability density of a point process over a bounded set D as the Radon-Nikodym derivative for the process measure with regard to an HPP with unit intensity over D. In particular, paralleling Chapter 4, we say that the point process model over D is a finite Gibbs process if, for n locations, its location density is

$$f(\mathbf{S}) = \exp(-Q(\mathbf{S})). \tag{8.15}$$

Here,

$$Q(\mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_n) = c_0 + \sum_{i=1}^{n} h_1(\mathbf{s}_i) + \sum_{i \neq j} h_2(\mathbf{s}_i, \mathbf{s}_j) + \cdots + h_n(\mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_n)\ . \tag{8.16}$$

In (8.16), the h's are potentials of order $1, 2, \ldots, n$, respectively, each symmetric in its arguments. Here, c_0 plays the role of a *normalizing* constant to make $f(\mathbf{S})$ integrate to 1 over $\times D^n$. With potentials only of order 1, we obtain a NHPP with $\lambda(\mathbf{s}) = e^{-h_1(\mathbf{s})}$. Higher order potentials capture/control interaction. As with Markov random fields, we only look at pairwise interactions, i.e., we only include h_1 and h_2. To guarantee integrability, we must take $h_2 \geq 0$. This implies that we can only capture inhibition. In other words, if we require $Q(\mathbf{s}_1, \mathbf{s}_2) \geq c_0 + h_1(\mathbf{s}_1) + h_1(\mathbf{s}_2)$, this means for pairs of points at a given distance, $f(\mathbf{s}_1, \mathbf{s}_2)$ puts less mass under the Gibbs specification than with the corresponding NHPP; we encourage inhibition. If $h_1(\mathbf{s})$ is constant, we have a homogeneous Gibbs process.

In this regard, the most common form for h_2 is $\phi(||\mathbf{s} - \mathbf{s}'||)$, e.g., $\phi(||\mathbf{s} - \mathbf{s}'||) = -||\mathbf{s} - \mathbf{s}'||^2/\tau^2$. Conveniently, the Papangelou conditional intensity has a simple form in this case,

$$\lambda(\mathbf{s}|\mathbf{S}) = \exp(-(h_1(\mathbf{s}) + \sum_{i=1}^{n} \phi(\|\mathbf{s} - \mathbf{s}_i\|))). \tag{8.17}$$

We have the intensity for the new point adjusted by its interaction with all of the points in the given $\mathbf{S}$. Attractively, the unknown normalizing constant cancels from the conditional intensity, which we leave as an exercise.

Forward simulation of realizations from a Gibbs process is straightforward using MCMC, with birth and death in order to have n random, as presented in Møller and Waagepetersen

(2004) (see, also, Illian et al., 2008). Intuitively, this should be the case since we have the full conditional intensities in (8.17) to use in order to develop a Gibbs sampler. However, Bayesian model fitting for Gibbs processes, using MCMC, requires this constant since it will be a function of the parameters in h_1 and h_2. Unfortunately, this constant will be intractable (as an n-dimensional integral over $\times D^n$). A clever auxiliary variables approach to dealing with this challenge has been developed by Møller and collaborators and is presented in Section 8.6.4.

We conclude this subsection by presenting some specific examples of Gibbs processes which have received attention in the literature. They are specified through $\phi(d)$ where, again, d is an interpoint distance. The Strauss process sets $\phi(d) = \beta, d \leq d_0, = 0, d > d_0$. We see that, when $\beta > 0$, $e^{-\phi(d)} \leq 1$ for all d implying inhibition or repulsion. That is, with $\beta > 0$, the interaction term downweights patterns with more points close to each other. Choosing $\beta < 0$ implies clustering but raises the integrability challenge; the resulting density for $\mathbf{S}$ will not be integrable. The so-called hardcore process is an extreme case, setting $\phi(d) = \infty, d \leq d_0, = 0, d > d_0$. Now, the density is 0 for all $\mathbf{S}$ having a pair of points less than d_0 apart. Given these two examples, we can imagine other choices for $\phi(d)$ (see, for example, Illian et al., 2010, Section 3.6) with the only constraint being that $\phi(d) \geq 0$.

8.6.4 Further Bayesian model fitting and inference

The challenge of Bayesian model fitting beyond NHPP's is often the difficulty in working with the likelihood. Some processes are specified only constructively, so that the likelihood may not be available explicitly. However, the challenge with the foregoing models is a bit more direct. For all of the Gibbs distributions we have considered the likelihood is available up to a normalizing constant. As we noted, this constant is computationally intractable but cannot be ignored since it is a function of the model parameters. Here, we show how clever ideas from Berthelsen and Møller (2003, 2004, 2006, 2008) enable us to handle MCMC model fitting in such cases.

The problem can be generically formulated as follows. We use the bracket notation, [], to simplify expressions. We seek to sample from the posterior, $[\boldsymbol{\theta}|\mathbf{y}] \propto [\mathbf{y}|\boldsymbol{\theta}][\boldsymbol{\theta}]$ where $[\mathbf{y}|\boldsymbol{\theta}] = q_{\boldsymbol{\theta}}(\mathbf{y})/\mathbf{C}_{\boldsymbol{\theta}}$ with $\mathbf{C}_{\boldsymbol{\theta}}$ intractable. The usual Metropolis-Hastings probability takes the form

$$\alpha(\boldsymbol{\theta}';\boldsymbol{\theta}) = \min\left(1, \frac{q_{\boldsymbol{\theta}'}(\mathbf{y})\mathbf{C}_{\boldsymbol{\theta}}[\boldsymbol{\theta} \mid \boldsymbol{\theta}'][\boldsymbol{\theta}']}{q_{\boldsymbol{\theta}}(\mathbf{y})\mathbf{C}_{\boldsymbol{\theta}'}[\boldsymbol{\theta}'|\boldsymbol{\theta}][\boldsymbol{\theta}]}\right). \tag{8.18}$$

Suppose we introduce a latent variable $\mathbf{x}$ with the same support as $\mathbf{y}$, with density $[\mathbf{x}|\boldsymbol{\theta}, \mathbf{y}]$. The Hastings ratio gets uglier, now

$$\alpha(\boldsymbol{\theta}', \mathbf{x}'; \boldsymbol{\theta}, \mathbf{x}) = \min\left(1, \frac{[\mathbf{x}'|\boldsymbol{\theta}', \mathbf{y}]q_{\boldsymbol{\theta}'}(\mathbf{y})[\boldsymbol{\theta}']\mathbf{C}_{\boldsymbol{\theta}}[\boldsymbol{\theta}, \mathbf{x}|\boldsymbol{\theta}', \mathbf{x}']}{[\mathbf{x}|\boldsymbol{\theta}, \mathbf{y}]q_{\boldsymbol{\theta}}(\mathbf{y})[\boldsymbol{\theta}]\mathbf{C}_{\boldsymbol{\theta}'}[\boldsymbol{\theta}', \mathbf{x}'|\boldsymbol{\theta}, \mathbf{x}]}\right). \tag{8.19}$$

But now we have the flexibility to specify $[\boldsymbol{\theta}', \mathbf{x}'|\boldsymbol{\theta}, \mathbf{x}] = [\mathbf{x}'|\boldsymbol{\theta}', \boldsymbol{\theta}, \mathbf{x}][\boldsymbol{\theta}'|\boldsymbol{\theta}, \mathbf{x}]$. First we set $[\mathbf{x}'|\boldsymbol{\theta}', \boldsymbol{\theta}, \mathbf{x}] = q_{\boldsymbol{\theta}'}(\mathbf{x}')/\mathbf{C}_{\boldsymbol{\theta}'}$. This simplifies (8.19) to

$$\alpha(\boldsymbol{\theta}', \mathbf{x}'; \boldsymbol{\theta}, \mathbf{x}) = \min\left(1, \frac{[\mathbf{x}'|\boldsymbol{\theta}', \mathbf{y}]q_{\boldsymbol{\theta}}(\mathbf{x})q_{\boldsymbol{\theta}'}(\mathbf{y})[\boldsymbol{\theta}'][\boldsymbol{\theta}|\boldsymbol{\theta}', \mathbf{x}']}{[\mathbf{x}|\boldsymbol{\theta}, \mathbf{y}]q_{\boldsymbol{\theta}'}(\mathbf{x}')q_{\boldsymbol{\theta}}(\mathbf{y})[\boldsymbol{\theta}][\boldsymbol{\theta}'|\boldsymbol{\theta}, \mathbf{x}]}\right). \tag{8.20}$$

We see that the normalizing constants have disappeared and what remains is to choose $[\mathbf{x}|\boldsymbol{\theta}, \mathbf{y}]$ and $[\boldsymbol{\theta}'|\boldsymbol{\theta}, \mathbf{x}]$ which we simplify to $[\boldsymbol{\theta}'|\boldsymbol{\theta},]$. Recalling that for us $\mathbf{y}$ is $\mathbf{S}_{obs}$, a simple choice is to use an HPP for $\mathbf{x}'$ with $\hat{\lambda}$ as a function of $\mathbf{y}$ taken as $N(D)/|D|$. Berthelsen and Møller (2003) recommend using perfect sampling to draw the proposal $\mathbf{x}'$ from $q_{\boldsymbol{\theta}'}(\mathbf{x}')/\mathbf{C}_{\boldsymbol{\theta}'}$, asserting that MCMC draws will be expensive and convergence will be slow. Sampling $[\boldsymbol{\theta}'|\boldsymbol{\theta},]$ is usually standard, using random walk or independence proposals.

The discussion above shows that, for clustering, we can fit a shot noise process model using the two-stage likelihood described there while for inhibition we can use a Gibbs process model employing the foregoing MCMC algorithm to deal with the normalizing constant. The challenge that remains for Bayesian model fitting is how to handle the case where the likelihood is not available explicitly.

8.6.5 Implementing fully Bayesian inference

The foregoing and, in addition, Section 8.5, naturally connect to the opportunity to implement full inference within the Bayesian framework. Using our customary notation, suppose we have a model of the form $[\mathbf{S}|\boldsymbol{\theta}][\boldsymbol{\theta}]$ where the first (likelihood) term can be written explicitly. Using the discussion above, we have shown how we can fit a variety of point pattern models, that is, we can now obtain posterior samples from $[\boldsymbol{\theta}|\mathbf{S}] \propto [\mathbf{S}|\boldsymbol{\theta}][\boldsymbol{\theta}]$. Then, for model features, we need to sample $[\mathbf{S}_{pred}|\mathbf{S}_{obs}] = \int [\mathbf{S}_{pred}|\boldsymbol{\theta}][\boldsymbol{\theta}|\mathbf{S}_{obs}]d\boldsymbol{\theta}$. This will mean drawing a $\mathbf{S}_b$ given a $\boldsymbol{\theta}_b$. For NHPP's and Cox processes, Section 8.5 describes how to do this. For Gibbs processes, Section 8.6.4 offers a route, i.e., drawing from $q_{\boldsymbol{\theta}_b}(\mathbf{S})/\mathbf{C}_{\boldsymbol{\theta}_b}$. Again, MCMC or, when available, perfect sampling, as noted above, can provide such a draw.

At this point, we earn our reward. Given we have obtained posterior predictive samples $\mathbf{S}_b, b = 1, 2, \ldots, B$, we can calculate sample analogs of any population functionals of interest. For instance, we can obtain posterior samples of $\sum_{\mathbf{s}_i \in \mathbf{S}, \mathbf{S} \in D} 1(N(\mathbf{0}, d; \mathbf{S} - \mathbf{s}_i) > 0)$ or $\sum_{\mathbf{s}_i \in \mathbf{S}, \mathbf{S} \in D} N(\mathbf{0}, d; \mathbf{S} - \mathbf{s}_i)$ with $N(\mathbf{0}, d, \mathbf{S} - \mathbf{s}_i)$, defined above (Section 8.3). Hence, using the left side of Campbell's theorem, we immediately obtain Monte Carlo integrations for say, $\lambda|D|G(d)$ or $\lambda^2|D|K(d)$. In fact, with rescaling, we can obtain Monte Carlo integrations for $G(d)$ and $K(d)$. That is, we can obtain model-based (parametric) inference for these analytically intractable quantities, rather than customary nonparametric (non-model-based) estimates such as those given in (8.7) and (8.9). As a result, we avoid the need for edge corrections.

We can also investigate the analogues of the sample residuals offered in Baddeley et al. (2005). For point patterns, they define analogues of raw residuals in linear models.[4] As an illustration, for a Gibbs process, they define the raw residual for a set $B \subset D$, $\hat{R}(B) = N(B|\mathbf{S}) - \int_B \hat{\lambda}(\mathbf{s}|\mathbf{S})d\mathbf{s}$ where $N(B|\mathbf{S})$ is the observed number of points in $\mathbf{S}$ that are in B and the integral is over the estimated Papangelou intensity as a function of the parameters in $Q(\mathbf{S})$. The realized residual here is $R(B) = N(B|\mathbf{S}) - \int_B \lambda(\mathbf{s}|\mathbf{S})$ where the integral, hence $R(B)$ is a function of the parameters in $Q(\mathbf{S})$. That is, we can obtain posterior samples of *realized* residuals to examine their posterior distributions.

With posterior samples for the objects above, we can also attach posterior uncertainty to these sums/integrals, in addition to a posterior mean. In summary, through posterior sampling of point patterns, we can explore model-based inference for other process features discussed in the literature, e.g., the spherical contact distribution function and other morphological functions (Illian et al., 2008).

8.6.6 An example

As an example, we consider the Japanese pine dataset consisting of the locations of 65 saplings in a 5.7×5.7 square meter sampling region in a natural stand (taken from "Analysing spatial point patterns in R" by A. Baddeley (http://www.csiro.au/content/pf16h). The point pattern is shown in Figure 8.13 (top left). The figure suggests that spacing between saplings may be larger than expected under CSR, perhaps reflecting competition for resources. The G and F functions in Figure 8.13 don't provide evidence of

[4]That is, $Y_i - \mathbf{X}_i^T\hat{\boldsymbol{\beta}}$. The Bayesian version of a raw residual is a *realized* residual, that is, $Y_i - \mathbf{X}_i^T\boldsymbol{\beta}$ which has a posterior distribution.

Figure 8.13 *Plots of the G, F and L functions for the Japanese pine data.*

inhibition but the L plot suggests fewer points than expected in neighborhoods of a given point. Hence, we fit a Strauss process to the data with Papangelou conditional intensity, $\log \lambda(\mathbf{s}|\mathbf{S}) = \gamma + \beta, d \leq d_0, = \gamma, d > 0$, using the approach of Section 8.6.4. The posterior distributions for λ and for β under two choices for d_0, 0.1 and 0.05, are shown in Figure 8.14. We see that γ is a bit smaller under the latter and in either case β is well away from 0, supporting inhibition.

Figure 8.14 *Figure showing the posterior distributions for λ and for β under two choices for d_0,* 0.1 *(top) and* 0.05 *(bottom), from fitting a Strauss process to the Japanese pine data.*

Figure 8.15 *Kernel intensity estimates for the three tree species in the Duke Forest dataset viewed as three different marks.*

8.7 Marked point processes

Marked point processes add considerable vitality to the investigation of point patterns. That is, each point carries the extra information of a mark which captures a feature of whatever object was observed at that point. Jointly, we may be better able to understand the process yielding the points. In other words, one could ignore the marks (essentially marginalizing over the marks) but it is evident that this will sacrifice potentially important information. Marks are often discrete such as providing labels for different types of cancers over a point pattern of cancer cases (see Section 8.7.4.2) or labels for different species of trees in a forest. In this case, the data is often referred to as *multi-type*. In fact they can even index time points, leading to space time point patterns (see Section 8.8). In each of these settings, we would be interested in seeing differences between the point patterns; aggregating them would lose this opportunity. Figure 8.15 shows kernel intensity estimates for the three tree species in the Duke Forest data viewed as three different marks. We clearly see differences in the intensities. With a fuller analysis, we might try to explain these differences using local environmental features such as soil type, soil moisture, or light availability (exposed canopy area).

Continuous marks may also be of interest. For trees, we may record a height or a basal area. For an earthquake, the mark may be its strength, say on the Richter scale. As noted in the introduction, the emphasis here is on looking at both the location and the mark as random, with appropriate modeling. This contrasts with the usual geostatistical analysis (Chapter 2), where only the feature at a location is viewed as random. (See Section 8.7.3 for further discussion in this regard.)

8.7.1 Model specifications

From a mathematical perspective, a mark is merely viewed as adding an extra coordinate to the observation, i.e., we observe $(\mathbf{s}, m)$ as a point over $D \times M$ where M is the support set for the marks. If the marks are continuous, M will be some subset of R^1 and the marked point pattern is equivalent to a point pattern over $D \times M$. If the marks are discrete, M will be a set of labels, say, $l, l = 1, 2, \ldots, L$ and the overall point pattern can be viewed as a set of L point patterns, each over D. We remark that the notation $(\mathbf{s}, m)$ is often modified according to interpretation of the marked point process. Sometimes we might write $\mathbf{S}$ with $m(\mathbf{s})), \mathbf{s} \in \mathbf{S}$, other times, $\mathbf{S}_m$. The former suggests drawing locations and then assigning labels, the latter suggests selecting labels and then drawing locations. (See Section 8.7.3.)

If we follow a product space representation for a marked point process, then a marked point process is really just a point process over this product space. So, we can adopt much of the earlier theory in this chapter. For instance, $N(B \times A)$ is the number of points with

locations in $B \subseteq D$ and marks in $A \subseteq M$. Defining a random counting measure leads us defining count random variables on a σ-algebra of sets over $D \times M$. In turn, this suggests a Poisson marked point process where $N(B \times A) \sim Po(\lambda(B \times A)$ for a suitable intensity measure $\lambda(B \times A)$, with independence of the count variables over disjoint product sets.

Stationarity assumes that, for any n, $\{\mathbf{s}_i, m_i, i = 1, 2, .., n\} \sim \{\mathbf{s}_i + \mathbf{h}, m_i, i = 1, 2, .., n\}$. This definition says that points are translated but marks remain the same. It may be a sensible assumption for marks that are size features. Applied to a Poisson marked point process, it simplifies the intensity measure to $\lambda(B \times A) = \lambda|B|\nu(A)$ where $Q\nu(\cdot)$ is a probability distribution over M. The marginal point process is an HPP with intensity λ. Conditional on locations, the marks are i.i.d. according to ν.

The process may have an intensity function, i.e., $\lambda(\mathbf{s}, m)$ such that $E(N(B \times A)) = \lambda(B \times A) = \int_B \int_A \lambda(\mathbf{s}, m) d\mu(m) d\mathbf{s}$. If the marks are continuous, we usually take $\mu(m)$ to be a Lebesgue measure. If the marks are discrete/categorical, we take $\mu(m)$ to be a counting measure and write $E(N(B \times A)) = \lambda(B \times A) = \int_B \sum_{l \in A} \lambda(\mathbf{s}, l) d\mathbf{s}$. In fact, more naturally, we would write $\lambda(\mathbf{s}, l)$ as $\lambda_l(\mathbf{s})$.

For continuous marks, integrating over m yields $\lambda(\mathbf{s}) = \int_M \lambda(\mathbf{s}, m) dm$, the intensity for the point process of locations. In fact, $f(m|\mathbf{s}) = \frac{\lambda(\mathbf{s},m)}{\lambda(\mathbf{s})}$ is the conditional density for the mark at location $\mathbf{s}$. For categorical marks, the marginal intensity is $\lambda(\mathbf{s}) = \sum_{l=1}^{L} \lambda_l(\mathbf{s})$. Now, the conditional probability for mark l at location $\mathbf{s}$ is $\frac{\lambda_l(\mathbf{s})}{\lambda(\mathbf{s})}$. In epidemiological settings, we would likely investigate the relative risk or relative intensity for mark l' to mark l, $\frac{\lambda_{l'}(\mathbf{s})}{\lambda_l(\mathbf{s})}$ with interest in how it varies over D.

More general marked point process models can be developed following the earlier ideas for unmarked point processes. One possibility would be thinning, using mark-dependent thinning (Section 8.5), e.g., with discrete marks, introducing a thinning surface $p_{l,D}$ for each mark l.

We can also envision random field mark processes. Such processes assume that the realization of the point pattern $\mathbf{S}$ over D is independent of the realization of the stochastic process $m_D = \{m(\mathbf{s}) : \mathbf{s} \in D\}$. This model would be appropriate if we wanted to assign marks that exhibited spatial structure, e.g., marks are more similar at locations closer to each other than at locations more distant from each other. The special case where m_D is white noise, i.e., random i.i.d. assignment of marks to locations, is a null model that we would hope to reject. Such processes can be readily sampled if both $\mathbf{S}$ and m_D can be sampled. Dependence structure for the joint field over $D \times M$ is very complicated even if m_D is deterministic.

Another option is to extend Gibbs process models. This can be specified most conveniently with discrete marks. In particular, following Section 8.6.3, we can introduce a Gibbs process for each l. Following Section 8.2.2, we can obtain a Papangelou conditional intensity for each l. Now, we might investigate relative conditional intensities. A special case could consider the pairwise interaction terms to be common over l with only the first order intensities dependent on l. It can be shown that such a process is equivalent to generating a realization from an unmarked Gibbs process with this common intensity and then labeling with probabilities according to the conditional mark probabilities above, $q_l(\mathbf{s}) = \frac{\lambda_l(\mathbf{s})}{\lambda(\mathbf{s})}$.

8.7.2 Bayesian model fitting for marked point processes

Model fitting for continuous marks will follow model fitting for unmarked point process models, working with the product space representation, employing one of the foregoing models over $D \times M$. In the literature, the predominance of such examples employ a NHPP with intensity $\lambda(\mathbf{s}, m)$ over $D \times M$. Extended to a LGCP, the foregoing approximate fitting using representative points can be directly carried over.

With discrete marks, we need to model an intensity for each mark. This raises the question of whether the point patterns are dependent. For instance, in ecological processes, we can imagine symbiotic relationships which will encourage approximate co-occurrence of points. We could also imagine competitive processes such that the presence of an individual from one species would discourage the presence of an individual from another species.

Formally, these ideas are different from modeling attraction/clustering or inhibition/repulsion associated with a single species, i.e., with a single intensity. Again, these latter objectives motivate pairwise interaction processes which were discussed in Section 8.6.3.

Returning to dependence between patterns, suppose our modeling assumes the points are conditionally independent given their intensity, as with a NHPP. So, to introduce dependence we must do it for the intensities which, in turn, will impart marginal dependence to the point patterns. If we work with parametric intensities, say, $\lambda(\mathbf{s};\boldsymbol{\theta}_1)$ and $\lambda(\mathbf{s};\boldsymbol{\theta}_2)$, then we would need to introduce a prior for $(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2)$ which makes them dependent. How to do this depends upon the nature of the specification for the λ's. For instance, with regression coefficients, we might assume them to be exchangeable and add another hierarchical level to the model. If we work with nonparametric forms, employing, say, log Gaussian Cox processes, specified through, say, $w_1(\mathbf{s})$ and $w_2(\mathbf{s})$ (Section 8.4.2), we can make these dependent using elementary coregionalization as developed in Chapter 9.

8.7.3 Modeling clarification

A brief clarification of some modeling issues alluded to in the introduction to this Section may prove useful here. Suppose we obtain data in the form $(\mathbf{s}_i, m_i), i = 1, 2, \ldots, n$, where the $\mathbf{s}_i$'s are observed locations and we think of $m_i = L(\mathbf{s}_i)$ as a discrete label, say, from $l = 1, 2, \ldots, M$. So, we think of the $L(\cdot)$'s as marks, indicating which mark was assigned to each of the observed points. If we ignored the marks, under a NHPP model, we know the joint distribution of $(n, \{s_1, s_2, \ldots, s_n\})|\lambda_D$ where $\lambda_D = \{\lambda(\mathbf{s}) : \mathbf{s} \in D\}$. From a Bayesian perspective, we only need to model λ_D to complete the specification. Adopting this perspective, with marks as above, we imagine a point pattern for each label/mark value and would extend to $\lambda_{l,D}$, the intensity associated with each label value. Then, assuming marks are also random, we would assign a prior on labels say $p_l, l = 1, 2, ...L$. In this fashion, we model the joint distribution of location and label as $[location|label][label]$ and we would assume the pairs $(\mathbf{s}_i, L(\mathbf{s}_i))$ are conditionally independent given the $\lambda_l(\mathbf{s})$'s. Under this modeling, we have specified $[\mathbf{S}|\ L = l; \{\lambda_l(\mathbf{s})\}]$.

As in Section 8.7.1, the cumulative intensity is $\lambda(\mathbf{s}) = \sum_l \lambda_l(\mathbf{s})$, which, we note, has nothing to do with the p_l's, and

$$f(\mathbf{s}) = \frac{\lambda(\mathbf{s})}{\lambda(D)} = \sum_l \frac{\lambda_l(\mathbf{s})}{\sum_l \lambda_l(D)}$$

is the marginal location density. Turning to the joint distribution, we have $f_l(\mathbf{s})p_l$ where $f_l(\mathbf{s}) = \lambda_l(\mathbf{s})/\lambda_l(D)$, the location density associated with mark l. We interpret $f_l(\mathbf{s})p_l$ as drawing a label $L = l$ and then locating the label at $\mathbf{s}$ given $L = l$. That is, the draw $(\mathbf{s}, L = l)$ creates the event $(\mathbf{s}, L(\mathbf{s}) = l)$. Note that this has nothing to do with the joint intensity $\lambda(\mathbf{s}, m)$ discussed in Section 8.7.1 which adopts a counting measure for m when m is discrete. It therefore yields $\lambda(\partial\mathbf{s}, \{l\}) \approx \lambda(\mathbf{s}, l)|\partial\mathbf{s}|$ and thus $\lambda(\mathbf{s}, l) = \lambda_l(\mathbf{s})$, again, free of the p_l's.

Now, consider conditioning in the opposite direction. That is, we draw a location and then assign a label to the location. Again, the draw $(\mathbf{s}, L = l)$ creates the event $(\mathbf{s}, L(\mathbf{s}) = l)$. Using Bayes' Theorem, $P(L = l|\mathbf{s}) = f_l(\mathbf{s})p_l/\tilde{f}(\mathbf{s})$. We have the familiar rescaling of the prior weights with $\tilde{f}(\mathbf{s}) = \sum_l p_l f_l(\mathbf{s})$, a mixture density having nothing to do with the foregoing marginal location density $f(\mathbf{s})$.

We can imagine that the modeling situation is reversed. The label is viewed as the response at a location; now we would be modeling the joint distribution as $[label|location][location]$. The model for location would now have a single λ_D and the distribution for label given location would be a multinomial trial with location-specific probabilities. In the case of two labels, we might adopt a logit model, i.e., a model for $\log\frac{P(L(\mathbf{s})=1)}{P(L(\mathbf{s})=2)}$. In general, the joint distribution becomes $P(L = l|\mathbf{s})f(\mathbf{s})$ where, as usual, $f(\mathbf{s}) = \lambda(\mathbf{s})/\lambda(D)$. Again, this means that we draw a location $\mathbf{s}$ and then assign a label $L = l$ to the location, creating the event $(\mathbf{s}, L(\mathbf{s}) = l)$. Turning to Bayes' Theorem, $f(\mathbf{s}|L = l) = P(L = l|\mathbf{s})f(\mathbf{s})/\int_D P(L = l|\mathbf{s})f(\mathbf{s})d\mathbf{s}$ and, in fact $f(\mathbf{s}|L = l) = f_l(\mathbf{s})$, the location density associated with mark l. Thus, $\lambda_l(\mathbf{s}) = c_l P(L = l|\mathbf{s})f(\mathbf{s})$ where the constant c_l cannot be identified; we can only learn about the location density for mark l but not the intensity for this mark. A last calculation shows that $\lambda(\mathbf{s}) = \sum_l \lambda_l(\mathbf{s}) = \frac{\sum_l c_l P(L=l|\mathbf{s})}{\lambda(D)}\lambda(\mathbf{s})$. Hence, $\sum_l c_l P(L = l|\mathbf{s}) = \lambda(D)$ but the c_l are not determined.

In summary, note the fundamental difference between the two joint modeling scenarios. In the first case, it is most natural that the L's have nothing to do with locations. There is a single distribution for them and given a realization (label), we have an associated intensity which provides the joint distribution for the points having that label. In the second case, we formalize an uncountable collection of $L(\mathbf{s})$'s with a single intensity for the observed points. In other words, conceptually, the joint distributions for the first case live in a different space from the joint distributions for the second case. See, also, the chapter by A. Baddeley in Gelfand et al. (2010, Chapter 21) in this regard.

This leads to the more general question of whether covariates are spatially referenced or not. Typically, variables such as sex or species type are not spatially referenced. They would more naturally be marks and we would find ourselves in the first modeling scenario, seeking to compare point patterns. If we turn them into response variables, this substantially changes the problem. It would presume that an individual can exist at every location whence we can imagine a label for every location. Expressed in different terms, we note that a point pattern can be expressed using an indicator variable $V(\mathbf{s})$ which takes the value 1 if there is a point in the pattern at $\mathbf{s}$ and takes the value 0 otherwise. That is, the pattern is $\{V(\mathbf{s}) = 1, \mathbf{s} \in D\}$. V is equal to 1 for only a finite set of locations and is 0 for an uncountable number of locations. This clarifies the impossibility of modeling $P(V(\mathbf{s}) = 1)$ from a point pattern and is another way of distinguishing the foregoing order of conditioning in the modeling. It is intimately connected to the distinction between modeling for presence-only data vs. modeling presence-absence data; this issue forms the basis for Section 8.9.2. For presence-only data, we see a point pattern of locations; for presence-absence data, we see a collection of Bernoulli trials at a given set of locations.

Other covariates are naturally spatially referenced such as elevation or aspect. In this case, using a NHPP, we would naturally insert them into the model for the intensity for the point pattern. That is, they may illuminate where points are more or less likely to be. Denoting such a covariate by $X(\mathbf{s})$, the intensity would become $\lambda(X(\mathbf{s}) : \boldsymbol{\theta})$, as discussed in Section 8.4.1. Alternatively, $X(\mathbf{s})$ can be a response variable. In this case, we would build a customary point-referenced spatial model (as in Chapter 2). Presumably we would be interested in interpolating $X(\mathbf{s})$ over D. For instance, with elevation, this would be the goal of stochastic modeling for a digital /elevation/terrain map.

8.7.4 Enriching intensities

So, we can imagine modeling the intensity using covariates of both types. For instance, with mark l and covariate $X(\mathbf{s})$, we might specify the $\lambda(\mathbf{s})$'s using the form:

$$\lambda_l(X(\mathbf{s}); \boldsymbol{\theta}) = \mu + \alpha_l + \beta_l X(\mathbf{s}). \tag{8.21}$$

Figure 8.16 *Jittered residential locations of colon (light circle) and rectum (dark circle) cancer cases, Twin Cities metro and exurban counties, 1998–2002.*

Here, we briefly present the development and extension of (8.21) as discussed in Shengde et al. (2009). To provide concrete motivation, we consider marks for colon and rectum cancer. Colon and rectum cancer share many risk factors, and are often tabulated together as "colorectal cancer" in published summaries. However, recent work indicating that exercise, diet, and family history may have differential impacts on the two cancers encourages analyzing them separately. The data is from the Minnesota Cancer Surveillance System from 1998–2002 over the 16-county Twin Cities (Minneapolis-St. Paul) metro and exurban area. The data consist of two marked point patterns, one for each cancer type and we expect association between the cancer types.

Figure 8.16 shows the seven counties comprising the Twin Cities metro area as those encircled by the dark boundary; also shown are nine adjacent, exurban counties. Within these 16 counties, we have 6544 individuals for analysis. Figure 8.16 plots the approximate locations of the cancers after the addition of a random "jitter" to protect patient confidentiality (explaining why some of the cases appear to lie outside of the spatial domain). The physiological adjacency of the colon and the rectum suggests positive dependence in these point patterns; persons with rectum cancer beyond stage 1 (i.e., regional or distant) are at risk for colon cancer due to metastasis. Moreover, the two cancers likely share unmodeled spatially-varying risk factors (such as local health care quality or availability), also suggesting positive dependence.

We assume a nonhomogeneous Poisson process with intensity function $\lambda(\mathbf{s})$ for all $\mathbf{s} \in D$. Let $\mathbf{X}(\mathbf{s})$ be a vector of location-specific covariates corresponding to a disease case observed at $\mathbf{s}$. For us, a key component of $\mathbf{X}(\mathbf{s})$ is the indicator of whether the case is in the metro area or not. However, in other contexts, we could envision information such as elevation, climate, exposure to pollutants, and so on to be relevant. We model $\lambda(\mathbf{s}) = r(\mathbf{s})\pi(\mathbf{s})$ where $r(\mathbf{s})$ is the population density surface at location $\mathbf{s}$. In practice, we may create such a surface using GIS tools and census data, or we may just work with areal unit population counts, letting $r(\mathbf{s}) = n(A)/|A|$ if $\mathbf{s} \in A$, where, as usual, $n(A)$ is the number of persons residing in A and $|A|$ is the area of A.

Returning to our framework, $r(\mathbf{s})$ serves as an offset and $\pi(\mathbf{s})$ is interpreted as a population adjusted (or *relative*) intensity, which we model on the log scale as

$$\pi(\mathbf{s}) = \exp(\mathbf{X}(\mathbf{s})'\boldsymbol{\beta} + w(\mathbf{s}))\ , \tag{8.22}$$

where $w(\mathbf{s})$ is a zero-centered stochastic process, and $\boldsymbol{\beta}$ is an unknown vector of regression coefficients. If $w(\mathbf{s})$ is taken to be a Gaussian process, we have the previously-discussed log Gaussian Cox process (LGCP). We follow the "representative points" approach resulting in replacement of w_D with a finite set, say $w^* = \{w(\mathbf{s}_j^*),\ j = 1, 2, \ldots, m\}$. Then we revise the NHPP likelihood to

$$L(\beta, w^*, w(\mathbf{s}_1), ...w(\mathbf{s}_n); S)p(w^*, w(\mathbf{s}_i), ...w(\mathbf{s}_n))p(\boldsymbol{\beta})\ . \tag{8.23}$$

Now, we only need to work with an $(n + m)$-dimensional random variable to handle the w's, hence their prior is just an $(n+m)$-dimensional multivariate normal distribution. Note that, in (8.23), we will require that $\mathbf{X}(\mathbf{s})$ be available at each of the above $\mathbf{s}$'s. To this point, we are essentially following Section 8.4.2.

8.7.4.1 Introducing non-spatial covariate information

Next, we introduce nonspatial covariates which we think of as being of two types (though the distinction will depend upon the application). One type of covariate provides marks. For us, this covariate is cancer type (colon vs. rectum), and we are interested in whether the two cancer intensity patterns differ. The second type of covariate we view as an "auxiliary" variable that provides additional information associated with intensity. For us, age and cancer stage are examples of such covariates. Clearly patient age is associated with cancer intensity, but the strength of this association may differ across cancers. We wish to adjust intensity to reflect patient age, analogous to the age standardization used in aggregated areal data settings.

In general, we view these latter covariates as continuous and introduce a second argument into the definition of the intensity, yielding a surface in $\mathbf{s}$ and $\mathbf{v}$ over the product space $D \times \mathcal{V}$ (i.e., the geographic space by the auxiliary covariate space).[5] Here, we are following the path of Section 8.7.1. Therefore, we generalize to

$$\pi(\mathbf{s}, \mathbf{v}) = \exp(\beta_0 + \mathbf{X}(\mathbf{s})'\boldsymbol{\beta} + \mathbf{v}'\boldsymbol{\alpha} + (\mathbf{v} \otimes \mathbf{X}(\mathbf{s}))'\boldsymbol{\gamma} + w(\mathbf{s}))\ , \tag{8.24}$$

where the Kronecker product $\mathbf{v} \otimes \mathbf{X}(\mathbf{s})$ denotes the set of all the first order multiplicative interaction terms between $\mathbf{X}(\mathbf{s})$ and $\mathbf{v}$. When a particular interaction term is not of interest, the corresponding coefficient in $\boldsymbol{\gamma}$ is set to zero. This development is essentially that of Section 8.7.2 while here we have made the intensity $\lambda(\mathbf{s}, m)$ explicit. That is, this expression envisions a conceptual intensity value at each $(\mathbf{s}, \mathbf{v})$ combination. The interaction terms between spatial and non-spatial covariates provide the ability to adjust the spatial intensity by individual risk factors. If we fix $\mathbf{v}$ in (8.24), we can view $\lambda(\mathbf{s}, \mathbf{v}) = r(\mathbf{s})\pi(\mathbf{s}, \mathbf{v})$ as the intensity associated with level $\mathbf{v}$. If we *integrate* over $\mathbf{v}$ (see below), we obtain the (cumulative) marginal intensity $\lambda(\mathbf{s})$ associated with $\pi(\mathbf{s}, \mathbf{v})$. Note that $\pi(\mathbf{s}, \mathbf{v})$ is a Cox process.

Now, introducing marks $k = 1, 2, \ldots, K$, a general additive form for the log relative intensity is

$$\log \pi_k(\mathbf{s}, \mathbf{v}) = \beta_{0k} + \mathbf{X}'(s)\boldsymbol{\beta}_k + \mathbf{v}'\boldsymbol{\alpha}_k + (\mathbf{v} \otimes \mathbf{X}(\mathbf{s}))'\boldsymbol{\gamma}_k + w_k(\mathbf{s})\ . \tag{8.25}$$

[5] In the case of a discrete valued covariate, any integrals over $\mathbf{v}$ in our development are replaced by sums.

We can immediately interpret the terms on the right side of (8.25). The global mark effect is captured with the β_{0k}. Therefore, there is no intercept in $\mathbf{X}(\mathbf{s})$ and we have mark-varying coefficients for the spatially-referenced covariates, reflecting the possibility that these covariates can differentially affect the intensity surfaces of the marks. Similarly, we have mark-varying coefficients for the nuisance variables. We also have mark-varying coefficients for the interaction terms, reflecting possibly different effects of the non-spatial covariates over spatial domains. Finally, we allow the spatial random effects to vary with mark, i.e., a different Gaussian process realization for each k. Dependence in the $w_k(\mathbf{s})$ surfaces may be expected (say, increased intensity at $\mathbf{s}$ for one marked outcome encourages increased intensity for another at that ($\mathbf{s}$), suggesting the need for a *multivariate* Gaussian process over the w_k. Both separable and nonseparable forms for the associated cross-covariance function can be conveniently specified through *coregionalization* as in Section 9.5.

Reduced models of (8.25) are immediately available, including e.g. $w_k(\mathbf{s}) = w(\mathbf{s})$, $\boldsymbol{\beta}_k = \boldsymbol{\beta}$, and $\boldsymbol{\alpha}_k = \boldsymbol{\alpha}$. Another interesting reduced model obtains by setting $\boldsymbol{\gamma}_k = 0$, leading to

$$\log \pi_k(\mathbf{s}, \mathbf{v}) = \beta_{0k} + \mathbf{X}'(\mathbf{s})\boldsymbol{\beta}_k + \mathbf{v}'\boldsymbol{\alpha}_k + w_k(\mathbf{s}) \ . \tag{8.26}$$

This separable form enables us to directly study the effect of the marks on spatial intensity. Specifically, the intensity associated with (8.25) is

$$\lambda_k(s, v) = \exp(\beta_{0k} + \mathbf{v}'\boldsymbol{\alpha}_k) \times r(\mathbf{s}) \exp(\mathbf{X}'(\mathbf{s})\boldsymbol{\beta}_k + w_k(\mathbf{s})) \ . \tag{8.27}$$

We see a factorization into nonspatial nuisance and spatial covariate terms. Presuming the former is integrable over $\mathbf{v}$, the latter, up to a constant, is the "marginal spatial intensity."

Integration of $\lambda_k(\mathbf{s}, \mathbf{v})$, based upon (8.25), can be computed analytically in most cases. When $\mathbf{v}$ is categorical, the likelihood integral involves only integration over the spatial domain D. When $\mathbf{v}$ is continuous, simple algebra shows

$$\begin{aligned} \int_{\mathcal{V}} \lambda_k(\mathbf{s}, \mathbf{v}) dv ds &= r(\mathbf{s}) \exp(\beta_{0k} + \mathbf{X}(\mathbf{s})'\boldsymbol{\beta}_k + w_k(\mathbf{s})) \\ &\quad \times \int_{\mathcal{V}} \exp(\mathbf{v}'\boldsymbol{\alpha}_k + (\mathbf{v} \otimes \mathbf{X}(\mathbf{s}))'\boldsymbol{\gamma}_k) d\mathbf{v} \ . \end{aligned}$$

Suppose, for instance, that there is only one component in $\mathbf{X}(\mathbf{s})$ and one component in $\mathbf{v}$ having range (v_l, v_u). Provided $\alpha_k + X(\mathbf{s})\gamma_k \neq 0$, the marginal intensity $\lambda_k(\mathbf{s})$ is then $int_{\mathcal{V}} \lambda_k(\mathbf{s}, v) dv d\mathbf{s}$, which is equal to

$$\begin{aligned} & r(\mathbf{s}) \exp(\beta_{0k} + \beta_k X(\mathbf{s}) + w_k(\mathbf{s})) \times \int_{\mathcal{V}} \exp(v(\alpha_k + X(\mathbf{s})\gamma_k)) dv \\ & = r(\mathbf{s}) \exp(\beta_{0k} + \beta_k X(\mathbf{s}) + w_k(\mathbf{s})) \\ & \quad \times \frac{1}{\alpha_k + X(\mathbf{s})\gamma_k} \left[\exp(v_u(\alpha_k + z(\mathbf{s})\gamma_k)) - \exp(v_l(\alpha_k + X(\mathbf{s})\gamma_k)) \right] \ . \end{aligned}$$

Turning to the revised likelihood associated with (8.25), let $\{(\mathbf{s}_{ki}, \mathbf{v}_{ki}), i = 1, 2, ...n_k\}$ be the locations and nuisance covariates associated with the n_k points having mark k. The likelihood becomes

$$\prod_k \exp\left(- \int_D \int_{\mathcal{V}} \lambda_k(\mathbf{s}, \mathbf{v}) d\mathbf{v} d\mathbf{s} \right) \times \prod_k \prod_{\mathbf{s}_{ki}, \mathbf{v}_{ki}} \lambda_k(\mathbf{s}_{ki}, \mathbf{v}_{ki}) \ . \tag{8.28}$$

Using the calculations above, the double integral becomes

$$\begin{aligned} & \int_D \int_{\mathcal{V}} \lambda_k(\mathbf{s}, v) dv d\mathbf{s} = \int_D \Big(r(\mathbf{s}) \exp(\beta_{0k} + \beta_k X(\mathbf{s}) + w_k(\mathbf{s})) \\ & \times \frac{1}{\alpha_k + X(\mathbf{s})\gamma_k} \big[\exp(v_u(\alpha_k + X(\mathbf{s})\gamma_k)) - \exp(v_l(\alpha_k + X(\mathbf{s})\gamma_k)) \big] \Big) d\mathbf{s} \ , \end{aligned}$$

provided that the set $\{\mathbf{s} : \alpha_k + X(\mathbf{s})\gamma_k = 0\}$ has Lebesgue measure zero. Hence the difficulty in the likelihood evaluation is the same as in the basic likelihood and can be treated with approximation, as above. In this regard, note that we bound the components of $\mathbf{v}$ in order to integrate explicitly over $\mathbf{v}$. We do not have a stochastic integration with regard to $\mathcal{V}$ as we have over D.

8.7.4.2 Results of the analysis

Previous studies suggest that covariates related to a patient's socioeconomic status (SES) may be related to the patient's risk factors through its impact on diet, health care quality, or propensity to seek care. While our dataset lacks any individual-level SES measures, from census data we have several related tract-level variables: percentage of farm population, percentage of rural population, percentage of people with less than high school education, percentage of minority population, and poverty rate. A preliminary population-adjusted nonspatial Poisson regression analysis of our data on these covariates revealed only poverty rate and the metro indicator as significant predictors.

We consider two location-specific covariates: $z_1(\mathbf{s})$, the metro area indicator, and $z_2(\mathbf{s})$, the poverty rate in the census tract containing $\mathbf{s}$. We also employ two non-location-specific covariates: v_1, cancer stage (set to 1 if the cancer is diagnosed "late" (regional or distant stage) and 0 otherwise), and v_2, the patient's age at diagnosis. The population density $r(\mathbf{s})$ we use for standardization is available at 2000 census tract level, meaning that we assume population density is constant across any tract.

Table 8.1 breaks down the data by stage and metro/non-metro area. We see that 38% of colon cancer cases were diagnosed at an early stage, while 44.5% of rectum cancer cases were. In total, colon cancer is nearly three times as prevalent as rectum cancer in both the metro and non-metro areas. A fact not revealed by the table is that there are 72 individuals who contribute *both* a colon and a rectum tumor. Since this is only around 1% of the total of 6544 individuals, we do not explicitly model this particular kind of dependence, but rather "lump it in" with the bivariate dependence modeled by ρ.

Figure 8.17 shows tract-level maps of population density, $r(\mathbf{s})$, and our two location-specific covariates, $z_1(\mathbf{s})$ and $z_2(\mathbf{s})$. Not surprisingly, the central metro areas are the most populated. The poverty rate is fairly uniform except for high rates in a concentrated portion of the central metro.

We fit our model, using independent Inverse Gamma$(2, 0.5)$ priors for σ_1^2 and σ_2^2, and a $Unif(-0.999, 0.999)$ prior for ρ. The scale of the spatial decay parameter ϕ is determined by the distance function employed. In this application, we started with a $Unif(130, 390)$ prior for ϕ, so that the effective range lies between one-fourth and three-fourths of the maximal distance between any two knots. As expected, ϕ is only weakly identified, so a fairly informative prior is needed for satisfactory MCMC behavior. For simplicity, we simply fix the range parameter at $\phi = 195$, so that the effective range is roughly half of the

	total	late=0	late=1	metro	non-metro
all	6544	2606(40%)	3938(60.2%)	5481(83.8%)	1063(16.2%)
colon	4857	1855(38%)	3002(61.8%)	4079(84%)	778(16%)
rectum	1687	751(44.5%)	751(55.5%)	1402(83.1%)	285(16.9%)
ratio	2.88	2.47	4.0	2.91	2.73

Table 8.1 *Table of colorectum cancer patients' characteristics in metro and adjacent area of Minnesota. Percentages across appropriate columns are given in parentheses, and "ratio" gives the ratio of colon to rectum cases.*

Figure 8.17 *Left, population density by tract; middle, metro/non-metro area; right, poverty rate by tract.*

model	p_D	DIC
GLM (no residuals)	11.8	1194.4
Univariate spatial residuals	72.0	692.4
Bivariate spatial residuals	80.2	688.8

Table 8.2 *Model comparison using effective model size p_D and DIC score. GLM refers to generalized linear model having no random effects.*

maximal distance. A random-walk Metropolis-Hastings algorithm is used to draw posterior samples.

Table 8.2 compares the effective model size and DIC score of three models. It can be seen that the no-random effect model (GLM) is unacceptably bad, and the model with a single set of spatial residuals is not much worse than the bivariate residual model. This suggests that the two sets of residuals are fairly similar, and that ρ is close to 1.

Table 8.3 shows parameter estimates from some of our models. We parameterize so that the top rows concern the fixed effects for colon cancers, $\boldsymbol{\beta}_1$, but the second set of rows give the *differential* effect in the rectum cancer group, $\boldsymbol{\Delta} \equiv \boldsymbol{\beta}_2 - \boldsymbol{\beta}_1$. Thus, any 95% Bayesian confidence intervals that exclude 0 in this part of the table suggest a variable that has a significantly different impact on the two cancers.

In general, the effects of the non-spatial covariates are fairly similar across the models considered. We find that in the metro area there are relatively fewer cases of both colon and rectum cancer. This is consistent with statewide patterns of colorectal cancer occurrence in Minnesota, where higher age-adjusted rates are often found in non-metro areas. However, there is no significant change in this relationship in the rectum group relative to the colon group. Turning to the non-location-specific covariates, age is significantly associated with increasing colon cancer, but a somewhat surprising relative *decrease* in rectum cancer. This difference (-0.18) is statistically significant, but not large enough in magnitude to make the overall age effect in the rectum group negative. A look at the data bears this out, with rectum cancers arising in a somewhat younger population; our preliminary Poisson regression also concurs, though here the relative decrease in the rectum group is not significant. Late detection provides another interesting difference between the colon and rectum groups: while there are significantly more cases diagnosed late than early, the effect of late diagnosis is significantly smaller in the rectum group (point estimate -0.26). Thus public health interventions to encourage screening and early detection of colorectal cancer will have significantly greater impact on prevention for colon than for rectum. The metro-age interaction shows that the effect of age on colon cancer is significantly less pronounced in the metro area; a smaller "age adjustment" to the colon cancer intensity process is needed in the metro area. This effect is largely absent for rectum cancer, but this difference is not quite statistically significant. Finally, the estimate of ρ is very close to 1, indicating

colon	BSR	USR	GLM
intercept	-8.76 (-9.12,-8.44)	-8.75 (-9.25,-8.40)	-8.91 (-8.99,-8.83)
metro	-0.23 (-0.49,0.04)	-0.19 (-0.42,0.06)	-0.21 (-0.29,-0.14)
poverty	-2.01 (-2.47,-1.55)	-1.90 (-2.36,-1.47)	-0.26 (-0.61,0.09)
age	0.36 (0.31,0.40)	0.36 (0.31,0.40)	0.32 (0.28,0.36)
late	0.48 (0.42,0.54)	0.48 (0.42,0.54)	0.48 (0.43,0.54)
metro*age	-0.06 (-0.11,-0.02)	-0.06 (-0.11,-0.02)	-0.06 (-0.11,-0.02)
rectum–colon	BSR	USR	GLM
intercept	-0.86 (-1.08,-0.65)	-0.84 (-1.00,-0.68)	-0.84 (-1.01,-0.69)
metro	0.02 (-0.21,0.26)	-0.07 (-0.22,0.08)	-0.07 (-0.22,0.09)
poverty	0.14 (-0.70,0.98)	-0.24 (-1.06,0.52)	-0.22 (-1.00,0.49)
age	-0.18 (-0.26,-0.10)	-0.18 (-0.26,-0.10)	-0.18 (-0.25,-0.11)
late	-0.26 (-0.37,-0.15)	-0.26 (-0.38,-0.15)	-0.26 (-0.37,-0.15)
metro*age	0.06 (-0.03,0.15)	0.05 (-0.03,0.15)	-0.01 (-0.08,0.07)
ρ	0.98(0.95,0.99)	–	–
ϕ	195	195	–
σ_1^2	0.95(0.57,1.48)	0.76(0.43,1.33)	–
σ_2^2	0.75(0.41,1.33)		–

Table 8.3 *Parameter estimates for the model with metro indicator and poverty rate as the spatial covariates, and stage and age as individual covariates. The estimates for rectum are relative effects to colon cancer. BSR=bivariate spatial residual model, USR=univariate spatial residual model, GLM=no random effects model.*

very similar spatial residual patterns. This is perhaps a surprisingly strong association, but believable given that these are *residual* surfaces, which account (at least conceptually) for important missing covariates, which could be spatial (e.g., local screening percentage, other sociodemographic factors) or nonspatial (e.g., the physiological adjacency of the colon and the rectum).

8.8 Space-time point patterns

As noted in the introduction to this chapter, we often encounter space-time point patterns. In practice, it would rarely be sensible to imagine that a point pattern occurs instantaneously. Rather, what occurs instantaneously are single events. That is, the time of occurrence of the event would be a continuous variable, e.g., when a cancer case was diagnosed, when a house was built. However, an interval of time is required for an entire point pattern to arise, e.g., the point pattern of disease cases over a week in a county or the point pattern of single family homes built in a city during a given year. This raises the additional issue of whether points are viewed as 'births' and remain in the pattern until "death." This perspective would be sensible for, say, trees or houses but not for cancer cases or locations of traffic accidents. This raises the question of whether we view the pattern cumulatively or differentially, with associated modeling implications. That is, in principle, we can move from one view to the other but model specification needs to reflect the choice.

Adding time to the investigation of point patterns can be critical with respect to adequate understanding of them. Point patterns can appear quite different over disjoint time windows and explaining these differences may be a vital aspect of the analysis. In different words, aggregating the patterns over time may remove this story. Furthermore, point pattern data collection is often very time consuming. Consider, for example, collecting complete inventories over reasonably large regions. Adding time can lead to infeasible demand in terms of number of hours and cost, especially for a slowly evolving process whence a long

time period may be needed. In Section 8.8.3.1, we consider one source of such data which is well-collected – new home construction over a specified geographic region. But, in general, methodology for space-time point patterns is underdeveloped and Bayesian analysis for such patterns even more so.

8.8.1 Space-time Poisson process models

So, when we consider adding time to the modeling, we have to determine whether we will view it as continuous or discrete. In the case of continuous time, we introduce a second argument to the intensity function, writing $\lambda(\mathbf{s}, t)$. We focus on space-time Poisson process models. Richer discussion is available in the book of Daley and Vere-Jones (2008). Under a NHPP specification over, say, $D \times (0, T]$, we observe $(\mathbf{s}, t)$'s and time is iewed as a continuous mark. Then, the intensity for a realization of a point pattern in the time interval $[t_1, t_2]$ is $\lambda_{[t_1,t_2]}(\mathbf{s}) = \int_{t_1}^{t_2} \lambda(\mathbf{s}, t)dt$. The intensity associated with, say, $B \times [t_1, t_2]$, with $B \in D$ and $0 \leq t_1 < t_2 \leq T$ is $\lambda(B \times [t_1, t_2]) = \int_B \int_{t_1}^{t_2} \lambda(\mathbf{s}, t)dtd\mathbf{s}$. Note that, though $\lambda(\cdot, \cdot)$ is a function over a subset of R^3, it makes no sense to consider Euclidean distance in three dimensions. The scale for time has nothing to do with the scale for space.

As with the spatial NHPP, we assume $N(B \times [t_1, t_2]) \sim Po(\lambda(B \times [t_1, t_2]))$ where $B \subseteq D$. In addition, given $N(D \times (0, T]$, we assume the points are scattered independently over $D \times (0, T]$ with density $\lambda(\mathbf{s}, t)/\lambda(D \times (0, T])$ where $\lambda(D \times (0, T]) = \int_D \int_0^T \lambda(\mathbf{s}, t)dtd\mathbf{s}$. Hence, $\lambda_{[t_1,t_2]}(\mathbf{s})$ is the intensity for a NHPP over D for events in the time window, $[t_1, t_2]$. Similarly, if we integrate over the set B, we get the intensity for a one-dimensional point pattern in time over $(0, T]$ for events restricted to B. Special cases of $\lambda(\mathbf{s}, t)$ lead to time stationarity with $\lambda(\mathbf{s}, t) = \lambda(\mathbf{s})$, spatial stationarity with $\lambda(\mathbf{s}, t) = \lambda(t)$, or space-time stationarity with $\lambda(\mathbf{s}, t) = \lambda$. With space-time stationarity, we can obtain the naive estimate of λ, i.e., $\hat{\lambda} = N(D \times (0, T]/T|D|$. If we retain stationarity but not an HPP, then we can develop analogues of the second order intensity, K-functions, etc. as in Daley and Vere-Jones (2008) or Diggle and Gabriel (2010). A relatively convenient class of Cox process models arises when $\lambda(\mathbf{s}, t)$ is a log space-time Gaussian process (as in Chapter 9). Also, the notion of a Gibbs process can be extended, allowing pairwise spatial interaction, scaled by a function in time. Extending Section 8.2.2, the Papangelou conditional intensity takes the form, $\lambda(\mathbf{s}, t|\mathbf{S}_t)$=exp$(\alpha(t) + \sum_{i=1}^{N_t} h(\mathbf{s}, \mathbf{s}_i))$ where $\mathbf{S}_t$ is the point pattern up to time t and N_t is the number of events in this pattern, i.e., in $(0, T]$.

More explicitly, if time is continuous, then we can imagine a parametric $\lambda(\mathbf{s}, t; \boldsymbol{\theta})$, say,

$$\log\lambda(\mathbf{s}, t; \boldsymbol{\theta}) = \mathbf{X}(\mathbf{s}, t)^T \boldsymbol{\theta}. \tag{8.29}$$

In (8.29), some components of the $\mathbf{X}$'s may be indexed only by space, e.g., elevation or by time, e.g., a global geographic covariate. Again, we can move to the nonparametric case, by introducing a spatio-temporal Gaussian process, $w(\mathbf{s}, t)$. That is,

$$\log\lambda(\mathbf{s}, t) = \mathbf{X}(\mathbf{s}, t)^T \boldsymbol{\theta} + w(\mathbf{s}, t). \tag{8.30}$$

As in Chapter 9, we have flexibility in the specification of $w(\mathbf{s}, t)$. For instance, $w(\mathbf{s}, t) = w(\mathbf{s}) + g(t)$, where $w(\mathbf{s})$ is a Gaussian process over D and $g(t)$ might be a specified function of time. Alternatively, $g(t)$ might be a stochastic process over $(0, T]$, e.g., Brownian motion, as in Section 8.8.3.

8.8.2 Dynamic models for discrete time data

In case of discrete t's, say, equally spaced, we introduce intensities $\lambda_t(\mathbf{s})$, $t = 1, 2, \ldots, T$. In this regard, time is viewed as a discrete mark, inviting comparison of intensities across

time. If time is discrete, then $\lambda_t(\mathbf{s})$ can be modeled dynamically, using parametric and/or nonparametric specifications. More precisely, we envision a customary dynamic model with conditionally independent first stage and time-evolving second stage (e.g., West and Harrison, 1997), in the spirit of Chapter 9. That is, the point pattern at time t

$$\mathbf{S}_t \sim \text{NHPP}(\lambda_t(\mathbf{s})), \tag{8.31}$$

and the patterns are independent over t with intensities, $\{\lambda_t(\mathbf{s}) : s \in D\}$. Then, we add a transition specification for $\{\lambda_t(\mathbf{s}) : s \in D\}|\{\lambda_{t-1}(\mathbf{s}) : s \in D\}$. Here, we have many modeling options. At the simplest level, we would have a parametric family for the λ_t's indexed by say, $\boldsymbol{\theta}_t$, e.g., $\boldsymbol{\gamma}_t$ with $\log\lambda_t(\mathbf{s}) = \mathbf{X}_t(\mathbf{s})^T\boldsymbol{\gamma}_t$, and the transition would take the form $[\boldsymbol{\theta}_t|\boldsymbol{\theta}_{t-1}]$ with an evolutionary equation for $\boldsymbol{\theta}_t$. A more complex transition could take the form of an integro-difference equation,

$$\lambda_t(\mathbf{s}) = \int_D K(\mathbf{s}, \mathbf{s}')\lambda_{t-1}(\mathbf{s}')d\mathbf{s}' \tag{8.32}$$

where $K(\cdot, \cdot)$ is a kernel or *propagator* function. Practically, the integral would have to be discretized over space in order to be evaluated.

With nonparametric specification, we find ourselves introducing log Gaussian processes, e.g.,

$$\log \lambda_t(\mathbf{s}) = \mathbf{X}(\mathbf{s})^T\boldsymbol{\gamma}_t + w_t(\mathbf{s}) \tag{8.33}$$

where $w_t(\mathbf{s})$ is a mean 0 Gaussian process. In addition to dynamics in $\boldsymbol{\gamma}_t$, we can also introduce dynamics in the $w_t(\mathbf{s})$, analogous to Section 11.4.

8.8.3 Space-time Cox process models using stochastic PDE's

Here, we turn to the use of a stochastic differential equation (SDE) to provide a Cox process model for space-time point patterns with continuous time, drawing upon ideas in Duan et al. (2009). Let D again be a fixed region and let $\mathbf{S}_T$ denote the observed space-time point pattern within D over the time interval $[0, T]$. Denote the stochastic intensity by $\lambda(\mathbf{s}, t), \mathbf{s} \in D, t \in [0, T]$. In practice, we may only know the spatial coordinates of all the points whereas the time coordinates are censored to time intervals in $[0, T]$.[6] Provided that $\lambda(\mathbf{s}, t)$ is integrable over $[0, T]$, the integrated process intensity is $\lambda(\mathbf{s}, T) = \int_0^T \lambda(t, \mathbf{s})\, dt$. As above, we may observe the point pattern in subintervals of $[0, T]$: $[t_1 = 0, t_2), \ldots, [t_{J-1}, t_J = T]$. These data constitute a series of discrete-time spatio-temporal point patterns, which we denote by $\mathbf{S}_{[t_1=0,t_2)}, \ldots, \mathbf{S}_{[t_{j-1},t_N=T]}$. The integrated process also provides stochastic intensities for these point patterns

$$\Delta\lambda_j(\mathbf{s}) = \lambda(\mathbf{s}, t_j) - \lambda(\mathbf{s}, t_{j-1}) = \int_{t_{j-1}}^{t_j} \lambda(\mathbf{s}, t)\, dt. \tag{8.34}$$

When the time intervals are sufficiently small, we may use the approximation

$$\Delta\lambda_j(\mathbf{s}) = \lambda(\mathbf{s}, t_j) - \lambda(\mathbf{s}, t_{j-1}) = \int_{t_{j-1}}^{t_j} \lambda(\mathbf{s}, \tau)\, d\tau \approx \lambda(\mathbf{s}, t_{j-1})(t_j - t_{j-1}) \ . \tag{8.35}$$

As a concrete example, below we consider a house construction dataset from Irving, TX, as discussed in Duan et al (2009). Let $\mathbf{S}_j = \mathbf{S}_{[t_{j-1},t_j)} = x_j$ be the observed set of locations

[6] For example, in our house construction data below, we only have the geo-coded locations of the newly constructed houses within a year. The exact time when the construction of a new house starts is not available.

of new houses built in region D and period $j=[t_{j-1}, t_j)$. We supply a Cox process model for the $\mathbf{S}_j$ by specifying $\lambda(\mathbf{s}, t)$ through a stochastic differential equation. We do this in order to introduce a mechanistic modeling scheme with parameters that convey physical meanings in the mechanism described by a stochastic differential equation. Here, we use the logistic equation

$$\frac{\partial \lambda(\mathbf{s}, t)}{\partial t} = r(\mathbf{s}, t)\lambda(\mathbf{s}, t)\left[1 - \frac{\lambda(\mathbf{s}, t)}{K(\mathbf{s})}\right], \tag{8.36}$$

where $K(\mathbf{s})$ is the "carrying capacity" (assuming it to be time-invariant) and $r(\mathbf{s}, t)$ is the "growth rate." Spatially and/or temporally varying *parameters*, such as growth rate and carrying capacity, can be modeled by spatio-temporal processes. In practice, the logistic equation finds applications in population growth in ecology, product and technology diffusion in economics, and urban development. The last is our context, with growth rate and carrying capacity being readily interpretable surfaces over space and time and over space, respectively.

Let the initial point pattern be $\mathbf{S}_0$ and the intensity be $\lambda_0(\mathbf{s}) = \lambda(\mathbf{s}, 0) = \int_{-\infty}^{0} \lambda(\tau, \mathbf{s})\, d\tau$. The hierarchical model for the space-time point patterns becomes

$$\begin{aligned} \mathbf{S}_j|\Delta\lambda_j &\sim \text{Poisson Process}(D, \Delta\lambda_j), \;\; j = 1, \ldots, J \\ \mathbf{S}_0|\lambda_0 &\sim \text{Poisson Process}(D, \lambda_0), \end{aligned} \tag{8.37}$$

where we suppress the indices $\mathbf{s}$ and t again for the periods $t_1, \ldots, t_J$. Note that the intensity $\Delta\lambda_j$ for $\mathbf{S}_j$ must be positive. Therefore, we model the growth rate r as a log-process, that is

$$\log r(\mathbf{s}, t) = \mu_r(\mathbf{s}; \beta_r) + \zeta(\mathbf{s}, t), \;\; \zeta \sim GP(0, \varrho(\mathbf{s} - \mathbf{s}', t - t'; \varphi_r)). \tag{8.38}$$

The J spatial point patterns are conditionally independent given the space-time intensity so the likelihood is

$$\begin{aligned} &\prod_{j=1}^{J}\left\{\exp\left(-\int_D \Delta\lambda_j(s)\, ds\right)\prod_{i=1}^{n_j}\Delta\lambda_j(x_{ji})\right\} \\ &\qquad \times \exp\left(-\int_D \lambda_0(s)\, ds\right)\prod_{i=1}^{n_0}\lambda_0(x_{0i}). \end{aligned} \tag{8.39}$$

This likelihood introduces the familiar stochastic integral, $\int_D \Delta\lambda_j(s)\, ds$, which we approximate in model fitting using a Riemann sum with representative points. As before, we divide the geographical region D into M cells and assume the intensity is homogeneous within each cell. Let $\Delta\lambda_j(m)$ and $\lambda_0(m)$ denote this average intensity in cell m. Let the area of cell m be $A(m)$. Then, the likelihood becomes

$$\begin{aligned} &\textstyle\prod_{j=1}^{J}\left[\exp\left(-\sum_{m=1}^{M}\Delta\lambda_j(m)A(m)\right)\prod_{m=1}^{M}\Delta\lambda_j(m)^{n_{jm}}\right] \\ &\times \exp\left(-\textstyle\sum_{m=1}^{M}\lambda_0(m)A(m)\right)\prod_{m=1}^{M}\lambda_0(m)^{n_{0m}} \end{aligned} \tag{8.40}$$

where n_{jm} is the number of point is in cell m in period j. We approximate the parameter processes $r(\mathbf{s}, t_j)$ and $K(\mathbf{s})$ accordingly as $r(m, t_j)$ and $K(m)$, which are homogeneous in each cell m.

8.8.3.1 Modeling the house construction data for Irving, TX

Our house construction dataset consists of the geo-coded locations and years of the newly constructed residential houses in Irving, TX, from 1901 to 2002. Irving started to develop in the early 1950's and the outline of the city was already in its current shape by the

Figure 8.18 *Growth of residential houses in Irving, TX.*

late 1960's. The city became almost fully developed by the early 1970's with little new construction afterward. We use the data from years 1951–1966 to fit our model and hold out the last three years (1967, 1968 and 1969) for prediction and model validation.

Figure 8.18 shows our study region D as a square of 5.6×5.6 square miles with Irving, TX, in the middle. This region is geographically disconnected from other major urban areas in Dallas County, which enables us to isolate Irving for analysis. In fact, the figure shows the growth of residential housing from 1951 to 1966. We divide the region into 100 (10×10) equally spaced grid cells as shown in Figure 8.19. Within each cell, we model the point pattern with a homogeneous Poisson process given $\Delta\lambda_j(m)$. The corresponding $\lambda_0(m)$, $K(m)$ and $r(m,j)$ are collected into vectors λ_0, K, and r which are modeled as follows:

$$\begin{aligned}
\log\lambda_0 &= \mu_\lambda + \theta_\lambda,\ \theta_\lambda \sim N(0, C_\lambda) \\
\log K &= \mu_K + \theta_K,\ \theta_K \sim N(0, C_K) \\
\log r &= \mu_r + \zeta,\ \zeta \sim N(0, C_r)
\end{aligned}$$

where the spatial covariance matrix C_λ and C_K are constructed using the Matérn class covariance function with distances between the centroids of the cells. The smoothness parameter ν is set to be 3/2. The variances σ_λ^2, σ_K^2 and range parameters ϕ_λ and ϕ_K are to be estimated. The spatio-temporal log growth rate r is assumed to have a separable covariance matrix $C_r = \sigma_r^2 \Sigma_{\mathbf{s}} \otimes \Sigma_t$, where the spatial correlation $\Sigma_{\mathbf{s}}$ is also constructed as a Matérn class function of the distances between cell centroids with smoothness parameter again being set to 3/2. The temporal correlation Σ_t is of exponential form. The variance σ_r^2, spatial and temporal correlation parameters ϕ_r and α_r are to be estimated.

We use vague priors for the parameters in the mean function: $\pi(\mu_\lambda)$, $\pi(\mu_K)$, $\pi(\mu_r) \overset{ind}{\sim} N\left(0, 10^8\right)$. We use natural conjugate priors for the precision parameters (inverse of variances) of r and λ_0: $\pi\left(1/\sigma_\lambda^2\right)$, $\pi\left(1/\sigma_K^2\right)$, $\pi\left(1/\sigma_r^2\right) \overset{ind}{\sim} Gamma(1,1)$. The temporal correlation parameter of r also has a vague log-normal prior: $\pi(\alpha_r) \sim log\text{-}N\left(0, 10^8\right)$. Again, the spatial range parameters ϕ_λ ϕ_K and ϕ_r are only weakly identified (Zhang, 2004) so we use

Figure 8.19 *The gridded study region encompassing Irving, TX.*

informative, discrete priors for them. In particular, we have chosen 40 values (from 1.1 to 5.0) and assume uniform priors on them for ϕ_λ ϕ_K and ϕ_r.

8.8.3.2 Results of the data analysis

Posterior inference (mean and 95% equal tail credible intervals) are presented in Table 8.8.3.2. Figure 8.20 shows the posterior mean growth curves and 95% Bayesian predictive bound for the intensity in the four blocks (marked as blocks 1, 2, 3 and 4). Comparing with the observed number of houses in the four blocks from 1951 to 1966, we can see the estimated curves fit the data very well.

In Figure 8.21 we display the posterior mean intensity surface for year 1966 and the predictive mean intensity surfaces for years 1967, 1968 and 1969. We also overlay the actual point patterns of the new homes construct in those four years on the intensity surfaces. Figure 8.21 shows that our model can forecast the major areas of high intensity, hence high growth, very well.

8.9 Additional topics

8.9.1 Measurement error in point patterns

Here, we consider the setting where the observed locations are measured with error and we seek to assess the resultant effect on the object of our interest, the intensity function, under a NHPP. Intuitively, adding noise will "blur" the intensity surface, making detection of its features more difficult. In fact, it is quite likely that, in recording locations, measurement error is introduced due to accuracy of the measuring instrument as well as factors influencing detection of event occurrences within the region.

Modeling point patterns using intensities requires restriction to a bounded subset of the plane. As a result, such noise can push locations in and out of the study domain. We are not

Figure 8.20 *Mean growth curves and their corresponding 95% predictive intervals (dotted lines) for the intensity for the four blocks marked in Figure 8.19.*

Figure 8.21 *Posterior and predictive mean intensity surfaces for the years 1966, 1967, 1968 and 1969.*

only observing a noisy version of the original realization, but it is also possible that we are missing some of the true events and also observing some which are not truly in the study region. For patterns having high event aggregation near the boundary of the region, this problem can be quite significant. Thus, measurement error results in a form of censoring to yield the actual dataset.

There is little previous literature on degraded point patterns. A general description is that the observed pattern is a random transformation of the true pattern, as in Diggle

Model Parameters	Posterior Mean	95% Equal-tail Interval
μ_λ	2.78	(2.15, 3.40)
σ_λ	1.77	(1.49, 2.11)
ϕ_λ	3.03	(2.70, 3.20)
μ_r	-2.76	(-3.24, -2.29)
σ_r	2.48	(2.32, 2.68)
ϕ_r	4.09	(3.70, 4.30)
α_r	0.52	(0.43, 0.62)
μ_K	6.49	(5.93, 7.01)
σ_K	1.17	(1.02, 1.44)
ϕ_K	1.91	(1.60, 2.20)

Table 8.4 *Posterior inference for the house construction data.*

(1993) who viewed the transformation as a conditionally independent random deformation of the true pattern. Chakraborty and Gelfand (2010) look at the problem as a two-stage specification — what is the model for the true pattern and given the true pattern, what is the model for the random transformation? Work in this spirit appears in Lund and Rudemo (2000) and Lund, Penttinen, and Rudemo (1999). There, the distinction is made between inferring about the properties of the true point pattern and reconstructing the true point pattern.

Within a fully hierarchical framework, there is no need to separate the point process parameter estimation and pattern reconstruction problems. Both can be addressed through suitable posterior inference for fairly general Cox processes. Here, we consider the scenario where events can only occur inside D so when we talk about shift of a location due to noise, it can only throw a point from D to D^c. But, since no event is allowed to take place outside D, each of the noisy locations observed corresponds to some true location in D. We term this setting an "island" model and employ an intensity surface which is a scaled mixture model where the scale parameter captures the expected number of points in D. We can remove this restriction by assuming that our region of interest is actually a subset of a bigger region of possible event findings (e.g., mapping tree locations in a specific part of a forest). Now events outside can also enter D because of noise; we refer to this case as a "subregion" model with details presented in Chakraborty and Gelfand (2010).

8.9.1.1 Modeling details

In order to simplify notation, in this subsection we label true locations by $\mathbf{x}$ and observed locations by $\mathbf{y}$. We consider measurement error in additive form. For a bounded study domain D and a true location $\mathbf{x}$, we assume the recorded location $\mathbf{y} = \mathbf{x} + \boldsymbol{\epsilon}$, where $\boldsymbol{\epsilon}$ is the measurement error. Further, we assume conditionally independent homogeneous displacements in the MEM scenario as employed in Diggle (1993). We note that, in some contexts, we might imagine that the error variability has spatial structure. Unfortunately, since the x's are latent, this specification produces a complicated posterior full conditional for x and overall model fitting that is unstable.

Under the island model we assume our study region D contains the support of the true point process i.e. $\mathrm{P}(\mathbf{x} \in D^c) = 0$ for any event location $\mathbf{x}$. So now we can only have (i) $\mathbf{x} \in D$, $\mathbf{y} \in D$, (ii) $\mathbf{x} \in D$, $\mathbf{y} \in D^c$. So, again, in a bounded region D, we assume n observed event locations $(\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_n)$, which are a noisy version of a set of m actual locations $(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_m)$ representing the complete realization of the point pattern in D. Now, m is unknown but $m \geq n$, that is, $(m - n)$ of these locations fell outside of D. For $\mathbf{x} \in D$ we adopt the Gaussian noise distribution, $\mathbf{y}|\mathbf{x}, \beta \sim N(\mathbf{x}, \Omega)$.

We model $\lambda(\mathbf{s}) = \lambda f(\mathbf{s})$, $\mathbf{s} \in D$, as discussed in Section 8.4.1. Here, for illustration and convenience, we choose f_D as Gaussian mixture distribution restricted to D. With regard to the number of mixture components, K, evidently more components can allow us to to tease out more features of the $\lambda(\cdot)$ surface but, within our measurement error framework will lead to poorly behaved computation. Practically, we can make an empirical choice based upon the observed point pattern or we can do model comparison across various choices for the number of components.

Relabeling the $\mathbf{x}'$s so that, for $i = 1, 2, \ldots, n$, $\mathbf{x}_i$ is the true location corresponding to $\mathbf{y}_i$ with the last $(m-n)$ $\mathbf{x}'$s corresponding to $\mathbf{y}$ locations outside D, we obtain the following model:

$$\begin{aligned} \mathbf{y}_i &\sim N_2(\mathbf{x}_i, \Omega),\ i = 1, 2, \ldots, n \\ f(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_m) &= NHPP(\lambda(.)) \\ \lambda(\mathbf{x}) &= \lambda f_D(\mathbf{x}) \end{aligned} \tag{8.41}$$

where $\lambda f_D(\mathbf{x}) = \lambda \sum_{k=1}^{K} q_k N_D(\mathbf{x}|\boldsymbol{\mu}_k, \Sigma_k)$, $N_{2,D}$ denotes the restriction of the bivariate normal density to D and q_k are the mixing weights.

Details on the MCMC model fitting and computation are supplied in Chakraborty and Gelfand (2010). We do note that the likelihood has two parts, one from the observed locations $(\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_n)$ (say, L_1), another from the unobserved $\mathbf{y}'$s known to be in D^c (say (L_2)). Upon associating the $\mathbf{x}_i$'s with $\mathbf{y}_i$'s, the likelihood takes the form $L = L_1 L_2$ (as in Lund and Rudemo, 2000, and in Lund, Penttinen, and Rudemo, 1999) where $L_1 = \prod_{i=1}^{n} \phi_2(\mathbf{y}_i|\mathbf{x}_i, \Omega)$ and $L_2 = \prod_{i=n+1}^{m} \bar{\Phi}_2(D; \mathbf{x}_i, \Omega)$, with ϕ_2 being the bivariate Gaussian pdf and $\bar{\Phi}_2$ its integral, in this case, over D, respectively, with appropriate parameters.

In writing the NHPP prior we assume that the first n of the $\mathbf{x}'$s are identified with the observed $\mathbf{y}'$s. In fact, there are $\frac{m!}{(m-n)!}$ possible matchings which have been collapsed into a single case. So, the prior density is, in fact,

$$\pi(\mathbf{x}_{1:n}, \mathbf{x}_{n+1:m}) = \frac{m!}{(m-n)!} \lambda^m \prod_{i=1}^{m} f_D(\mathbf{x}_i) \frac{e^{-\lambda}}{m!} \tag{8.42}$$

In the sequel we assume Ω for the measurement error process is known, obtained in some fashion, following our earlier discussion.

Hence, the full posterior for the model parameters becomes

$$\pi(m, \mathbf{x}_{1:m}, \lambda, \boldsymbol{\mu}_{1:K}, \Sigma_{1:K}, q_{1:k} \mid \mathbf{y}_{1:n}) \tag{8.43}$$

$$\propto \binom{m}{n} e^{-\lambda} \frac{\lambda^m}{m!} \prod_{i=1}^{m} f_D(\mathbf{x}_i|\boldsymbol{\mu}_{1:K}, \Sigma_{1:K}, q_{1:K}) \prod_{i=1}^{n} \phi_2(\mathbf{y}_i \mid \mathbf{x}_i) \prod_{i=n+1}^{m} \bar{\Phi}_2(D \mid \mathbf{x}_i) \pi(\lambda, \boldsymbol{\mu}_{1:K}, \Sigma_{1:k}, q_{1:K}) \tag{8.44}$$

We offer a simulation example to illustrate our methodology. We take f to be a 2-component normal mixture distribution (see Table 8.5) within a unit square and contaminate it with Gaussian noise having dispersion matrix as $\begin{pmatrix} 0.023 & 0.002 \\ 0.002 & 0.019 \end{pmatrix}$. In Figure 8.22 we show the original and perturbed datasets. There were 199 points initially in the window,

Parameters	Simulated Model	Island Model	Noiseless NHPP
$\mu_1^{(1)}$	0.64	0.6291, (0.5855, 0.6689)	**0.6053**, (0.5697, 0.6389)
$\mu_2^{(1)}$	0.61	0.5965, (0.5656, 0.6291)	0.5821, (0.5524, 0.6123)
$\mu_1^{(2)}$	0.25	0.2454, (0.1713, 0.3243)	0.2546, (0.2002, 0.3069)
$\mu_2^{(2)}$	0.14	0.1575, (0.0889, 0.2292)	0.1694, (0.1282, 0.2125)
q	0.71	0.7238, (0.6138, 0.8140)	**0.8150**, (0.7461, 0.8771)
λ	200	200.7096, (168.4630, 238.9672)	**177.7389**, (152.5286, 204.9620)
$\Sigma_{11}^{(1)}$	0.016	0.0153, (0.0068, 0.0275)	**0.0339**, (0.0263, 0.0432)
$\Sigma_{12}^{(1)}$	0.0007	0.0003, (−0.0060, 0.0081)	0.0037, (−0.0019, 0.0098)
$\Sigma_{22}^{(1)}$	0.018	0.0116, (0.0040, 0.0206)	**0.0271**, (0.0207, 0.0352)
$\Sigma_{11}^{(2)}$	0.007	0.0105, (0.0038, 0.0176)	**0.0175**, (0.0091, 0.0306)
$\Sigma_{12}^{(2)}$	0.0005	0.0004, (−0.0050, 0.0072)	−0.0033, (−0.0093, 0.0017)
$\Sigma_{22}^{(2)}$	0.002	0.0037, (0.0015, 0.0060)	**0.0096**, (0.0051, 0.0172)

Table 8.5 *Comparison of models with and without measurement error in case of bivariate Gaussian mixture intensity. Point estimates with* 95% *interval estimates in parentheses.*

Figure 8.22 *Original (left) and perturbed (right) point patterns.*

but after noise addition only 177 are left, roughly 11% loss of the points. Apart from the increased spread in the noisy pattern, the bimodality of the intensity essentially disappears. We fit the island model as well as a noiseless NHPP. Included in Table 8.5 is the comparison between the models while Figure 8.23 provides comparison of the estimated intensities. We see the benefit of the measurement error model. As expected, estimation of the Σ's along with q and λ was severely affected by the noise. The effect on the μ's is noteworthy. Fitting a mixture model directly to that data has likely caused the μ's to shift a bit in order to adjust for the overlap. In Table 8.5 we can see that the 95% credible interval for $\mu_1^{(1)}$ produced by the noiseless NHPP excludes the true value (parameters that noticeably differ are in ***bold***).

8.9.2 Presence-only data application

Learning about species distributions is a long-standing issue in ecology with, by now, an enormous literature. The focus of the work here is on the so-called *presence-only* setting, drawing on the work of Chakraborty et al. (2010). Analysis of presence-only data has seen growing popularity in recent years due to increased availability of such records from museum databases and other non-systematic surveys. One model-based strategy for presence-only data has attempted to implement a presence/absence approach. All of this work depends upon drawing so-called *background samples*, a random sample of locations in the region with known environmental features. Early work here characterized these samples as pseudo-absences and fitted a logistic regression to the observed presences and these pseudo-absences. Since presence/absence is unknown for these samples, recent work (Ward et al., 2009) shows how to adjust the resulting logistic regression to account for this. All of this work is non-spatial and requires the choice of an *arbitrary* number of background samples. Perhaps, most importantly, as we argue below, this approach conditions in the wrong direction. We assert that the observed presences should be viewed as a *marked point pattern*, with the mark indicating presence (see the recent work of Warton and Shepherd, 2010, in this regard). We

Figure 8.23 *Model Analysis: (a) actual intensity surface, (b) its estimate based on noiseless NHPP, (c) posterior intensity estimate from Island model, (d) uncertainty of estimated intensity.*

do not have a point pattern of absences; pseudo-absences create an unobserved and artificial pattern of absences.

We model presence-only data as a point pattern under a Cox process specification with associated intensity given as a regression in terms of the available environments across the region. We employ a hierarchical model to introduce spatial structure for the intensity surface through spatial random effects. We do not assume any background or pseudo-absence samples; rather, we assume that the covariates we employ are available as surfaces over the region in order to interpolate an intensity over the entire region. We acknowledge that the observed point pattern is degraded/biased through anthropogenic processes, e.g., human intervention to transform the landscape and non-uniform (in fact, often very irregular) sampling effort.

We work with presence-only data collected from the Cape Floristic Region (CFR) in South Africa (Figure 8.24). The region is divided into approximately $37,000$ grid cells, each one minute by one minute (roughly $1.55\ km \times 1.85\ km$). Covariate information is only available at grid cell level so we model the intensity as a tiled surface over these cells. We illustrate with potential and degraded intensities for six species.

8.9.2.1 Probability model for presence locations

Again, we assume a Cox process model for the set of presence locations. We have to introduce degradation caused by sampling bias as well as by land transformation. As a result, we conceptualize a *potential* intensity, i.e., the intensity in the absence of degradation, as well as a *realized* (or effective) intensity that operates in the presence of degradation. Further, we tile the intensity to reflect our inability to explain it at spatial resolution finer than our grid cells.

Figure 8.24 *Cells within the CFR that have at least one observation from the Protea Atlas dataset are shown in light grey, while cells with no observations are shown in dark grey.*

We imagine three surfaces over D. Let $\lambda(\mathbf{s})$ be the "potential" intensity surface, i.e., a positive function which is integrable over D. $\lambda(\mathbf{s})$ is the intensity in the absence of degradation. Let $\int_D \lambda(\mathbf{s})d\mathbf{s} = \lambda(D)$. Then, $f(s) = \lambda(\mathbf{s})/\lambda(D)$ gives the potential density over D. Modeling for $\lambda(\mathbf{s})$ is given below. Next, we envision an availability surface, $U(\mathbf{s})$, a binary surface over D such that $U(\mathbf{s}) = 1$ or 0 according to whether location $\mathbf{s}$ is untransformed by land use or not. That is, assuming no sampling bias, $\lambda(\mathbf{s})U(\mathbf{s})$ can only be $\lambda(\mathbf{s})$ or 0 according whether $\mathbf{s}$ is available or not. Let A_i denote the geographical region corresponding to cell i. Then, if we average $U(s)$ over A_i, we obtain $u_i = \int_{A_i} U(\mathbf{s})d\mathbf{s}/|A_i|$, where u_i is the proportion of cell i that is transformed and $|A_i|$ is the area of cell i. In our setting u_i is known, through remote sensing, for all grid cells. Similarly, we envision a sampling effort surface over D which we denote as $T(\mathbf{s})$. $T(\mathbf{s})$ is also a binary surface and $T(\mathbf{s})U(\mathbf{s}) = 1$ indicates that location $\mathbf{s}$ is both available and sampled. Now, we can set $q_i = \int_{A_i} T(\mathbf{s})U(\mathbf{s})d\mathbf{s}/|A_i|$ and interpret q_i as the probability that a randomly selected location in A_i was available and sampled. Thus, we can capture availability and sampling effort at areal unit scale.

Hence, $\lambda(\mathbf{s})U(\mathbf{s})T(\mathbf{s})$ becomes the degradation at location $\mathbf{s}$. This implies that in regions where no locations were sampled, the operating intensity for the species is 0. If $T(\mathbf{s})$ is viewed as random, with $p(\mathbf{s}) = P(T(\mathbf{s}) = 1) \in [0, 1]$, then $p(s)$ gives the local probability of sampling. This is analogous to $p(\mathbf{s})$ thinning discussed in Section 8.5.

To go forward, we assume that $\lambda(\mathbf{s})$ is independent of $T(\mathbf{s})U(\mathbf{s})$. That is, the potential intensity for a species is independent of the degradation process. Then, omitting the details, we can write $\int_{A_i} \lambda(\mathbf{s})T(\mathbf{s})U(\mathbf{s})d\mathbf{s} = \lambda_i q_i$ where $\lambda_i = \int_{A_i} \lambda(\mathbf{s})d\mathbf{s}$ is the cumulative intensity associated with cell A_i and, again, $q_i = \frac{1}{|A_i|}\int_{A_i} T(\mathbf{s})U(\mathbf{s})d\mathbf{s}$. It is not sensible to imagine that sampling effort is independent of land transformation. For instance, if $U(\mathbf{s}) = 0$ then $T(\mathbf{s}) = 0$. Hence, if we define $q_i = u_i p_i$, then $p_i = \frac{\int_{A_i} T(\mathbf{s})U(\mathbf{s})d\mathbf{s}}{\int_{A_i} U(\mathbf{s})d\mathbf{s}}$, i.e., p_i is the conditional probability that a randomly selected location in cell i is sampled given it is available. In our application below we set p_i equal to 1 or 0 which we interpret as $T(\mathbf{s}) = U(\mathbf{s})\ \forall \mathbf{s} \in A_i$ or $T(\mathbf{s}) = 0\ \forall \mathbf{s} \in A_i$, respectively. In particular, we set $p_i = 1$ if cell i was sampled for any species in our dataset; otherwise, we set $p_i = 0$. For the CFR, this sets $p_i = 1$ for the $10,158$ grid cells (28%) that have been visited.

To model the potential intensity surface $\lambda(\cdot)$, we employ a log Gaussian Cox process (Section 8.4.2). We expect the environmental covariates, say, $\mathbf{x}(\mathbf{s})$ to influence the intensity and model the mean of the GP as a linear combination of them. Then for any location

$\mathbf{s} \in D$, we have

$$log\ \lambda(\mathbf{s}) \quad = \quad \mathbf{x}^T(\mathbf{s})\beta + w(\mathbf{s}) \tag{8.45}$$

with $w(\cdot)$, a zero-mean stationary, isotropic GP over D; in the sequel we use the exponential covariance function.

Suppose we have n_i presence locations $(\mathbf{s}_{i,1}, \mathbf{s}_{i,2}, \ldots, \mathbf{s}_{i,n_i})$ within cell i for $i = 1, 2, \ldots, I$. Following the discussion above, $U(\mathbf{s}_{i,j})T(\mathbf{s}_{i,j}) \equiv 1,\ 0 \leq j \leq n_i, 1 \leq i \leq I$. Then the likelihood function becomes

$$L(\lambda(\cdot); \{\mathbf{s}_{i,j}\}) \quad \propto \quad e^{-\int_D \lambda(\mathbf{s})U(\mathbf{s})T(\mathbf{s})\ d\mathbf{s})} \prod_{i=1}^{I}\prod_{j=1}^{n_i} \lambda(\mathbf{s}_{i,j}) \tag{8.46}$$

Fortunately, we have a natural approximation to handle the stochastic integral in (8.46) by recalling that the dataset is gathered at the scale of grid cells in the CFR. That is, though we have geo-coded locations for the observed sites, with covariate information at grid cell level, we only attempt to explain the point pattern at grid cell level. In particular, let D denote our CFR study domain where D is divided into $I = 36,907$ grid cells of equal area. For each cell $i = 1, 2, 3, \ldots, I$, we are given information on l covariates as $x_i = (x_{i1}, x_{i2}, \ldots, x_{il})$. We also have cell level information about land availability across D, as a proportion of of the area of the cell (Figure 8.24). Following the previous subsection, we denote this by u_i. For many cells $n_i = 0$ primarily because 72% were actually unsampled. Additionally, a computational advantage accrues to working at grid cell level; we can work with a product Poisson likelihood approximation rather than the point pattern likelihood in (8.46). That is, we assume $\lambda(\cdot)$ is a tiled surface such that for cell i, the height is $\Delta\lambda(\mathbf{s}_i)$ where Δ is the area of the cell and $\mathbf{s}_i$ is the centroid. Then, given the set $\{\lambda(\mathbf{s}_i), i = 1, 2, \ldots, I$, the n_i are independent and $n_i \sim \text{Po}(\Delta\lambda(s_i)q_i)$. For any cell with $q_i = 0$ (which, by definition, can happen if either $p_i = 0$ or $u_i = 0$) there is no contribution from A_i in the product Poisson likelihood. Altogether, the posterior distribution takes the form

$$\begin{aligned}\pi(\lambda(\mathbf{s}_{1:m}), \beta, \theta|\mathbf{n}, \mathbf{x}, \mathbf{u}, \mathbf{q}) \propto \exp\left\{-\sum_{i=1}^{I} \lambda(\mathbf{s}_i)\Delta_i q_i\right\} \prod_{i=1}^{m} \lambda^{n_i}(\mathbf{s}_i) \\ \times \phi_m(\log\ \lambda(\mathbf{s}_{1:m})|\beta, \mathbf{x}, \theta)\pi(\beta)\pi(\theta)\end{aligned} \tag{8.47}$$

where ϕ_m denotes the m dimensional Gaussian density and θ the parameters in the covariance function of $w(\cdot)$ in (8.45).

The primary computational challenge in working with the CFR is handling the model fitting for $37,000$ grid cells, the familiar "large n" problem for Gaussian process, see Chapter 12. We employ the predictive process approximation for Gaussian random fields as in Section 12.4, with bias correction. With grid cells, an alternative is a parallelization scheme in conjunction with a CAR model as described in Chakraborty et al. (2010). The joint set of locations, $(\mathbf{s}_{1:I}, \mathbf{s}^0_{1:r})$, partition the spatial covariance matrix as $\sigma^2 R_{n+r}(\phi) = \sigma^2 \begin{pmatrix} R_I(\phi) & R_{r,I}(\phi) \\ R_{I,r}(\phi) & R_r(\phi) \end{pmatrix}$, where the entries of R_{r+I} are exponential correlation terms with decay parameter ϕ. We rewrite $\lambda_{0,i} = \lambda(\mathbf{s}_i)\Delta$, which denotes the expected species count in cell i under the potential intensity. Now the hierarchical model becomes

$$\begin{aligned} n_i|\lambda_{0,i} &\overset{ind}{\sim} \text{Poi}(\lambda_{0,i}q_i), i = 1, 2, ...I \\ log\ \lambda(\mathbf{s}_i) &= x_i^T\beta + \tilde{w}(\mathbf{s}_i) + \epsilon_i^* \\ \tilde{w}(\mathbf{s}_{1:I}) &= R_{I,r}(\phi)R_r^{-1}(\phi)\ w(\mathbf{s}^{\mathbf{0}}_{1:r}) \\ w(\mathbf{s}^{\mathbf{0}}_{1:r}) &\sim N_r(\mathbf{0}_r, R_r(\phi)) \\ \epsilon_i^* &\overset{ind}{\sim} N(0, \sigma^2(1 - (R_{I,r}(\phi)R_r^{-1}(\phi)R_{r,I}(\phi))_{ii})) \\ \pi(\beta, \phi, \sigma^2) &= \pi(\beta)\pi(\phi)\pi(\sigma^2) \end{aligned} \tag{8.48}$$

Species	EVAP	MAX01	MIN07	MAP	SMDSUM	FERT
PRAURE	−4.909	2.702	−0.301	−1.222	−0.049	0.501
	(-6.506,-3.057)	(1.574,3.678)	(-0.967,0.425)	(-1.975,-0.423)	(-0.816,0.711)	(0.034,0.967)
PRCYNA	−2.447	1.268	−1.032	−0.833	0.552	0.802
	(-2.981,-1.859)	(0.853,1.619)	(-1.314,-0.469)	(-1.021,-0.626)	(0.255,0.830)	(0.642,0.957)
LDSG	0.721	−0.420	0.137	−0.376	0.488	0.099
	(0.373,1.085)	(-0.658,-0.181)	(-0.011,0.295)	(-0.513,-0.237)	(0.304,0.673)	(0.045, 0.152)
PRMUND	−0.219	0.028	−0.609	−0.199	1.082	0.507
	(-2.724,1.429)	(-1.163,1.702)	(-1.039,-0.055)	(-1.024,0.510)	(0.277,1.809)	(0.101,0.929)
PRPUNC	2.076	−1.590	−1.722	0.363	0.535	0.186
	(1.031,3.096)	(-2.290,-0.921)	(-2.048,-1.409)	(0.082,0.662)	(0.052,1.079)	(-0.014,0.381)
PRREPE	1.690	−1.205	−0.275	0.124	0.094	0.224
	(1.243,2.124)	(-1.498,-0.907)	(-0.431,-0.110)	(-0.011,0.278)	(-0.112,0.320)	(0.152,0.295)

Table 8.6 *Posterior mean of covariate effects with central 95% credible interval in parenthesis.*

Posterior inference addresses two principal objectives: to understand the effect of environmental variables on species distribution and to construct maps of the potential and realized intensities over the entire study region. Posterior samples of β help us to infer whether a particular factor has a significant impact (positive or negative) on species intensity. With regard to displays of intensity surfaces, since, in our CFR application, $p_i = 1$ (i.e., $T(\mathbf{s}) = U(\mathbf{s})$ for all $\mathbf{s}$ in cell i) or $p_i = 0$ (i.e., $T(\mathbf{s}) = 0$ for all $\mathbf{s}$ in cell i) and since only 28% of cells were sampled, the $\lambda_i p_i$ surface will be 0 for 72% of cells, primarily capturing the (lack of) sampling effort. The $\lambda_i u_i$ surface reveals the effect of transformation. Since few cells are completely transformed, most $\lambda_i u_i > 0$. Of course, the $\lambda(\mathbf{s})$ surface is most interesting since it offers insight into the expected pattern of presences over all of D. Posterior draws of $\lambda_{1:I}$ can be used to infer about the potential intensity, displaying say the posterior mean surface. We can also learn about the potential density f in this discretized setting as $f_i = \lambda_i / \sum_{k=1}^{I} \lambda_k$, and the corresponding density under transformation as $f_{u,i} = \lambda_i u_i / \sum_{k=1}^{I} \lambda_k u_k$.

We consider six species within the Proteaceae family, ranging from prevalent to somewhat rare. Our point pattern for each species is drawn from the Protea Atlas data set (Rebelo et al., 2002). They are: *Protea aurea* (PRAURE) at 603 locations, *Protea cynaroides* (PRCYNA) at 8172 locations, *Leucadendron salignum* (LDSG) at 22949 locations, *Protea mundii* (PRMUND) at 764 locations, *Protea punctata* (PRPUNC) at 2148 locations, *Protea repens* (PRREPE) at 14574 locations.

In earlier work (Gelfand et al., 2005a; Gelfand, et al. 2005b), 18 environmental explanatory variables were considered, available at the 1 minute by 1 minute resolution. Based upon these analyses, the six most important were selected as covariates for the intensity function. They are: mean annual precipitation (MAP), July (winter) minimum temperature (MIN07), January (summer) maximum temperature (MAX01), potential evapotranspiration (EVAP), summer soil moisture days (SMDSUM), and percent of the grid cell with low fertility soil (FERT). There is considerable spatial variation in these covariates across the region

Table 8.6 provides the posterior mean covariate effects for all species along with the associated 95% equal tail credible intervals in parentheses. Most of the coefficients are significantly different from 0 and also the direction of significance varies with species.

Together, Figures 8.25 and 8.26 show the posterior mean intensity surfaces (potential and transformed) for the six species. Evidently there is strong spatial pattern and the pattern varies with species, i.e., the nature of local adjustment to the regression is species dependent. Comparison between the transformed and potential for each species is illuminating. Differentials of multiple orders of magnitude in expected cell counts are seen across many grid cells.

Figure 8.25 *Intensity maps for (a)* Protea aurea*, (b)* Protea cynaroides*, and (c)* Leucadendron salignum*, potential (left) and transformed (right). Values are cellwise expected frequency for the corresponding species.*

Figure 8.26 *Intensity maps for (a)* Protea mundii*, (b)* Protea punctata*, and (c)* Protea repens*, potential (left) and transformed (right). Values are cellwise expected frequency for the corresponding species.*

8.9.3 Scan statistics

Scan statistics are widely used to detect hotspots or local clustering in point patterns. They find application to epidemiology where there is interest in disease clustering but also in surveillance problems, e.g., syndromic, environmental or military settings. A primary reason for their wide usage is the availability of a free user-friendly software package, SaTScan, developed by Martin Kulldorff (http://www.satscan.org/). See also Kulldorff (1997). However, this material does not usually find its way into point pattern textbooks as it is ad hoc in its nature, because it is not clearly an analysis of the pattern of the points, and because it is primarily algorithmic in its implementation. We offer just a brief treatment here.

The basic idea is as follows: define a window (subregion of the region of interest); scan the region using this window; compute the value of some statistic for each stop of the window; determine which windows (if any) give significantly large values of the statistic. Hence, the

ad hoc aspect of the problem emerges. First, the notion of a cluster is not well-defined. How many points, how concentrated, what shape? Also, the choice of window is arbitrary. What shape, what size, what orientation? And, there is also a substantial computational challenge. With a large region and with a relatively small window, the statistic must be computed over many windows. Significance testing is usually done through Monte Carlo tests (recall the Monte Carlo tests in conjunction wit the significance of Moran's I and Geary's C in Section 3.1) so random datasets must be generated over the region of interest and the scanning process must be conducted for each Monte Carlo replicate. Such computation does not scale for large datasets while data mining settings are natural candidates for such procedures.

To provide a bit more detail, the setting is, in fact, one of a marked point pattern where each point has an an observed mark and a baseline mark. That is, we have $\mathbf{s}$ with mark $c(\mathbf{s})$, an observed count for that location and $b(\mathbf{s})$, a baseline or expected count (based, say, on the population at risk at that location). Hence, it emerges that the points are, in fact, small areal units relative to the region D and the randomness may be viewed primarily with regard to the observed mark at the location rather than with regard to the observed location. In fact, the customary model assumes, at $\mathbf{s}_i$, that $c(\mathbf{s}_i) \sim Po(q_i b(\mathbf{s}_i))$, independently over i. So, of interest are large q's and we see similarity with the modeling for the disease mapping problem in Chapters 3 and 5. There r_i played the role of a relative risk or rate for an areal unit, with interest in elevated r's. Here, q_i is a relative risk, with interest as well in large q's. In Chapter 3, we modeled spatially dependent r_i's using Markov random fields (CAR models). Here, there is no spatial structure attached to the q_i's.

Now, consider the spatial scan window, say, W. Assume q_i is constant over all points $\mathbf{s}_i$ in W, say, equal to q_{in} and that it is constant for all points $\mathbf{s}_i$ in D but outside of W, say, q_{out}. Under these clearly dubious assumptions along with the independent Poisson assumption, a likelihood ratio test can be created to test the null hypothesis that $q_{in} = q_{out}$ against the alternative that $q_{in} > q_{out}$ (elevated risk). It takes a simple form. Let $c(W) = \sum_{\mathbf{s}_i \in W} c(\mathbf{s}_i)$, similarly $b(W) = \sum_{\mathbf{s}_i \in W} b(\mathbf{s}_i)$, with analogous statistics $c(W^C)$ and $b(W^C)$ where W^C is the complement of W relative to D. Then the log of the likelihood ratio is $c(W)\log(\frac{c(W)}{b(W)}) + c(W^C)\log(\frac{c(W^C)}{b(W^C)})$. Again, we seek large values of this statistic. Evidently, there is the issue of multiple testing over many W's. Clearly there will be no tractable test statistic when implementing repeated calculations over overlapping regions, hence the foregoing Monte Carlo testing. There is a Bayesian version of this. See Neill et al. (2006). They introduce conjugate gamma priors and for the q's and obtain posterior probabilities for the two hypotheses. Scan statistics can be created for variables other than counts, using alternative specifications to the Poisson, creating different likelihood ratio statistics.

8.9.4 Preferential sampling

The choice of the sampling locations in a spatial network is often guided by practical demands such as the need to monitor air pollution levels near their most likely sources and in areas of high population density. Air pollution surfaces constructed solely on the basis of data obtained from these networks are likely to be biased if they are not adjusted for the effects of the choice of the monitoring sites. For example, if, due to locations, monitors tend to record high levels of exposure, interpolation of levels for low population density areas or locations away from sources such as power stations may be upwardly biased. That is, if the sampling locations are preferentially chosen to capture high (or low) values of a response, for example, air pollution levels, then subsequent model estimation and prediction of the exposure surface can become biased due to the selective sampling. Gelfand et al. (2012) use the term "bias" informally but with the intention of capturing departure from what the

exposure surface would look like if we interpolated given that the locations were selected under complete spatial randomness.

As a simple motivating example consider a model where space, denoted by t, is one dimensional. The model is written as $Z(t) = a + b\cos(t \bmod 2\pi) + \epsilon(t)$ where a and b are unknown parameters and $\epsilon(t)$ is a mean zero Gaussian process. Information contained in the data $Z(t)$ for $t = 1, \ldots, n$, regarding a and b, can vary between two very different functions of a and b depending on the set of t's where we observe the $Z(t)$ process. If we only observed the process at $t \approx 2\pi k$, we would only see observed values near $a + b$. On the other hand, if we take all the observations near $t \approx \pi(2k + 1)$, we would only see values near $a - b$.

Recent discussion on preferential sampling has been sparked by the work of Diggle et al (2010) who proposed a joint hierarchical model for the response and the locations. In particular, they adopt a model for the intensity that drives the locations which is assumed to be a spatial Gaussian process realization. Then, they employ this same Gaussian process realization to explain the responses. This may not be a sensible practical specification. Pati et al. (2011) generalize this approach in a Bayesian hierarchical setting, introducing common covariates into both the intensity for locations and the mean of the response model with two spatial Gaussian processes, one for the intensity and one for the response. It is unclear how well the use of these *informative* covariates in the regression model corrects for the preferential sampling bias introduced by these covariates in the location model.

We note that while preferential sampling often operates in practice, it is rare that sampling sites would be drawn randomly, using an explicit intensity function. In fact, there is a substantial literature on spatial design. See, e.g., Müller et al (2001) or, from a Bayesian perspective, Pilz and Spöck (2008). As a result, we doubt that complete spatial randomness ever operates in practice. Rather, geometric ideas like space filling designs (Nychka and Saltzman, 1998) or spatially-balanced designs (Theobald et al, 2007) offer non-model based, non-preferential, deterministic strategies. With regard to preferential sampling, if interest is in levels at certain locations, then it would be inappropriate to discourage sampling at those locations. Furthermore, if the available data is preferentially sampled but is the only data that can be expected, then, presumably it would be analyzed. The primary point is only that one might not feel comfortable with the potential bias in predictions made from it.

Since prediction is often the main utility of the modeling, arguably, the effect of preferential sampling lies more importantly in the resulting predictive surface than in parameter estimation. In this regard, Gelfand et al. (2013) take a direct simulation approach to assess the effect. Their basic idea is to compare two predictive surfaces. One originates from the notion of an 'operating' intensity driving the selection of monitoring sites. The other considers what would have been predicted had the sampling intensity been uniform, i.e., complete spatial randomness, over the study region. Using stylized models with extensive simulation they demonstrate the nature of the bias that can be incurred under preferential sampling. Using a real ozone exposure dataset, they show that, within a given network of locations, there can be substantial differences in the spatial prediction using preferentially chosen locations vs. roughly randomly selected locations and that the latter provide much improved predictive validation.

8.10 Exercises

1. If $\mathcal{S}_1$ is a realization from a NHPP with intensity $\lambda_1(\mathbf{s})$ and $\mathcal{S}_2$ is a realization from a NHPP with intensity $\lambda_2(\mathbf{s})$, independent of $\mathcal{S}_1$, show that the joint realization comes from a NHPP with intensity $\lambda_1(\mathbf{s}) + \lambda_2(\mathbf{s})$.
2. If $f(\mathbf{s}_1, \ldots, \mathbf{s}_n) = \Pi_i f(\mathbf{s}_i)$, show that the second order intensity satisfies $\gamma(\mathbf{s}, \mathbf{s}') = \lambda(\mathbf{s})\lambda(\mathbf{s}')$.

3. (a) **Stationary spatial point process**: For a spatial point process, suppose $N(B + \mathbf{h}) \sim N(B), \forall B \in \mathcal{B}$ and $\forall \mathbf{h}$. Show that the second-order intensity, $\gamma(\mathbf{s}, \mathbf{s}')$ satisfies $\gamma(\mathbf{s}, \mathbf{s} + \mathbf{h}) = \gamma(\mathbf{h})$.
   (b) **Isotropic spatial point process**: For a spatial point process, suppose $N(B) \sim N(PB), \forall B \in \mathcal{B}$ where P is an orthogonal matrix and $PB = \{\mathbf{s}^* = P\mathbf{s} : \mathbf{s} \in B\}$. Then $\gamma(\mathbf{s}, \mathbf{s}') = \gamma(||\mathbf{s} - \mathbf{s}'||)$.
4. Prove Campbell's Theorem. That is, if $g(\mathbf{s})$ is a point feature and we are interested in $\sum_{\mathbf{s}_i \in \mathbf{S}} g(\mathbf{s}_i)$, then $E_{\mathbf{S}}(\sum_{\mathbf{s}_i \in \mathbf{S}} g(\mathbf{s}_i) = \int g(\mathbf{s})\lambda(\mathbf{s})d\mathbf{s}$. (*Hint*: Show it is true for an indicator function).
5. Show that, for $\mathbf{s} \in D$, a bounded region, for the conditional intensity, $\lambda(\mathbf{s}|\mathbf{S})$, $E_{\mathbf{S}}(\lambda(\mathbf{s}|\mathbf{S}) = \lambda(\mathbf{s})$. (*Hint*: If you can not prove this in general, show it for the NHPP.)
6. Show that all nonstationary point processes that arise from $p(\mathbf{s})$ thinning of a stationary point process are such that the second order reweighted intensity functions (or pair correlation functions) for the original stationary and the reweighted nonstationary process are identical.
7. For the homogeneous Gibbs process in Section 8.6.3, show that the Papangelou conditional intensity is given by (8.17). What does it become for the Strauss process? For the hardcore process?
8. Recalling the discussion of the F and G functions from Section 8.3.2, show that, if $X \sim G$, then $Z = \pi X^2 \sim \texttt{Exp}(\lambda)$. Hence, obtain the mean and variance of X.
9. For a log Gaussian Cox process with log intensity, $\log\lambda(\mathbf{s}) = X^T(\mathbf{s})\beta + z(\mathbf{s})$ where $z(\mathbf{s})$ is a mean 0 Gaussian process with covariance function $C(\mathbf{s}, \mathbf{s}')$, obtain the marginal first and second order product densities.
10. For a Matérn process with restriction of offspring to a circle of radius R, consider the distribution of the distance between two random points in the same cluster. Show that it is $f_{interpoint}(d) = \frac{4d}{\pi R^2}\left(\cosh\frac{d}{2R} - \frac{d}{2R}\sqrt{1 - \frac{d^2}{4R^2}}\right), 0 \leq d \leq 2R$.
11. Justify the *rejection* method approach for obtaining samples of point patterns under a NHPP. That is, let $\lambda_{max} = \max_s \lambda(s), s \in D$. Generate a point pattern from an HPP using the constant intensity λ_{max}. For each point generated, do a rejection step, i.e., for s_i, draw $U_i \sim U(0, 1)$ and retain s_i if $U_i < \lambda(s_i)/\lambda_{max}$. Argue that the remaining points come from a NHPP with intensity $\lambda(s)$. (*Hint*: Obtain the distribution of the retained number of points in A, $N(A)$ for any $A \subset D$.)
12. Suppose we have a NHPP with intensity $\lambda(\mathbf{s})$ and we implement p-thinning to a point pattern from this process (Section 8.5). Show that the thinned pattern is a realization from a NHPP with intensity $p\lambda(\mathbf{s})$.
13. Generate a random point pattern over the unit square from the intensity $\lambda(s) = \lambda(x, y) = \{200\exp(-3(x - \frac{1}{3})^2 - 4(y - \frac{2}{3})^2) + 300\exp(-2(x - \frac{2}{3})^2 - 2(y - \frac{1}{3})^2)\}I(x \in (0, 1), y \in (0, 1))$
    (a) How many points do you expect in the unit square?
    (b) Test for complete spatial randomness using a quadrat based χ-square test.
    (c) Obtain $\hat{G}(w)$. Obtain $\hat{F}(w)$ for $m = 100$. Obtain the theoretical $Q - Q$ plot of $\hat{G}(w)$ vs. G for complete spatial randomness. Obtain the theoretical $Q - Q$ plot of $\hat{F}(w)$ vs. G for complete spatial randomness.
    (d) Obtain $\hat{K}(d)$. Plot $\hat{L}(d) = \sqrt{\frac{\hat{K}(d)}{\pi}} - d$ vs. d.
14. For the Lansing Woods data, available at `http://www.spatstat.org/`, you will find six species. Choose a couple and, for each, assume a stationary, in fact, isotropic model, and provide the exploratory data analysis from Section 8.3. That is, estimate $G(d)$, $F(d)$,

$K(d)$, and $\gamma(d)$. Compare with corresponding plots assuming CSR. Plot $J(d)$ and $L(d)$ vs. d. Fit NHPP and LGCP models to these choices within a Bayesian framework.

15. For the Strauss-Ripley redwood saplings dataset (redwood) available at `http://www.spatstat.org/`, fit a Gibbs process model, in fact, a Strauss process within a Bayesian framework.

Chapter 9

Multivariate spatial modeling for point-referenced data

In this chapter we take up the problem of multivariate spatial modeling. Spatial data is often *multivariate* in the sense that multiple (i.e. more than one) outcomes are measured at each spatial unit. As in the univariate case, the spatial units can be referenced by points or by areal units. Examples of multivariate point-referenced data abound in the sciences. For example, at a particular environmental monitoring station, levels of several pollutants would typically be measured (e.g., ozone, nitric oxide, carbon monoxide, $PM_{2.5}$, etc.). In atmospheric modeling, at a given site we may observe surface temperature, precipitation, and wind speed. In examining commercial real estate markets, for an individual property we may observe both selling price and total rental income. In forestry, investigators seek to produce spatially explicit predictions of multiple forest attributes (e.g., abundance and basal area) using a multi-source forest inventory approach. In each of these illustrations, we anticipate both dependence between measurements at a particular location, and association between measurements across locations. Multivariate areal data are conspicuous in public health where each county or administrative unit supplies counts or rates for a number of diseases. Again, we expect dependence between diseases within each county as well as across counties.

We first treat multivariate point-referenced data and turn to multivariate areal data in the next chapter. Analysis of multivariate point-referenced data can proceed using either a *conditioning approach*, along the lines of the way misalignment was treated in Chapter 7 (e.g., X followed by $Y \mid X$) or a *joint approach* that directly models the joint distribution of the outcomes variables. Both these approaches are based upon extensions of univariate kriging to multivariate contexts, where the conditional approach is also referred to as *kriging with external drift* (see, e.g., Royle and Berliner, 1999), while the joint approach is referred to as *co-kriging* (see Wackernagel, 2003).

9.1 Joint modeling in classical multivariate geostatistics

Classical multivariate geostatistics begins, as with much of geostatistics, with early work of Matheron (1973, 1979). The basic ideas here include cross-variograms and cross-covariance functions, intrinsic coregionalization, and co-kriging. The emphasis is on prediction. A thorough discussion of the work in this area is provided in Wackernagel (2003). To add generality to the conditional modeling approach, one could envision a latent multivariate spatial process defined over locations in a region. For example, in ambient air quality assessment, we seek to jointly model multiple contaminants at a fixed set of monitoring sites. Inference focuses upon three major aspects: (i) estimate associations among the contaminants, (ii) estimate the strength of spatial association for each contaminant, and (iii) predict the contaminants at arbitrary locations.

Let $\mathbf{Y}(\mathbf{s}) = (Y_1(\mathbf{s}), Y_2(\mathbf{s}), \ldots, Y_p(\mathbf{s}))^T$ be a $p \times 1$ vector, where each $Y_i(\mathbf{s})$ represents an outcome of interest referenced by $\mathbf{s} \in \mathcal{D}$. We seek to capture the association both within components of $\mathbf{Y}(\mathbf{s})$ and across $\mathbf{s}$. The joint second order (weak) stationarity hypothesis defines the cross-variogram as

$$\gamma_{ij}(\mathbf{h}) = \frac{1}{2}E(Y_i(\mathbf{s}+\mathbf{h}) - Y_i(\mathbf{s}))(Y_j(\mathbf{s}+\mathbf{h}) - Y_j(\mathbf{s})). \tag{9.1}$$

Implicitly, we assume $E(Y(\mathbf{s}+\mathbf{h}) - Y(\mathbf{s})) = 0$ for all $\mathbf{s}$ and $\mathbf{s}+\mathbf{h} \in \mathcal{D}$. Clearly, $\gamma_{ij}(\mathbf{h})$ is an even function, i.e. $\gamma_{ij}(\mathbf{h}) = \gamma_{ij}(-\mathbf{h})$ and, as a consequence of the Cauchy-Schwarz inequality, satisfies $|\gamma_{ij}(\mathbf{h})|^2 \leq \gamma_{ii}(\mathbf{h})\gamma_{jj}(\mathbf{h})$.

The cross-covariance function is defined as

$$C_{ij}(\mathbf{h}) = E[(Y_i(\mathbf{s}+\mathbf{h}) - \mu_i)(Y_j(\mathbf{s}) - \mu_j)] \tag{9.2}$$

where, for each i, a constant mean μ_i is assumed for $Y_i(\mathbf{s})$. The associated $p \times p$ matrix $C(\mathbf{h})$ is called the cross-covariance matrix. Note that it need not be symmetric (think of a setting where you might expect $C_{ij}(\mathbf{h}) \neq C_{ji}(\mathbf{h})$). Of course, $C(\mathbf{h})$ need not be positive definite but, in the limit, as $\|h\| \to 0$, it becomes positive definite (see Section 9.3 below).

Using standard properties of covariances, we see that the cross-covariance function must satisfy $|C_{ij}(\mathbf{h})|^2 \leq C_{ii}(\mathbf{0})C_{jj}(\mathbf{0})$. However, $|C_{ij}(\mathbf{h})|$ need not be bounded above by $C_{ij}(\mathbf{0})$ because the maximum value of $C_{ij}(\mathbf{h})$ need not occur at $\mathbf{0}$. A particular example is the so-called *spatial delay models* (Wackernagel, 2003). Consider the bivariate setting with $p = 2$ and assume that

$$Y_2(\mathbf{s}) = aY_1(\mathbf{s}+\mathbf{h}_0) + \epsilon(\mathbf{s}) \ ,$$

where $Y_1(\mathbf{s})$ is a spatial process with stationary covariance function $C(\mathbf{h})$, and $\epsilon(\mathbf{s})$ is a pure error process with variance τ^2. Then, the associated cross-covariance function has $C_{11}(\mathbf{h}) = C(\mathbf{h}), C_{22}(\mathbf{h}) = a^2C(\mathbf{h})$ and $C_{12} = C(\mathbf{h}+\mathbf{h}_0)$. Similarly, $|C_{ij}(\mathbf{h})|^2$ need not be bounded above by $C_{ii}(\mathbf{h})C_{jj}(\mathbf{h})$. The matrix $C(\mathbf{h})$ of direct and cross-covariances with $C_{ij}(\mathbf{h})$ as its (i,j)-th entry is called the cross-covariance matrix. It need not be positive definite at any $\mathbf{h}$ though as $\mathbf{h} \to \mathbf{0}$, it converges to a positive definite matrix, the (local) covariance matrix associated with $\mathbf{Y}(\mathbf{s})$. We will discuss cross-covariance matrices in greater detail in Section 9.2.

Analogous to the relationship between the variogram and the covariance function, we find a familiar connection between the cross-variogram and the cross-covariance function. The latter determines the former and it is easy to show that

$$\gamma_{ij}(\mathbf{h}) = C_{ij}(\mathbf{0}) - \frac{1}{2}(C_{ij}(\mathbf{h}) + C_{ij}(-\mathbf{h})). \tag{9.3}$$

If we decompose $C_{ij}(\mathbf{h})$ as $\frac{1}{2}(C_{ij}(\mathbf{h}) + C_{ij}(-\mathbf{h})) + \frac{1}{2}(C_{ij}(\mathbf{h}) - C_{ij}(-\mathbf{h}))$, then the cross-variogram only captures the even term of the cross-covariance function, suggesting that it may be inadequate in certain modeling situations. Such concerns led to the proposal of the pseudo cross-variogram (Clark et al., 1989; Myers, 1991; Cressie, 1993). In particular, Clark et al. (1989) proposed $\pi^c_{ij}(\mathbf{h}) = E(Y_i(\mathbf{s}+\mathbf{h}) - Y_j(\mathbf{h}))^2$ and Myers (1991) subsequently offered a mean-corrected version, $\pi^m_{ij}(\mathbf{h}) = var(Y_i(\mathbf{s}+\mathbf{h}) - Y_j(\mathbf{h}))$. It is easy to show that $\pi^c_{ij}(\mathbf{h}) = \pi^m_{ij}(\mathbf{h}) + (\mu_i - \mu_j)^2$. The psuedo cross-variogram is not constrained to be an even function. However, the assumption of stationary cross-increments is unrealistic, certainly with variables measured on different scales and even with rescaling of the variables. A further limitation is the restriction of the pseudo cross-variogram to be positive. Despite the unattractiveness of "apples and oranges" comparison across components, Cressie and Wikle (1998) demonstrate successful multivariate kriging. or co-kriging, using $\pi^m_{ij}(\mathbf{h})$.

9.1.1 Co-kriging

Co-kriging in classical multivariate geostatistics refers to spatial prediction at a new location that uses not only information from observations of the outcome variable being considered, say $Y_i(\mathbf{s})$, but also information from the measurements of the other outcome variables, i.e., the other components in $\mathbf{Y}(\mathbf{s})$. Early presentations of co-kriging are available in Journel and Huijbregts (1978) and Matheron (1979), while Myers (1982) presents a general and more comprehensive treatment using matrices. Corsten (1989) and Stein and Corstein (1991) develop co-kriging within a linear regression framework. Again, a detailed review can be found in Wackernagel (2003).

It is instructive to distinguish between prediction of a single variable as above and joint prediction of several variables at a new location (Myers, 1982). Consider the joint second order stationarity model in (9.1) above. Suppose we seek to predict $Y_1(\mathbf{s}_0)$, i.e., the first element in $\mathbf{Y}(\mathbf{s})$, at a new location $\mathbf{s}_0$. An unbiased estimator based upon $\mathbf{Y} = (\mathbf{Y}(\mathbf{s}_1), \mathbf{Y}(\mathbf{s}_2), ..., \mathbf{Y}(\mathbf{s}_n))^T$ would assume the form $\hat{Y}_1(\mathbf{s}_0) = \sum_{i=1}^n \sum_{l=1}^p \lambda_{il} Y_l(\mathbf{s}_i)$ where we have the constraints that $\sum_{i=1}^n \lambda_{il} = 0, l \neq 1, \sum_{i=1}^n \lambda_{i1} = 1$. On the other hand, if we were interested in predicting $\mathbf{Y}(\mathbf{s}_0)$, we would now write $\hat{\mathbf{Y}}(\mathbf{s}_0) = \sum_{i=1}^n \Lambda_i \mathbf{Y}(\mathbf{s}_i)$. The unbiasedness condition is $\sum_{i=1}^n \Lambda_i = I$. Moreover, now, what should we take as the "optimality" condition? One choice is to choose the set $\{\Lambda_{0i}, 1 = 1, 2, ..., n\}$ with associated estimator $\hat{\mathbf{Y}}_0(\mathbf{s}_0)$ such that for any other unbiased estimator, $\tilde{\mathbf{Y}}(\mathbf{s}_0)$, $E(\tilde{\mathbf{Y}}(\mathbf{s}_0) - \mathbf{Y}(\mathbf{s}_0))(\tilde{\mathbf{Y}}(\mathbf{s}_0) - \mathbf{Y}(\mathbf{s}_0))^T - E(\hat{\mathbf{Y}}_0(\mathbf{s}_0) - \mathbf{Y}(\mathbf{s}_0))(\hat{\mathbf{Y}}_0(\mathbf{s}_0) - \mathbf{Y}(\mathbf{s}_0))^T$ is non-negative definite (Ver-Hoef and Cressie, 1993). Alternatively, one could minimize $\mathrm{tr}E(\hat{\mathbf{Y}}(\mathbf{s}_0) - \mathbf{Y}(\mathbf{s}_0))(\hat{\mathbf{Y}}(\mathbf{s}_0) - \mathbf{Y}(\mathbf{s}_0))^T = E(\hat{\mathbf{Y}}(\mathbf{s}_0) - \mathbf{Y}(\mathbf{s}_0))^T(\hat{\mathbf{Y}}(\mathbf{s}_0) - \mathbf{Y}(\mathbf{s}_0))$ (Myers, 1982).

Returning to the individual prediction case, minimization of predictive mean square error, $E(Y_1(\mathbf{s}_0) - \hat{Y}_1(\mathbf{s}_0))^2$ amounts to a quadratic optimization subject to linear constraints and the solution can be obtained using Lagrange multipliers. As in the case of univariate kriging, the solution can be written as a function of a cross-variogram specification. In fact, Ver-Hoef and Cressie (1993) show that $\pi_{ij}(\mathbf{h})$ above emerges in computing predictive means square error, which suggests that it is a natural cross-variogram for co-kriging. But, altogether, given the concerns noted regarding $\gamma_{ij}(\mathbf{h})$ and $\pi_{ij}(\mathbf{h})$, it seems preferable to assume the existence of second moments for the multivariate process, captured through a *valid* cross-covariance function (as defined in (9.2)). In fact, to introduce a likelihood, to implement full inference, and to do prediction with accurate uncertainty assessment, in the sequel we will work exclusively with Gaussian process models defined through cross-covariance functions.

We next turn to a brief discussion of *valid* cross-variograms. The matter is not as clear as in the univariate setting. Wackernagel (2003) induces valid cross-variograms from valid cross-covariance functions (see Section 9.2). Myers (1982) and Ver-Hoef and Cressie (1993) assume second order stationarity and also a finite cross-covariance in order to bring $\gamma_{ij}(\mathbf{h})$ into the optimal co-kriging equations. Rehman and Shapiro (1996) define a *permissible* cross-variogram $\gamma_{ij}(\mathbf{h})$ as one that meets the following conditions: (i) the $\gamma(\mathbf{h})$ are continuous except possibly at the origin, (ii) $\gamma_{ij}(\mathbf{h}) \geq 0, \forall\, \mathbf{h} \in \mathcal{D}$, (iii) $\gamma_{ij}(\mathbf{h}) = \gamma(-\mathbf{h}), \forall\, \mathbf{h} \in \mathcal{D}$, and (iv) the functions, $-\gamma_{ij}(\mathbf{h})$, are conditionally non-negative definite, the usual condition for individual variograms.

In fact, this directly renders an explicit solution to the individual co-kriging problem if we assume a multivariate Gaussian spatial process. The preceding developments show that such a process specification only requires supplying mean surfaces for each component of the outcome $\mathbf{Y}(\mathbf{s})$ and a valid cross-covariance function. For simplicity, assume $\mathbf{Y}(\mathbf{s})$ is centered to have mean $\mathbf{0}$. The cross-covariance function provides $\Sigma_{\mathbf{Y}}$, the $np \times np$ covariance matrix for the data $\mathbf{Y} = (\mathbf{Y}(\mathbf{s}_1)^T, \mathbf{Y}(\mathbf{s}_2)^T, ..., \mathbf{Y}(\mathbf{s}_n)^T)^T$. In addition, it provides the $np \times 1$ vector, $\mathbf{c}_0$ which is blocked as vectors $\mathbf{c}_{0j}, j = 1, 2, .., n$ with l-th element $c_{0j,l} = cov(Y_1(\mathbf{s}_0), Y_l(\mathbf{s}_j))$. Then, from the multivariate normal distribution of $\mathbf{Y}, Y_1(\mathbf{s}_0)$, we obtain the co-kriging

estimate,

$$\mathrm{E}(Y_1(\mathbf{s}_0) \mid \mathbf{Y}) = \mathbf{c}_0^T \Sigma_{\mathbf{Y}}^{-1} \mathbf{Y}. \tag{9.4}$$

The associated variance, $\mathrm{var}(Y_1(\mathbf{s}_0) \mid \mathbf{Y})$ is also immediately available, i.e., $\mathrm{var}(Y_1(\mathbf{s}_0) \mid \mathbf{Y}) = C_{11}(\mathbf{0}) - \mathbf{c}_0^T \Sigma_{\mathbf{Y}}^{-1} \mathbf{c}_0$.

In particular, consider the special case of the $p \times p$ cross-covariance matrix, $C(\mathbf{h}) = \rho(\mathbf{h})T$, where $\rho(\cdot)$ is a valid correlation function and T is the local positive definite covariance matrix. Then, $\Sigma_{\mathbf{Y}} = R \otimes T$, where R is the $n \times n$ matrix with (i,j)-th entry $\rho(\mathbf{s}_i - \mathbf{s}_j)$ and $\otimes$ denotes the Kronecker product. This specification also yields $\mathbf{c}_0 = \mathbf{r}_0 \otimes \mathbf{t}_{*1}$, where $\mathbf{r}_0$ is $n \times 1$ with entries $\rho(\mathbf{s}_0 - \mathbf{s}_j)$ and $\mathbf{t}_{*1}$ is the first column of T. Then, (9.4) becomes $t_{11}\mathbf{r}_0^T R^{-1} \tilde{\mathbf{Y}}_1$ where t_{11} is the $(1,1)$-th element of T and $\tilde{\mathbf{Y}}_1$ is the vector of observations associated with the first component of the $\mathbf{Y}(\mathbf{s}_j)$'s. This specification is known as the *intrinsic* multivariate correlation and is discussed in greater generality in Section 9.1.2. In other words, under an intrinsic specification, only observations on the first component are used to predict the first component at a new location. We leave this as an exercise but also see Helterbrand and Cressie (1994) and Wackernagel (2003) in this regard.

In all of the foregoing work, inference assumes the cross-covariance or the cross-variogram to be known. In practice, a parametric model is adopted and data-based estimates of the parameters are plugged in. A related issue here is whether the data is available for each variable at all sampling points (so-called *isotopy* — not to be confused with "isotropy"), some variables share some sample locations (partial *heterotopy*), or the variables have no sample locations in common (entirely *heterotopic*). Similarly, in the context of prediction, if any of the $Y_l(\mathbf{s}_0)$ are available to help predict $Y_1(\mathbf{s}_0)$, we refer to this as "collocated co-kriging." The challenge with heterotopy in classical work is that the empirical cross-variograms cannot be computed and empirical cross-covariances, though they can be computed, do not align with the sampling points used to compute the empirical direct covariances. Furthermore, the value of the cross-covariances at $\mathbf{0}$ can notbe computed.[1]

9.1.2 Intrinsic multivariate correlation and nested models

A somewhat different approach to multivariate spatial models is based upon the so called *structural analysis* approach, where typically more than one variogram model is used to accommodate different spatial scales. For example, following Grzebyk and Wackernagel (1994), we might deploy a *nested* variogram model written as

$$\gamma(\mathbf{h}) = t_1\gamma_1(\mathbf{h}) + t_2\gamma_2(\mathbf{h}) + \cdots + t_r\gamma_r(\mathbf{h}) \ .$$

For instance, with three spatial scales, corresponding to a nugget, fine scale dependence, and long range dependence, respectively, we might write $\gamma(\mathbf{h}) = t_1\gamma_1(\mathbf{h}) + t_2\gamma_2(\mathbf{h}) + t_3\gamma_3(\mathbf{h})$ where $\gamma_1(\mathbf{h}) = 0$ if $\|\mathbf{h}\| = 0$, $= 1$ if $\|\mathbf{h}\| > 0$, while $\gamma_2(\cdot)$ reaches a sill equal to 1 very rapidly and $\gamma_3(\cdot)$ reaches a sill equal to 1 much more slowly.

The nested variogram model corresponds to the spatial process $\sqrt{t_1}w_1(\mathbf{s}) + \sqrt{t_2}w_2(\mathbf{s}) + \sqrt{t_3}w_3(\mathbf{s})$ — a linear combination of independent processes. Why not, then, use this idea to build a multivariate version of a nested variogram model? Journel and Huijbregts (1978) propose to do this using the specification

$$w_l(\mathbf{s}) = \sum_{r=1}^{m} \sum_{j=1}^{p} a_{rj}^{(l)} w_{rj}(\mathbf{s}) \ \text{ for } \ l = 1, \ldots, p \ ,$$

[1]The empirical cross-variogram imitates the usual variogram (Chapter 3), creating bins and computing averages of cross-products of differentials within the bins. Similar words apply to the empirical cross-covariance.

where the $w_{rj}(\mathbf{s})$ are independent process replicates across j and, for each r, the process has correlation function $\rho_r(\mathbf{h})$ and variogram $\gamma_r(\mathbf{h})$ (with sill 1). In the case of isotropic ρ's, this implies that we have a different range for each r but a common range for all components given r.

The representation in terms of independent processes can now be given in terms of the $p \times 1$ vector process $\mathbf{w}(\mathbf{s}) = [w_l(\mathbf{s})]_{l=1}^p$, formed by collecting the $w_l(\mathbf{s})$'s into a column for $l = 1, \ldots, p$. We write the above linear specification as $\mathbf{w}(\mathbf{s}) = \sum_{r=1}^m A_r \mathbf{w}_r(\mathbf{s})$, where each A_r is a $p \times p$ matrix with (l, j)-th element $a_{rj}^{(l)}$ and $\mathbf{w}_r(\mathbf{s}) = (w_{r1}(\mathbf{s}), \ldots, w_{rp}(\mathbf{s}))^T$ are $p \times 1$ vectors that are independent replicates from a spatial process with correlation function $\rho_r(\mathbf{h})$ and variogram $\gamma_r(\mathbf{h})$ for $r = 1, 2, .., p$.

Letting $C_r(\mathbf{h})$ be the $p \times p$ cross covariance matrix and $\Gamma_r(\mathbf{h})$ denote the $p \times p$ matrix of direct and cross variograms associated with $\mathbf{w}(\mathbf{s})$, we have $C_r(\mathbf{h}) = \rho_r(\mathbf{h})T_r$ and $\Gamma_r(\mathbf{h}) = \gamma_r(\mathbf{h})T_r$. Here, T_r is positive definite with $T_r = A_r A_r^T = \sum_{j=1}^p \mathbf{a}_{rj}\mathbf{a}_{rj}^T$, where $\mathbf{a}_{rj}$ is the j-th column vector of A_r. Finally, the cross covariance and cross variogram nested model representations take the form $C(\mathbf{h}) = \sum_{r=1}^m \rho_r(\mathbf{h})T_r$ and $\Gamma(\mathbf{h}) = \sum_{r=1}^m \gamma_r(\mathbf{h})T_r$.

The case $m = 1$ is called the intrinsic correlation model, the case $m > 1$ is called the intrinsic multivariate correlation model. In more recent work, Vargas-Guzmán, Warrick and Myers (2002) allow the $w_{rj}(\mathbf{s})$ to be dependent. Such modeling is natural when scaling is the issue, i.e., we want to introduce spatial effects to capture dependence at different scales (and, thus, m has nothing to do with p). When we have prior knowledge about these scales, such modeling will be successful. However, to find datasets that inform about such scaling may be challenging.

9.2 Some theory for cross-covariance functions

Using the generic notation $\mathbf{Y}(s)$ to denote a $p \times 1$ vector of random variables at location $\mathbf{s}$, we seek flexible, interpretable, and computationally tractable models to describe the process $\{\mathbf{Y}(\mathbf{s}) : \mathbf{s} \in D\}$. As in the univariate setting, a well-defined multivariate process must ensure that for every finite set of locations $\mathcal{S} = \{\mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_n\}$, the vector $\mathbf{Y} = (\mathbf{Y}(\mathbf{s}_1)^T, \mathbf{Y}(\mathbf{s}_2)^T, \ldots, \mathbf{Y}(\mathbf{s}_n)^T)^T$ has a valid joint distribution.

The key ingredient to ensure a well-defined multivariate spatial process is the *cross-covariance* function associated with $\mathbf{Y}(\mathbf{s})$. Given its importance in multivariate spatial modeling, we provide some formal theory regarding the validity and properties of these functions. Let $\mathcal{D} \subset \Re^d$ be a connected subset of the d-dimensional Euclidean space and let $\mathbf{s} \in \mathcal{D}$ represent a generic point in $\mathcal{D}$. Consider a vector-valued spatial process $\{\mathbf{w}(\mathbf{s}) \in \Re^p : \mathbf{s} \in \mathcal{D}\}$, where $\mathbf{w}(\mathbf{s})$ is $p \times 1$ with components $w_j(\mathbf{s})$. For convenience, assume that $\mathrm{E}[\mathbf{w}(\mathbf{s})] = \mathbf{0}$.

The *cross-covariance function* is a matrix-valued function, say $\mathbf{C}(\mathbf{s}, \mathbf{s}')$, defined for any pair of locations $(\mathbf{s}, \mathbf{s}') \in \mathcal{D} \times \mathcal{D}$ and yielding the $p \times p$ matrix whose (i, j)-th element is the cross-covariance function

$$C_{ij}(\mathbf{s}, \mathbf{s}') = \mathrm{Cov}(w_i(\mathbf{s}), w_j(\mathbf{s}')) = \mathrm{E}[w_i(\mathbf{s})w_j(\mathbf{s}')] \,. \tag{9.5}$$

Note that the cross-covariance function in (9.2) arises as a special case of (9.5) when each $w_i(\mathbf{s}) = Y_i(\mathbf{s}) - \mu_i$ and $C_{ij}(\mathbf{s}, \mathbf{s}')$ is a function purely of $\mathbf{h} = \mathbf{s} - \mathbf{s}'$. Using more compact notation, as is customary in multivariate statistics, we write the cross-covariance matrix as

$$C(\mathbf{s}, \mathbf{s}') = \mathrm{Cov}(\mathbf{w}(\mathbf{s}), \mathbf{w}(\mathbf{s}')) = \mathrm{E}[\mathbf{w}(\mathbf{s})\mathbf{w}^T(\mathbf{s}')] \,. \tag{9.6}$$

For example, if $\mathbf{w}(\mathbf{s})$ is a bivariate process with components $w_1(\mathbf{s})$ and $w_2(\mathbf{s})$, then the cross-covariance matrix for $\mathbf{w}(\mathbf{s})$ is 2×2:

$$C(\mathbf{s}, \mathbf{s}') = \begin{pmatrix} \mathrm{Cov}(w_1(\mathbf{s}), w_1(\mathbf{s}')) & \mathrm{Cov}(w_1(\mathbf{s}), w_2(\mathbf{s}')) \\ \mathrm{Cov}(w_2(\mathbf{s}), w_1(\mathbf{s}')) & \mathrm{Cov}(w_2(\mathbf{s}), w_2(\mathbf{s}')) \end{pmatrix} .$$

In general, a cross-covariance matrix $C(\mathbf{s}, \mathbf{s}')$ is not required to be symmetric (hence, positive definite) since $\text{Cov}(w_i(\mathbf{s}), w_j(\mathbf{s}'))$ is not necessarily equal to $\text{Cov}(w_j(\mathbf{s}), w_i(\mathbf{s}'))$. Neither is it required that $C(\mathbf{s}, \mathbf{s}') = C(\mathbf{s}', \mathbf{s})$. However, since $\text{Cov}(w_i(\mathbf{s}), w_j(\mathbf{s}')) = \text{Cov}(w_j(\mathbf{s}'), w_i(\mathbf{s}))$, the cross-covariance matrix must satisfy

$$C(\mathbf{s}, \mathbf{s}') = C(\mathbf{s}', \mathbf{s})^T \tag{9.7}$$

In other words, the cross-covariance matrix evaluated at $(\mathbf{s}, \mathbf{s}')$ is the transpose of the cross-covariance matrix evaluated at $(\mathbf{s}', \mathbf{s})$. A second condition for $C(\mathbf{s}, \mathbf{s}')$ is necessitated by the fact that for any finite collection of locations $\mathcal{S}$, we must have

$$\text{Var}\left\{\sum_{i=1}^{n} \mathbf{a}_i^T \mathbf{w}(\mathbf{s}_i)\right\} = \sum_{i=1}^{n}\sum_{j=1}^{n} \mathbf{a}_i^T C(\mathbf{s}_i, \mathbf{s}_j)\mathbf{a}_j > 0 \tag{9.8}$$

for every nonzero $\mathbf{a}_i \in \Re^d$.

The cross-covariance matrix completely determines the joint dispersion structure implied by the spatial process. To be precise, for any n and any arbitrary collection of sites $\mathcal{S} = \{\mathbf{s}_1, \ldots, \mathbf{s}_n\}$ the $np \times 1$ vector of realizations $\mathbf{w} = (\mathbf{w}(\mathbf{s}_1)^T, \mathbf{w}(\mathbf{s}_2)^T, \ldots, \mathbf{w}(\mathbf{s}_n)^T)^T$ will have the variance-covariance matrix $\Sigma_{\mathbf{w}}$, which is an $nm \times nm$ block matrix whose (i, j)-th block is precisely the $p \times p$ cross-covariance matrix $\mathbf{C}(\mathbf{s}_i, \mathbf{s}_j)$. The conditions (9.7) and (9.8) ensure that $\Sigma_{\mathbf{w}}$ is symmetric and positive-definite. The conditions (9.7) and (9.8) also imply that $C(\mathbf{s}, \mathbf{s})$ is always symmetric and positive definite. In fact, it is precisely the variance-covariance matrix for the elements of $\mathbf{w}(\mathbf{s})$ within site $\mathbf{s}$. Again, since our primary focus in this text is Gaussian process models (or mixtures of such processes), specification of $C(\mathbf{s}, \mathbf{s}')$ is all we need to provide all finite dimensional distributions.

We say that $\mathbf{w}(\mathbf{s})$ is *stationary* if $C(\mathbf{s}, \mathbf{s}') = C(\mathbf{s}' - \mathbf{s})$, i.e., the cross-covariance function depends only upon the separation of the sites, while we say that the cross-covariance matrix is *isotropic* if $\mathbf{C}(\mathbf{s}, \mathbf{s}') = \mathbf{C}(\|\mathbf{s}' - \mathbf{s}\|)$, i.e., it depends only upon the distance between the sites. Note that for stationary processes we write the cross-covariance matrix as $C(\mathbf{h}) = C(\mathbf{s}, \mathbf{s} + \mathbf{h})$. From (9.7) it is immediate that

$$C(-\mathbf{h}) = C(\mathbf{s} + \mathbf{h}, \mathbf{s}) = C^T(\mathbf{s}, \mathbf{s} + \mathbf{h}) = C^T(\mathbf{h}).$$

Thus, for a stationary process, a symmetric cross-covariance functions is equivalent to having $C(-\mathbf{h}) = C(\mathbf{h})$ (i.e., even function). For isotropic functions,

$$C(\mathbf{h}) = C(\|\mathbf{h}\|) = C(\|-\mathbf{h}\|) = C(-\mathbf{h}) = C^T(\mathbf{h}),$$

hence the cross-covariance function is even and the matrix is necessarily symmetric.

What more can we say about functions that will satisfy (9.7) and (9.8)? The primary characterization theorem for cross-covariance functions (Cramér, 1940; Yaglom, 1987) says that real-valued functions, say $C_{ij}(\mathbf{h})$, will form the elements of a valid cross-covariance matrix $C(\mathbf{h})$ if and only if each $C_{ij}(\mathbf{h})$ has the cross-spectral representation

$$C_{ij}(\mathbf{h}) = \int \exp(2\pi i \mathbf{t}^T \mathbf{h}) d(F_{ij}(\mathbf{t})) \ , \quad \text{where} \quad i = \sqrt{-1} \ , \tag{9.9}$$

with respect to a positive definite measure $F(\cdot)$, i.e., where the cross-spectral matrix $M(B) = [F_{ij}(B)]_{i,j=1}^p$ is positive definite for any Borel subset $B \subseteq \Re^d$. The representation in (9.9) can be considered the most general representation theorem for cross-covariance functions. It is the analogue of Bochner's Theorem for covariance functions and has been employed by several authors to construct classes of cross-covariance functions.

Essentially, one requires a choice of the $F_{ij}(\mathbf{t})$'s. Matters simplify when $F_{ij}(\mathbf{t})$ is assumed to be square-integrable ensuring that a spectral density function $f_{ij}(\mathbf{t})$ exists such

that $d(F_{ij}(\mathbf{t})) = f_{ij}(\mathbf{t})d\mathbf{t}$. Now, one simply needs to ensure that the $p \times p$ matrix constructed with $f_{ij}(\mathbf{t})$ as its (i,j)-th entry is positive definite for all $\mathbf{t} \in \Re^d$. Corollaries of the above representation lead to the approaches proposed in Gaspari and Cohn (1999) and in Majumdar and Gelfand (2007) for constructing valid cross-covariance functions as convolutions of covariance functions of stationary random fields (see Section 9.7 later). For isotropic settings we use the notation $||\mathbf{s}' - \mathbf{s}||$ for the distance between sites $\mathbf{s}$ and $\mathbf{s}'$. The representation in (9.9) can be viewed more broadly in the sense that, working in the complex plane, if the matrix valued measure $M(\cdot)$ is Hermitian non-negative definite, then we obtain a valid cross-covariance matrix in the complex plane. Rehman and Shapiro (1996) use this broader definition to obtain permissible cross-variograms. Grzebyk and Wackernagel (1994) employ the induced complex covariance function to create a bilinear model of coregionalization.

As in the univariate case, it is evident that not every matrix $C(\mathbf{s}, \mathbf{s}')$ which we might propose will be *valid*. Consistent with our objective of using multivariate spatial process models in an applied context, we prefer constructive approaches for such cross-covariance functions. The next three sections describe approaches based upon separability, coregionalization, moving averages, and convolution. Nonstationarity can be introduced following the univariate approaches in Section 3.2; however, no details are presented here (see, e.g., Gelfand, Schmidt, Banerjee and Sirmans, 2004, Sec. 4).

9.3 Separable models

Perhaps the most obvious specification of a valid cross-covariance function for a p-dimensional $\mathbf{Y}(\mathbf{s})$ is to let ρ be a valid correlation function for a univariate spatial process, let T be a $p \times p$ positive definite matrix, and let

$$C(\mathbf{s}, \mathbf{s}') = \rho(\mathbf{s}, \mathbf{s}') \cdot T \ . \tag{9.10}$$

In (9.10), $T \equiv (T_{ij})$ is interpreted as the covariance matrix associated with $\mathbf{Y}(\mathbf{s})$, and ρ attenuates association as $\mathbf{s}$ and $\mathbf{s}'$ become farther apart. The covariance matrix for $\mathbf{Y}$ resulting from (9.10) is easily shown to be

$$\Sigma_{\mathbf{Y}} = H \otimes T \ , \tag{9.11}$$

where $(H)_{ij} = \rho(\mathbf{s}_i, \mathbf{s}_j)$ and $\otimes$ denotes the Kronecker product. $\Sigma_{\mathbf{Y}}$ is evidently positive definite since H and T are. In fact, $\Sigma_{\mathbf{Y}}$ is convenient to work with since $|\Sigma_{\mathbf{Y}}| = |H|^p \, |T|^n$ and $\Sigma_{\mathbf{Y}}^{-1} = H^{-1} \otimes T^{-1}$. This means that updating $\Sigma_{\mathbf{Y}}$ requires working with a $p \times p$ and an $n \times n$ matrix, rather than an $np \times np$ one. Moreover, if we permute the rows of $\mathbf{Y}$ to $\widetilde{\mathbf{Y}}$ where $\widetilde{\mathbf{Y}}^T = (Y_1(\mathbf{s}_1), \ldots, Y_1(\mathbf{s}_n), Y_2(\mathbf{s}_1), \ldots, Y_2(\mathbf{s}_n), \ldots, Y_p(\mathbf{s}_1), \ldots, Y_p(\mathbf{s}_n))$, then $\Sigma_{\widetilde{\mathbf{Y}}} = T \otimes H$.

In fact, working in the fully Bayesian setting, additional advantages accrue to (9.10). With $\boldsymbol{\phi}$ and T *a priori* independent and an inverse Wishart prior for T, the full conditional distribution for T, that is, $p(T|\mathbf{W}, \boldsymbol{\phi})$, is again an inverse Wishart (e.g., Banerjee, Gelfand, and Polasek, 2000). If the Bayesian model is to be fitted using a Gibbs sampler, updating T requires a draw of a $p \times p$ matrix from a Wishart distribution, substantially faster than updating the $np \times np$ matrix $\Sigma_{\mathbf{Y}}$.

What limitations are associated with (9.10)? Clearly $C(\mathbf{s}, \mathbf{s}')$ is symmetric, i.e., $cov(Y_\ell(\mathbf{s}_i), Y_{\ell'}(\mathbf{s}_{i'})) = cov(Y_{\ell'}(\mathbf{s}_i), Y_\ell(\mathbf{s}_{i'}))$ for all i, i', ℓ, and ℓ'. Moreover, it is easy to check that if ρ is stationary, the *generalized* correlation, also referred to as the *coherence* in the time series literature (see, e.g., Wei, 1990), is such that

$$\frac{\mathrm{cov}(Y_\ell(\mathbf{s}), Y_{\ell'}(\mathbf{s}+\mathbf{h}))}{\sqrt{\mathrm{cov}(Y_\ell(\mathbf{s}), Y_\ell(\mathbf{s}+\mathbf{h}))\mathrm{cov}(Y_{\ell'}(\mathbf{s}), Y_{\ell'}(\mathbf{s}+\mathbf{h}))}} = \frac{T_{\ell\ell'}}{\sqrt{T_{\ell\ell}T_{\ell'\ell'}}} \ , \tag{9.12}$$

regardless of $\mathbf{s}$ and $\mathbf{h}$. Also, if ρ is isotropic and strictly decreasing, then the spatial range (see Section 2.1.3) is identical for each component of $\mathbf{Y}(\mathbf{s})$. This must be the case since only one correlation function is introduced in (9.10). This seems the most unsatisfying restriction, since if, e.g., $\mathbf{Y}(\mathbf{s})$ is a vector of levels of different pollutants at $\mathbf{s}$, then why should the range for all pollutants be the same? In any event, some preliminary marginal examination of the $Y_\ell(\mathbf{s}_i)$ for each ℓ, $\ell = 1, \ldots, p$, might help to clarify the feasibility of a common range.

Additionally, (9.10) implies that, for each component of $\mathbf{Y}(\mathbf{s})$, correlation between measurements tends to 1 as distance between measurements tends to 0. For some variables, including those in our illustration, such an assumption is appropriate. For others it may not be, in which case microscale variability (captured through a nugget) is a possible solution. Formally, suppose independent $\boldsymbol{\epsilon}(\mathbf{s}) \sim N\left(\mathbf{0}, Diag(\boldsymbol{\tau}^2)\right)$, where $Diag(\boldsymbol{\tau}^2)$ is a $p \times p$ diagonal matrix with (i,i) entry τ_i^2, are included in the modeling. That is, we write $\mathbf{Y}(\mathbf{s}) = \mathbf{V}(\mathbf{s}) + \boldsymbol{\epsilon}(\mathbf{s})$ where $\mathbf{V}(\mathbf{s})$ has the covariance structure in (9.10). An increased computational burden results, since the full conditional distribution for T is no longer an inverse Wishart, and likelihood evaluation requires working with an $np \times np$ matrix.

In a sequence of papers by Le and Zidek and colleagues (mentioned in the next subsection), it was proposed that $\Sigma_{\mathbf{Y}}$ be taken as a random covariance matrix drawn from an inverse Wishart distribution centered around (9.11). In other words, an extra hierarchical level is added to the modeling for $\mathbf{Y}$. In this fashion, we are not specifying a spatial process for $\mathbf{Y}(\mathbf{s})$; rather, we are creating a joint distribution for $\mathbf{Y}$ with a flexible covariance matrix. Indeed, the resulting $\Sigma_{\mathbf{Y}}$ will be nonstationary. In fact, the entries will have no connection to the respective $\mathbf{s}_i$ and $\mathbf{s}_j$. This may be unsatisfactory since we expect to obtain many inconsistencies with regard to distance between points and corresponding association across components. We may be able to obtain the posterior distribution of, say, $Corr(Y_\ell(\mathbf{s}_i), Y_\ell(\mathbf{s}_j))$, but there will be no notion of a range.

The form in (9.10) was presented in Mardia and Goodall (1993) who used it in conjunction with maximum likelihood estimation. Banerjee and Gelfand (2002) discuss its implementation in a fully Bayesian context, as we outline in the next subsection.

9.4 Spatial prediction, interpolation, and regression

Multivariate spatial process modeling is required when we are analyzing several point-referenced data layers, when we seek to explain or predict for one layer given the others, or when the layers are not all collected at the same locations. The last of these is a type of spatial misalignment that can also be viewed as a missing data problem, in the sense that we are missing observations to completely align all of the data layers. For instance, in monitoring pollution levels, we may observe some pollutants at one set of monitoring sites, and other pollutants at a diffferent set of sites. Alternatively, we might have data on temperature, elevation, and wind speed, but all at different locations.

More formally, suppose we have a conceptual response $Z(\mathbf{s})$ along with a conceptual vector of covariates $\mathbf{x}(\mathbf{s})$ at each location $\mathbf{s}$. However, in the sampling, the response and the covariates are observed at possibly different locations. To set some notation, let us partition our set of sites into three mutually disjoint groups: let S_Z be the sites where only the response $Z(\mathbf{s})$ has been observed, S_X the set of sites where only the covariates have been observed, S_{ZX} the set where both $Z(\mathbf{s})$ and the covariates have been observed, and finally S_U the set of sites where no observations have been taken.

In this context we can formalize three types of inference questions. One concerns $Y(\mathbf{s})$ when $\mathbf{s} \in S_X$, which we call *interpolation*. The second concerns $Y(\mathbf{s})$ for $\mathbf{s}$ belonging to S_U, which we call *prediction*. Evidently, prediction and interpolation are similar but interval estimates will be at least as tight for the latter compared with the former. The last concerns the functional relationship between $X(\mathbf{s})$ and $Y(\mathbf{s})$ at an arbitrary site $\mathbf{s}$, along with other covariate information at $\mathbf{s}$, say, $\mathbf{U}(\mathbf{s})$. We capture this through $E[Y(\mathbf{s})|X(\mathbf{s}), \mathbf{U}(\mathbf{s})]$, and

Figure 9.1 *A graphical representation of the S sets. Interpolation applies to locations in S_X, prediction applies to locations in S_U, and regression applies to all locations.* $\mathbf{X}_{aug} = (\mathbf{X}, \mathbf{X}')$, $\mathbf{Y}_{aug} = (\mathbf{Y}, \mathbf{Y}')$.

refer to it as *spatial regression.* Figure 9.1 offers a graphical clarification of the foregoing definitions.

In the usual stochastic regressors setting one is interested in the relationship between $Y(\mathbf{s})$ and $X(\mathbf{s})$ where the pairs $(X(\mathbf{s}_i), Y(\mathbf{s}_i))$, $i = 1, \ldots, n$ (suppressing $\mathbf{U}(\mathbf{s}_i)$) are independent. For us, they are dependent with the dependence captured through a spatial characterization. Still, one may be interested in the regression of $Y(\mathbf{s})$ on $X(\mathbf{s})$ at an arbitrary $\mathbf{s}$. Note that there is no conditional spatial process, $Y(\mathbf{s}) \mid X(\mathbf{s})$, associated with the bivariate spatial process $(X(\mathbf{s}), Y(\mathbf{s}))$; how would one define the joint distribution of $Y(\mathbf{s}_i) \mid X(\mathbf{s}_i)$ and $Y(\mathbf{s}_{i'}) \mid X(\mathbf{s}_{i'})$?

We also note that our modeling structure here differs considerably from that of Diggle, Tawn, and Moyeed (1998). These authors specify a univariate spatial process in order to introduce unobserved spatial effects (say, $\mathrm{V}(\mathbf{s})$) into the modeling, after which the $Y(\mathbf{s})$'s are conditionally independent given the $V(\mathbf{s})$'s. In other words, the $V(\mathbf{s})$'s are intended to capture spatial association in the means of the $Y(\mathbf{s})$'s. For us, the $X(\mathbf{s})$'s are also modeled through a spatial process, but they are observed and introduced as an explanatory variable with a regression coefficient. Hence, along with the $Y(\mathbf{s})$'s, we require a bivariate spatial process.

Here we provide a fully Bayesian examination of the foregoing questions. In Subsection 9.4.1 we study the case where $Y(\mathbf{s})$ is Gaussian, but in Subsection 9.4.3 we allow the response to be binary.

The Gaussian interpolation problem is addressed from an empirical Bayes perspective in a series of papers by Zidek and coworkers. For instance, Le and Zidek (1992) and Brown, Le, and Zidek (1994) develop a Bayesian interpolation theory (both spatial and temporal) for multivariate random spatial data. Le, Sun, and Zidek (1997) extend this methodology to account for misalignment, i.e., where possibly not all monitored sites measured the same set of pollutants (data missing by design). Their method produces the joint predictive distribution for several locations and different time points using all available data, thus allowing for simultaneous temporal and spatial interpolation without assuming the random field to be stationary. Their approach provides a first-stage multivariate normal distribution for the observed data. However, this distribution does not arise from a spatial Gaussian process.

Framing multivariate spatial prediction (often referred to as *co-kriging*) in the context of linear regression dates at least to Corsten (1989) and Stein and Corsten (1991). In this work,

the objective is to carry out predictions for a possible future observation. Stein and Corsten (1991) advocate looking at the prediction problem under a regression setup. They propose trend surface modeling of the point source response using polynomials in the coordinates. Typically in trend surface analysis (Cressie, 1993), spatial structure is modeled through the mean but observations are assumed to be independent. Instead, Stein and Corsten (1991) retain familiar spatial dependence structure but assume the resultant covariances and cross-covariances (and hence the dispersion matrix) are known. In this context, Stein et al. (1991) use restricted maximum likelihood to estimate unknown spatial dependence structure parameters.

9.4.1 Regression in the Gaussian case

The regression problem posed here is solely to learn about the conditional distribution for $Y(\mathbf{s}_0)|X(\mathbf{s}_0)$. We are not interested in kriging. In this regard, the reader might ask why we do not proceed as in Chapter 6, specifying a univariate Gaussian process model for $Y(\mathbf{s})$, our usual geostatistical model; why do we consider a bivariate process model here? Two reasons are as follows. First, we might be interested in an inverse regression problem, to learn about $X(\mathbf{s}_0)$ for a given $Y(\mathbf{s}_0)$. To do so, requires modeling $X(\mathbf{s})$ to be random, hence introducing a process model for the X's as well as for the Y's, a bivariate process model. Second, and perhaps more importantly, we may have missingness, equivalently, misalignment. That is, we have some X's without associated Y's and vice versa. We would like to use all of the data rather than just the *complete* cases. We need a model for the missing data; we need a bivariate model in order to do model based imputation (see, e.g., Little and Rubin, 2002).

So, to start, assume a single covariate with no misalignment, and let $\mathbf{X} = (X(\mathbf{s}_1), \ldots, X(\mathbf{s}_n))^T$ and $\mathbf{Y} = (Y(\mathbf{s}_1), \ldots, Y(\mathbf{s}_n))^T$. be the measurements on the covariates and the response, respectively. Supposing that $X(\mathbf{s})$ is continuous and that is it meaningful to model it in a spatial fashion, our approach is to envision (perhaps after a suitable transformation) a bivariate Gaussian spatial process, $\mathbf{W}(\mathbf{s}) = (X(\mathbf{s}), Y(\mathbf{s}))^T$ with mean $\boldsymbol{\mu}(\mathbf{s}) = (\mu_X(\mathbf{s}), \mu_Y(\mathbf{s}))^T$ and a separable cross-covariance function as in (9.10). Since $\rho(\mathbf{s}, \mathbf{s}) \equiv 1$, this specification implies that the joint distribution of $Y(\mathbf{s})$ and $X(\mathbf{s})$ at any location $\mathbf{s}$ is

$$\mathbf{W}(\mathbf{s}) = \begin{pmatrix} X(\mathbf{s}) \\ Y(\mathbf{s}) \end{pmatrix} \sim \mathrm{N}(\boldsymbol{\mu}(\mathbf{s}), T) \ . \tag{9.13}$$

With misalignment, let $\mathbf{X}$ be the vector of observed $X(\mathbf{s})$'s at the sites in $S_{XY} \cup R_X$, while $\mathbf{Y}$ will be the vector of $Y(\mathbf{s})$'s the sites in $S_{XY} \cup S_Y$. If we let $\mathbf{X}'$ denote the vector of missing X observations in S_Y and $\mathbf{Y}'$ the vector of missing Y observations in S_X, then in the preceding discussion we can replace $\mathbf{X}$ and $\mathbf{Y}$ by the augmented vectors $\mathbf{X}_{aug} = (\mathbf{X}, \mathbf{X}')$ and $\mathbf{Y}_{aug} = (\mathbf{Y}, \mathbf{Y}')$; see Figure 9.1 for clarification. After permutation to line up the X's and Y's, they can be collected into a vector $\mathbf{W}_{aug}$. In the Bayesian model specification, $\mathbf{X}'$ and $\mathbf{Y}'$ are viewed as latent (unobserved) vectors. In implementing a Gibbs sampler for model-fitting, we update the model parameters given $\mathbf{X}'$ and $\mathbf{Y}'$ (i.e., given $\mathbf{W}_{aug}$), and then update $(\mathbf{X}', \mathbf{Y}')$ given $\mathbf{X}$, $\mathbf{Y}$, and the model parameters.

The latter updating is routine since the associated full conditional distributions are normal. Such augmentation proves computationally easier with regard to bookkeeping since we retain the convenient Kronecker form for $\Sigma_{\mathbf{W}}$. That is, it is easier to marginalize over $\mathbf{X}'$ and $\mathbf{Y}'$ after simulation than before. For convenience of notation, we suppress the augmentation in the sequel.

In what we have called the prediction problem, it is desired to predict the outcome of the response variable at some unobserved site. Thus we are interested in the posterior predictive distribution $p(y(\mathbf{s}_0)|\mathbf{y}, \mathbf{x})$. We note that $x(\mathbf{s}_0)$ is also not observed here. On the other hand, the interpolation problem may be regarded as a method of imputing missing data. Here the

covariate $x(\mathbf{s}_0)$ is observed but the response is "missing." Thus our attention shifts to the posterior predictive distribution For the regression problem, the distribution of interest is $p(E[Y(\mathbf{s}_0)|x(\mathbf{s}_0)] \mid x(\mathbf{s}_0), \mathbf{y}, \mathbf{x})$.

For simplicity, suppose $\boldsymbol{\mu}(\mathbf{s}) = (\mu_1, \mu_2)^T$, independent of the site coordinates. (With additional fixed site-level covariates for $Y(\mathbf{s})$, say, $\mathbf{U}(\mathbf{s})$, we would replace μ_2 with $\mu_2(\mathbf{s}) = \boldsymbol{\alpha}^T\mathbf{U}(\mathbf{s})$.) Then, from (9.13), for the pair $(X(\mathbf{s}), Y(\mathbf{s}))$, $p(y(\mathbf{s})|x(\mathbf{s}), \beta_0, \beta_1, \sigma^2)$ is $N\left(\beta_0 + \beta_1 x(\mathbf{s}), \sigma^2\right)$. That is, $E[Y(\mathbf{s})|x(\mathbf{s})] = \beta_0 + \beta_1 x(\mathbf{s})$, where

$$\beta_0 = \mu_2 - \frac{T_{12}}{T_{11}}\mu_1,\ \beta_1 = \frac{T_{12}}{T_{11}},\ \text{and}\ \sigma^2 = T_{22} - \frac{T_{12}^2}{T_{11}}\ . \tag{9.14}$$

So, given samples from the joint posterior distribution of $(\mu_1, \mu_2, T, \boldsymbol{\phi})$, we directly have samples from the posterior distributions for the parameters in (9.14), and thus from the posterior distribution of $E[Y(\mathbf{s})|x(\mathbf{s})]$.

Rearrangement of the components of $\mathbf{W}$ as below (9.11) yields

$$\begin{pmatrix} \mathbf{X} \\ \mathbf{Y} \end{pmatrix} \sim N\left(\begin{pmatrix} \mu_1\mathbf{1} \\ \mu_2\mathbf{1} \end{pmatrix},\ T \otimes H(\boldsymbol{\phi})\right), \tag{9.15}$$

which simplifies calculation of the conditional distribution of $\mathbf{Y}$ given $\mathbf{X}$.

Assuming an inverse Wishart prior for T, completing the Bayesian specification requires a prior for μ_1, μ_2, and $\boldsymbol{\phi}$. For (μ_1, μ_2), for convenience we would take a vague but proper bivariate normal prior. A suitable prior for $\boldsymbol{\phi}$ depends upon the choice of $\rho(h; \boldsymbol{\phi})$. Then we use a Gibbs sampler to simulate the necessary posterior distributions. The full conditionals for μ_1 and μ_2 are in fact Gaussian distributions, while that of the T matrix is inverted Wishart as already mentioned. The full conditional for the $\boldsymbol{\phi}$ parameter finds $\boldsymbol{\phi}$ arising in the entries in H, and so is not available in closed form. Metropolis or slice sampling can be employed for its updating.

Under the above framework, interpolation presents no new problems. Let $\mathbf{s}_0$ be a new site at which we would like to predict the variable of interest. We first modify the $H(\boldsymbol{\phi})$ matrix forming the new matrix H^* as follows:

$$H^*(\boldsymbol{\phi}) = \begin{pmatrix} H(\boldsymbol{\phi}) & \mathbf{h}(\boldsymbol{\phi}) \\ \mathbf{h}(\boldsymbol{\phi})^T & \rho(0;\, \boldsymbol{\phi}) \end{pmatrix}, \tag{9.16}$$

where $\mathbf{h}(\boldsymbol{\phi})$ is the vector with components $\rho(\mathbf{s}_0 - \mathbf{s}_j; \boldsymbol{\phi})$, $j = 1, 2, \ldots, n$. It then follows that

$$\mathbf{W}^* \equiv (\mathbf{W}(\mathbf{s}_0), \ldots, \mathbf{W}(s_n))^T \sim \mathrm{N}\left(\mathbf{1}_{n+1} \otimes \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix},\ H^*(\boldsymbol{\phi}) \otimes T\right). \tag{9.17}$$

Once again a simple rearrangement of the above vector enables us to arrive at the conditional distribution $p(\mathbf{y}(\mathbf{s}_0)|\mathbf{x}(\mathbf{s}_0), \mathbf{y}, \mathbf{x}, \boldsymbol{\mu}, T, \boldsymbol{\phi})$ as a Gaussian distribution. The predictive distribution for the interpolation problem, $p(\mathbf{y}(\mathbf{s}_0)|\mathbf{y}, \mathbf{x})$, can now be obtained by marginalizing over the parameters, i.e.,

$$p(y(\mathbf{s}_0)|\mathbf{y}, \mathbf{x}) = \int p(y(\mathbf{s}_0)|x(\mathbf{s}_0), \mathbf{y}, \mathbf{x}, \boldsymbol{\mu}, T, \boldsymbol{\phi})\, p(\boldsymbol{\mu}, T, \boldsymbol{\phi}|x(\mathbf{s}_0), \mathbf{y}, \mathbf{x})\ . \tag{9.18}$$

For prediction, we do not have $x(\mathbf{s}_0)$. But this does not create any new problems, as it may be treated as a latent variable and incorporated into $\mathbf{x}'$. This only results in an additional draw within each Gibbs iteration, and is a trivial addition to the computational task.

9.4.2 Avoiding the symmetry of the cross-covariance matrix

In the spirit of Le and Zidek (1992), we can avoid the symmetry in separable cross-covariances noted above (9.12). Instead of directly modeling $\Sigma_{\mathbf{W}} = H(\boldsymbol{\phi}) \otimes T$, we can add a further hierarchical level, by assuming that $\Sigma_{\mathbf{W}} \mid \boldsymbol{\phi}, T$ follows an inverted Wishart distribution with mean $H(\boldsymbol{\phi}) \otimes T$. All other specifications remain as before. Note that the marginal model (i.e., marginalizing over $\Sigma_{\mathbf{W}}$) is no longer Gaussian. However, using standard calculations, the resulting cross-covariance matrix is a function of $\rho(\mathbf{s} - \mathbf{s}'; \boldsymbol{\phi})$, retaining desirable spatial interpretation. Once again we resort to the Gibbs sampler to arrive at the posteriors, although in this extended model the number of parameters has increased substantially, since the elements of $\Sigma_{\mathbf{W}}$ are being introduced as new parameters.

The full conditionals for the means μ_1 and μ_2 are still Gaussian and it is easily seen that the full conditional for $\Sigma_{\mathbf{W}}$ is inverted Wishart. The full conditional distribution for $\boldsymbol{\phi}$ is now proportional to $p(\Sigma_{\mathbf{W}}|\boldsymbol{\phi}, T)p(\boldsymbol{\phi})$; a Metropolis step may be employed for its updating. Also, the full conditional for T is no longer inverted Wishart and a Metropolis step with an inverted Wishart proposal is used to sample the T matrix. All told, this is indeed a much more computationally demanding proposition since we now have to deal with the $2n \times 2n$ matrix $\Sigma_{\mathbf{W}}$ with regard to sampling, inversion, determinants, and so on.

9.4.3 Regression in a probit model

Now suppose we have binary response from a point-source spatial dataset. At each site, $Z(\mathbf{s})$ equals 0 or 1 according to whether we observed "failure" or "success" at that particular site. Thus, a realization of the process can be partitioned into two disjoint subregions, one for which $Z(\mathbf{s}) = 0$, the other $Z(\mathbf{s}) = 1$, and is called a *binary map* (DeOliveira, 2000). Again, the process is only observed at a finite number of locations. Along with this binary response we have a set of covariates observed at each site. We follow the latent variable approach for probit modeling as in, e.g., DeOliveira (2000). Let $Y(\mathbf{s})$ be a latent spatial process associated with the sites and let $X(\mathbf{s})$ be a process that generates the values of a particular covariate, in particular, one that is misaligned with $Z(\mathbf{s})$ and is sensible to model in a spatial fashion. For the present we assume $X(\mathbf{s})$ is univariate but extension to the multivariate case is apparent. Let $Z(\mathbf{s}) = 1$ if and only if $Y(\mathbf{s}) > 0$. We envision our bivariate process $\mathbf{W}(\mathbf{s}) = (X(\mathbf{s}), Y(\mathbf{s}))^T$ distributed as in (9.13), but where now $\boldsymbol{\mu}(\mathbf{s}) = (\mu_1, \mu_2 + \boldsymbol{\alpha}^T \mathbf{U}(\mathbf{s}))^T$, with $\mathbf{U}(\mathbf{s})$ regarded as a $p \times 1$ vector of fixed covariates. Note that the conditional variance of $Y(\mathbf{s})$ given $X(\mathbf{s})$ is not identifiable. Thus, without loss of generality, we set $T_{22} = 1$, so that the T matrix has only two parameters.

Now, we formulate a probit regression model as follows:

$$\begin{aligned} &P(Z(\mathbf{s}) = 1 \mid x(\mathbf{s}), \mathbf{U}(\mathbf{s}), \boldsymbol{\alpha}, \mu_1, \mu_2, T_{11}, T_{12}) \\ &\qquad = \Phi\left(\left[\beta_0 + \beta_1 X(\mathbf{s}) + \boldsymbol{\alpha}^T \mathbf{U}(\mathbf{s})\right] / \sqrt{1 - \tfrac{T_{12}^2}{T_{11}}} \right) . \end{aligned} \tag{9.19}$$

Here, as in (9.14), $\beta_0 = \mu_2 - (T_{12}/T_{11})\mu_1$, and $\beta_1 = T_{12}/T_{11}$.

The posterior of interest is $p(\mu_1, \mu_2, \boldsymbol{\alpha}, T_{11}, T_{12}, \boldsymbol{\phi}, \mathbf{y} \mid \mathbf{x}, \mathbf{z})$, where $\mathbf{z} = (z(\mathbf{s}_1), \ldots, z(\mathbf{s}_n))^T$ is a vector of 0's and 1's. The fitting again uses MCMC. Here, $\mathbf{X} = (X(\mathbf{s}_1), \ldots, X(\mathbf{s}_n))^T$ and $\mathbf{Y} = (Y(\mathbf{s}_1), \ldots, Y(\mathbf{s}_n))^T$ as in Subsection 9.4.1, except that $\mathbf{Y}$ is now unobserved, and introduced only for computational convenience. Analogous to (9.15),

$$\begin{pmatrix} \mathbf{X} \\ \mathbf{Y} \end{pmatrix} \sim \mathrm{N}\left(\begin{pmatrix} \mu_1 \mathbf{1} \\ \mu_2 \mathbf{1} + \mathbf{U}\boldsymbol{\beta} \end{pmatrix}, T \otimes H(\boldsymbol{\phi}) \right) , \tag{9.20}$$

where $\mathbf{U} = (U(\mathbf{s}_1), \ldots, U(\mathbf{s}_n))^T$.

From (9.20), the full conditional distribution for each latent $Y(\mathbf{s}_i)$ is a univariate normal truncated to a set of the form $\{Y(\mathbf{s}_i) > 0\}$ or $\{Y(\mathbf{s}_i) < 0\}$. The full conditionals for μ_1 and μ_2 are both univariate normal, while that of $\boldsymbol{\beta}$ is multivariate normal with the appropriate dimension. For the elements of the T matrix, we may simulate first from a Wishart distribution (as mentioned in Subsection 9.4.1) and then proceed to scale it by T_{22}, or we may proceed individually for T_{12} and T_{11} using Metropolis-Hastings over a restricted convex subset of a hypercube (Chib and Greenberg, 1998). Finally, $\boldsymbol{\phi}$ can be simulated using a Metropolis step, as in Subsection 9.4.1. Misalignment is also treated as in Subsection 9.4.1, introducing appropriate latent $\mathbf{X}'$ and $\mathbf{Y}'$.

With posterior samples from $p(\mu_1, \mu_2, \boldsymbol{\alpha}, T_{11}, T_{12}, \boldsymbol{\phi} \mid \mathbf{x}, \mathbf{z})$, we immediately obtain samples from the posterior distributions for β_0 and β_1. Also, given $x(\mathbf{s}_0)$, (9.19) shows how to obtain samples from the posterior for a particular probability, such as $p(P(Z(\mathbf{s}_0) = 1 \mid x(\mathbf{s}_0), \mathbf{U}(\mathbf{s}_0), \boldsymbol{\alpha}, \mu_1, \mu_2, T_{11}, T_{12}) \mid x(\mathbf{s}_0), \mathbf{x}, \mathbf{z})$ at an unobserved site $\mathbf{s}_0$, clarifying the regression structure. Were $x(\mathbf{s}_0)$ not observed, we could still consider the chance that $Z(\mathbf{s}_0)$ equals 1. This probability, $P\left(Z(\mathbf{s}_0) = 1 \mid \mathbf{U}(\mathbf{s}_0), \boldsymbol{\alpha}, \mu_1, \mu_2, T_{11}, T_{12}\right)$, arises by averaging over $X(\mathbf{s}_0)$, i.e.,

$$\int P\left(Z(\mathbf{s}_0) = 1 \mid x(\mathbf{s}_0), \mathbf{U}(\mathbf{s}_0), \boldsymbol{\alpha}, \mu_1, \mu_2, T_{11}, T_{12}\right)\, p(x(\mathbf{s}_0)|\mu_1, T_{11})\, dx(\mathbf{s}_0)\ . \tag{9.21}$$

In practice, we would replace the integration in (9.21) by a Monte Carlo integration. Then, plugging into this Monte Carlo integration, the foregoing posterior samples would yield essentially posterior realizations of (9.21).

Both the prediction problem and the interpolation problem may be viewed as examples of *indicator kriging* (e.g., Solow, 1986; DeOliveira, 2000). For the prediction case we seek $p(z(\mathbf{s}_0)|\mathbf{x}, \mathbf{z})$; realizations from this distribution arise if we can obtain realizations from $p(y(\mathbf{s}_0)|\mathbf{x}, \mathbf{z})$. But

$$p(y(\mathbf{s}_0) \mid \mathbf{x}, \mathbf{z}) = \int p(y(\mathbf{s}_0) \mid \mathbf{x}, \mathbf{y})\, p(\mathbf{y}|\mathbf{x}, \mathbf{z}) d\,\mathbf{y}\ . \tag{9.22}$$

Since the first distribution under the integral in (9.22) is a univariate normal, as in Subsection 9.4.1, the posterior samples of $\mathbf{Y}$ immediately provide samples of $Y(\mathbf{s}_0)$. For the interpolation case we seek $p(z(\mathbf{s}_0) \mid x(\mathbf{s}_0), \mathbf{x}, \mathbf{z})$. Again we only need realizations from $p(y(\mathbf{s}_0) \mid x(\mathbf{s}_0), \mathbf{x}, \mathbf{z})$, but

$$p(y(\mathbf{s}_0) \mid x(\mathbf{s}_0), \mathbf{x}, \mathbf{z}) = \int p(y(\mathbf{s}_0) \mid x(\mathbf{s}_0), \mathbf{x}, \mathbf{y})\, p(\mathbf{y} \mid x(\mathbf{s}_0), \mathbf{x}, \mathbf{z}) d\,\mathbf{y}\ . \tag{9.23}$$

As with (9.22), the first distribution under the integral in (9.23) is a univariate normal.

9.4.4 Examples

Example 9.1 *(Gaussian model).* Our examples are based upon an ecological dataset collected over a west-facing watershed in the Negev Desert in Israel. The species under study is called an isopod, and builds its residence by making burrows. Some of these burrows thrive through the span of a generation while others do not. We study the following variables at each of 1129 sites. The variable "dew" measures time in minutes (from 8 a.m.) to evaporation of the morning dew. The variables "shrub" and "rock" density are percentages (the remainder is sand) characterizing the environment around the burrows. In our first example we try to explain shrub density (Y) through dew duration (X). In our second example we try to explain burrow survival (Z) through shrub density, rock density, and dew duration, treating only the last one as random and spatial. We illustrate the Gaussian case for the first example with 694 of the sites offering both measurements, 204 sites providing only the shrub density, and 211 containing only the dew measurements.

Figure 9.2 *Spatial locations of the isopod burrows data. The axes represent the eastings and the northings on a UTM projection.*

Parameter	Quantiles 2.5%	50%	97.5%
μ_1	73.118	73.885	74.665
μ_2	5.203	5.383	5.572
T_{11}	95.095	105.220	117.689
T_{12}	–4.459	–2.418	–0.528
T_{22}	5.564	6.193	6.914
$T_{12}/\sqrt{T_{11}T_{22}}$ (nonspatial corr. coef.)	–0.171	-0.095	–0.021
β_0 (intercept)	5.718	7.078	8.463
β_1 (slope)	–0.041	–0.023	–0.005
σ^2	5.582	6.215	6.931
ϕ	0.0091	0.0301	0.2072

Table 9.1 *Posterior quantiles for the shrub density/dew duration example.*

The spatial locations are displayed in Figure 9.2 using rescaled planar coordinates after UTM projection. The rectangle in Figure 9.2 is roughly 300 km by 250 km. Hence the vector $\mathbf{X}$ consists of $694 + 211 = 905$ measurements, while the vector $\mathbf{Y}$ consists of $694 + 204 = 898$ measurements. For these examples we take the exponential correlation function, $\rho(h; \boldsymbol{\phi}) = e^{-\phi h}$. We assign a vague inverse gamma specification for the parameter ϕ, namely an IG(2, 1/0.024). This prior has infinite variance and suggests a range $(3/\phi)$ of 125 km, which is roughly half the maximum pairwise distance in our region. We found little inference sensitivity to the mean of this prior. The remaining prior specifications are all rather noninformative, i.e., a $N\left(\mathbf{0}, Diag(10^5, 10^5)\right)$ prior for (μ_1, μ_2) and an $IW\left(2, Diag(0.001, 0.001)\right)$ for T. That is, $E(T_{11}) = E(T_{22}) = 0.001$, $E(T_{12}) = 0$, and the variances of the T_{ij}'s do not exist.

Table 9.1 provides the 95% credible intervals for the regression parameters and the decay parameter ϕ. The significant negative association between dew duration and shrub density

Figure 9.3 *Posterior distributions for inference at the location* s_0*, denoted by "×" in the previous figure. Line legend: dotted line denotes* $p\left(E[Y(\mathbf{s}_0)|x(\mathbf{s}_0)] \mid \mathbf{x}, \mathbf{y}\right)$ *(regression); solid line denotes* $p\left(y(\mathbf{s}_0) \mid x(\mathbf{s}_0), \mathbf{x}, \mathbf{y}\right)$ *(prediction); and dashed line denotes* $p\left(y(\mathbf{s}_0) \mid \mathbf{x}, \mathbf{y}\right)$ *(interpolation).*

Figure 9.4 *For the Gaussian analysis case, a three-dimensional surface plot of* $E\left(Y(\mathbf{s})|\mathbf{x}, \mathbf{y}\right)$ *over the isopod burrows regional domain.*

is unexpected but is evident on a scatterplot of the 714 sites having both measurements. The intercept β_0 is significantly high, while the slope β_1 is negative. The maximum distance in the sample is approximately 248.1 km, so the spatial range, computed from the point estimate of $3/\phi$ from Table 9.1, is approximately 99.7 km, or about 40% of the maximum distance.

In Figure 9.3(a) we show the relative performances (using posterior density estimates) of prediction, interpolation, and regression at a somewhat central location $\mathbf{s}_0$, indicated by an "×" in Figure 9.2. The associated $X(\mathbf{s}_0)$ has the value 73.10 minutes. Regression (dotted line), since it models the means rather than predicting a variable, has substantially smaller variability than prediction (solid line) or interpolation (dashed line). In Figure 9.3(b), we "zoom in" on the latter pair. As expected, interpolation has less variability due to the specification of $x(\mathbf{s}_0)$. It turns out that in all cases, the observed value falls within the associated

Parameter	Quantiles 2.5%	50%	97.5%
μ_1	75.415	76.095	76.772
μ_2	0.514	1.486	2.433
T_{11}	88.915	99.988	108.931
T_{12}	0.149	0.389	0.659
ϕ	0.0086	0.0302	0.2171
β_0 (intercept)	0.310	1.256	2.200
β_1 (dew slope)	0.032	0.089	0.145
α_1 (shrub)	–0.0059	–0.0036	–0.0012
α_2 (rock)	–0.00104	–0.00054	–0.00003

Table 9.2 *Posterior quantiles for the burrow survival example.*

intervals. Finally, in Figure 9.4 we present a three-dimensional surface plot of $E(Y(\mathbf{s})|\mathbf{x},\mathbf{y})$ over the region. This plot reveals the spatial pattern in shrub density over the watershed. Higher measurements are expected in the eastern and particularly the southeastern part of the region, while relatively fewer shrubs are found in the northern and western parts.

Example 9.2 *(Probit model).* Our second example uses a smaller data set, from the same region as Figure 9.2, which has 246 burrows of which 43 do not provide the dew measurements. Here the response is binary, governed by the success ($Y = 1$) or failure ($Y = 0$) of a burrow at a particular site. The explanatory variables (dew duration, shrub density, and rock density) relate, in some fashion, to water retention. Dew measurements are taken as the X's in our modeling with shrub and rock density being U_1 and U_2, respectively. The prior specifications leading to the probit modeling again have vague bivariate normal priors for (μ_1, μ_2) and also for $\boldsymbol{\beta}$, which is two-dimensional in this example. For ϕ we again assign a noninformative inverse gamma specification, the IG(2, 0.024). We generate T_{11} and T_{12} through scaling a Wishart distribution for T with prior $IW\,(2, Diag(0.001, 0.001))$.

In Table 9.2, we present the 95% credible intervals for the parameters in the model. The positive coefficient for dew is expected. It is interesting to note that shrub and rock density seem to have a negative impact on the success of the burrows. This leads us to believe that although high shrub and rock density may encourage the hydrology, it is perhaps not conducive to the growth of food materials for the isopods, or encourages predation of the isopods. The spatial range parameter again explains about 40% of the maximum distance. Figure 9.5 presents the density estimates for the posteriors $p\,(P(Z(\mathbf{s}_0) = 1 \mid x(s_0), \mathbf{U}(\mathbf{s}_0), \boldsymbol{\alpha}, \beta_0, \beta_1) \mid x(s_0), \mathbf{x}, \mathbf{z})$ and $p\,(P(Z(\mathbf{s}_0) = 1 \mid \mathbf{U}(\mathbf{s}_0), \boldsymbol{\alpha}, \mu_1, \mu_2, T_{11}, T_{12}) \mid \mathbf{x}, \mathbf{z})$, with $\mathbf{s}_0$ being a central location and $x(\mathbf{s}_0) = 74.7$ minutes (after 8 a.m.), to compare performance of interpolation and prediction. As expected, interpolation provides a slightly tighter posterior distribution.

9.4.5 Conditional modeling

The spatial regression model in Section 9.4.1 uses a bivariate Gaussian process to model the variables $Y(\mathbf{s})$ and $X(\mathbf{s})$ *jointly.* Alternatively, one could proceed using a *conditional approach*, where we first model one variable and, conditional upon the first, model the second. Again, consider two spatial variables $Y(\mathbf{s})$ and $X(\mathbf{s})$ observed over a finite set of locations $\mathcal{S} = \{\mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_n\}$. Let $\mathbf{Y}$ and $\mathbf{X}$ represent $n \times 1$ vectors of the observed $Y(\mathbf{s}_i)$'s and $X(\mathbf{s}_i)$'s respectively. The conditional approach models specifies the distribution of $\mathbf{X}$

Figure 9.5 *Estimated posterior densities for the probit data analysis: solid line indicates* $P(Z(s_0) = 1 \mid X(s_0), \mathbf{U}(s_0), \alpha, \beta_0, \beta_1)$*, while dashed line indicates* $P(Z(s_0) = 1 \mid \mathbf{U}(s_0), \alpha, \mu_1, \mu_2, T_{11}, T_{12})$*.*

and, subsequently, the conditional distribution of $\mathbf{Y}$ given $\mathbf{X}$. This is attractive in that valid specification of these two distributions yields a legitimate joint distribution.

How do we model $\mathbf{Y} \mid \mathbf{X}$? It would be attractive to imagine a conditional *process* $Y(\mathbf{s}) \mid X(\mathbf{s})$ but this, in general, is not well defined for an arbitrary collection of variables. In fact, recall that, practically, we attempt such definition through finite dimensional distributions. Yet, it is meaningless to talk about the joint distribution of $Y(\mathbf{s}_i) \mid X(\mathbf{s}_i)$ and $Y(\mathbf{s}_j) \mid X(\mathbf{s}_j)$ for two distinct locations $\mathbf{s}_i$ and $\mathbf{s}_j$. This reveals the impossibility of specifying conditioning that would yield a consistent (in the sense of Section 3.1) definition of a process. In fact, this point reveals a flaw in the discussion of conditioning for multivariate spatial process modeling in Cressie and Wikle (2011). They consider multivariate normal distributions associated with, say, m locations for $Y(\mathbf{s})$ and, say, n locations for $X(\mathbf{s})$. It is evident that their specification does not create a bivariate spatial process and, thus, it is not possible to develop kriging in their context.

A special case of the conditional approach does produce a valid bivariate process model. Suppose that $X(\mathbf{s})$ is a univariate Gaussian spatial process with mean $\mu_X(\mathbf{s})$ and covariance function $C_X(\cdot; \boldsymbol{\theta}_X)$ indexed by process parameters $\boldsymbol{\theta}_X$. Therefore, $\mathbf{X} \sim N(\boldsymbol{\mu}_X\, \Sigma_X(\boldsymbol{\theta}_X))$, where $\boldsymbol{\mu}_X$ is $n \times 1$ with $\mu_X(\mathbf{s}_i)$ as its i-th entry and $\Sigma_X(\boldsymbol{\theta}_X)$ is an $n \times n$ spatial covariance matrix with entries $C_X(\mathbf{s}_i, \mathbf{s}_j; \boldsymbol{\theta}_X)$. Let $e(\mathbf{s})$ be another Gaussian process, independent of $X(\mathbf{s})$, with zero mean and covariance function $C_e(\cdot; \boldsymbol{\theta}_e)$ indexed by process parameters $\boldsymbol{\theta}_e$. Then, for any finite collection of n locations, suppose that

$$Y(\mathbf{s}_i) = \beta_0 + \beta_1 X(\mathbf{s}_i) + e(\mathbf{s}_i)\ , \quad \text{for} \quad i = 1, 2, \ldots, n\ . \tag{9.24}$$

The joint distribution between $\mathbf{X}$ and $\mathbf{Y}$ is

$$\begin{pmatrix} \mathbf{X} \\ \mathbf{Y} \end{pmatrix} \sim N\left(\begin{pmatrix} \boldsymbol{\mu}_X \\ \boldsymbol{\mu}_Y \end{pmatrix}, \begin{pmatrix} \Sigma_X(\boldsymbol{\theta}_X) & \beta_1 \Sigma_X(\boldsymbol{\theta}_X) \\ \beta_1 \Sigma_X(\boldsymbol{\theta}_X) & \Sigma_e(\boldsymbol{\theta}_e) + \beta_1^2 \Sigma_X(\boldsymbol{\theta}_X) \end{pmatrix} \right)\ , \tag{9.25}$$

where $\boldsymbol{\mu}_Y = \beta_0 \mathbf{1} + \beta_1 \boldsymbol{\mu}_X$ and $\Sigma_e(\boldsymbol{\theta}_e)$ is the $n \times n$ variance-covariance matrix for the $e(\mathbf{s}_i)$'s. Note that the above joint distribution arises from a legitimate bivariate spatial process

$\mathbf{W}(\mathbf{s}) = (X(\mathbf{s}), Y(\mathbf{s}))^T$, with mean $\boldsymbol{\mu}_{\mathbf{W}}(\mathbf{s}) = \begin{pmatrix} \mu_X(\mathbf{s}) \\ \beta_0 + \beta_1 \mu_X(\mathbf{s}) \end{pmatrix}$ and cross-covariance

$$C_{\mathbf{W}}(\mathbf{s}, \mathbf{s}') = \begin{pmatrix} C_X(\mathbf{s}, \mathbf{s}') & \beta_1 C_X(\mathbf{s}, \mathbf{s}') \\ \beta_1 C_X(\mathbf{s}, \mathbf{s}') & \beta_1^2 C_X(\mathbf{s}, \mathbf{s}') + C_e(\mathbf{s}, \mathbf{s}') \end{pmatrix}, \tag{9.26}$$

where we have suppressed the dependence of $C_X(\mathbf{s}, \mathbf{s}')$ and $C_e(\mathbf{s}, \mathbf{s}')$ on $\boldsymbol{\theta}_X$ and $\boldsymbol{\theta}_e$ respectively. Equation (9.24) implies that $\mathrm{E}[Y(\mathbf{s}) \mid X(\mathbf{s})] = \beta_0 + \beta_1 X(\mathbf{s})$ for any arbitrary location $\mathbf{s}$, thereby specifying a well-defined spatial regression model for an arbitrary $\mathbf{s}$.

An adaptation of (9.24) produces asymmetric cross-covariance matrices. For clarity, suppose that $X(\mathbf{s})$ is a Gaussian process with mean $\mu_X(\mathbf{s})$ and a stationary covariance function $C_X(\mathbf{s}, \mathbf{s}') = C_X(\mathbf{h})$, where $\mathbf{h} = \mathbf{s} - \mathbf{s}'$, and consider the following regression model

$$Y(\mathbf{s}) = \beta_0 + \beta_1 X(\mathbf{s} + \mathbf{r}) + \epsilon(\mathbf{s}), \tag{9.27}$$

where $\mathbf{r}$ is some *fixed* location. This model, often called a *spatial delay*, again defines a legitimate bivariate process with a stationary but *asymmetric* cross-covariance function

$$C_{\mathbf{W}}(\mathbf{h}) = \begin{pmatrix} C_X(\mathbf{h}) & \beta_1 C_X(\mathbf{r} + \mathbf{h}) \\ \beta_1 C_X(\mathbf{r} - \mathbf{h}) & \beta_1^2 C_X(\mathbf{h}) + C_e(\mathbf{h}) \end{pmatrix}.$$

While the introduction of the asymmetry may seem attractive, it is unclear when a spatial delay arises in practice. Put differently, why would we want to regress $Y(\mathbf{s})$ on $X(\mathbf{s} + \mathbf{r})$? Wackernagel (2003) draws an analogy with time-series, where it has been observed that in some cases that the effect of one variable on another is not instantaneous. The time for the second variable to react to the first causes a lag or "delay" in the correlations between the time-series. Here is another explanation: the Cauchy-Schwarz inequality ensures that $\mathrm{Cov}(X(\mathbf{s}), Y(\mathbf{s} + \mathbf{h})) = C_{XY}(\mathbf{h})$ attains an extremum which is proportional to that of $C_X(\mathbf{h})$. It may happen that the extremum (maximum in the case of a positively correlated variable pair and minimum otherwise) is shifted away in a direction $\mathbf{r}$. In that case, $C_{XY}(\mathbf{h})$ attains an extremum which is proportional to that of $C_{XX}(\mathbf{r} + \mathbf{h})$. Wackernagel (2003) suggests plotting the empirical cross-covariance function over appropriately constructed bins to diagnose spatial lag and, hence, ascertain $\mathbf{r}$. This, however, is awkward in practice. In any case, the modeling is rather restrictive and is perhaps feasible only in the bivariate setting. Later in the chapter we discuss richer and more flexible nonstationary spatially-varying cross-covariance modeling for jointly modeling several spatial outcomes, which accommodate assymetric cross-covariances without requiring one to estimate spatial lags or delays.

The conditional approach is not bereft of problems. As already mentioned, it will not generally produce multivariate process models, which precludes predictions at arbitrary locations for all the outcomes. The example above is one of the few instances where a legitimate bivariate process is obtained (see Royle and Berliner, 1999, for other examples). Another issue with the conditional approach is the order of the hierarchy. Specifying $p(\mathbf{X})p(\mathbf{Y} \mid \mathbf{X})$ $p(\mathbf{Y})$ yields a joint distribution different from specifying $p(\mathbf{Y})p(\mathbf{X} \mid \mathbf{Y})$. How do we decide upon the ordering? In certain applications, there may be a natural ordering based upon a causal relationship. When that is not the case, this ambiguity with regard to the ordering is undesirable and practically inefficient as we move to more than two outcomes. For example, with three outcomes, we will have six models emerging from the order of the outcomes in the hierarchy and with four outcomes we have 24 models. Clearly, when the information regarding the order of the hierarchy is lacking, we have an explosion in the number of models.

Finally, we note that the specification associated with $X(\mathbf{s})$ and $Y(\mathbf{s})$ in (9.24) can obviously be extended so that both $\mathbf{X}(\mathbf{s})$ and $\mathbf{Y}(\mathbf{s})$ are vectors. We can create a joint process model through such conditioning. However, in this case, we would need multivariate process models for both $\mathbf{X}(\mathbf{s})$ and $\mathbf{Y}(\mathbf{s})$; we are back to our original joint modeling problem.

9.4.6 Spatial regression with kernel averaged predictors

In formulating a spatial regression model, we have argued that is can be useful to assume that $X(\mathbf{s})$ is random and in the previous sections, we have shown how to implement spatial regression under a bivariate process model for $\mathbf{Z}(\mathbf{s}) = (Y(\mathbf{s}), X(\mathbf{s}))$. However, in spatial applications, *neighboring* $X(\mathbf{s}')$ can be expected to inform about $Y(\mathbf{s})$ particularly when the distance between $\mathbf{s}$ and $\mathbf{s}'$ is small. For instance, precipitation can affect the water table (the depth at which soil and pour spaces become completely saturated with water) not only where the precipitation fell but also at surrounding locations due to run off and changes in slope from uneven ground surfaces. Similarly, the concentration of ozone is affected by pollutants, ultraviolet rays, and temperature within a neighborhood of location $\mathbf{s}$. Thus, a mean specification which only includes $X(\mathbf{s})$ and not neighboring $X(\mathbf{s}')$ may not adequately capture the process. Here, drawing upon Heaton and Gelfand (2011), we develop a kernel averaged predictor regression model to address this issue.

A mean specification which incorporates neighboring predictors might take the form $E(Y(\mathbf{s})) = \beta_0(\mathbf{s}) + \int_D \boldsymbol{\beta}(\mathbf{s}, \mathbf{u}; \boldsymbol{\theta}) X(\mathbf{u}) d\mathbf{u}$ where $\boldsymbol{\beta}(\mathbf{s}, \mathbf{s}'; \boldsymbol{\theta})$ is a coefficient model capturing the effect of $X(\mathbf{s}')$ on $Y(\mathbf{s})$ with parameters $\boldsymbol{\theta}$. With, say n observations $\mathbf{Z} = (\mathbf{Z}(\mathbf{s}_1)^T, \ldots, \mathbf{Z}(\mathbf{s}_n)^T)^T$, we might consider the conditional distribution, $Y(\mathbf{s}_i) \mid X(\mathbf{s}_1), \ldots, X(\mathbf{s}_n)$. In the Gaussian setting, $[Y(\mathbf{s}_i) \mid X(\mathbf{s}_1), \ldots, X(\mathbf{s}_n)]$ contains n coefficients $\boldsymbol{\beta}(\mathbf{s}_i, \mathbf{s}_1), \ldots, \boldsymbol{\beta}(\mathbf{s}_i, \mathbf{s}_n)$ which describe the loading of $X(\mathbf{s}_j)$ on $Y(\mathbf{s}_i)$ for $j = 1, \ldots, n$. Such a choice is unattractive because the explained relationship between the response and predictor depends on the number of sampling locations and the arrangement of those locations in D. It also fails to explicitly capture the idea of local $X(\mathbf{s}')$ informing about $Y(\mathbf{s})$.

Rather, we seek a *kernel averaged* predictor, $\tilde{X}(\mathbf{s})$ based upon $\{X(\mathbf{s}') : \mathbf{s}' \in D\}$ under an appropriate choice for $\boldsymbol{\beta}(\mathbf{s}, \mathbf{s}'); \boldsymbol{\theta})$. What we are doing differs from what is usually referred to as employing functional covariates (see Baillo and Grane, 2009). Functional covariates envision a function at location $\mathbf{s}$, say, $X(\mathbf{s}, t)$ for $t \in (0, T]$ and seek to reduce this to a single covariate at $\mathbf{s}$ to explain $Y(\mathbf{s})$. Usually, this is achieved through some integration of the function wherein the integration is over t rather than over $\mathbf{s}$ as we seek. Also, we do not seek to build process models for $\boldsymbol{\beta}(\mathbf{s}, \mathbf{s}'; \boldsymbol{\theta})$. We provide process modeling for $(Y(\mathbf{s}), X(\mathbf{s}))$ but specify $\boldsymbol{\beta}$ as a parametric function. Specifying the latter using processes will provide poorly-identified, over-fitted models.

Let $Y(\mathbf{s})$ and $X(\mathbf{s})$ denote a univariate response variable and a single covariate at location $\mathbf{s} \in D$. Furthermore, assume $X(\mathbf{s})$ follows a Gaussian process of the form,

$$X(\mathbf{s}) = \mu_X(\mathbf{s}) + \sigma_X w_X(\mathbf{s}), \tag{9.28}$$

where $\mu_X(\mathbf{s})$ is the mean surface at location $\mathbf{s}$ and $w_X(\mathbf{s})$ is a mean 0, variance 1 GP with correlation function $\rho_X(\mathbf{s}, \mathbf{s}'; \boldsymbol{\phi}_X)$. Here, we imitate the specification above. Notice that (9.28) defines a purely spatial covariate process. In some applications, however, covariates are measured with error. In these cases, $X(\mathbf{s})$ can be thought of as the "true" underlying covariate process and the observed covariate is $H(\mathbf{s}) = X(\mathbf{s}) + \epsilon_X(\mathbf{s})$ where $\epsilon_X(\mathbf{s})$ is an *i.i.d.* white noise process. Distributional results for $H(\mathbf{s})$ will differ from the those of $X(\mathbf{s})$ by an additive nugget variance term only, so, for simplicity, below, we assume $X(\mathbf{s})$ is observed. Moreover, we still want to use $X(\mathbf{s})$ to explain $Y(\mathbf{s})$, i.e., we want the regression of $Y(\mathbf{s})$ to be on the true $X(\mathbf{s})$.

We define the unobserved kernel-averaged local covariate at $\mathbf{s}$ as

$$\tilde{X}(\mathbf{s}) \equiv \frac{1}{K(\mathbf{s}; \boldsymbol{\xi})} \int_D K(\mathbf{s}, \mathbf{u}; \boldsymbol{\xi}) X(\mathbf{u}) d\mathbf{u}, \tag{9.29}$$

where $K(\mathbf{s}, \mathbf{s}'; \boldsymbol{\xi})$ is a kernel defining a weight on the distance between $\mathbf{s}$ and $\mathbf{s}'$ with parameters $\boldsymbol{\xi}$ and $0 < K(\mathbf{s}; \boldsymbol{\xi}) = \int_D K(\mathbf{s}, \mathbf{u}; \boldsymbol{\xi}) d\mathbf{u} < \infty$. The parameter $\boldsymbol{\xi}$, most commonly, consists

of *scale* parameters such as the entries of a covariance matrix but can also include *location* parameters as in Higdon (1998) and Xu et al. (2005). Choices for K are discussed below.

Because a valid GP was defined for $X(\mathbf{s})$, $\tilde{X}(\mathbf{s})$ also is a valid GP with mean $\tilde{\mu_X}(\mathbf{s}) = K(\mathbf{s};\boldsymbol{\xi})^{-1}\int_D K(\mathbf{s},\mathbf{u};\boldsymbol{\xi})\mu_X(\mathbf{u})d\mathbf{u}$ and

$$\begin{aligned}\mathrm{Cov}(\tilde{X}(\mathbf{s}),\tilde{X}(\mathbf{s}')) &= \sigma_X^2\rho_{\tilde{X}}(\mathbf{s},\mathbf{s}') \text{ where} \\ \rho_{\tilde{X}}(\mathbf{s},\mathbf{s}') &= \frac{1}{K(\mathbf{s};\boldsymbol{\xi})K(\mathbf{s}';\boldsymbol{\xi})}\int_D\int_D K(\mathbf{s},\mathbf{u};\boldsymbol{\xi})K(\mathbf{s}',\mathbf{v};\boldsymbol{\xi})\rho_X(\mathbf{u},\mathbf{v})d\mathbf{v}d\mathbf{u}\ .\end{aligned} \tag{9.30}$$

Not only are $X(\mathbf{s})$ and $\tilde{X}(\mathbf{s})$ marginally Gaussian processes, but a valid bivariate GP is induced for the pair $(X(\mathbf{s}),\tilde{X}(\mathbf{s}))^T$. Specifically, $(X(\mathbf{s}),\tilde{X}(\mathbf{s}))^T$ follows a bivariate GP with mean $(\mu_X(\mathbf{s}),\tilde{\mu_X}(\mathbf{s}))^T$ and

$$\mathrm{Cov}\left(\begin{pmatrix} X(\mathbf{s}) \\ \tilde{X}(\mathbf{s})\end{pmatrix},\begin{pmatrix} X(\mathbf{s}') \\ \tilde{X}(\mathbf{s}')\end{pmatrix}\right) = \sigma_X^2\begin{pmatrix}\rho_X(\mathbf{s},\mathbf{s}') & \rho_{X,\tilde{X}}(\mathbf{s},\mathbf{s}') \\ \rho_{\tilde{X},X}(\mathbf{s},\mathbf{s}') & \rho_{\tilde{X}}(\mathbf{s},\mathbf{s}')\end{pmatrix}, \tag{9.31}$$

where

$$\begin{aligned}\rho_{X,\tilde{X}}(\mathbf{s},\mathbf{s}') &= \frac{1}{K(\mathbf{s}';\boldsymbol{\xi})}\int_D K(\mathbf{s}',\mathbf{u};\boldsymbol{\xi})\rho_X(\mathbf{s},\mathbf{u})d\mathbf{u} \text{ and} \\ \rho_{\tilde{X},X}(\mathbf{s},\mathbf{s}') &= \frac{1}{K(\mathbf{s};\boldsymbol{\xi})}\int_D K(\mathbf{s},\mathbf{u};\boldsymbol{\xi})\rho_X(\mathbf{u},\mathbf{s}')d\mathbf{u}\end{aligned}$$

for any other location $\mathbf{s}' \in D$.

Next, consider the linear model defined by,

$$Y(\mathbf{s}) \mid X(\mathbf{s}),\tilde{X}(\mathbf{s}) = \beta_0 + \beta_1\tilde{X}(\mathbf{s}) + \sigma_Y w_Y(\mathbf{s}) + \epsilon_Y(\mathbf{s}), \tag{9.32}$$

where $w_Y(\mathbf{s})$ is defined analogously to $w_X(\mathbf{s})$ in (9.28) but with correlation function ρ_Y and $\epsilon_Y(\mathbf{s})$ is a Gaussian white noise process with variance τ_Y^2. Intuitively, the kernel $K(\mathbf{s},\mathbf{u};\boldsymbol{\xi})$ describes how the effect of the covariate $X(\mathbf{s}), \mathbf{s} \in D$ propagates to the response.

Notice that the bivariate GP for $(X(\mathbf{s}),\tilde{X}(\mathbf{s}))^T$ along with (9.32) provide a joint specification of a trivariate GP for $\mathbf{Z}(\mathbf{s}) = (X(\mathbf{s}),\tilde{X}(\mathbf{s}),Y(\mathbf{s}))^T$. Specifically, $\mathbf{Z}(\mathbf{s})$ follows a valid trivariate GP with mean

$$\boldsymbol{\mu}(\mathbf{s}) = (\mu_X(\mathbf{s}),\tilde{\mu}_X(\mathbf{s}),\beta_0 + \beta_1\tilde{\mu}_X(\mathbf{s}))^T$$

and cross-covariance,

$$\sigma_X^2\begin{pmatrix}\rho_X(\mathbf{s},\mathbf{s}') & \rho_{X,\tilde{X}}(\mathbf{s},\mathbf{s}') & \beta_1\rho_{X,\tilde{X}}(\mathbf{s},\mathbf{s}') \\ \rho_{\tilde{X},X}(\mathbf{s},\mathbf{s}') & \rho_{\tilde{X}}(\mathbf{s},\mathbf{s}') & \beta_1\rho_{\tilde{X}}(\mathbf{s},\mathbf{s}') \\ \beta_1\rho_{\tilde{X},X}(\mathbf{s},\mathbf{s}') & \beta_1\rho_{\tilde{X}}(\mathbf{s},\mathbf{s}') & \frac{\sigma_Y^2}{\sigma_{\tilde{X}}^2}\rho_Y(\mathbf{s},\mathbf{s}') + \beta_1^2\rho_{\tilde{X}}(\mathbf{s},\mathbf{s}')\end{pmatrix} \tag{9.33}$$

The joint distribution of $\mathbf{Z}(\mathbf{s})$ is useful for evaluating the properties of (9.32) such as induced correlations between $Y(\mathbf{s})$ and $(X(\mathbf{s}),\tilde{X}(\mathbf{s}))$. Note also that the induced bivariate process model for $(X(\mathbf{s}),Y(\mathbf{s}))$ is not a usual bivariate process specification; the kernel appears in the covariance structure.

We focus on the *local* spatial regression in terms of the conditional distribution of $Y(\mathbf{s})$ given $\tilde{X}(\mathbf{s})$ in the form of the conditional mean

$$E[Y(\mathbf{s}) \mid \tilde{X}(\mathbf{s})] = \beta_0 + \beta_1\tilde{X}(\mathbf{s})\ .$$

Thus, given $\tilde{X}(\mathbf{s})$, concepts such as R^2, mean square error, variable selection, shrinkage, etc. are applicable. Of course, all are random and would be averaged over the distribution

Kernel	$K(\mathbf{s}, \mathbf{s}'; \boldsymbol{\xi})$	Parameters ($\boldsymbol{\xi}$)
Uniform	$\mathbf{I}_{\{\|\mathbf{s}-\mathbf{s}'\| \leq \xi\}}$	ξ
Epanechnikov	$(\xi^2 - \|\mathbf{s}-\mathbf{s}'\|^2)\mathbf{I}_{\{\|\mathbf{s}-\mathbf{s}'\| \leq \xi\}}$	ξ
Component Wise Gaussian	$\prod_{i=1}^d \xi_i^{-1} \exp\left\{-(s_i - s_i')^2/(2\xi_i^2)\right\}$	$\xi_i, i = 1, \ldots, d$
Oriented Gaussian	$\|\Xi\|^{-1/2} \exp\left\{-\frac{(\mathbf{s}'-\mathbf{s})^T \Xi^{-1} (\mathbf{s}'-\mathbf{s})}{2}\right\}$	$\Xi = \{\xi_{ij}\}$

Table 9.3 *Examples of kernel functions $K(\mathbf{s}, \mathbf{s}'; \boldsymbol{\xi})$ and their parameters. $\mathbf{I}_A$ is an indicator for the set A.*

of $\tilde{X}(\mathbf{s})$ in order to interpret them. Evidently, the potentially complex relationship between $\{X(\mathbf{s}) : \mathbf{s} \in D\}$ and $Y(\mathbf{s})$ is captured through a single parameter (β_1). However, from (9.32), we see that, effectively, we are introducing a coefficient weighting of the entire surface $X(\mathbf{s})$ to explain $Y(\mathbf{s})$. That is, the coefficient of $X(\mathbf{s}')$ is $\beta_1 K(\mathbf{s}, \mathbf{s}'; \boldsymbol{\xi})/K(\mathbf{s}; \boldsymbol{\xi})$. The normalization by $K(\mathbf{s}; \boldsymbol{\xi})$ identifies β_1.

The choice of K with the data informing about $\boldsymbol{\xi}$ enables a fairly rich regression specification while attractively reducing to a simple linear regression model in $\tilde{X}(\mathbf{s})$. Examples of kernels are given in Table 9.3. We note that computation of $K(\mathbf{s}; \boldsymbol{\xi})$ for general regions D varies with $\mathbf{s}$ and can be computationally expensive when done repeatedly over MCMC iterations. Below, the kernel is taken to be $K(\mathbf{s}, \mathbf{s}'; \xi) = I(\{||\mathbf{s} - \mathbf{s}'|| \leq \xi g(\boldsymbol{\phi}_X)\})$ where $I(A)$ is an indicator for the set A, $||\cdot||$ denotes Euclidean distance, $\xi \in (0, 1)$ and $g(\boldsymbol{\phi}_X)$ is the effective spatial range associated with ρ_X; for example, from Chapter 2, the exponential correlation function has $g(\boldsymbol{\phi}_X) \approx 3/\phi_X$ where ϕ_X is the spatial decay parameter. Intuitively, the kernel $K(\mathbf{s}, \mathbf{s}'; \xi)$ is a disk centered at location $\mathbf{s}$ with radius $r = \xi g(\boldsymbol{\phi}_X)$. The scale parameter r can be loosely interpreted as a hard threshold "decay" parameter in that the effect of $X(\mathbf{s}')$ on $Y(\mathbf{s})$ is negligible if $||\mathbf{s} - \mathbf{s}'|| \geq r$. For very small r, $\tilde{X}(\mathbf{s}) \approx X(\mathbf{s})$. For large r, $\tilde{X}(\mathbf{s}) \approx \tilde{X}(\mathbf{s}')$ for all $\mathbf{s}, \mathbf{s}'$ yielding a highly collinear regression.

With the choice of $K(\mathbf{s}, \mathbf{s}'; \xi)$ above, r and $\boldsymbol{\phi}_X$ are strongly associated parameters; this is evident from the forms in (9.33). Plausible values for scale parameters of kernels depend on ϕ_X as well as D. For the K above, parameterizing the scale parameter as $r = \xi g(\boldsymbol{\phi}_X)$ where $\xi \in (0, 1)$ removes this dependency such that, a priori, ξ and ϕ_X can be taken as independent and also restricts K to be within the effective spatial range of ρ_X.

Next, consider what (9.32) implies about using $X(\mathbf{s})$ instead of $\tilde{X}(\mathbf{s})$ in the conditional mean for $Y(\mathbf{s}) \mid X(\mathbf{s}), \tilde{X}(\mathbf{s})$ assuming $\mathrm{E}(Y(\mathbf{s}) \mid X(\mathbf{s}), \tilde{X}(\mathbf{s})) = \beta_0 + \beta_1 \tilde{X}(\mathbf{s})$. Using the fact that $\mathbf{Z}(\mathbf{s}) = (X(\mathbf{s}), \tilde{X}(\mathbf{s}), Y(\mathbf{s}))^T$ follows a trivariate Gaussian process with known mean and covariance given by (9.33), multivariate normal theory gives that $Y(\mathbf{s}) \mid X(\mathbf{s})$ is also normally distributed with mean,

$$\mathrm{E}[Y(\mathbf{s}) \mid X(\mathbf{s})] = \beta_0 + \beta_1(\tilde{\mu}_X(\mathbf{s}) + \rho_{X,\tilde{X}}(\mathbf{s}, \mathbf{s})(X(\mathbf{s}) - \mu_X(\mathbf{s}))), \tag{9.34}$$

variance,

$$\mathrm{Var}(Y(\mathbf{s}) \mid X(\mathbf{s})) = \tau_Y^2 + \sigma_Y^2 + \beta_1^2 \sigma_X^2 (\rho_{\tilde{X}}(\mathbf{s}, \mathbf{s}) - \rho_{X,\tilde{X}}^2(\mathbf{s}, \mathbf{s})), \tag{9.35}$$

and covariance,

$$\mathrm{Cov}(Y(\mathbf{s}), Y(\mathbf{s}') \mid X(\mathbf{s}), X(\mathbf{s}')) = \sigma_Y^2 \rho_Y(\mathbf{s}, \mathbf{s}') + \beta_1^2 \sigma_X^2 \rho_{\tilde{X}}(\mathbf{s}, \mathbf{s}') \,. \tag{9.36}$$

So, when the model given by (9.32) holds, then the change in $Y(\mathbf{s})$ as a result of a unit change in $X(\mathbf{s})$ is $\beta_1 \rho_{X,\tilde{X}}(\mathbf{s}, \mathbf{s})$ as opposed to β_1 when using $\tilde{X}(\mathbf{s})$. Hence, under the trivariate Gaussian process model, i.e., when the assumptions of (9.32) hold, using $X(\mathbf{s})$ implies that the effect of the covariate on $Y(\mathbf{s})$ is shrunk towards zero; $|\beta_1 \rho_{X,\tilde{X}}(\mathbf{s}, \mathbf{s})| \leq |\beta_1|$ because $0 \leq \rho_{X,\tilde{X}}(\mathbf{s}, \mathbf{s}) \leq 1$. This result is not surprising in that if other $X(\mathbf{s}')$ in D besides $X(\mathbf{s})$ affect $Y(\mathbf{s})$ then the effect due to $X(\mathbf{s})$ is expected to diminish. The amount

of shrinkage is determined by the kernel parameters, $\boldsymbol{\xi}$, as well as $\boldsymbol{\phi}_X$. For example, if $Y(\mathbf{s})$ is ozone and $X(\mathbf{s})$ is temperature, the implication would be that using $X(\mathbf{s})$ in the model could, potentially, underestimate the change in $Y(\mathbf{s})$ as a result from a unit change in temperature, leading to underestimation of the production of ozone.

A second consequence of using $X(\mathbf{s})$ instead of $\tilde{X}(\mathbf{s})$ when (9.32) holds is that the percent of variation in $Y(\mathbf{s})$ explained by $X(\mathbf{s})$ is less than the percent of variation in $Y(\mathbf{s})$ explained by $\tilde{X}(\mathbf{s})$ for many common covariance functions as detailed by the following result from Heaton and Gelfand (2011):

Let $\rho^2_{Y|X}$ and $\rho^2_{Y|\tilde{X}}$ be the population coefficient of determination for the linear model defined by (9.34) and (9.32), respectively. If ρ_X is an isotropic, log-concave correlation function then $\rho^2_{Y\,|\,X} \leq \rho^2_{Y\,|\,\tilde{X}}$.

The class of log-concave covariance functions includes the powered exponential, $\rho(\mathbf{s}, \mathbf{s}') = \exp\{-\phi\|\mathbf{s} - \mathbf{s}'\|^\alpha$ $0 \leq \alpha \leq 2$, by direct calculation. Also included are closed form Matérn models, i.e., those with smoothness parameter ν of the form $\nu = k + 1/2$ for $k \in \{0, 1, 2, \ldots\}$ again by direct calculation with an indication that this is the case for arbitrary ν (see Majumdar and Gelfand, 2007). Also, it is easy to argue that convolution of covariance functions produces valid covariance functions. In fact, convolution of log-concave functions produces log-concave functions. Thus, for such covariance functions, using $X(\mathbf{s})$ instead of $\tilde{X}(\mathbf{s})$ results in less variation in $Y(\mathbf{s})$ explained. Additionally, notice that marginalizing over $\tilde{X}(\mathbf{s})$ adds extra spatial variation to $Y(\mathbf{s})$. To see this, notice that the covariance given by (9.36) has the added variance term $\beta_1^2\sigma_X^2\rho_{\tilde{X}}(\mathbf{s})$.

Model fitting, efficient computation, kernel selection, and illustrative examples are presented in Heaton and Gelfand (2011) and we encourage the interested reader to look at this paper for further discussion and details.

9.5 Coregionalization models ⋆

9.5.1 Coregionalization models and their properties

We now consider a constructive modeling strategy to add flexibility to (9.10) while retaining interpretability and computational tractability. Our approach is through the *linear model of coregionalization* (LMC), as, for example, in Grzebyk and Wackernagel (1994) and Wackernagel (1998). The term "coregionalization" is intended to denote a model for measurements that covary jointly over a region.

The most basic coregionalization model, the so-called *intrinsic specification*, dates at least to Matheron (1982). It arises as $\mathbf{Y}(\mathbf{s}) = A\mathbf{w}(\mathbf{s})$ where the components of $\mathbf{w}(\mathbf{s})$ are i.i.d. spatial processes. If the $w_j(\mathbf{s})$ have mean 0 and are stationary with variance 1 and correlation function $\rho(h)$, then $E(\mathbf{Y}(\mathbf{s}))$ is $\mathbf{0}$ and the cross-covariance matrix, $\Sigma_{\mathbf{Y}(\mathbf{s}),\mathbf{Y}(\mathbf{s}')} \equiv C(\mathbf{s} - \mathbf{s}') = \rho(\mathbf{s} - \mathbf{s}')AA^T$. Letting $AA^T = T$ this immediately reveals the equivalence between this simple intrinsic specification and the separable covariance specification as in Section 9.3 above. As in Subsection 2.1.2, the term "intrinsic" is taken to mean that the specification only requires the first and second moments of differences in measurement vectors and that the first moment difference is $\mathbf{0}$ and the second moments depend on the locations only through the separation vector $\mathbf{s} - \mathbf{s}'$. In fact here $E(\mathbf{Y}(\mathbf{s}) - \mathbf{Y}(\mathbf{s}')) = \mathbf{0}$ and $\frac{1}{2}\Sigma_{\mathbf{Y}(\mathbf{s})-\mathbf{Y}(\mathbf{s}')} = G(\mathbf{s} - \mathbf{s}')$ where $G(\mathbf{h}) = C(\mathbf{0}) - C(\mathbf{h}) = T - \rho(\mathbf{s} - \mathbf{s}')T = \gamma(\mathbf{s} - \mathbf{s}')T$ where γ is a valid variogram. Of course, as in the $p = 1$ case, we need not begin with a covariance function but rather just specify the process through γ and T. A more insightful interpretation of "intrinsic" is that given in equation (9.12).

We assume A is full rank and, for future reference, we note that A can be assumed to be the lower triangular. This, lower-triangular specification for A does impose certain conditional independence constraints. To see how, consider the bivariate case, two locations $\mathbf{s}_1$ and $\mathbf{s}_2$ and suppose that the elements of A are fixed. Then, for $i = 1, 2$, $Y_1(\mathbf{s}_i) = a_{11}w_1(\mathbf{s}_i)$

and $Y_2(\mathbf{s}_i) = a_{21}w_1(\mathbf{s}_i) + a_{22}w_2(\mathbf{s}_i)$. Since the process $w_1(\mathbf{s})$ completely determines $Y_1(\mathbf{s})$, we can write

$$\begin{aligned}\mathrm{Cov}\{&Y_1(\mathbf{s}_1), Y_2(\mathbf{s}_2) \mid Y_1(\mathbf{s}_2)\}\\ &= \mathrm{Cov}\{a_{11}w_1(\mathbf{s}_1), a_{21}w_1(\mathbf{s}_2) + a_{22}w_2(\mathbf{s}_2) \mid w_1(\mathbf{s}_2)\}\\ &= a_{11}a_{22}\mathrm{Cov}\{w_1(\mathbf{s}_1), w_2(\mathbf{s}_2) \mid w_1(\mathbf{s}_2)\} = 0\ ,\end{aligned} \tag{9.37}$$

where the last equality follows because the process $w_1(\cdot)$ is independent of $w_2(\cdot)$. This reveals that $Y_1(\mathbf{s}_1)$ and $Y_2(\mathbf{s}_2)$ will be independent conditional upon $Y_1(\mathbf{s}_2)$. In terms of the dispersion matrix, this conditional independence implies that the Σ_Y^{-1} will have redundant zeroes. In theory, we can easily obviate restrictions such as in (9.37) by specifying A to be a non-triangular square root of T obtained by spectral decomposition. One example is a spectrally decomposed matrix, which is thereafter modeled using a set of eigenvalues and *Given* angles. One could set $A = P\Lambda^{1/2}$, or the symmetric square-root $A = P\Lambda^{1/2}P'$, where $T = P\Lambda P'$ is the spectral decomposition for T. This requires further parametrization for the orthogonal matrix P, such as in terms of the $p(p-1)/2$ *Given's* angles $\theta_{ij}(\mathbf{s})$ for $i = 1, \ldots, p-1$ and $j = i+1, \ldots, p$ (e.g., Daniels and Kass, 1999). Specifically, $P = \prod_{i=1}^{p-1}\prod_{j=i+1}^{p} G_{ij}(\theta_{ij})$ where i and j are distinct and $G_{ij}(\theta_{ij})$ is almost the $p \times p$ identity matrix except that its i-th and j-th diagonal elements are replaced by $\cos(\theta_{ij})$ and $\pm\sin(\theta_{ij})$ respectively. Given P for any $\mathbf{s}$, the θ_{ij}'s are unique within range $(-\pi/2, \pi/2)$. These may be further modeled by means of Gaussian processes on a suitably transformed function, say, $\tilde{\theta}_{ij} = \log(\frac{\pi/2+\theta_{ij}}{\pi/2-\theta_{ij}})$. For further reference, see Daniels and Kass, 1999; Kang and Cressie, 2011.

While the conditional independence in (9.37) from the triangular specification for A may seem theoretically unnecessary, it is only an *a priori* assumption that does not carry over to posterior inference. Our ultimate interest resides with the cross-covariance function, which is robustly estimated using any bijective mapping with some square-root matrix. Furthermore, the number of parameters to be estimated in the *Given's* angle specification is the same as that for the triangular Cholesky. In practical settings, these specifications matter little but Cholesky decompositions are numerically more stable than the spectral decomposition. The former is also less expensive, requiring $O(m^3/3)$ flops as compared to more than $O(4m^3/3)$ flops required by the latter. Hence, we opt for the Cholesky decomposition in our subsequent data analysis. No additional richness accrues, at least from a practical standpoint, to a more general A.

A more general LMC arises if again $\mathbf{Y}(\mathbf{s}) = A\mathbf{w}(\mathbf{s})$ but now the $w_j(\mathbf{s})$ are independent but no longer identically distributed. In fact, let the $w_j(\mathbf{s})$ process have mean μ_j, variance 1, and correlation function $\rho_j(h)$. Then $E(\mathbf{Y}(\mathbf{s})) = A\boldsymbol{\mu}$ where $\boldsymbol{\mu}^T = (\mu_1, \cdots, \mu_p)$ and the cross-covariance matrix associated with $\mathbf{Y}(\mathbf{s})$ is now

$$\Sigma_{\mathbf{Y}(\mathbf{s}),\mathbf{Y}(\mathbf{s}')} \equiv C(\mathbf{s}-\mathbf{s}') = \sum_{j=1}^{p} \rho_j(\mathbf{s}-\mathbf{s}')T_j\ , \tag{9.38}$$

where $T_j = \mathbf{a}_j\mathbf{a}_j^T$ with $\mathbf{a}_j$ the jth column of A. Note that $\sum_j T_j = T$. More importantly, we note that such linear combination produces stationary spatial processes. We return to this point in Section 9.6.3.

The one-to-one relationship between T and lower triangular A is standard. For future use, when $p = 2$ we have

$$a_{11} = \sqrt{T_{11}}\ , \quad a_{21} = \frac{T_{12}}{\sqrt{T_{11}}} \ \text{ and } \ a_{22} = \sqrt{T_{22} - \frac{T_{12}^2}{T_{11}}}\ .$$

When $p = 3$ we add

$$a_{31} = \frac{T_{13}}{\sqrt{T_{11}}}\ , \quad a_{32} = \frac{T_{11}T_{23} - T_{12}T_{13}}{\sqrt{T_{11}T_{22} - T_{12}^2}\sqrt{T_{11}}} \text{ and}$$
$$a_{33} = \sqrt{T_{33} - \frac{T_{13}^2}{T_{11}} - \frac{(T_{11}T_{23} - T_{12}T_{13})^2}{T_{11}(T_{11}T_{22} - T_{12}^2)}}.$$

Lastly, if we introduce monotonic isotropic correlation functions, we will be interested in the range associated with $Y_j(\mathbf{s})$. An advantage to (9.38) is that each $Y_j(\mathbf{s})$ has its own range. In particular, for $p = 2$ the range for $Y_1(\mathbf{s})$ solves $\rho_1(d) = 0.05$, while the range for $Y_2(\mathbf{s})$ solves the weighted average correlation,

$$\frac{a_{21}^2\rho_1(d) + a_{22}^2\rho_2(d)}{a_{21}^2 + a_{22}^2} = 0.05\ . \tag{9.39}$$

Since ρ_1 and ρ_2 are monotonic the left side of (9.39) is decreasing in d. Hence, solving (9.39) is routine. If we have $p = 3$, we need in addition the range for $Y_3(\mathbf{s})$. We require the solution of

$$\frac{a_{31}^2\rho_1(d) + a_{32}^2\rho_2(d) + a_{33}^2\rho_3(d)}{a_{31}^2 + a_{32}^2 + a_{33}^2} = 0.05\ . \tag{9.40}$$

The left side of (9.40) is again decreasing in d. The form for general p is clear.

In practice, the ρ_j are parametric classes of functions. Hence the range d is a parametric function that is not available explicitly. However, within a Bayesian context, when models are fitted using simulation-based methods, we obtain posterior samples of the parameters in the ρ_j's, as well as A. Each sample, when inserted into the left side of (9.39) or (9.40), enables solution for a corresponding d. In this way, we obtain posterior samples of each of the ranges, one-for-one with the posterior parameter samples.

Extending in a different fashion, we can define a process having a general *nested* covariance model (see, e.g., Wackernagel, 1998) as

$$\mathbf{Y}(\mathbf{s}) = \sum \mathbf{Y}^{(u)}(\mathbf{s}) = \sum_{u=1}^{r} A^{(u)}\mathbf{w}^{(u)}(\mathbf{s})\ , \tag{9.41}$$

where the $\mathbf{Y}^{(u)}$ are independent intrinsic LMC specifications with the components of $\mathbf{w}^{(u)}$ having correlation function ρ_u. The cross-covariance matrix associated with (9.41) takes the form

$$C(\mathbf{s} - \mathbf{s}') = \sum_{u=1}^{r} \rho_u(\mathbf{s} - \mathbf{s}')T^{(u)}\ , \tag{9.42}$$

with $T^{(u)} = A^{(u)}(A^{(u)})^T$. The $T^{(u)}$ are full rank and are referred to as *coregionalization matrices*. Expression (9.42) can be compared to (9.38). Note that r need not be equal to p, but $\mathbf{\Sigma}_{\mathbf{Y}(\mathbf{s})} = \sum_u T^{(u)}$. Also, recent work of Vargas-Guzmán et al. (2002) allows the $\mathbf{w}^{(u)}(\mathbf{s})$, hence the $\mathbf{Y}^{(u)}(\mathbf{s})$ in (9.41), to be dependent.

Returning to the more general LMC, in applications we introduce (9.38) as a spatial random effects component of a general multivariate spatial model for the data. That is, we assume

$$\mathbf{Y}(\mathbf{s}) = \boldsymbol{\mu}(\mathbf{s}) + \mathbf{v}(\mathbf{s}) + \boldsymbol{\epsilon}(\mathbf{s})\ , \tag{9.43}$$

where $\boldsymbol{\epsilon}(\mathbf{s})$ is a white noise vector, i.e., $\boldsymbol{\epsilon}(\mathbf{s}) \sim N(\mathbf{0}, D)$ where D is a $p \times p$ diagonal matrix with $(D)_{jj} = \tau_j^2$. In (9.43), $\mathbf{v}(\mathbf{s}) = A\mathbf{w}(\mathbf{s})$ following (9.38) as above, but further assuming

that the $w_j(\mathbf{s})$ are mean-zero Gaussian processes. Lastly $\boldsymbol{\mu}(\mathbf{s})$ arises from $\mu_j(\mathbf{s}) = \mathbf{X}_j^T(\mathbf{s})\boldsymbol{\beta}_j$. Each component can have its own set of covariates with its own coefficient vector.

As in Section 6.1, (9.43) can be viewed as a hierarchical model. At the first stage, given $\{\boldsymbol{\beta}_j, j = 1, \cdots, p\}$ and $\{\mathbf{v}(\mathbf{s}_i)\}$, the $\mathbf{Y}(\mathbf{s}_i)$, $i = 1, \cdots, n$ are conditionally independent with $\mathbf{Y}(\mathbf{s}_i) \sim N(\boldsymbol{\mu}(\mathbf{s}_i) + \mathbf{v}(\mathbf{s}_i), D)$. At the second stage, the joint distribution of $\mathbf{v} \equiv (\mathbf{v}(\mathbf{s}_1), \cdots, \mathbf{v}(\mathbf{s}_n))^T$ is $N(\mathbf{0}, \sum_{j=1}^p H_j \otimes T_j)$, where H_j is $n \times n$ with $(\mathbf{H}_j)_{ii'} = \rho_j(\mathbf{s}_i - \mathbf{s}_{i'})$. Concatenating the $\mathbf{Y}(\mathbf{s}_i)$ into an $np \times 1$ vector $\mathbf{Y}$ (and similarly $\boldsymbol{\mu}(\mathbf{s}_i)$ into $\boldsymbol{\mu}$), we can marginalize over $\mathbf{v}$ to obtain

$$p(\mathbf{Y} \mid \{\boldsymbol{\beta}_j\}, D, \{\rho_j\}, T) = N\left(\boldsymbol{\mu}\,,\, \sum_{j=1}^p (H_j \otimes T_j) + I_{n\times n} \otimes D\right) . \tag{9.44}$$

Prior distributions on $\{\boldsymbol{\beta}_j\}$, $\{\tau_j^2\}$, T, and the parameters of the ρ_j complete the Bayesian hierarchical model specification.

9.5.2 Unconditional and conditional Bayesian specifications

9.5.2.1 Equivalence of likelihoods

In Section 9.4.5 we briefly discussed bivariate spatial modeling using a conditional specification. Having discussed the LMC in the previous section in a fairly general context, it is worth exploring the connections between the conditional approach and the LMC in more general settings. The LMC of the previous section can be developed through a conditional approach rather than a joint modeling approach. This idea has been elaborated in, e.g., Royle and Berliner (1999) and Berliner (2000), who refer to it as a hierarchical modeling approach to multivariate spatial modeling and prediction.

In the context of say $\mathbf{v}(\mathbf{s}) = A\mathbf{w}(\mathbf{s})$ where the $w_j(\mathbf{s})$ are mean-zero Gaussian processes, by taking A to be lower triangular the equivalence and associated reparametrization are easy to see. Upon permutation of the components of $\mathbf{v}(\mathbf{s})$ we can, without loss of generality, write

$$p(\mathbf{v}(\mathbf{s})) = p(v_1(\mathbf{s})) \times p(v_2(\mathbf{s}) \mid v_1(\mathbf{s})) \times \cdots \times p(v_p(\mathbf{s}) \mid v_1(\mathbf{s}), \cdots, v_{p-1}(\mathbf{s})) .$$

In the case of $p = 2$, $p(v_1(\mathbf{s}))$ is clearly $N(0, T_{11})$, i.e. $v_1(\mathbf{s}) = \sqrt{T_{11}}w_1(\mathbf{s}) = a_{11}w_1(\mathbf{s})$, $a_{11} > 0$. But

$$v_2(\mathbf{s}) \mid v_1(\mathbf{s})) \sim N\left(\frac{T_{12}v_1(\mathbf{s})}{T_{11}}, T_{22} - \frac{T_{12}^2}{T_{11}}\right) \;,\;\; \text{i.e.}\;\; N\left(\frac{a_{21}}{a_{11}}v_1(\mathbf{s}), a_{22}^2\right) .$$

In fact, from the previous section we have $\Sigma_{\mathbf{v}} = \sum_{j=1}^p H_j \otimes T_j$. If we permute the rows of $\mathbf{v}$ to $\tilde{\mathbf{v}} = \left(\mathbf{v}^{(1)}, \mathbf{v}^{(2)}\right)^T$, where $\mathbf{v}^{(l)} = (v_l(\mathbf{s}_1), \ldots, v_l(\mathbf{s}_n))^T$ for $l = 1, 2$, then $\Sigma_{\mathbf{v}} = \sum_{j=1}^p T_j \otimes H_j$. Again with $p = 2$ we can calculate $E(\mathbf{v}^{(2)}|\mathbf{v}^{(1)}) = \frac{a_{21}}{a_{11}}\mathbf{v}^{(1)}$ and $\Sigma_{\mathbf{v}^{(2)}|\mathbf{v}^{(1)}} = a_{22}^2 H_2$. But this is exactly the mean and covariance structure associated with variables $\{v_2(\mathbf{s}_i)\}$ given $\{v_1(\mathbf{s}_i)\}$, i.e., with $v_2(\mathbf{s}_i) = \frac{a_{21}}{a_{11}}v_1(\mathbf{s}_i) + a_{22}w_2(\mathbf{s}_i)$. Note that as in Subsection 9.4, there is no notion of a *conditional* process here. Again there is only a joint distribution for $\mathbf{v}^{(1)}, \mathbf{v}^{(2)}$ given any n and any $\mathbf{s}_1, \cdots, \mathbf{s}_n$, hence a conditional distribution for $\mathbf{v}^{(2)}$ given $\mathbf{v}^{(1)}$.

Suppose we write $v_1(\mathbf{s}) = \sigma_1 w_1(\mathbf{s})$ where $\sigma_1 > 0$ and $w_1(\mathbf{s})$ is a mean 0 spatial process with variance 1 and correlation function ρ_1 and we write $v_2(\mathbf{s})|v_1(\mathbf{s}) = \alpha v_1(\mathbf{s}) + \sigma_2 w_2(\mathbf{s})$ where $\sigma_2 > 0$ and $w_2(\mathbf{s})$ is a mean 0 spatial process with variance 1 and correlation function ρ_2. The parametrization $(\alpha, \sigma_1, \sigma_2)$ is obviously equivalent to (a_{11}, a_{12}, a_{22}), i.e., $a_{11} = \sigma_1$, $a_{21} = \alpha\sigma_1$, $a_{22} = \sigma_2$ and hence to T, i.e., to (T_{11}, T_{12}, T_{22}), that is, $T_{11} = \sigma_1^2$, $T_{12} = \alpha\sigma_1^2$, $T_{22} = \alpha^2\sigma_1^2 + \sigma_2^2$.

Extension to general p is straightforward but notationally messy. We record the transformations for $p = 3$ for future use. First, $v_1(\mathbf{s}) = \sigma_1 w_1(\mathbf{s})$, $v_2(\mathbf{s})|v_1(\mathbf{s}) = \alpha^{(2|1)} v_1(\mathbf{s}) + \sigma_2 w_2(\mathbf{s})$ and $v_3(\mathbf{s})|v_1(\mathbf{s}), v_2(\mathbf{s}) = \alpha^{(3|1)} v_1(\mathbf{s}) + \alpha^{(3|2)} v_2(\mathbf{s}) + \sigma_3 w_3(\mathbf{s})$. Then $a_{11} = \sigma_1$, $a_{21} = \alpha^{(2|1)}\sigma_1$, $a_{22} = \sigma_2$, $a_{31} = \alpha^{(3|1)}\sigma_1$, $a_{32} = \alpha^{(3|2)}\sigma_2$ and $a_{33} = \sigma_3$. But also $a_{11} = \sqrt{T_{11}}$, $a_{21} = \dfrac{T_{12}}{\sqrt{T_{11}}}$, $a_{22} = \sqrt{T_{22} - \dfrac{T_{12}^2}{T_{11}}}$, $a_{31} = \dfrac{T_{13}}{\sqrt{T_{11}}}$, $a_{32} = \sqrt{\dfrac{T_{11}T_{23} - T_{12}T_{13}}{T_{11}(T_{11}T_{22} - T_{12}^2)}}$, and $a_{33} = \sqrt{T_{33} - \dfrac{T_{13}^2}{T_{11}} - \dfrac{(T_{11}T_{23} - T_{12}T_{13})^2}{T_{11}(T_{11}T_{12} - T_{12}^2)}}$.

Advantages to working with the conditional form of the model are certainly computational and possibly mechanistic or interpretive. For the former, with the "σ, α" parametrization, the likelihood factors and thus, with a matching prior factorization, models can be fitted componentwise. Rather than the $pn \times pn$ covariance matrix involved in working with $\mathbf{v}$ we obtain p covariance matrices each of dimension $n \times n$, one for $\mathbf{v}^{(1)}$, one for $\mathbf{v}^{(2)}|\mathbf{v}^{(1)}$, etc. Since likelihood evaluation with spatial processes is more than an order n^2 calculation, there can be substantial computational savings in using the conditional model. If there is some natural chronology or perhaps causality in events, then this would determine a natural order for conditioning and hence suggest natural conditional specifications. For example, in the illustrative commercial real estate setting of Example 9.3, we have the income (I) generated by an apartment block and the selling price (P) for the block. A natural modeling order here is I, then P given I.

9.5.2.2 Equivalence of prior specifications

Working in a Bayesian context, it is appropriate to ask about choice of parametrization with regard to prior specification. Suppose we let $\boldsymbol{\phi}_j$ be the parameters associated with the correlation function ρ_j. Let $\boldsymbol{\phi}^T = (\boldsymbol{\phi}_1, \cdots, \boldsymbol{\phi}_p)$. Then the distribution of $\mathbf{v}$ depends upon T and $\boldsymbol{\phi}$. Suppose we assume *a priori* that $p(T, \boldsymbol{\phi}) = p(T)p(\boldsymbol{\phi}) = p(T)\prod_j p(\boldsymbol{\phi}_j)$. Then reparametrization, using obvious notation, to the $(\boldsymbol{\sigma}, \boldsymbol{\alpha})$ space results on a prior $p(\boldsymbol{\sigma}, \boldsymbol{\alpha}, \boldsymbol{\phi}) = p(\boldsymbol{\sigma}, \boldsymbol{\alpha}) \prod_j p(\boldsymbol{\phi}_j)$.

Standard prior specification for T would of course be an inverse Wishart, while standard modeling for $(\boldsymbol{\sigma}^2, \boldsymbol{\alpha})$ would be a product inverse gamma by normal form. In the present situation, when will they agree? We present the details for the $p = 2$ case. The Jacobian from $T \to (\sigma_1, \sigma_2, \alpha)$ is $|\mathbf{J}| = \sigma_1^2$, hence in the reverse direction it is $1/T_{11}$. Also $|T| = T_{11}T_{22} - T_{12}^2 = \sigma_1^2\sigma_2^2$ and

$$T^{-1} = \frac{1}{T_{11}T_{22} - T_{12}^2} \begin{pmatrix} T_{22} & -T_{12} \\ -T_{12} & T_{11} \end{pmatrix} = \frac{1}{\sigma_1^2\sigma_2^2} \begin{pmatrix} \alpha^2\sigma_1^2 + \sigma_2^2 & -\alpha\sigma_1^2 \\ -\alpha\sigma_1^2 & \sigma_1^2 \end{pmatrix} .$$

After some manipulation we have the following result:

Result 1: $T \sim IW_2(\nu, (\nu' D)^{-1})$; that is,

$$p(T) \propto |\mathbf{T}|^{-\frac{\nu+3}{2}} \exp\left\{ -\frac{1}{2} tr(\nu' D \mathbf{T}^{-1}) \right\} ,$$

where $D = \mathrm{Diag}(d_1, d_2)$ and $\nu' = \nu - 3$ if and only if

$$\sigma_1^2 \sim IG\left(\frac{\nu-1}{2}, \frac{d_1}{2}\right), \ \sigma_2^2 \sim IG\left(\frac{\nu+1}{2}, \frac{d_2}{2}\right), \ \text{and} \ \alpha|\sigma_2^2 \sim N\left(0, \frac{\sigma_2^2}{d_1}\right) .$$

Note also that the prior in $(\boldsymbol{\sigma}, \boldsymbol{\alpha})$ space factors into $p(\sigma_1^2)p(\sigma_2^2, \alpha)$ to match the likelihood factorization.

This result is obviously order dependent. If we condition in the reverse order, σ_1^2, σ_2^2, and α no longer have the same meanings. In fact, writing this parametrization as $(\tilde{\sigma}_1^2, \tilde{\sigma}_2^2, \tilde{\alpha})$, we

obtain equivalence to the above inverse Wishart prior for T if and only if $\tilde{\sigma}_1^2 \sim IG\left(\frac{\nu+1}{2}, \frac{d_1}{2}\right)$, $\tilde{\sigma}_2^2 \sim IG\left(\frac{\nu-1}{2}, \frac{d_2}{2}\right)$, and $\tilde{\alpha}|\tilde{\sigma}_1^2 \sim N\left(0, \frac{\tilde{\sigma}_1^2}{d_1}\right)$.

The result can be extended to $p > 2$ but the expressions become messy. However, if $p = 3$ we have:

Result 2: $T \sim IW_3(\nu, (\nu' D)^{-1}$, that is,

$$p(T) \propto |\mathbf{T}|^{-\frac{\nu+4}{2}} \exp\left\{-\frac{1}{2} tr(\nu' D \mathbf{T}^{-1})\right\} ,$$

where now where $D = Diag(d_1, d_2, d_3)$ and $\nu' = \nu - 3 + 1$ if and only if $\sigma_1^2 \sim IG\left(\frac{\nu-2}{2}, \frac{d_1}{2}\right)$, $\sigma_2^2 \sim IG\left(\frac{\nu}{2}, \frac{d_2}{2}\right)$, $\sigma_3^2 \sim IG\left(\frac{\nu+2}{2}, \frac{d_3}{2}\right)$, $\alpha^{(2|1)}|\sigma_2^2 \sim N\left(0, \frac{\sigma_2^2}{d_1}\right)$, $\alpha^{(3|1)}|\sigma_3^2 \sim N\left(0, \frac{\sigma_3^2}{d_1}\right)$, and $\alpha^{(3|2)}|\sigma_3^2 \sim N\left(0, \frac{\sigma_3^2}{d_2}\right)$. Though there is a one-to-one transformation from T-space to $(\boldsymbol{\sigma}, \boldsymbol{\alpha})$-space, a Wishart prior with nondiagonal D implies a nonstandard prior on $(\boldsymbol{\sigma}, \boldsymbol{\alpha})$-space. Moreover, it implies that the prior in $(\boldsymbol{\sigma}, \boldsymbol{\alpha})$-space will not factor to match the likelihood factorization.

Returning to the model in (9.43), the presence of white noise in (9.43) causes difficulties with the attractive factorization of the likelihood under conditioning. Consider again the $p = 2$ case. If

$$\begin{aligned} Y_1(\mathbf{s}) &= \mathbf{X}_1^T(\mathbf{s})\boldsymbol{\beta}_1 + v_1(\mathbf{s}) + \epsilon_1(\mathbf{s}) \\ \text{and } Y_2(\mathbf{s}) &= \mathbf{X}_2^T(\mathbf{s})\boldsymbol{\beta}_1 + v_2(\mathbf{s}) + \epsilon_2(\mathbf{s}) , \end{aligned} \tag{9.45}$$

then the conditional form of the model writes

$$\begin{aligned} Y_1(\mathbf{s}) &= \mathbf{X}_1^T(\mathbf{s})\boldsymbol{\beta}_1 + \sigma_1 w_1(\mathbf{s}) + \tau_1 u_1(\mathbf{s}) \\ \text{and } Y_2(\mathbf{s})|Y_1(\mathbf{s}) &= \mathbf{X}_2^T(\mathbf{s})\boldsymbol{\beta}_2 + \alpha Y_1(\mathbf{s}) + \sigma_2 w_2(\mathbf{s}) + \tau_2 u_2(\mathbf{s}) . \end{aligned} \tag{9.46}$$

In (9.46), $w_1(\mathbf{s})$ and $w_2(\mathbf{s})$ are as above with $u_1(\mathbf{s})$, $u_2(\mathbf{s}) \sim N(0, 1)$, independent of each other and the $w_l(\mathbf{s})$. But then, unconditionally, $Y_2(\mathbf{s})$ equals

$$\begin{aligned} &\mathbf{X}_2^T(\mathbf{s})\widetilde{\boldsymbol{\beta}}_2 + \alpha\left(\mathbf{X}_1^T(\mathbf{s})\boldsymbol{\beta}_1 + \sigma_1 w_1(\mathbf{s}) + \tau_1 u_1(\mathbf{s})\right) + \sigma_2 w_2(\mathbf{s}) + \tau_2 u_2(\mathbf{s}) \\ &= \mathbf{X}_2^T(\mathbf{s})\widetilde{\boldsymbol{\beta}}_2 + \mathbf{X}_1^T(\mathbf{s})\alpha\boldsymbol{\beta}_1 + \alpha\sigma_1 w_1(\mathbf{s}) + \sigma_2 w_2(\mathbf{s}) + \alpha\tau_1 u_1(\mathbf{s})) + \tau_2 u_2(\mathbf{s}) . \end{aligned} \tag{9.47}$$

In attempting to align (9.47) with (9.45) we require $\mathbf{X}_2(\mathbf{s}) = \mathbf{X}_1(\mathbf{s})$, whence $\boldsymbol{\beta}_2 = \widetilde{\boldsymbol{\beta}}_2 + \alpha\boldsymbol{\beta}_1$. We also see that $v_2(\mathbf{s}) = \alpha\sigma_1 w_1(\mathbf{s}) + \sigma_2 w_2(\mathbf{s})$. But, perhaps most importantly, $\epsilon_2(\mathbf{s}) = \alpha\tau_1 u_1(\mathbf{s}) + \tau_2 u_2(\mathbf{s})$. Hence $\epsilon_1(\mathbf{s})$ and $\epsilon_2(\mathbf{s})$ are not independent, violating the white noise modeling assumption associated with (10.3). If we have a white noise component in the model for $Y_1(\mathbf{s})$ and also in the conditional model for $Y_2(\mathbf{s})|Y_1(\mathbf{s})$ we do not have a white noise component in the unconditional model specification. Obviously, the converse is true as well.

If $\mathbf{u}_1(\mathbf{s}) = 0$, i.e., the $Y_1(\mathbf{s})$ process is purely spatial, then, again with $\mathbf{X}_2(\mathbf{s}) = \mathbf{X}_1(\mathbf{s})$, the conditional and marginal specifications agree up to reparametrization. More precisely, the parameters for the unconditional model are $\boldsymbol{\beta}_1$, $\boldsymbol{\beta}_2$, τ_2^2 with T_{11}, T_{12}, T_{22}, $\boldsymbol{\phi}_1$, and $\boldsymbol{\phi}_2$. For the conditional model we have $\boldsymbol{\beta}_1$, $\boldsymbol{\beta}_2$, τ_2^2 with σ_1, σ_2, α, $\boldsymbol{\phi}_1$, and $\boldsymbol{\phi}_2$. We can appeal to the equivalence of (T_{11}, T_{12}, T_{22}) and $(\sigma_1, \sigma_2, \alpha)$ as above. Also note that if we extend (9.45) to $p > 2$, in order to enable conditional and marginal specifications to agree, we will require a common covariate vector and that $u_1(\mathbf{s}) = u_2(\mathbf{s}) = \cdots = u_{p-1}(\mathbf{s}) = 0$, i.e., that all but one of the processes is purely spatial.

9.6 Spatially varying coefficient models

In Section 11.1 we introduce a spatially varying coefficient process in the evolution equation of the spatiotemporal dynamic model. Similarly, in Section 14.4 we consider multiple spatial

frailty models with regression coefficients that were allowed to vary spatially. Here, we develop this topic more fully to amplify the scope of possibilities for such modeling. In particular, in the spatial-only case, we denote the value of the coefficient at location $\mathbf{s}$ by $\beta(\mathbf{s})$. This coefficient can be resolved at either areal unit or point level. With the former, the $\beta(\mathbf{s})$ surface consists of "tiles" at various heights, one tile per areal unit. The former are, perhaps, more natural with areal data (see Section 10.1). For the latter, we achieve a more flexible spatial surface.

Using tiles, concern arises regarding the arbitrariness of the scale of resolution, the lack of smoothness of the surface, and the inability to interpolate the value of the surface to individual locations. When working with point-referenced data it will be more attractive to allow the coefficients to vary by location, to envision for a particular coefficient, a spatial surface. For instance, in our example below we also model the (log) selling price of single-family houses. Customary explanatory variables include the age of the house, the square feet of living area, the square feet of other area, and the number of bathrooms. If the region of interest is a city or greater metropolitan area, it is evident that the capitalization rate (e.g., for age) will vary across the region. In some parts of the region older houses will be more valued than in other parts. By allowing the coefficient of age to vary with location, we can remedy the foregoing concerns. With practical interest in mind (say, real estate appraisal), we can predict the coefficient for arbitrary properties, not just for those that sold during the period of investigation. Similar issues arise in modeling environmental exposure to a particular pollutant where covariates might include temperature and precipitation.

One possible approach would be to model the spatial surface for the coefficient parametrically. In the simplest case this would require the rather arbitrary specification of a polynomial surface function; surfaces too limited or inflexible might result. More flexibility could be introduced using a spline surface over two-dimensional space; see, e.g., Luo and Wahba (1998) and references therein. However, this requires selection of a spline function and determination of the number of and locations of the knots in the space. Also, with multiple coefficients, a multivariate specification of a spline surface is required.

The approach we adopt here is arguably more natural and at least as flexible. We model the spatially varying coefficient surface as a realization from a spatial process. For multiple coefficients we employ a multivariate spatial process model.

To clarify interpretation and implementations, we first develop our general approach in the case of a single covariate, hence two spatially varying coefficient processes, one for the "intercept" and one for the "slope." We then turn to the case of multiple covariates. Recall that, even when fitting a simple linear regression, the slope and intercept are almost always strongly (and, usually, inversely) correlated. (This is intuitive if one envisions overlaying random lines that are likely relative to a fixed scatterplot of the data points.) So, if we extend to a process model for each, it seems clear that we would want the processes to be dependent (and the same reasoning would extend to the case of multiple covariates). Hence, we employ a multivariate process model. Indeed we present a further generalization to build a spatial analogue of a multilevel regression model (see, e.g., Goldstein, 1995). We also consider flexible spatiotemporal possibilities (anticipating Chapter 10). The previously mentioned real estate setting provides site level covariates whose coefficients are of considerable practical interest and a data set of single-family home sales from Baton Rouge, LA, enables illustration. Except for regions exhibiting special topography, we anticipate that a spatially varying coefficient model will prove more useful than, for instance, a trend surface model. That is, incorporating a polynomial in latitude and longitude into the mean structure would not be expected to serve as a surrogate for allowing a coefficient for, say, age or living area of a house to vary across the region.

9.6.1 Approach for a single covariate

Recall the usual Gaussian stationary spatial process model as in (6.1),

$$Y(\mathbf{s}) = \mu(\mathbf{s}) + w(\mathbf{s}) + \epsilon(\mathbf{s}) \ , \tag{9.48}$$

where $\mu(\mathbf{s}) = \mathbf{x}(\mathbf{s})^T\beta$ and $\epsilon(\mathbf{s})$ is a white noise process, i.e., $\mathrm{E}(\epsilon(\mathbf{s})) = 0$, $\mathrm{Var}(\epsilon(\mathbf{s})) = \tau^2$, $\mathrm{cov}(\epsilon(\mathbf{s}), \epsilon(\mathbf{s}')) = 0$, and $w(\mathbf{s})$ is a second-order stationary mean-zero process independent of the white noise process, i.e., $\mathrm{E}(w(\mathbf{s})) = 0$, $\mathrm{Var}(w(\mathbf{s})) = \sigma^2$, $\mathrm{cov}(w(\mathbf{s}), w(\mathbf{s}')) = \sigma^2\rho(\mathbf{s}, \mathbf{s}'; \phi)$, where ρ is a valid two-dimensional correlation function.

Letting $\mu(\mathbf{s}) = \beta_0 + \beta_1 x(\mathbf{s})$, write $w(\mathbf{s}) = \beta_0(\mathbf{s})$ and define $\tilde{\beta}_0(\mathbf{s}) = \beta_0 + \beta_0(\mathbf{s})$. Then $\beta_0(\mathbf{s})$ can be interpreted as a random spatial adjustment at location $\mathbf{s}$ to the overall intercept β_0. Equivalently, $\tilde{\beta}_0(\mathbf{s})$ can be viewed as a random intercept process. For an observed set of locations $\mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_n$ given $\beta_0, \beta_1, \{\beta_0(\mathbf{s}_i)\}$ and τ^2, the $Y(\mathbf{s}_i) = \beta_0 + \beta_1 x(\mathbf{s}_i) + \beta_0(\mathbf{s}_i) + \epsilon(\mathbf{s}_i), i = 1, \ldots, n$, are conditionally independent. Then $L(\beta_0, \beta_1, \{\beta_0(\mathbf{s}_i)\}, \tau^2; \mathbf{y})$, the first-stage likelihood, is

$$(\tau^2)^{-\frac{n}{2}} \exp\left\{-\frac{1}{2\tau^2}\sum(Y(\mathbf{s}_i) - (\beta_0 + \beta_1 x(\mathbf{s}_i) + \beta_0(\mathbf{s}_i))^2\right\} \ . \tag{9.49}$$

In obvious notation, the distribution of $\mathbf{B}_0 = (\beta_0(\mathbf{s}_1), \ldots, \beta_0(\mathbf{s}_n))^T$ is

$$f(\mathbf{B}_0 \mid {\sigma_0}^2, \phi_0) = N(\mathbf{0}, {\sigma_0}^2 H_0(\phi_0)) \ , \tag{9.50}$$

where $(H_0(\phi_0))_{ij} = \rho_0(\mathbf{s}_i - \mathbf{s}_j; \phi_0)$. For all of the discussion and examples below, we adopt the Matérn correlation function, (2.8). With a prior on $\beta_0, \beta_1, \tau^2, \sigma_0^2$, and ϕ_0, specification of the Bayesian hierarchical model is completed. Under (9.49) and (9.50), we can integrate over $\mathbf{B}_0$, obtaining $L(\beta_0, \beta_1, \tau^2, {\sigma_0}^2, \phi_0; \mathbf{y})$, the marginal likelihood, as

$$|\sigma_0^2 H_0(\phi_0) + \tau^2 I|^{-\frac{1}{2}} \times \exp\left\{-\frac{1}{2}Q\right\} \ , \tag{9.51}$$

where $Q = (\mathbf{y} - \beta_0\mathbf{1} - \beta_1\mathbf{x})^T \left(\sigma_0^2 H_0(\phi_0) + \tau^2 I\right)^{-1} (\mathbf{y} - \beta_0\mathbf{1} - \beta_1\mathbf{x})$ and $\mathbf{x} = (x(\mathbf{s}_1), \ldots, x(\mathbf{s}_n))^T$.

Following Gelfand, Kim, Sirmans and Banerjee (2003), the foregoing development immediately suggests how to formulate a spatially varying coefficient model. Suppose we write

$$Y(\mathbf{s}) = \beta_0 + \beta_1 x(\mathbf{s}) + \beta_1(\mathbf{s})x(\mathbf{s}) + \epsilon(\mathbf{s}) \ . \tag{9.52}$$

In (9.52), $\beta_1(\mathbf{s})$ is a second-order stationary mean-zero Gaussian process with variance σ_1^2 and correlation function $\rho_1(\cdot; \phi_1)$. Also, let $\tilde{\beta}_1(\mathbf{s}) = \beta_1 + \beta_1(\mathbf{s})$. Now $\beta_1(\mathbf{s})$ can be interpreted as a random spatial adjustment at location $\mathbf{s}$ to the overall slope β_1. Equivalently, $\tilde{\beta}_1(\mathbf{s})$ can be viewed as a random slope process. In effect, we are employing an uncountable dimensional function to explain the relationship between $x(\mathbf{s})$ and $Y(\mathbf{s})$. This model might be characterized as *locally linear*; however, it is difficult to imagine a more flexible specification for the relationship between $x(\mathbf{s})$ and $Y(\mathbf{s})$.

Expression (9.52) yields obvious modification of (9.49) and (9.50). In particular, the resulting marginalized likelihood becomes

$$L(\beta_0, \beta_1, \tau^2, {\sigma_1}^2, \phi_1; \mathbf{y}) = |{\sigma_1}^2 D_x H_1(\phi_1) D_x + \tau^2 I|^{-\frac{1}{2}} \times \exp\left\{-\frac{1}{2}Q\right\} \ , \tag{9.53}$$

where $Q = (\mathbf{y} - \beta_0\mathbf{1} - \beta_1\mathbf{x})^T(\sigma_1^2 D_x H_1(\phi_1) D_x + \tau^2 I)^{-1}(\mathbf{y} - \beta_0\mathbf{1} - \beta_1\mathbf{x})$ and D_x is diagonal with $(D_x)_{ii} = x(\mathbf{s}_i)$. With $\mathbf{B}_1 = (\beta_1(\mathbf{s}_1), \ldots, \beta_1(\mathbf{s}_n))^T$ we can sample $f(\mathbf{B}_1|\mathbf{y})$ and $f(\beta_1(\mathbf{s}_{new})|\mathbf{y})$ via composition.

Note that (9.52) provides a heterogeneous, nonstationary process for the data regardless of the choice of covariance function for the $\beta_1(\mathbf{s})$ process, since $\mathrm{Var}(Y(\mathbf{s}) \mid \beta_0, \beta_1, \tau^2, {\sigma_1}^2, \phi_1) = x^2(\mathbf{s}){\sigma_1}^2 + \tau^2$ and $\mathrm{cov}(Y(\mathbf{s}), Y(\mathbf{s}') \mid \beta_0, \beta_1, \tau^2, {\sigma_1}^2, \phi_1) = {\sigma_1}^2 x(\mathbf{s})x(\mathbf{s}')\rho_1(\mathbf{s}-\mathbf{s}'; \phi_1)$. As a result, we observe that in practice, (9.52) is sensible only if we have $x(\mathbf{s}) > 0$. In fact, centering and scaling, which is usually advocated for better behaved model fitting, is inappropriate here. With centered $x(\mathbf{s})$'s we would find the likely untenable behavior that $\mathrm{Var}(Y(\mathbf{s}))$ decreases and then increases in $x(\mathbf{s})$. Worse, for an essentially central $x(\mathbf{s})$ we would find $Y(\mathbf{s})$ essentially independent of $Y(\mathbf{s}')$ for any $\mathbf{s}'$. Also, scaling the $x(\mathbf{s})$'s accomplishes nothing. $\beta_1(\mathbf{s})$ would be inversely rescaled since the model only identifies $\beta_1(\mathbf{s})x(\mathbf{s})$.

This leads to concerns regarding possible approximate collinearity of $\mathbf{x}$, the vector of $x(\mathbf{s}_i)$'s, with the vector $\mathbf{1}$. Expression (9.53) shows that a badly behaved likelihood will arise if $\mathbf{x} \approx c\mathbf{1}$. But, we can reparametrize (9.52) to $Y(\mathbf{s}) = {\beta_0}' + {\beta_1}'\tilde{x}(\mathbf{s}) + \beta_1(\mathbf{s})x(\mathbf{s}) + \epsilon(\mathbf{s})$ where $\tilde{x}(\mathbf{s})$ is centered and scaled with obvious definitions for ${\beta_0}'$ and ${\beta_1}'$. Now $\tilde{\beta}_1(\mathbf{s}) = \beta_1'/s_x + \beta_1(\mathbf{s})$ where s_x is the sample standard deviation of the $x(\mathbf{s})$'s.

As below (9.51), we can draw an analogy with standard longitudinal linear growth curve modeling, where $Y_{ij} = \beta_0 + \beta_1 x_{ij} + \beta_{1i} x_{ij} + \epsilon_{ij}$, i.e., a random slope for each individual. For growth curve models, we consider both a population level growth curve (marginalizing over the random effects) as well as these individual level growth curves. In this regard, here, $Y(\mathbf{s}) = \beta_0 + \beta_1 x(\mathbf{s}) + \epsilon(\mathbf{s})$ provides the global growth curve while (9.54) below provides the local growth curves.

Specifically, the general specification incorporating both $\beta_0(\mathbf{s})$ and $\beta_1(\mathbf{s})$ would be

$$Y(\mathbf{s}) = \beta_0 + \beta_1 x(\mathbf{s}) + \beta_0(\mathbf{s}) + \beta_1(\mathbf{s})x(\mathbf{s}) + \epsilon(\mathbf{s}). \tag{9.54}$$

Expression (9.54) parallels the usual linear growth curve modeling by introducing both an intercept process and a slope process. The model in (9.54) requires a bivariate process specification in order to determine the joint distribution of $\mathbf{B}_0$ and $\mathbf{B}_1$. It also partitions the total error into intercept process error, slope error, and pure error. A noteworthy remark is that we can fit the bivariate spatial process model in (9.54) without ever observing it. That is, we only observe the $Y(\mathbf{s})$ process. This demonstrates the power of hierarchical modeling with structured dependence.

9.6.2 Multivariate spatially varying coefficient models

For the case of a $p \times 1$ multivariate covariate vector $\mathbf{X}(\mathbf{s})$ at location $\mathbf{s}$ where, for convenience, $\mathbf{X}(\mathbf{s})$ includes a 1 as its first entry to accommodate an intercept, we generalize (9.54) to

$$Y(\mathbf{s}) = \mathbf{X}^T(\mathbf{s})\tilde{\boldsymbol{\beta}}(\mathbf{s}) + \epsilon(\mathbf{s}) \ , \tag{9.55}$$

where $\tilde{\boldsymbol{\beta}}(\mathbf{s})$ is assumed to follow a p-variate spatial process model. With observed locations $\mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_n$, let X be $n \times np$ block diagonal having as block for the ith row $\mathbf{X}^T(\mathbf{s}_i)$. Then we can write $\mathbf{Y} = X^T\tilde{\mathbf{B}} + \boldsymbol{\epsilon}$ where $\tilde{\mathbf{B}}$ is $np \times 1$, the concatenated vector of the $\tilde{\boldsymbol{\beta}}(s)$, and $\epsilon \sim N(0, \tau^2 I)$.

As above, in practice, to assume that the component processes of $\tilde{\boldsymbol{\beta}}(\mathbf{s})$ are independent is likely inappropriate. The dramatic improvement in model performance when dependence is incorporated is shown in Example 9.4. To formulate a multivariate Gaussian process for $\tilde{\boldsymbol{\beta}}(\mathbf{s})$ we require the mean and the cross-covariance function. For the former, following Subsection 9.6.1, we take this to be $\boldsymbol{\mu}_\beta = (\beta_1, \ldots, \beta_p)^T$. For the latter we require a valid p-variate choice. In the following paragraphs we work with a separable form (Section 9.3), yielding

$$\tilde{\mathbf{B}} \sim N(\mathbf{1}_{n\times 1} \otimes \boldsymbol{\mu}_\beta \ , \ H(\phi) \otimes T) \ . \tag{9.56}$$

If if $\tilde{\mathbf{B}} = \mathbf{B} + \mathbf{1}_{n\times 1} \otimes \boldsymbol{\mu}_\beta$, then we can write (9.55) as

$$Y(\mathbf{s}) = \mathbf{X}^T(\mathbf{s})\boldsymbol{\mu}_\beta + \mathbf{X}^T(\mathbf{s})\boldsymbol{\beta}(\mathbf{s}) + \epsilon(\mathbf{s}) \ . \tag{9.57}$$

In (9.57) the total error in the regression model is partitioned into $p+1$ pieces, each with an obvious interpretation. Following Subsection 9.6.1, using (9.55) and (9.56) we can integrate over $\boldsymbol{\beta}$ to obtain

$$L(\boldsymbol{\mu}_\beta, \tau^2, T, \phi; \mathbf{y}) = |X(H(\phi) \otimes T)X^T + \tau^2 I|^{-\frac{1}{2}} \times \exp\{-\tfrac{1}{2}Q\} \ , \tag{9.58}$$

where $Q = (\mathbf{y} - X(\mathbf{1} \otimes \boldsymbol{\mu}_\beta))^T (X(H(\phi) \otimes T)X^T + \tau^2 I)^{-1}(\mathbf{y} - X(\mathbf{1} \otimes \boldsymbol{\mu}_\beta))$. This apparently daunting form still involves only $n \times n$ matrices.

The Bayesian model is completed with a prior $p\left(\boldsymbol{\mu}_\beta, \tau^2, T, \phi\right)$, which we assume to take the product form $p(\boldsymbol{\mu}_\beta)p(\tau^2)p(T)p(\phi)$. Below, these components will be normal, inverse gamma, inverse Wishart, and gamma, respectively.

With regard to prediction, $p(\tilde{\mathbf{B}}|\mathbf{y})$ can be sampled one for one with the posterior samples from $f\left(\boldsymbol{\mu}_\beta, \tau^2, T, \phi|\mathbf{y}\right)$ using $f\left(\tilde{\mathbf{B}}|\boldsymbol{\mu}_\beta, \tau^2, T, \phi, y\right)$, which is $N\left(A\mathbf{a}, A\right)$ where $A = (X^T X/\tau^2 + H^{-1}(\phi) \otimes T^{-1})^{-1}$ and $\mathbf{a} = X^T\mathbf{y}/\tau^2 + \left(H^{-1}(\phi) \otimes T^{-1}\right)\left(\mathbf{1} \otimes \boldsymbol{\mu}_\beta\right)$. Here A is $np \times np$ but, for sampling $\tilde{\boldsymbol{\beta}}$, only a Cholesky decomposition of A is needed, and only for the retained posterior samples. Prediction at a new location, say, $\mathbf{s}_{new}$, requires samples from $f\left(\tilde{\boldsymbol{\beta}}(\mathbf{s}_{new})|\tilde{\mathbf{B}}, \boldsymbol{\mu}_\beta, \tau^2, T, \phi\right)$. Defining $\mathbf{h}_{new}(\phi)$ to be the $n \times 1$ vector with ith row entry $\rho(\mathbf{s}_i - \mathbf{s}_{new}; \phi)$, this distribution is normal with mean

$$\begin{aligned} &\boldsymbol{\mu}_\beta + \left(\mathbf{h}_{new}^T(\phi) \otimes T\right)\left(H^{-1}(\phi) \otimes T^{-1}\right)\left(\tilde{\mathbf{B}} - \mathbf{1}_{nx1} \otimes \boldsymbol{\mu}_\beta\right) \\ &\quad = \boldsymbol{\mu}_\beta + \left(\mathbf{h}_{new}^T(\phi) H^{-1}(\phi) \otimes I\right)\left(\tilde{\mathbf{B}} - \mathbf{1}_{nx1} \otimes \boldsymbol{\mu}_\beta\right) \ , \end{aligned}$$

and covariance matrix

$$T - \left(\mathbf{h}_{new}^T(\phi) \otimes T\right)\left(H^{-1}(\phi) \otimes T^{-1}\right)\left(\mathbf{h}_{new}(\phi) \otimes T\right) = \left(I - \mathbf{h}_{new}^T(\phi) H^{-1}(\phi)\mathbf{h}_{new}(\phi)\right)T \ .$$

Finally, the predictive distribution for $Y(\mathbf{s}_{new})$, namely $f\left(Y(\mathbf{s}_{new}) \mid \mathbf{y}\right)$, is sampled by composition, as usual.

We conclude this subsection by noting an extension of (9.55) when we have repeated measurements at location s. That is, suppose we have

$$Y(\mathbf{s}, l) = \mathbf{X}^T(\mathbf{s}, l)\boldsymbol{\beta}(\mathbf{s}) + \epsilon(\mathbf{s}, l) \ , \tag{9.59}$$

where $l = 1, \ldots, L_{\mathbf{s}}$ with $L_{\mathbf{s}}$ the number of measurements at $\mathbf{s}$ and the $\epsilon(\mathbf{s}, l)$ still white noise. As an illustration, in the real estate context, $\mathbf{s}$ might denote the location for an apartment block and l might index apartments in this block that have sold, with the lth apartment having characteristics $\mathbf{X}(\mathbf{s}, l)$. Suppose further that $\mathbf{Z}(\mathbf{s})$ denotes an $r \times 1$ vector of site-level characteristics. For an apartment block, these characteristics might include amenities provided or distance to the central business district. Then (9.59) can be extended to a multilevel model in the sense of Goldstein (1995) or Raudenbush and Bryk (2002). In particular we can write

$$\boldsymbol{\beta}(s) = \begin{pmatrix} \mathbf{Z}^T(\mathbf{s})\boldsymbol{\gamma}_1 \\ \vdots \\ \mathbf{Z}^T(\mathbf{s})\boldsymbol{\gamma}_p \end{pmatrix} + \mathbf{w}(\mathbf{s}) \ . \tag{9.60}$$

In (9.60), $\boldsymbol{\gamma}_j, j = 1, \ldots, p$, is an $r \times 1$ vector associated with $\tilde{\beta}_j(\mathbf{s})$, and $\mathbf{w}(\mathbf{s})$ is a mean-zero multivariate Gaussian spatial process, for example, as above. In (9.60), if the $\mathbf{w}(\mathbf{s})$ were independent we would have a usual multilevel model specification. In the case where $\mathbf{Z}(\mathbf{s})$ is a scalar capturing just an intercept, we return to the initial model of this subsection.

9.6.3 Spatially varying coregionalization models

A possible extension of the LMC would replace A by $A(\mathbf{s})$ and thus define

$$\mathbf{Y}(\mathbf{s}) = A(\mathbf{s})\mathbf{w}(\mathbf{s}) \ . \tag{9.61}$$

We refer to the model in (9.61) as a *spatially varying LMC*. Following the notation in Section 9.5.1, let $\mathbf{T}(\mathbf{s}) = A(\mathbf{s})A(\mathbf{s})^T$. Again $A(\mathbf{s})$ can be taken to be lower triangular for convenience. Now $C(\mathbf{s}, \mathbf{s}')$ is such that

$$C(\mathbf{s}, \mathbf{s}') = \sum \rho_j(\mathbf{s} - \mathbf{s}')\mathbf{a}_j(\mathbf{s})\mathbf{a}_j(\mathbf{s}') \ , \tag{9.62}$$

with $\mathbf{a}_j(\mathbf{s})$ the jth column of $A(\mathbf{s})$. Letting $T_j(\mathbf{s}) = \mathbf{a}_j(\mathbf{s})\mathbf{a}_j^T(\mathbf{s})$, again, $\sum T_j(\mathbf{s}) = T(\mathbf{s})$. We see from (9.62) that $\mathbf{Y}(\mathbf{s})$ is no longer stationary. Extending the intrinsic specification for $\mathbf{Y}(\mathbf{s})$, $C(\mathbf{s}, \mathbf{s}') = \rho(\mathbf{s}-\mathbf{s}')\mathbf{T}(\mathbf{s})$, which is a multivariate version of the case of a spatial process with a spatially varying variance.

This motivates a natural definition of $A(\mathbf{s})$ through its one-to-one correspondence with $T(\mathbf{s})$ (again from Section 9.5.1) since $T(\mathbf{s})$ is the covariance matrix for $\mathbf{Y}(\mathbf{s})$. In the univariate case choices for $\sigma^2(\mathbf{s})$ include $\sigma^2(\mathbf{s}, \theta)$, i.e., a parametric function of location; $\sigma^2(x(\mathbf{s})) = g(x(\mathbf{s}))\sigma^2$ where $x(\mathbf{s})$ is some covariate used to explain $\mathbf{Y}(\mathbf{s})$ and $g(\cdot) > 0$ (then $g(x(\mathbf{s}))$ is typically $x(\mathbf{s})$ or $x^2(\mathbf{s})$); or $\sigma^2(\mathbf{s})$ is itself a spatial process (e.g., $\log \sigma^2(\mathbf{s})$ might be a Gaussian process). In practice, $T(\mathbf{s}) = g(x(\mathbf{s}))T$ will likely be easiest to work with.

Note that all of the discussion in Section 9.5.2 regarding the relationship between conditional and unconditional specifications is applicable here. Particularly, if $p = 2$ and $T(\mathbf{s}) = g(x(\mathbf{s}))T$ then (T_{11}, T_{12}, T_{22}) is equivalent to $(\sigma_1, \sigma_2, \alpha)$, and we have $a_{11}(\mathbf{s}) = \sqrt{g(x(\mathbf{s})}\sigma_1$, $a_{22}(\mathbf{s}) = \sqrt{g(x(\mathbf{s})}\sigma_2$, and $a_{21} = \sqrt{g(x(\mathbf{s})}\alpha\sigma_1$. See Gelfand et al. (2004) for further discussion.

9.6.4 Model-fitting issues

This subsection starts by discussing the computational issues in fitting the joint multivariate model presented in Subsection 9.5.1. It will be shown that it is a challenging task to fit this joint model. On the other hand, making use of the equivalence of the joint and conditional models, as discussed in Section 9.5.2, we demonstrate that it is much simpler to fit the latter.

9.6.4.1 Fitting the joint model

Different from previous approaches that have employed the coregionalization model, our intent is to follow the Bayesian paradigm. For this purpose, the model specification is complete only after assigning prior distributions to all unknown quantities in the model. The posterior distribution of the set of parameters is obtained after combining the information about them in the likelihood (see Equation (9.44)) with their prior distributions.

Observing Equation (9.44), we see that the parameter vector defined as $\boldsymbol{\theta}$ consists of $\{\boldsymbol{\beta}_j\}, D, \{\rho_j\}, \mathbf{T}$, $j = 1, \cdots, p$. Adopting a prior that assumes independence across j we take $p(\boldsymbol{\theta}) = \prod_j p(\boldsymbol{\beta}_j)\, p(\rho_j)\, p(\tau_j^2)\, p(T)$. Hence $p(\boldsymbol{\theta}|\mathbf{y})$ is given by

$$p(\boldsymbol{\theta}|\mathbf{y}) \propto p(\mathbf{y} \mid \{\boldsymbol{\beta}_j\}, D, \{\rho_j\}, T)\, p(\boldsymbol{\theta}) \ .$$

For the elements of $\boldsymbol{\beta}_j$, a normal mean-zero prior distribution with large variance can be assigned, resulting in a full conditional distribution that will also be normal. Inverse gamma distributions can be assigned to the elements of D, the variances of the p white noise processes. If there is no information about such variances, the means of these inverse gammas could be based on the least squares estimates of the independent models with

large variances. Assigning inverse gamma distributions to τ_j^2 will result in inverse gamma full conditionals. The parameters of concern are the elements of ρ_j and T. Regardless of what prior distributions we assign, the full conditional distributions will not have a standard form. For example, if we assume that ρ_j is the exponential correlation function, $\rho_j(h) = \exp(-\phi_j h)$, a gamma prior distribution can be assigned to the ϕ_j's. In order to obtain samples of the ϕ_j's we can use the Metropolis-Hastings algorithm with, for instance, log-normal proposals centered at the current $\log \phi_j$.

We now consider how to sample T, the covariance matrix among the responses at each location $\mathbf{s}$. Due to the one-to-one relationship between T and the lower triangular A, one can either assign a prior to the elements of A, or set a prior on the matrix T. The latter seems to be more natural, since T is interpreted as the covariance matrix of the elements of $\mathbf{Y}(\mathbf{s})$. As T must be positive definite, we use an inverse Wishart prior distribution with ν degrees of freedom and mean D^*, i.e., the scale matrix is $(\nu - p - 1)(D^*)^{-1}$. If there is no information about the prior mean structure of T, rough estimates of the elements of the diagonal of D^* can be obtained using ordinary least squares estimates based on the independent spatial models for each $Y_j(\mathbf{s})$, $j = 1, \cdots, p$. A small value of $\nu (> p + 1)$ would be assigned to provide high uncertainty in the resulting prior distribution.

To sample from the full conditional for T, Metropolis-Hastings updates are a place to start. In our experience, random walk Wishart proposals do not work well, and importance sampled Wishart proposals have also proven problematic. Instead, we recommend updating the elements of T individually. In fact, it is easier to work in the unconstrained space of the components of A, so we would reparametrize the full conditional from T to A. Random walk normal proposals for the $\mathbf{a}$'s with suitably tuned variances will mix well, at least for $p = 2$ or 3. For larger p, repeated decomposition of T to A may prove too costly.

9.6.4.2 Fitting the conditional model

Section 9.5.2 showed the equivalence of conditional and unconditional specifications in terms of $\mathbf{v}(\mathbf{s})$. Here we write the multivariate model for $\mathbf{Y}(\mathbf{s})$ in its conditional parametrization and see that the inference procedure is simpler than for the multivariate parametrization. Following the discussion in Section 9.5.2, for a general p, the conditional parametrization is

$$\begin{aligned} Y_1(\mathbf{s}) &= \mathbf{X}_1^T(\mathbf{s})\boldsymbol{\beta}_1 + \sigma_1 w_1(\mathbf{s}) \\ Y_2(\mathbf{s}) \mid Y_1(\mathbf{s}) &= \mathbf{X}_2^T(\mathbf{s})\boldsymbol{\beta}_2 + \alpha^{2|1} Y_1(\mathbf{s}) + \sigma_2 w_2(\mathbf{s}) \\ &\vdots \\ Y_p(\mathbf{s}) \mid Y_1(\mathbf{s}), \cdots, Y_p(\mathbf{s}) &= \mathbf{X}_p^T(\mathbf{s})\boldsymbol{\beta}_p + \alpha^{p|1} Y_1(\mathbf{s}) \\ &\qquad + \cdots + \alpha^{p|p-1} Y_{p-1}(\mathbf{s}) + \sigma_p w_p(\mathbf{s}) \,. \end{aligned} \tag{9.63}$$

In (9.63), the set of parameters to be estimated is $\boldsymbol{\theta}_c = \{\boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{\sigma}^2, \boldsymbol{\phi}\}$, where $\boldsymbol{\alpha}^T = (\alpha^{2|1}, \alpha^{3|1}, \alpha^{3|2}, \cdots, \alpha^{p|p-1})$, $\boldsymbol{\beta}^T = (\boldsymbol{\beta}_1, \cdots, \boldsymbol{\beta}_p)$, $\boldsymbol{\sigma}^2 = (\sigma_1^2, \cdots, \sigma_p^2)$, and $\boldsymbol{\phi}$ is as defined in Subsection 9.5.2. The likelihood is given by

$$f_c\left(\mathbf{Y}|\boldsymbol{\theta}_c\right) = f(\mathbf{Y}_1|\boldsymbol{\theta}_{c_1})\, f(\mathbf{Y}_2|\mathbf{Y}_1, \boldsymbol{\theta}_{c_2}) \cdots f(\mathbf{Y}_p|\mathbf{Y}_1, \cdots, \mathbf{Y}_{p-1}, \boldsymbol{\theta}_{c_p}) \,.$$

If $\pi(\boldsymbol{\theta}_c)$ is taken to be $\prod_{j=1}^p \pi(\boldsymbol{\theta}_{c_j})$ then this equation implies that the conditioning yields a factorization into p models each of which can be fitted separately. Prior specification of the parameters was discussed in Subsection 9.5.2.2. With those forms, standard univariate spatial models that can be fit using the `GeoBUGS` package arise.

Example 9.3 *(Commercial real estate example).* The selling price of commercial real estate, for example an apartment property, is theoretically the expected income capitalized at some (risk-adjusted) discount rate. (See Kinnard, 1971, and Lusht, 1997, for general discussions of the basics of commercial property valuation theory and practice.) Here we consider

Figure 9.6 *Locations of the 78 sites* (x) *used to fit the (price, income) model, and the 20 sites used for prediction* (p).

a data set consisting of 78 apartment buildings, with 20 additional transactions held out for prediction of the selling price based on four different models. The locations of these buildings are shown in Figure 9.6. The aim here is to fit a joint model for selling price and net income and obtain a spatial surface associated with the risk, which, for any transaction, is given by net income/price. For this purpose we fit a model using the following covariates: average square feet of a unit within the building (sqft), the age of the building (age), the number of units within the building (unit), the selling price of the transaction (P), and the net income (I). Figure 9.7 shows the histograms of these variables on the log scale. Using the conditional parametrization, the model is

$$\begin{aligned} I(\mathbf{s}) &= sqft(\mathbf{s})\beta_{I1} + age(\mathbf{s})\beta_{I2} + unit(\mathbf{s})\beta_{I3} + \sigma_1 w_1(\mathbf{s}) \\ P(\mathbf{s})|I(\mathbf{s}) &= sqft(\mathbf{s})\beta_{P1} + age(\mathbf{s})\beta_{P2} + unit(\mathbf{s})\beta_{P3} \\ &\qquad + I(\mathbf{s})\alpha^{(2|1)} + \sigma_2 w_2(\mathbf{s}) + \epsilon(\mathbf{s}) \ . \end{aligned} \tag{9.64}$$

Notice that $I(\mathbf{s})$ is considered to be purely spatial since, adjusted for building characteristics, we do not anticipate a microscale variability component. The need for white noise in the price component results from the fact that two identical properties at essentially the same location need not sell for the same price due to the motivation of the seller, the buyer, the brokerage process, etc. (If a white noise component for $I(\mathbf{s})$ were desired, we would fit the joint model as described near the beginning of Subsection 9.6.4.) The model in (9.64) is in accordance with the conditional parametrization in Subsection 9.5.2.2. The prior distributions were assigned as follows. For all the coefficients of the covariates, including $\alpha^{(2|1)}$, we assigned a normal 0 mean distribution with large variance. For σ_1^2 and σ_2^2 we used inverse gammas with infinite variance. We use exponential correlation functions and the decay parameters ϕ_j, $j = 1, 2$ have a gamma prior distribution arising from a mean range of one half the maximum interlocation distance, with infinite variance. Finally, τ_2^2, the variance of $\epsilon(\cdot)$, has an inverse gamma prior centered at the ordinary least squares variance estimate obtained from an independent model for log selling price given log net income.

Figure 9.7 *Histograms of the logarithm of the variables.*

Table 9.4 presents the posterior summaries of the parameters of the model. For the income model the age coefficient is significantly negative, the coefficient for number of units is significantly positive. Notice further that the correlation between net income and price is very close to 1. Nevertheless, for the conditional price model age is still significant. Also we see that price shows a bigger range than net income. Figure 9.8 shows the spatial surfaces associated with the three processes: net income, price, and risk. It is straightforward to show that the logarithm of the spatial surface for risk is obtained through $(1-\alpha^{(2|1)})\sigma_1 w_1(\mathbf{s}) - \sigma_2 w_2(\mathbf{s})$. Therefore, based on the posterior samples of $\alpha^{(2|1)}$, $w_1(\mathbf{s})$, σ_1, σ_2, and $w_2(\mathbf{s})$ we are able to obtain samples for the spatial surface for risk. From Figure 9.8(c), we note that the spatial risk surface tends to have smaller values than the other surfaces. Since $\log R(\mathbf{s}) = \log I(\mathbf{s}) - \log P(\mathbf{s})$ with $R(\mathbf{s})$ denoting the risk at location $\mathbf{s}$, the strong association between $I(\mathbf{s})$ and $P(\mathbf{s})$ appears to result in some cancellation of spatial effect for log risk. Actually, we can obtain the posterior distribution of the variance of the spatial process for $\log R(\mathbf{s})$. It is $(1-\alpha^{(2|1)})^2\sigma_1^2 + \sigma_2^2$. The posterior mean of this variance is 0.036 and the 95% credible interval is given by $(0.0087, 0.1076)$ with median equal 0.028. The posterior variance of the noise term is given by τ_2^2, which is in Table 9.4. If we compare the medians of the posteriors of the variance of the spatial process of the risk and the variance of the white noise, we see that the spatial process presents a smaller variance; the variability of the risk process is being more explained by the residual component.

In order to examine the comparative performance of the model proposed above we decided to run four different models for the selling price using each one to predict at the locations marked with p in Figure 9.6. For all these models we used the same covariates as described before. Model 1 comprises an independent model for price, i.e., without a spatial component or net income. Model 2 has a spatial component and is not conditioned on net income. In Model 3 the selling price is conditioned on the net income but without a spatial component, and Model 4 has net income as a covariate and also a spatial component.

Parameter	Mean	2.50%	Median	97.50%
β_{I1}	0.156	-0.071	0.156	0.385
β_{I2}	-0.088	-0.169	-0.088	-0.008
β_{I3}	0.806	0.589	0.804	1.014
β_{P1}	0.225	0.010	0.229	0.439
β_{P2}	-0.092	-0.154	-0.091	-0.026
β_{P3}	-0.150	-0.389	-0.150	0.093
$\alpha^{(2\|1)}$	0.858	0.648	0.856	1.064
σ_1^2	0.508	0.190	0.431	1.363
σ_2^2	0.017	0.006	0.014	0.045
τ_2^2	0.051	0.036	0.051	0.071
ϕ_I	3.762	1.269	3.510	7.497
ϕ_P	1.207	0.161	1.072	3.201
range_I	0.969	0.429	0.834	2.291
range_P	1.2383	0.554	1.064	2.937
$\text{corr}(I, P)$	0.971	0.912	0.979	0.995
T_{II}	0.508	0.190	0.431	1.363
T_{IP}	0.435	0.158	0.369	1.136
T_{PP}	0.396	0.137	0.340	1.000

Table 9.4 *Posterior summaries, joint model of price and income.*

Figure 9.8 *Image plots of the spatial processes of (a) net income, (b) price, and (c) risk.*

Table 9.5 shows both $\sum_{j=1}^{20} e_j^2$, where $e_j = P(\mathbf{s}_j) - E(P(\mathbf{s}_j)|\mathbf{y}, \text{model})$ and $P(\mathbf{s}_j)$ is the observed log selling price for the jth transaction, and $\sum_{j=1}^{20} Var\,(P(\mathbf{s}_j)|\mathbf{y}, \text{model})$. Recall from Equation (5.14) that the former is a measurement of predictive goodness of fit, while the latter is a measure of predictive variability. It is clear from the table that the model conditioned on net income and with a spatial component is best, both in terms of fit and predictive variability.

Model	$\sum_{j=1}^{20} e_j^2$	$\sum_{j=1}^{20} Var(P(\mathbf{s}_j)\|\mathbf{y})$
Independent, nonspatial	2.279	3.277
Independent, spatial	1.808	2.963
Conditional, nonspatial	0.932	1.772
Conditional, spatial	0.772	1.731

Table 9.5 *Squared error and sum of the variances of the predictions for the 20 sites left out in the fitting of the model.*

9.7 Other constructive approaches ⋆

Here we consider two additional constructive strategies for building valid cross-covariance functions. The first is referred to as a moving average approach in Ver Hoef and Barry (1998). It is a multivariate version of the kernel convolution development of Subsection 3.2.2 that convolves process variables to produce a new process. The second approach convolves valid covariance functions to produce a valid cross-covariance function.

For the first approach, expressions (3.9) and (3.10) suggest several ways to achieve multivariate extension. Again with $\mathbf{Y}(\mathbf{s}) = (Y_1(\mathbf{s}), \ldots, Y_p(\mathbf{s}))^T$, define

$$Y_\ell(\mathbf{s}) = \int_{\Re^2} k_\ell(\mathbf{u}) Z(\mathbf{s}+\mathbf{u}) d\mathbf{u}\,, \quad \ell = 1, \ldots, p. \tag{9.65}$$

In this expression, $Z(\cdot)$ is a mean 0 stationary process with correlation function $\rho(\cdot)$, and k_ℓ is a kernel associated with the ℓth component of $\mathbf{Y}(\mathbf{s})$. In practice, $k_\ell(\mathbf{u})$ would be parametric, i.e., $k_\ell(\mathbf{u}; \boldsymbol{\theta}_\ell)$. The resulting cross-covariance matrix for $\mathbf{Y}(\mathbf{s})$ has entries

$$C_{\ell,\ell'}(\mathbf{s},\mathbf{s}') = \sigma^2 \int_{\Re^2}\int_{\Re^2} k_\ell(\mathbf{s}-\mathbf{s}'+\mathbf{u}) k_{\ell'}(\mathbf{u}') \rho(\mathbf{u}-\mathbf{u}') d\mathbf{u} d\mathbf{u}'\,. \tag{9.66}$$

This cross-covariance matrix is necessarily valid. It is stationary and, as may be easily verified, is symmetric, i.e., $cov(Y_\ell(\mathbf{s}), Y_{\ell'}(\mathbf{s}')) = C_{\ell\ell'}(\mathbf{s}-\mathbf{s}') = C_{\ell'\ell}(\mathbf{s}-\mathbf{s}') = cov(Y_{\ell'}(\mathbf{s}), Y_\ell(\mathbf{s}'))$. Since the integration in (9.66) will not be possible to do explicitly except in certain special cases, finite sum approximation of (9.65), analogous to (3.15), is an alternative.

An alternative extension to (3.10) introduces *lags* $\mathbf{h}_\ell$, defining

$$Y_\ell(\mathbf{s}) = \int_{\Re^2} k(\mathbf{u}) Z(\mathbf{s}+\mathbf{h}_\ell+\mathbf{u}) d\mathbf{u}\,,\ \ell = 1, \ldots, p.$$

Now

$$C_{\ell,\ell'}(\mathbf{s},\mathbf{s}') = \sigma^2 \int_{\Re^2}\int_{\Re^2} k(\mathbf{s}-\mathbf{s}'+\mathbf{u}) k(\mathbf{u}') \rho(\mathbf{h}_\ell - \mathbf{h}_{\ell'} + \mathbf{u}-\mathbf{u}') d\mathbf{u} d\mathbf{u}'.$$

Again the resulting cross-covariance matrix is valid; again the process is stationary. However now it is easy to verify that the cross-covariance matrix is not symmetric. Whether a lagged relationship between the variables is appropriate in a purely spatial specification would depend upon the application. However, in practice the $\mathbf{h}_\ell$ would be unknown and would be considered as model parameters. A fully Bayesian treatment of such a model has not yet been discussed in the literature.

For the second approach, suppose $C_\ell(\mathbf{s})$, $\ell = 1, ..., p$ are each squared integrable stationary covariance functions valid in two-dimensional space. We now show $C_{\ell\ell}(\mathbf{s}) = \int_{R^2} C_\ell(\mathbf{s}-\mathbf{u}) C_\ell(\mathbf{u}) d\mathbf{u}$, the convolution of C_ℓ with itself, is again a valid covariance function. Writing $\widehat{C}_\ell(\mathbf{w}) = \int e^{-i\mathbf{w}^T\mathbf{h}} C_\ell(\mathbf{h}) d\mathbf{h}$, by inversion, $C_\ell(\mathbf{s}) = \int e^{i\mathbf{w}^T\mathbf{s}} \frac{\widehat{C}_\ell(\mathbf{w})}{(2\pi)^2} d\mathbf{w}$. But also, from (3.2), $\widehat{C}_{\ell\ell}(\mathbf{w}) \equiv \int e^{-i\mathbf{w}^T\mathbf{s}} C_{\ell\ell}(\mathbf{s}) d\mathbf{s} = \int e^{-i\mathbf{s}^T\mathbf{s}} \int C_\ell(\mathbf{s}-\mathbf{u}) C_\ell(\mathbf{u}) d\mathbf{u} d\mathbf{s} = \int\int e^{i\mathbf{w}^T(\mathbf{s}-\mathbf{u})} C_\ell(\mathbf{s} -$

$\mathbf{u})e^{i\mathbf{w}^T\mathbf{u}}C_\ell(\mathbf{u})d\mathbf{u}d\mathbf{s} = (\widehat{C}_\ell(\mathbf{w}))^2$. Self-convolution of C_ℓ produces the square of the Fourier transform. However, since $C_\ell(\cdot)$ is valid, Bochner's Theorem (Subsection 3.1.2) tells us that $\widehat{C}_\ell(\mathbf{w})/(2\pi)^2C(0)$ is a spectral density symmetric about 0. But then due to the squared integrability assumption, up to proportionality, so is $(\widehat{C}_\ell(w))^2$, and thus $C_{\ell\ell}(\cdot)$ is valid.

The same argument ensures that

$$C_{\ell\ell'}(\mathbf{s}) = \int_{R^2} C_\ell(\mathbf{s}-\mathbf{u})C_{\ell'}(\mathbf{u})d\mathbf{u} \tag{9.67}$$

is also a valid stationary covariance function; cross-convolution provides a valid covariance function. (Now $\widehat{C}_{\ell\ell'}(w) = \widehat{C}_\ell(w)\widehat{C}_{\ell'}(w)$.) Moreover, it can be shown that $C(\mathbf{s}-\mathbf{s}')$ defined by $(C(\mathbf{s}-\mathbf{s}'))_{\ell\ell'} = C_{\ell\ell'}(\mathbf{s}-\mathbf{s}')$ is a valid $p\times p$ cross-covariance function (see Majumdar and Gelfand, 2003). It is also the case that if each C_ℓ is isotropic, then so is $C(\mathbf{s}-\mathbf{s}')$. To see this, suppose $\|\mathbf{h}_1\| = \|\mathbf{h}_2\|$. We need only show that $C_{\ell\ell'}(\mathbf{h}_1) = C_{\ell\ell'}(\mathbf{h}_2)$. But $\mathbf{h}_1 = P\mathbf{h}_2$ where P is orthogonal. Hence, $C_{\ell\ell'}(\mathbf{h}_1) = \int C_\ell(\mathbf{h}_1-\mathbf{u})C_{\ell'}(\mathbf{u})d\mathbf{u} = \int C_\ell(P(\mathbf{h}_1-\mathbf{u}))C_{\ell'}(P\mathbf{u})d\mathbf{u} = \int C_\ell(\mathbf{h}_2-\widetilde{\mathbf{u}})C_{\ell'}(\widetilde{\mathbf{u}})d\widetilde{\mathbf{u}} = C_{\ell\ell'}(\mathbf{h}_2)$.

We note that the range associated with $C_{\ell\ell}$ is not the same as that for C_ℓ but that if the C_ℓ's have distinct ranges then so will the components, $Y_\ell(\mathbf{s})$. Computational issues associated with using $C(s-s')$ in model-fitting are also discussed in Majumdar and Gelfand (2003). We note that (9.67) can in most cases be conveniently computed by transformation to polar coordinates and then using Monte Carlo integration. We leave this calculation to an exercise.

As a concluding remark here, we note the recent work of Apanasovich and Genton (2010) (see also Apanasovich, Genton and Sun, 2012), which builds extremely flexible classes of cross-covariance functions, drawing upon the constructions of Gneiting (2002), as mentioned briefly in Section 11.3. The basic idea is to build valid forms using covariance functions involving latent dimensions. General expressions are complex, involving many parameters, suggesting potential identifiability problems in model fitting. The interested reader is encouraged to consult this work for more detail. Here, we offer a simple illustration. Recall that, under coregionalization, in the two-dimensional case, with exponential covariance functions, we create $C_{11}(\mathbf{h}) = a_{11}^2\mathtt{exp}(-\alpha_1||\mathbf{h}||)$, $C_{22}(\mathbf{h}) = a_{21}^2\mathtt{exp}(-\alpha_1||\mathbf{h}||)+a_{22}^2\exp(-\alpha_2||\mathbf{h}||)$, and $C_{12}(\mathbf{h}) = a_{11}a_{21}\exp(-\alpha_1||\mathbf{h}||)$. Apanasovich and Genton (2010) show that, if we retain $C_{11}(\mathbf{h})$ and $C_{22}(\mathbf{h})$ as they are but generalize $C_{12}(\mathbf{h})$ to $C_{12}(\mathbf{h}) = \frac{a_{11}a_{21}}{\delta_{12}+1}\mathtt{exp}(-\frac{\alpha_1||\mathbf{h}||}{(\delta_{12}+1)^{\beta/2}})$, we still obtain a valid cross-covariance function with added flexibility of the parameters $\delta_{12}\geq 0$ and $\beta\geq 0$.

Example 9.4 *(Baton Rouge housing prices).* We analyze a sample from a database of real estate transactions in Baton Rouge, LA, during the eight-year period 1985–1992. Here, we focus upon the static case. In particular, we focus on modeling the log selling price of single-family homes. In real estate modeling it is customary to work with log selling price in order to achieve better approximate normality. A range of house characteristics are available. We use four of the most common choices: age of house, square feet of living area, square feet of other areas (e.g., garages, carports, storage), and number of bathrooms. For the static spatial case, a sample of 237 transactions was drawn from 1992. Figure 9.9 shows the parish of Baton Rouge and the locations contained in an encompassing rectangle within the parish.

We fit a variety of models, where in all cases the correlation function is from the Matérn class. We used priors that are fairly noninformative and comparable across models as sensible. First, we started with a spatially varying intercept and one spatially varying slope coefficient (the remaining coefficients do not vary), requiring a bivariate process model. There are four such models, and using D_K, the Gelfand and Ghosh (1998) criterion (5.14), the model with a spatially varying living area coefficient emerges as best. Next, we introduced two spatially varying slope coefficient processes along with a spatially varying intercept, requiring a trivariate process model. There are six models here; the one with spatially varying

Figure 9.9 *Locations sampled within the parish of Baton Rouge for the static spatial models.*

Model	Fit	Variance penalty	D_K
Five-dimensional	42.21	36.01	78.22
Three-dimensional (best)	61.38	47.83	109.21
Two-dimensional (best)	69.87	46.24	116.11
Independent process	94.36	59.34	153.70

Table 9.6 *Values of posterior predictive model choice criterion (over all models).*

age and living area is best. Finally, we allowed five spatially varying processes: an intercept and all four coefficients, using a five-dimensional process model. We also fit a model with five independent processes. From Table 9.6 the five-dimensional dependent process model is far superior and the independent process model is a dismal last, supporting our earlier intuition.

The prior specification used for the five-dimensional dependent process model is as follows. We take vague $N\left(\mathbf{0}, 10^5 I\right)$ for $\boldsymbol{\mu}_\beta$, a five-dimensional inverse Wishart, $IW\left(5, Diag\left(0.001\right)\right)$, for T, and an inverse gamma $IG\left(2,1\right)$ for τ^2 (mean 1, infinite variance). For the Matérn correlation function parameters ϕ and ν we assume gamma priors $G\left(2, 0.1\right)$ (mean 20 and variance 200). For all the models three parallel chains were run to assess convergence. Satisfactory mixing was obtained within 3000 iterations for all the models; 2000 further samples were generated and retained for posterior inference.

The resulting posterior inference summary is provided in Table 9.7. We note a significant negative overall age coefficient with significant positive overall coefficients for the other three covariates, as expected. The contribution to spatial variability from the components of $\boldsymbol{\beta}$ is captured through the diagonal elements of the T matrix scaled by the corresponding

Parameter	2.5%	50%	97.5%
β_0 (intercept)	9.908	9.917	9.928
β_1 (age)	–0.008	–0.005	–0.002
β_2 (living area)	0.283	0.341	0.401
β_3 (other area)	0.133	0.313	0.497
β_4 (bathrooms)	0.183	0.292	0.401
T_{11}	0.167	0.322	0.514
$\overline{x}_1^2 T_{22}$	0.029	0.046	0.063
$\overline{x}_2^2 T_{33}$	0.013	0.028	0.047
$\overline{x}_3^2 T_{44}$	0.034	0.045	0.066
$\overline{x}_4^2 T_{55}$	0.151	0.183	0.232
$T_{12}/\sqrt{T_{11}T_{22}}$	–0.219	–0.203	–0.184
$T_{13}/\sqrt{T_{11}T_{33}}$	–0.205	–0.186	–0.167
$T_{14}/\sqrt{T_{11}T_{44}}$	0.213	0.234	0.257
$T_{15}/\sqrt{T_{11}T_{55}}$	–0.647	–0.583	–0.534
$T_{23}/\sqrt{T_{22}T_{33}}$	–0.008	0.011	0.030
$T_{24}/\sqrt{T_{22}T_{44}}$	0.061	0.077	0.098
$T_{25}/\sqrt{T_{22}T_{55}}$	–0.013	0.018	0.054
$T_{34}/\sqrt{T_{33}T_{44}}$	–0.885	–0.839	–0.789
$T_{35}/\sqrt{T_{33}T_{55}}$	–0.614	–0.560	–0.507
$T_{45}/\sqrt{T_{44}T_{55}}$	0.173	0.232	0.301
ϕ (decay)	0.51	1.14	2.32
ν (smoothness)	0.91	1.47	2.87
range (in km)	2.05	4.17	9.32
τ^2	0.033	0.049	0.077

Table 9.7 *Inference summary for the five-dimensional multivariate spatially varying coefficients model.*

covariates following the discussion at the end of Subsection 9.6.1. We see that the spatial intercept process contributes most to the error variability with, perhaps surprisingly, the "bathrooms" process second. Clearly spatial variability overwhelms the pure error variability τ^2, showing the importance of the spatial model.

The dependence between the processes is evident in the posterior correlation between the components. We find the anticipated negative association between the intercept process and the slope processes (apart from that with the "other area" process). Under the Matérn correlation function, by inverting $\rho(\cdot;\phi) = 0.05$ for a given value of the decay parameter γ and the smoothing parameter ν, we obtain the range, i.e., the distance beyond which spatial association becomes negligible. Posterior samples of (γ, ν) produce posterior samples for the range. The resulting posterior median is roughly 4 km over a somewhat sprawling parish that is roughly 22 km $\times$ 33 km. The smoothness parameter suggests processes with mean square differentiable realizations ($\nu > 1$). Contour plots of the posterior mean spatial surfaces for each of the processes (not shown) are quite different.

9.7.1 Generalized linear model setting

We briefly consider a generalized linear model version of (9.55), replacing the Gaussian first stage with

$$f\left(y\left(\mathbf{s}_i\right) \mid \theta\left(\mathbf{s}_i\right)\right) = h\left(y\left(\mathbf{s}_i\right)\right) \exp\left(\theta\left(\mathbf{s}_i\right) y\left(\mathbf{s}_i\right) - b\left(\theta\left(\mathbf{s}_i\right)\right)\right) , \tag{9.68}$$

where, using a canonical link, $\theta\left(\mathbf{s}_i\right) = \mathbf{X}^T\left(\mathbf{s}_i\right) \tilde{\boldsymbol{\beta}}\left(\mathbf{s}_i\right)$. In (9.68) we could include a dispersion parameter with little additional complication.

The resulting first-stage likelihood becomes

$$L\left(\tilde{\boldsymbol{\beta}};\mathbf{y}\right) = \exp\left\{\sum y\left(\mathbf{s}_i\right)\mathbf{X}^T\left(\mathbf{s}_i\right)\tilde{\boldsymbol{\beta}}\left(\mathbf{s}_i\right) - b\left(\mathbf{X}^T\left(\mathbf{s}_i\right)\tilde{\boldsymbol{\beta}}\left(\mathbf{s}_i\right)\right)\right\}. \tag{9.69}$$

Taking the prior on $\tilde{\boldsymbol{\beta}}$ in (9.56), the Bayesian model is completely specified with a prior on on ϕ, T and $\boldsymbol{\mu}_\beta$.

This model can be fit using a conceptually straightforward Gibbs sampling algorithm, which updates the components of $\boldsymbol{\mu}_{\boldsymbol{\beta}}$ and $\tilde{\boldsymbol{\beta}}$ using adaptive rejection sampling. With an inverse Wishart prior on T, the resulting full conditional of T is again inverse Wishart. Updating ϕ is usually very awkward because it enters in the Kronecker form in (9.56). Slice sampling is not available here since we cannot marginalize over the spatial effects; Metropolis updates are difficult to design but offer perhaps the best possibility. Also problematic is the repeated componentwise updating of $\tilde{\boldsymbol{\beta}}$. This hierarchically centered parametrization (Gelfand, Sahu, and Carlin, 1995, 1996) is preferable to working with $\boldsymbol{\mu}_{\boldsymbol{\beta}}$ and $\boldsymbol{\beta}$, but in our experience the algorithm still exhibits serious autocorrelation problems.

9.8 Illustrating multivariate spatial modeling with spBayes

We motivate this session with soil nutrient data which was collected at the La Selva Biological Station, Costa Rica, analyzed in greater detail by Guhaniyogi et al. (2013). Here, $n = 80$ soil cores were sampled over a sparse grid centered on a more intensively sampled transect. Soil nutrient concentrations of calcium (Ca), potassium (K) and magnesium (Mg) were measured for each sample. These nutrient concentrations show a high positive correlation as seen in (9.70):

$$\begin{pmatrix} 1 & & \\ 0.7 & 1 & \\ 0.7 & 0.8 & 1 \end{pmatrix} \tag{9.70}$$

suggesting that we might build a richer model by explicitly accounting for spatial association among the $q = 3$ outcome variables. Our objective is to predict these nutrients at a fine resolution over the study plot. Ultimately, posterior predictive samples will serve as input to a vegetation competition model. We begin by log transforming the response variables and taking a look at sample location across the study plot and then interpolating the outcome variables using the `mba.surf` function in the `MBA` package.

```
> dat <- read.table("CostaRica/T4.csv", header=T, sep=",")
> coords <- as.matrix(dat[,c("X","Y")])
> nut.names <- c("Ca","K","Mg")
> log.nut <- log(dat[,nut.names])
> par(mfrow=c(2,2))
> for(i in 1:length(nut.names)){
+    surf <- mba.surf(cbind(coords,data=log.nut[,i]),
+ no.X=100, no.Y=100)$xyz.est
+    image.plot(surf, main=paste("Log ",nut.names[i],sep=""))
+  points(coords)
+ }
```

Figure 9.10 depicts the nutrient concentration surfaces using the `mba.surf` function in the `MBA` package. These patterns can be more formally examined using empirical semivariograms. In the code block below, we fit an exponential variogram model to each of the soil nutrients. The resulting variogram estimates are offered in Figure 9.11. Here the upper and lower horizontal lines are the *sill* and *nugget*, respectively, and the vertical line is the effective range (i.e., that distance at which the correlation drops to 0.05). Despite the patterns

Figure 9.10 *Soil nutrient concentrations and sample array.*

of spatial dependence seen in Figure 9.10, the variograms do not show much of a spatial process. Changing the number of bins (`bins`) and maximum distance considered (`max`) will produce effective spatial ranges of less than 20m for each of the nutrients; however, the signal is weak, likely due to the paucity of samples.

```
> max <- 0.25*max(as.matrix(dist(dat[,c("X","Y")])))
> bins <- c(9,8,9)
> par(mfrow=c(3,1))
> for(i in 1:length(nut.names)){
+  vario <- variog(coords=coords,data=log.nut[,i],
+      uvec=(seq(0,max, length=bins[i])))
+  fit <-variofit(vario, ini.cov.pars=c(0.3, 20/-log(0.05)),
+         cov.model="exponential",
+ minimisation.function="nls",
+ weights="equal")
+  plot(vario, pch=19,
+ main=paste("Log ",nut.names[i],sep=""))
+  lines(fit)
+  abline(h=fit$nugget, col="blue")
+  abline(h=fit$cov.pars[1]+fit$nugget, col="green")
+  abline(v=-log(0.05)*fit$cov.pars[2], col="red3")
+ }
```

We continue with fitting a multivariate regression that allows for spatial ($AA^T = K$) and non-spatial (Ψ) cross-covariance matrices. In the following code block we define the model parameters' starting, tuning, and prior distribution, then call the `spMvLm` function in `spBayes` to estimate a simple Bayesian LMC (not spatially-varying). The sampler took ~30 minutes to collect 50,000 samples. Recall, the sampler must compute Cholesky factorizations for 240×240 dispersion matrices for each iteration.

```
> q <- 3
```

Figure 9.11 *Isotropic semivariograms for log nutrient concentrations.*

```
> n.samples <- 10000
> n.ltr <- q*(q+1)/2
> nut.spMvLM <- spMvLM(list(Ca~1,K~1,Mg~1), coords=coords,
+ data=log.nut,
+ starting=list("beta"=rep(1,q), "phi"=rep(3/20,q),
+ "A"=rep(0.1,n.ltr), "Psi"=rep(0.1,q)),
+ tuning=list("phi"=rep(0.3,q),
+ "A"=rep(0.001,n.ltr),
+ "Psi"=rep(0.01,q)),
+ priors=list("phi.Unif"=list(rep(3/100,q),
+                                     rep(3/10,q)),
+                        "K.IW"=list(q+1, diag(0.1,q)),
+ "Psi.IG"=list(rep(2,q),rep(0.1,q))),
+                        cov.model="exponential",
+ n.samples=n.samples,
+ verbose=TRUE, n.report=2500)
```

The posterior summaries of the entries of $AA^T = K$ are given by

```
          2.5%   50% 97.5%
K[1,1]   0.299 0.448 0.614
K[2,1]   0.136 0.216 0.301
K[3,1]   0.209 0.320 0.430
K[2,2]   0.092 0.134 0.195
K[3,2]   0.112 0.169 0.231
K[3,3]   0.179 0.255 0.355
Psi[1,1] 0.022 0.048 0.099
Psi[2,2] 0.023 0.042 0.063
Psi[3,3] 0.010 0.024 0.055
```

Figure 9.12 *Interpolated surface of the non-spatial model residuals and the mean of the random spatial effects posterior distribution.*

```
phi[1]    0.100 0.199 0.294
phi[2]    0.045 0.156 0.279
phi[3]    0.036 0.133 0.279
```

We recover the residual spatial effects using the `spRecover` function (and round it up to 3 decimal places) as shown in the block of code below:

```
> burn.in <- floor(0.75*n.samples)
> nut.spMvLM <- spRecover(nut.spMvLM, start=burn.in,
+ verbose=FALSE)
> nut.summary = summary(nut.spMvLM$p.theta.recover.samples)
> round(nut.summary$quantiles[,c(1,3,5)],3)
```

Figure 9.12 shows the nutrient concentration random spatial effects and compares them with the residual image plots from a non-spatial regression.

With a sparse sample array, an estimated mean effective range of ∼20, and no predictor variables, we cannot expect our prediction to differ much from a constant mean concentration over the domain. In the code block below, we define our prediction grid, construct the prediction design matrix using the `mkMvX` function in `spBayes`, and subsequently invoke `spPredict`.

```
> x.range <- range(coords[,1])
> y.range <- range(coords[,2])
> pred.coords <- expand.grid(seq(x.range[1], x.range[2], by=4),
+ seq(y.range[1], y.range[2], by=4))
> m <- nrow(pred.coords)
> pred.X <- mkMvX(list(matrix(1,m,1),
+ matrix(1,m,1), matrix(1,m,1)))
> nut.pred <- spPredict(nut.spMvLM, start=burn.in,
+ pred.coords=pred.coords,
```

```
+ pred.covars=pred.X, verbose=FALSE)
```

We can produce an interpolated image plot for the posterior predictive mean and standard deviation for calcium through the following steps. We first extract the posterior predictive means and standard deviations for each location. These are stored in "`Ca.pred.mu`" and "`Ca.pred.sd`," respectively.

```
> y.pred.mu <- apply(nut.pred$p.y.predictive.samples, 1, mean)
> y.pred.sd <- apply(nut.pred$p.y.predictive.samples, 1, sd)
> Ca.pred.mu <- y.pred.mu[seq(1,length(y.pred.mu),q)]
> Ca.pred.sd <- y.pred.sd[seq(1,length(y.pred.sd),q)]
```

We can similarly store the posterior predictive means and standard deviations for the other nutrients and store them in "`K.pred.mu`", "`Mg.pred.mu`", "`K.pred.sd`" and "`Mg.pred.sd`".

In the next block of code, `nut.pred` list object holds the posterior predictive samples for the spatial effects `w.pred` and response `y.pred`. Again, like with the random spatial effect in the `spMvLM` object, the posterior samples are stacked by location and can be unstacked as detailed in the code block below. We also convert our prediction grid into an `sp` data frame of type `SpatialGridDataFrame` and, subsequently, to a format that can be plotted by the `image` or `image.plot` function in the `fields` package.

```
> nut.pred.grid <- as.data.frame(list(x=pred.coords[,1],
+ y=pred.coords[,2],
+ Ca.mu=Ca.pred.mu, K.mu=K.pred.mu, Mg.mu=Mg.pred.mu,
+ Ca.sd=Ca.pred.sd, K.sd=K.pred.sd, Mg.sd=Mg.pred.sd))
> coordinates(nut.pred.grid) <- c("x", "y")
> gridded(nut.pred.grid) <- TRUE # promote to SpatialGridDataFrame
> toImage <- function(x){as.image.SpatialGridDataFrame(x)}
```

The interpolated image plot for the posterior predictive means can then be produced using the `mba.surf` function as below:

```
> res <- 100
> image.plot(toImage(nut.pred.grid["Ca.mu"]),
+ xaxs = "r", yaxs = "r",
+ zlim=z.lim, main="Mean of Ca prediction")
> points(coords)
```

Similarly, we can obtain the corresponding plots for the other nutrients. These are displayed in Figure 12.8. To obtain interpolated maps for the predictive standard deviations for calcium, we simply replace "`Ca.mu`" by "`Ca.sd`" in the above code. The computation for the other nutrients is analogous and are all plotted in Figure 9.13. With such a small spatial range, increased precision does not extend far from the sample locations.

9.9 Exercises

1. Compute the coherence (generalized correlation) in (9.12):
   (a) for the cross-covariance in (9.38), and
   (b) for the cross-covariance in (9.42).
2. Show that the product of separable cross-covariance functions is a valid cross-covariance function.
3. Suppose $Z_j(\mathbf{s}) \sim \texttt{Poi}(\lambda_j(\mathbf{s})), j = 1, 2$ where $\lambda_j(\mathbf{s}) = N_j(\mathbf{s})e^{\beta_j + w_j(\mathbf{s})}$ and $\mathbf{w}(\mathbf{s}) = \begin{pmatrix} w_1(\mathbf{s}) \\ w_2(\mathbf{s}) \end{pmatrix} = A \begin{pmatrix} v_1(\mathbf{s}) \\ v_2(\mathbf{s}) \end{pmatrix}$ with $v_1(\mathbf{s})$ and $v_2(\mathbf{s})$ independent mean 0 Gaussian processes. Obtain $\text{cov}(Z_1(\mathbf{s}), Z_2(\mathbf{s}'))$.

Figure 9.13 *Interpolated surface of observed log nutrient concentrations and standard deviation of each pixel's posterior predictive distribution.*

4. Let $Y(\mathbf{s}) = (Y_1(\mathbf{s}), Y_2(\mathbf{s}))^T$ be a bivariate process with a stationary cross-covariance matrix function
$$C(\mathbf{s}-\mathbf{s}') = \begin{pmatrix} c_{11}(\mathbf{s}-\mathbf{s}') & c_{12}(\mathbf{s}-\mathbf{s}') \\ c_{12}(\mathbf{s}'-\mathbf{s}) & c_{22}(\mathbf{s}-\mathbf{s}') \end{pmatrix},$$
and a set of covariates $\mathbf{x}(\mathbf{s})$. Let $\mathbf{y} = \left(\mathbf{y}_1^T, \mathbf{y}_2^T\right)^T$ be the $2n \times 1$ data vector, with $\mathbf{y}_1 = (y_1(\mathbf{s}_1), \ldots, y_1(\mathbf{s}_n))^T$ and $\mathbf{y}_2 = (y_2(\mathbf{s}_1), \ldots, y_2(\mathbf{s}_n))^T$.
   (a) Show that the cokriging predictor has the form
$$E[Y_1(\mathbf{s}_0)|\mathbf{y}] = \mathbf{x}^T(\mathbf{s}_0)\boldsymbol{\beta} + \boldsymbol{\gamma}^T\Sigma^{-1}(\mathbf{y} - X\boldsymbol{\beta}),$$
   i.e., as in (2.15), but with appropriate definitions of $\boldsymbol{\gamma}$ and Σ.
   (b) Show further that if $\mathbf{s}_k$ is a site where $y_l(\mathbf{s}_k)$ is observed, then for $l = 1, 2$, $E[Y_l(\mathbf{s}_k)|\mathbf{y}] = y_l(\mathbf{s}_k)$ if and only if $\tau_l^2 = 0$.
5. Suppose $\mathbf{Y}(\mathbf{s}) = (Y_1(\mathbf{s}), Y_2(\mathbf{s}))$ is a constant mean, bivariate Gaussian process with separable cross-covariance function. Suppose we observe data $\mathbf{Y} = \{\mathbf{Y}(\mathbf{s}_i), i = 1, 2, ..., n\}$. Show that the co-kriging predictor for $Y_1(\mathbf{s}_0)$ at a new location $\mathbf{s}_0$, $E(Y_1(\mathbf{s}_0)|\mathbf{Y})$ depends only upon the set of $Y_1(\mathbf{s}_i)$. Is the result still true if $E(Y_j(\mathbf{s}) = \mathbf{X}^T(\mathbf{s})\boldsymbol{\beta}_j, j = 1, 2$?
6. Suppose $\mathbf{Y}(\mathbf{s}) = (Y_1(\mathbf{s}), Y_2(\mathbf{s}))$ is a mean 0, bivariate Gaussian process with separable cross covariance function, $\rho(\mathbf{s}-\mathbf{s}'; \phi)T$ where $T = \begin{pmatrix} 1 & \gamma \\ \gamma & 1 \end{pmatrix}$. Suppose we observe data $\mathbf{Y} = \{\mathbf{Y}(\mathbf{s}_i), i = 1, 2, ..., n\}$ and we wish to predict $Z(\mathbf{s}_0) = Y_1(\mathbf{s}_0) + Y_2(\mathbf{s}_0)$ at the new location $\mathbf{s}_0$. Show that the predictive distribution $Z(\mathbf{s}_0) \mid \mathbf{Y}$ is the same as the predictive distribution $Z(\mathbf{s}_0)|\{Z(\mathbf{s}_i) = Y_1(\mathbf{s}_i) + y_2(\mathbf{s}_i)\}$. Does this result hold for a general T?
7. Suppose $\mathbf{Y}(\mathbf{s})$ is a bivariate spatial process as in Exercise 4 In fact, suppose $\mathbf{Y}(\mathbf{s})$ is a Gaussian process. Let $Z_1(\mathbf{s}) = I(Y_1(\mathbf{s}) > 0)$, and $Z_2(\mathbf{s}) = I(Y_2(\mathbf{s}) > 0)$. Approximate the cross-covariance matrix of $\mathbf{Z}(\mathbf{s}) = (Z_1(\mathbf{s}), Z_2(\mathbf{s}))^T$.

8. The data in `www.biostat.umn.edu/~brad/data/ColoradoLMC.dat` record maximum temperature (in tenths of a degree Celsius) and precipitation (in cm) during the month of January 1997 at 50 locations in the U.S. state of Colorado.
   (a) Let X denote temperature and Y denote precipitation. Following the model of Example 9.3, fit an LMC model to these data using the conditional approach, fitting X and then $Y|X$.
   (b) Repeat this analysis, but this time fitting Y and then $X|Y$. Show that your new results agree with those from part (a) up to simulation variability.
9. If C_l and $C_{l'}$ are isotropic, obtain $C_{ll'}(\mathbf{s})$ in (9.67) by transformation to polar coordinates.
10. More on the coregionalization asymmetry. Following up on the discussion surrounding (9.37), suppose we have the usual bivariate coregionalization model $\mathbf{W}(\mathbf{s}) = AV(\mathbf{s})$ with A lower triangular and say, $V_1(\mathbf{s})$ and $V_2(\mathbf{s})$ are independent mean 0 Gaussian processes. Clarify that $W_1(\mathbf{s})$ and $W_2(\mathbf{s}')$ are conditionally independent given $W_1(\mathbf{s}')$ but $W_1(\mathbf{s}')$ and $W_2(\mathbf{s})$ are not conditionally independent given $W_2(\mathbf{s}')$. (Of course, the asymmetry disappears if A is not lower triangular.)

Chapter 10

Models for multivariate areal data

In this chapter we explore the extension of univariate CAR methodology (Sections 4.3 and 6.4.3) to the multivariate setting. Such models can be employed to introduce multiple, dependent spatial random effects associated with areal units (as standard CAR models do for a single set of random effects). In this regard, Kim et al. (2001) presented a "twofold CAR" model to model counts for two different types of disease over each areal unit. Similarly, Knorr-Held and Best (2000) have developed a "shared component" model for the above purpose, but their methodology too seems specific to the bivariate situation. Knorr-Held and Rue (2002) illustrate sophisticated MCMC blocking approaches in a model placing three conditionally independent CAR priors on three sets of spatial random effects in a shared component model setting.

Multivariate CAR (MCAR) models can also provide coefficients in a multiple regression setting that are dependent and spatially varying at the areal unit level. For example, Gamerman et al. (2002) investigate a Gaussian Markov random field (GMRF) model (a multivariate generalization of the pairwise difference IAR model) and compare various MCMC blocking schemes for sampling from the posterior that results under a Gaussian multiple linear regression likelihood. They also investigate a "pinned down" version of this model that resolves the impropriety problem by centering the $\boldsymbol{\phi}_i$ vectors around some mean location. These authors also place the spatial structure on the spatial regression coefficients themselves, instead of on extra intercept terms (that is, in (6.27) we would drop the ϕ_i, and replace β_1 by β_{1i}, which would now be assumed to have a CAR structure). Assunção et al. (2002) refer to these models as *space-varying coefficient* models, and illustrate in the case of estimating fertility schedules. Assunção (2003) offers a nice review of the work up to that time in this area. Also working with areal units, Sain and Cressie (2007) offer multivariate GMRF models, proposing a generalization that permits asymmetry in the spatial conditional cross-correlation matrix. They use this approach to jointly model the counts of white and minority persons residing in the census block groups of St. James Parish, LA, a region containing several hazardous waste sites.

We point out that multivariate CAR models are not the only option available for analyzing multivariate areal data. Zhang, Hodges and Banerjee (2009) develop an arguably much simpler alternative approach building upon the techniques of smoothed ANOVA (SANOVA) (see Hodges, Cui, Sargent and Carlin, 2007). Instead of simply shrinking effects without any structure, these authors propose SANOVA to smooth spatial random effects by taking advantage of the spatial structure. The underlying idea is to extend SANOVA to cases in which one factor is a spatial lattice, which is smoothed using a CAR model, and a second factor is, for example, type of disease. Datasets routinely lack enough information to identify the additional structure of MCAR. SANOVA offers a simpler and more intelligible structure than the MCAR while performing as well. Nevertheless, the MCAR and more general CAR-based approaches provide a diverse and rich class of models suitable for capturing complex spatial associations. We focus upon these approaches in the remainder of this section.

10.1 The multivariate CAR (MCAR) distribution

For a vector of univariate variables $\boldsymbol{\phi} = (\phi_1, \phi_2, \ldots, \phi_n)$, zero-centered CAR specifications were detailed in Section 4.3. For the MCAR model we instead let $\boldsymbol{\phi}^T = (\boldsymbol{\phi}_1, \boldsymbol{\phi}_2, \ldots, \boldsymbol{\phi}_n)$ where each $\boldsymbol{\phi}_i = (\phi_{i1}, \phi_{i2}, \ldots, \phi_{ip})^T$ is $p \times 1$. Most multivariate CAR models are members of the family developed by Mardia (1988). Analogous to the univariate case, the joint distribution is derived from the full conditional distributions. Under the MRF assumption, we can specify these conditional distributions as

$$p(\boldsymbol{\phi}_i | \boldsymbol{\phi}_{j \neq i}, \Gamma_i) \;=\; N\left(\sum_{i \sim j} B_{ij}\boldsymbol{\phi}_j \;,\; \Gamma_i\right) \;,\; i, j = 1, \ldots, n, \tag{10.1}$$

where Γ_i and B_{ij} are $p \times p$ matrices. Mardia (1988) proved, using a multivariate analogue of Brook's Lemma, that the full conditional distributions in (10.1) yield a joint distribution of the form

$$p(\boldsymbol{\phi} \mid \{\Gamma_i\}) \propto \exp\left\{-\frac{1}{2}\boldsymbol{\phi}^T \Gamma^{-1}(I - \tilde{B})\boldsymbol{\phi}\right\} \;,$$

where Γ is block-diagonal with blocks Γ_i, and $\tilde{B}$ is $np \times np$ with (i, j)-th block B_{ij}.

As in the univariate case, symmetry of $\Gamma^{-1}(I - \tilde{B})$ is required. A convenient special case sets $B_{ij} = b_{ij} I_{p \times p}$, yielding the symmetry condition $b_{ij}\Gamma_j = b_{ji}\Gamma_i$, analogous to (4.14). If as in Subsection 4.3.1 we take $b_{ij} = w_{ij}/w_{i+}$ and $\Sigma_i = w_{i+}^{-1}\Sigma$, then the symmetry condition is satisfied.

Kronecker product notation simplifies the form of $\Gamma^{-1}(I - \tilde{B})$. That is, setting $\tilde{B} = B \otimes I$ with B as in (4.13) and $\Gamma = D^{-1} \otimes \Sigma$ so

$$\Gamma^{-1}(I - \tilde{B}) = (D \otimes \Sigma^{-1})(I - B \otimes I) = (D - W) \otimes \Sigma^{-1} \;.$$

Again, the singularity of $D - W$ implies that $\Gamma^{-1}(I - B)$ is singular. We denote this distribution by $MCAR(1, \Sigma)$.

To consider remedies to the impropriety, Mardia (1988) proposed rewriting (10.1) as

$$p(\boldsymbol{\phi}_i | \boldsymbol{\phi}_{j \neq i}, \Gamma_i) \;=\; N\left(R_i \sum_{i \sim j} B_{ij}\boldsymbol{\phi}_j \;,\; \Gamma_i^{-1}\right) \;,\; i, j = 1, \ldots, n,$$

where R_i is $p \times p$. Now $\Gamma^{-1}(I - \tilde{B})$ is revised to $\Gamma^{-1}(I - \tilde{B}_R)$ where $\tilde{B}_R$ has (i, j)th block $R_i B_{ij}$. In general, then, the symmetry condition becomes

$$(\Gamma_i^{-1} R_i B_{ij})^T = \Gamma_j^{-1} R_j B_{ji} \;\text{ or }\; \Gamma_j B_{ij}^T R_i^T = R_j B_{ji} \Gamma_i \;.$$

See Mardia, (1988), Expression (2.4) in this regard. If, in addition, $\Gamma^{-1}(I - \tilde{B}_R)$ is positive definite, then the conditional distributions uniquely determine the joint distribution

$$\boldsymbol{\phi} \sim N\left(\mathbf{0} \;,\; [\Gamma(I - \tilde{B}_R)]^{-1}\right) \;. \tag{10.2}$$

In particular, if $B_{ij} = b_{ij} I_{p \times p}$ and $b_{ij} = w_{ij}/w_{i+}$, the symmetry condition simplifies to

$$w_{j+}\Gamma_j R_i^T = w_{i+} R_j \Gamma_i \;.$$

Finally, if in addition we take $\Gamma_i = w_{i+}^{-1}\Lambda$, we obtain $\Lambda R_i^T = R_j \Lambda$, which reveals that we must have $R_i = R_j = R$, and thus

$$\Lambda R^T = R\Lambda \;. \tag{10.3}$$

For any arbitrary positive definite Λ, a generic solution to (10.3) is $R = \rho\Lambda^t$. Hence, regardless of t, (10.3) introduces a total of $\binom{p+1}{2} + 1$ parameters. Thus, without loss of generality, we can set $t = 0$, hence $R = \rho I$. Calculations as above yield

$$\Sigma^{-1} = \Gamma^{-1}(I - \tilde{B}_R) = (D - \rho W) \otimes \Lambda^{-1} \ , \tag{10.4}$$

where Σ is the $np \times np$ variance-covariance matrix of $\boldsymbol{\phi}$. Therefore, Σ has a *separable* structure and is nonsingular under the same restriction to ρ as in the univariate case, and we have

$$\boldsymbol{\phi} \sim N\left(0,\ [(D - \rho W) \otimes \Lambda]^{-1}\right) \ . \tag{10.5}$$

Following Gelfand and Vounatsou (2003) and Carlin and Banerjee (2003), we denote this model by $MCAR(\rho, \Sigma)$.

The separable MCAR model in (10.5) is obtained under the assumption that $R_i = \rho I_{p\times p}, i = 1, \ldots, n$. While the positive definiteness condition of the covariance matrix follows immediately from its Kronecker product form, which also has computational benefits, the assumption of a common ρ for $j = 1, \ldots, p$ may well be too strong. To elucidate further, suppose $p = 2$ (e.g., two diseases in each county), and define $\boldsymbol{\phi}_1' = (\phi_{11}, \ldots, \phi_{n1})$ and $\boldsymbol{\phi}_2' = (\phi_{12}, \ldots, \phi_{n2})$. Then the MCAR formulation (10.5) can be written as

$$\begin{pmatrix} \boldsymbol{\phi}_1 \\ \boldsymbol{\phi}_2 \end{pmatrix} \sim N\left(\begin{pmatrix} \mathbf{0} \\ \mathbf{0} \end{pmatrix}, \begin{pmatrix} (D - \rho W)\Lambda_{11} & (D - \rho W)\Lambda_{12} \\ (D - \rho W)\Lambda_{12} & (D - \rho W)\Lambda_{22} \end{pmatrix}^{-1} \right) \ , \tag{10.6}$$

where $\Lambda_{ij},\ i = 1, 2,\ j = 1, 2$ are the elements of Λ. More generally, we may need three different ρ_i parameters in (10.6) to explain the correlation between the two types of cancer and across the counties that neighbor each other (Kim et al., 2001). The covariance matrix Σ would then be revised to

$$\Sigma = \begin{pmatrix} (D - \rho_1 W)\Lambda_{11} & (D - \rho_3 W)\Lambda_{12} \\ (D - \rho_3 W)\Lambda_{12} & (D - \rho_2 W)\Lambda_{22} \end{pmatrix}^{-1} \ , \tag{10.7}$$

where ρ_1 and ρ_2 are the smoothing parameters for the two cancer types, and ρ_3 is the "bridging" or "linking" parameter associating ϕ_{i1} with ϕ_{j2}, $i \neq j$. Unfortunately, with this general covariance matrix, it is difficult to derive the conditions for positive definiteness as they depend upon the unknown Λ matrix.

Carlin and Banerjee (2003) and Gelfand and Vounatsou (2003) generalize the separable MCAR model by allowing two different ρ parameters (say, ρ_1 and ρ_2), and denote this model as $MCAR(\rho_1, \rho_2, \Lambda)$. They write the precision matrix Σ^{-1} as

$$\begin{pmatrix} U_1^T U_1 \Lambda_{11} & U_1^T U_2 \Lambda_{12} \\ U_2^T U_1 \Lambda_{12} & U_2^T U_2 \Lambda_{22} \end{pmatrix} = \begin{pmatrix} U_1^T & 0 \\ 0 & U_2^T \end{pmatrix} (\Lambda \otimes I_{n\times n}) \begin{pmatrix} U_1 & 0 \\ 0 & U_2 \end{pmatrix} \ , \tag{10.8}$$

where $U_k' U_k = D - \rho_k W,\ k = 1, 2$. Carlin and Banerjee (2003) take U_k to be the Cholesky decomposition of $D - \rho_k W$ so that U_k is an upper-triangular matrix, while Gelfand and Vounatsou (2003) employ a spectral decomposition, i.e., $U_k = \mathrm{Diag}(1 - \rho_k \lambda_i)^{\frac{1}{2}} P' D^{\frac{1}{2}} P$, where the λ_i are the eigenvalues of $D^{-\frac{1}{2}} W D^{-\frac{1}{2}}$ and P is an orthogonal matrix with the corresponding eigenvectors as its columns. Either way, this generalization of the MCAR model permits different smoothing parameters ρ_k for each k (e.g., different strengths of spatial correlation for each type of cancer). As before, Λ controls the nonspatial correlation among cancers at any given location.

The conditions for the covariance matrix to be positive definite are easy to find as long as the Cholesky or spectral decompositions exist and Λ is positive definite. For the $p = 2$ case, these reduce to $|\rho_1| < 1$ and $|\rho_2| < 1$. The spectral approach may be better in terms

of Bayesian computing, since it does not requires calculation of a Cholesky decomposition at each MCMC iteration, a substantial burden particularly for a data set with many spatial regions. Neither of these MCAR structures allows a smoothing parameter ρ on the off-diagonal of the precision matrix as in (10.7); we cannot model the off-diagonal, since it is determined by the diagonal. Finally, since the decomposition of $D - \rho_k W$ is not unique, we can have different MCAR models with the covariance structure (10.8).

Note that the covariance structure (10.8) easily generalizes to an arbitrary number, say, p, of diseases. We write

$$\boldsymbol{\phi} \sim N_{np}\left(\mathbf{0},\ \left[\mathrm{Diag}(U_1^T, \ldots, U_p^T)(\Lambda \otimes I_{n\times n})\mathrm{Diag}(U_1, \ldots, U_p)\right]^{-1}\right), \tag{10.9}$$

where $U_j^T U_j = D - \rho_j W$, $j = 1, \ldots, p$. We denote the distribution in (10.9) by $MCAR(\rho_1, \ldots, \rho_p, \Lambda)$. Note that the off-diagonal block matrices (the U_i's) in the precision matrix in (10.9) are completely determined by the diagonal blocks. Thus, the spatial precision matrices for each disease induce the cross-covariance structure in (10.9).

Kim, Sun, and Tsutakawa (2001) proposed a multivariate CAR model in the bivariate ($p = 2$) case, which they dub the "twofold conditionally autoregressive" model, and which we notate as $2fCAR(\rho_0, \rho_1, \rho_2, \rho_3, \tau_1, \tau_2)$. They specify the moments of the full conditional distributions as

$$\mathrm{E}(\phi_{ik}|\phi_{il}, \phi_{jk}, \phi_{jl}) = \frac{1}{2m_i + 1}\left(\rho_k \sum_{j\sim i} \phi_{jk} + \rho_3\sqrt{\frac{\tau_l}{\tau_k}} \sum_{j\sim i} \phi_{jl} + \rho_0\sqrt{\frac{\tau_l}{\tau_k}}\phi_{il}\right)$$

and

$$\mathrm{Var}(\phi_{ik}|\phi_{il}, \phi_{jk}, \phi_{jl}) = \frac{\tau_k^{-1}}{2m_i + 1}, \quad i, j = 1, \ldots n, \quad l, k = 1, 2, \quad l \neq k,$$

where $j \sim i$ again means that region j is a neighbor of region i. Adding the Gaussian MRF structure, they derive the joint distribution arising from these full conditional distributions as

$$\begin{pmatrix}\boldsymbol{\phi}_1\\ \boldsymbol{\phi}_2\end{pmatrix} \sim N\left(\begin{pmatrix}\mathbf{0}\\ \mathbf{0}\end{pmatrix}, \begin{pmatrix} (2D + I - \rho_1 W)\tau_1 & -(\rho_0 I + \rho_3 W)\sqrt{\tau_1\tau_2} \\ -(\rho_0 I + \rho_3 W)\sqrt{\tau_1\tau_2} & (2D + I - \rho_2 W)\tau_2 \end{pmatrix}^{-1}\right), \tag{10.10}$$

where again $\boldsymbol{\phi}_1' = (\phi_{11}, \ldots, \phi_{n1})$, $\boldsymbol{\phi}_2' = (\phi_{12}, \ldots, \phi_{n2})$, $D = Diag(m_i)$, and W is the adjacency matrix. This model has the same number of parameters in the covariance structure (six) as the general formulation (10.7) in the bivariate case, so they are related to each other. In (10.10), ρ_1 and ρ_2 are the smoothing parameters, while ρ_0 and ρ_3 are the bridging parameters associating ϕ_{i1} with ϕ_{i2} and ϕ_{j2}, $j \neq i$, respectively. Unfortunately, this MCAR model is only designed for the bivariate case ($p = 2$), and seems difficult to generalize to higher dimensions. Also, under this approach it is hard to find conditions that guarantee a positive definite covariance matrix in (10.10). The conditions $|\rho_l| < 1$, $l = 0, 1, 2, 3$ given by Kim et al. (2001) are sufficient but not necessary, and may be overly restrictive for some data sets since they restrict the correlation of ϕ_{i1} with ϕ_{i2} and ϕ_{j2}, $j \neq i$. Finally, this generalization comes at a significant price in terms of computing, since it requires many matrix multiplications, determinant evaluations, and inverses at each MCMC iteration, so can be very time-consuming even when working on a relatively small spatial domain.

10.2 Modeling with a proper, non-separable MCAR distribution

As in Chapter 6, the MCAR specifications of the previous section are employed in models for spatial random effects arising in a hierarchical model. For instance, suppose we have

a linear model with continuous data $\mathbf{Y}_{ik}$, $i = 1, \ldots, n$, $k = 1, \ldots, m_i$, where $\mathbf{Y}_{ik}$ is a $p \times 1$ vector denoting the kth response at the ith areal unit. The mean of the $\mathbf{Y}_{ik}$ is $\boldsymbol{\mu}_{ik}$ where $\mu_{ikj} = (\mathbf{X}_{ik})_j \boldsymbol{\beta}^{(j)} + \phi_{ij}$, $j = 1, \ldots, p$. Here $\mathbf{X}_{ik}$ is a $p \times s$ matrix with covariates associated with $\mathbf{Y}_{ik}$ having jth row $(\mathbf{X}_{ik})_j$, $\boldsymbol{\beta}^{(j)}$ is an $s \times 1$ coefficient vector associated with the jth component of the $\mathbf{Y}_{ik}$'s, and ϕ_{ij} is the jth component of the $p \times 1$ vector $\boldsymbol{\phi}_i$. Given $\left\{\boldsymbol{\beta}^{(j)}\right\}, \{\boldsymbol{\phi}_i\}$ and V, the $\mathbf{Y}_{ik}$ are conditionally independent $N(\boldsymbol{\mu}_{ik}, V)$ variables. Adding a prior for $\left\{\boldsymbol{\beta}^{(j)}\right\}$ and V and one of the MCAR models from Subsection 10.1 for the $\boldsymbol{\phi}_i$ completes the second stage of the specification. Finally, a hyperprior on the MCAR parameters completes the model.

Alternatively, we might change the first stage to a multinomial. Here k disappears and $\mathbf{Y}_i$ is assumed to follow a multinomial distribution with sample size n_i and with $(p+1) \times 1$ probability vector $\boldsymbol{\pi}_i$. Working on the logit scale, using cell $p+1$ as the baseline, we could set $\log\left(\frac{\pi_{ij}}{\pi_{i,p+1}}\right) = \mathbf{X}_i^T \boldsymbol{\beta}^{(j)} + \phi_{ij}$, $j = 1, \ldots, p$, with $\mathbf{X}$'s, $\boldsymbol{\beta}$'s and $\boldsymbol{\phi}$'s interpreted as in the previous paragraph. Many other multivariate first stages could also be used, such as other multivariate exponential family models.

Regardless, model-fitting is most easily implemented using a Gibbs sampler with Metropolis updates where needed. The full conditionals for the $\boldsymbol{\beta}$'s will typically be normal (under a normal first-stage model) or else require Metropolis, slice, or adaptive rejection sampling (Gilks and Wild, 1992). For the $MCAR(1, \Sigma)$ and $MCAR(\rho, \Sigma)$ models, the full conditionals for the $\boldsymbol{\phi}_i$'s will be likelihood-adjusted versions of the conditional distributions that define the MCAR, and are updated as a block. For the $MCAR(\boldsymbol{\rho}, \Sigma)$ model, we can work with either the $\boldsymbol{\phi}$ or the $\boldsymbol{\psi}$ parametrization. With a non-Gaussian first stage, it will be awkward to pull the transformed effects out of the likelihood in order to do the updating. However, with a Gaussian first stage, it may well be more efficient to work on the transformed scale. Under the Gaussian first stage, the full conditional for V will be seen to follow an inverse Wishart, as will Σ. The ρ's do not follow standard distributions; in fact, discretization expedites computation, avoiding Metropolis steps.

We have chosen an illustrative prior for ρ in the ensuing example following three criteria. First, we insist that $\rho < 1$ to ensure propriety but allow $\rho = 0.99$. Second, we do not allow $\rho < 0$ since this would violate the similarity of spatial neighbors that we seek. Third, since even moderate spatial dependence requires values of ρ near 1 (recall the discussion in Subsection 4.3.1) we place prior mass that favors the upper range of ρ. In particular, we put equal mass on the following 31 values: $0, 0.05, 0.1, \ldots, 0.8, 0.82, 0.84, \ldots, 0.90, 0.91, 0.92, \ldots, 0.99$.

Finally, model choice arises here only in selecting among MCAR specifications. That is, we do not alter the mean vector in these investigations; our interest here lies solely in comparing the spatial explanations. Multivariate versions of the Gelfand and Ghosh (1998) criterion (5.14) for multivariate Gaussian data are employed.

Example 10.1 *(Analysis of the child growth data).* Child growth is usually monitored using anthropometric indicators such as height adjusted for age (HAZ), weight adjusted for height (WHZ), and weight adjusted for age (WAZ). Independent analysis of each of these indicators is normally carried out to identify factors influencing growth that may range from genetic and environmental factors (e.g., altitude, seasonality) to differences in nutrition and social deprivation. Substantial variation in growth is common within as well as between populations. Recently, geographical variation in child growth has been thoroughly investigated for the country of Papua New Guinea in Mueller et al. (2001). Independent spatial analyses for each of the anthropometric growth indicators identified complex geographical patterns of child growth finding areas where children are taller but skinnier than average, others where they are heavier but shorter, and areas where they are both short and thin. These geographical patterns could be linked to differences in diet and subsistence agriculture, leading to the analysis presented here; see Gelfand and Vounatsou (2003) for further discussion.

The data for our illustration comes from the 1982–1983 Papua New Guinea National Nutrition Survey (NNS) (Heywood et al., 1988). The survey includes anthropometric measures (age, height, weight) of approximately 28,000 children under five years of age, as well as dietary, socioeconomic, and demographic data about those children and their families. Dietary data include the type of food that respondents had eaten the previous day. Subsequently, the data were coded to 14 important staples and sources of protein. Each child was assigned to a village and each village was assigned to one of 4566 environmental zones (resource mapping units, or RMUs) into which Papua New Guinea has been divided for agriculture planning purposes. A detailed description of the data is given in Mueller et al. (2001).

The nutritional scores, height adjusted for age (HAZ), and weight adjusted for age (WAZ) that describe the nutritional status of a child were obtained using the method of Cole and Green (1992), which yields age-adjusted standard normal deviate Z-scores. The data set was collected at 537 RMUs. To overcome sparseness and to facilitate computation, we collapsed to 250 spatial units. In the absence of digitized boundaries, Delaunay tessellations were used to create the neighboring structure in the spatial units.

Because of the complex, multidimensional nature of human growth, a bivariate model that considers differences in height and weight jointly might be more appropriate for analyzing child growth data in general and to identify geographical patterns of growth in particular. We propose the use of Bayesian hierarchical spatial models and with *multivariate CAR* (MCAR) specifications to analyze the bivariate pairs of indicators, HAZ and WAZ, of child growth. Our modeling reveals bivariate spatial random effects at RMU level, justifying the MCAR specification.

Recalling the discussion of Subsection 10.1, it may be helpful to provide explicit expressions, with obvious notation, for the modeling and the resulting association structure. We have, for the jth child in the ith RMU,

$$\mathbf{Y}_{ij} = \begin{pmatrix} (HAZ)_{ij} \\ (WAZ)_{ij} \end{pmatrix} = \mathbf{X}_{ij}^T \begin{pmatrix} \boldsymbol{\beta}^{(H)} \\ \boldsymbol{\beta}^{(W)} \end{pmatrix} + \begin{pmatrix} \phi_i^{(H)} \\ \phi_i^{(W)} \end{pmatrix} + \begin{pmatrix} \epsilon_{ij}^{(H)} \\ \epsilon_{ij}^{(W)} \end{pmatrix} .$$

In this setting, under say the $MCAR(\rho, \Sigma)$ model,

$$\begin{aligned} &\mathrm{Cov}((HAZ)_{ij}, (HAZ)_{i'j'} \mid \boldsymbol{\beta}^{(H)}, \boldsymbol{\beta}^{(W)}, \rho, \Sigma, V) \\ &\quad = \mathrm{Cov}(\phi_i^{(H)}, \phi_{i'}^{(H)}) + V_{11} I_{i=i',j=j'} , \\ &\mathrm{Cov}((WAZ)_{ij}, (WAZ)_{i'j'} \mid \boldsymbol{\beta}^{(H)}, \boldsymbol{\beta}^{(W)}, \rho, \Sigma, V) \\ &\quad = \mathrm{Cov}(\phi_i^{(W)}, \phi_{i'}^{(W)}) + V_{22} I_{i=i',j=j'} , \\ \text{and} \quad &\mathrm{Cov}((HAZ)_{ij}, (WAZ)_{i'j'} \mid \boldsymbol{\beta}^{(H)}, \boldsymbol{\beta}^{(W)}, \rho, \Sigma, V) \\ &\quad = \mathrm{Cov}(\phi_i^{(H)}, \phi_{i'}^{(W)}) + V_{12} I_{i=i',j=j'} , \end{aligned}$$

where $cov(\phi_i^{(H)}, \phi_{i'}^{(H)}) = (D_W - \rho W)_{ii'} \Sigma_{11}$, $cov(\phi_i^{(W)}, \phi_{i'}^{(W)}) = (D_W - \rho W)_{ii'} \Sigma_{22}$, and $cov(\phi_i^{(H)}, \phi_{i'}^{(W)}) = (D_W - \rho W)_{ii'} \Sigma_{12}$. The interpretation of the components of Σ and V (particularly Σ_{12} and V_{12}) is now clarified.

We adopted noninformative uniform prior specifications on $\boldsymbol{\beta}^{(H)}$ and $\boldsymbol{\beta}^{(W)}$. For Σ and V we use inverse Wishart priors, i.e., $\Sigma^{-1} \sim W(\Omega_1, c_1)$, $V^{-1} \sim W(\Omega_2, c_2)$ where Ω_1, Ω_2 are $p \times p$ matrices and c_1, c_2 are shape parameters. Since we have no prior knowledge regarding the nature or extent of dependence, we choose Ω_1 and Ω_2 diagonal; the data will inform about the dependence *a posteriori*. Since the $\mathbf{Y}_{ij}$'s are centered and scaled on each dimension, setting $\Omega_1 = \Omega_2 = I$ seems appropriate. Finally, we set $c_1 = c_2 = 4$ to provide low precision for these priors. We adopted for ρ_1 and ρ_2 the prior discussed in the previous section. Simulation from the full conditional distributions of the $\boldsymbol{\beta}$'s and the $\boldsymbol{\psi}_i, i = 1, \ldots, n$ is straightforward

Model	G	P	D_∞
$MCAR(1, \Sigma)$	34300.69	33013.10	67313.79
$MCAR(\rho, \Sigma)$	34251.25	33202.86	67454.11
$MCAR(\boldsymbol{\rho}, \Sigma)$	34014.46	33271.97	67286.43

Table 10.1 *Model comparison for child growth data.*

	Height (HAZ)			Weight (WAZ)		
Covariate	2.5%	50%	97.5%	2.5%	50%	97.5%
Global mean	–0.35	–0.16	–0.01	–0.48	–0.25	–0.15
Coconut	0.13	0.20	0.29	0.04	0.14	0.24
Sago	–0.16	–0.07	–0.00	–0.07	0.03	0.12
Sweet potato	–0.11	–0.03	0.05	–0.08	0.01	0.12
Taro	–0.09	0.01	0.10	–0.19	–0.09	0.00
Yams	–0.16	–0.04	0.07	–0.19	–0.05	0.08
Rice	0.30	0.40	0.51	0.26	0.38	0.49
Tinned fish	0.00	0.12	0.24	0.04	0.17	0.29
Fresh fish	0.13	0.23	0.32	0.08	0.18	0.28
Vegetables	–0.08	0.08	0.25	0.02	0.19	0.35
V_{11}, V_{22}	0.85	0.87	0.88	0.85	0.87	0.88
V_{12}	0.60	0.61	0.63			
Σ_{11}, Σ_{22}	0.30	0.37	0.47	0.30	0.39	0.52
Σ_{12}	0.19	0.25	0.35			
ρ_1, ρ_2	0.95	0.97	0.97	0.10	0.80	0.97

Table 10.2 *Posterior summaries of the dietary covariate coefficients, covariance components, and autoregression parameters for the child growth data using the most complex MCAR model.*

as they are standard normal distributions. Similarly, the full conditionals for V^{-1} and Σ^{-1} are Wishart distributions. We implemented the Gibbs sampler with 10 parallel chains.

Table 10.1 offers a comparison of three MCAR models using (5.14), the Gelfand and Ghosh (1998) criterion. The most complex model is preferred, offering sufficient improvement in goodness of fit to offset the increased complexity penalty. Summaries of the posterior quantities under this model are shown in Table 10.2. These were obtained from a posterior sample of size 1,000, obtained after running a 10-chain Gibbs sampler for 30,000 iterations with a burn-in of 5,000 iterations and a thinning interval of 30 iterations. Among the dietary factors, high consumption of sago and taro are correlated with thinner and shorter children, while high consumption of rice, fresh fish, and coconut are associated with both heavier and taller children. Children from villages with high consumption of vegetables or tinned fish are heavier.

The posterior for the correlation associated with Σ, $\Sigma_{12}/\sqrt{\Sigma_{11}\Sigma_{22}}$, has mean 0.67 with 95% credible interval $(0.57, 0.75)$, while the posterior for the correlation associated with V, $V_{12}/\sqrt{V_{11}V_{22}}$, has mean 0.71 with 95% credible interval $(0.70, 0.72)$. In addition, ρ_1 and ρ_2 differ.

10.3 Conditionally specified Generalized MCAR (GMCAR) distributions

Jin, Carlin and Banerjee (2005) expand upon this idea by building the joint distribution for a multivariate Markov random field (MRF) through specifications of simpler conditional

and marginal models. The approach can be regarded as the analogue of the conditioning approach of Royle and Berliner (1999); see Section 9.4.5, for areal models.

This approach is best elucidated in the bivariate setting. We now assume the joint distribution of $\boldsymbol{\phi}_1$ and $\boldsymbol{\phi}_2$ is

$$\begin{pmatrix}\boldsymbol{\phi}_1\\ \boldsymbol{\phi}_2\end{pmatrix} \sim N\left(\begin{pmatrix}\mathbf{0}\\ \mathbf{0}\end{pmatrix}, \begin{pmatrix}\Sigma_{11} & \Sigma_{12}\\ \Sigma_{12}^T & \Sigma_{22}\end{pmatrix}\right) ,$$

where the Σ_{kl}, $k, l = 1, 2$ are $n \times n$ covariance matrices. From standard multivariate normal theory, we have $\mathrm{E}(\boldsymbol{\phi}_1|\boldsymbol{\phi}_2) = \Sigma_{12}\Sigma_{22}^{-1}\boldsymbol{\phi}_2$ and $\mathrm{Var}(\boldsymbol{\phi}_1|\boldsymbol{\phi}_2) = \Sigma_{11\cdot 2} = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{12}^T$. Now writing $A = \Sigma_{12}\Sigma_{22}^{-1}$, we can rewrite the joint distribution of $\boldsymbol{\phi}_1$ and $\boldsymbol{\phi}_2$ as

$$\begin{pmatrix}\boldsymbol{\phi}_1\\ \boldsymbol{\phi}_2\end{pmatrix} \sim N\left(\begin{pmatrix}\mathbf{0}\\ \mathbf{0}\end{pmatrix}, \begin{pmatrix}\Sigma_{11\cdot 2} + A\Sigma_{22}A^T & A\Sigma_{22}\\ (A\Sigma_{22})^T & \Sigma_{22}\end{pmatrix}\right) . \tag{10.11}$$

According to Harville (1997, Corollary 14.8.5), the conditions that ensure the propriety of (10.11) are that Σ_{22} and $\Sigma_{11\cdot 2}$ are positive definite. Since $\boldsymbol{\phi}_1|\boldsymbol{\phi}_2 \sim N(A\boldsymbol{\phi}_2, \Sigma_{11\cdot 2})$ and $\boldsymbol{\phi}_2 \sim N(0, \Sigma_{22})$, we can construct $p(\boldsymbol{\phi}) = p(\boldsymbol{\phi}_1|\boldsymbol{\phi}_2)p(\boldsymbol{\phi}_2)$ where $\boldsymbol{\phi}^T = (\boldsymbol{\phi}_1^T, \boldsymbol{\phi}_2^T)$. For the joint distribution of $\boldsymbol{\phi}$, then, we need to specify the matrices $\Sigma_{11\cdot 2}$, Σ_{22}, and A.

Jin et al. (2005) propose specifying the conditional distribution for $\boldsymbol{\phi}_1 \mid \boldsymbol{\phi}_2$ as $\boldsymbol{\phi}_1|\boldsymbol{\phi}_2 \sim N\left(A\boldsymbol{\phi}_2\,,\, [(D-\rho_1 W)\tau_1]^{-1}\right)$, and the marginal distribution of $\boldsymbol{\phi}_2$ as $\boldsymbol{\phi}_2 \sim N\left(\mathbf{0}\,,\, [(D-\rho_2 W)\tau_2]^{-1}\right)$, where ρ_1 and ρ_2 are the smoothing parameters associated with the conditional distribution of $\boldsymbol{\phi}_1|\boldsymbol{\phi}_2$ and the marginal distribution of $\boldsymbol{\phi}_2$ respectively, and τ_1 and τ_2 scale the precision of $\boldsymbol{\phi}_1|\boldsymbol{\phi}_2$ and $\boldsymbol{\phi}_2$, respectively. The induced joint distribution will always be proper as long as these two CAR distributions are valid, so the positive definiteness of the covariance matrix in (10.11) is easily verified. If $D = Diag(m_i)$ and W be the adjacency matrix, then the positive definiteness conditions require only that $|\rho_1| < 1$ and $|\rho_2| < 1$. Further restricting these parameters between 0 and 1 $0 < \rho_1 < 1$ avoid negative spatial autocorrelation.

Regarding the A matrix, since $\mathrm{E}(\boldsymbol{\phi}_1 \mid \boldsymbol{\phi}_2) = A\boldsymbol{\phi}_2$, we assume its elements are of the form

$$a_{ij} = \begin{cases} \eta_0 & \text{if } j = i\\ \eta_1 & \text{if } j \in N_i \text{ (i.e., if region } j \text{ is a neighbor of region } i)\\ 0 & \text{otherwise}\end{cases} .$$

Thus $A = \eta_0 I + \eta_1 W$ and $E(\boldsymbol{\phi}_1|\boldsymbol{\phi}_2) = (\eta_0 I + \eta_1 W)\boldsymbol{\phi}_2$. Here η_0 and η_1 are the bridging parameters associating ϕ_{i1} with ϕ_{i2} and ϕ_{j2}, $j \neq i$. One could easily augment A with another bridging parameter η_2 associated with the *second-order* neighbors (neighbors of neighbors) in each region, but we do not pursue this generalization here.) Under these assumptions, the covariance matrix in the joint distribution (10.11) can be written as $\Sigma = \begin{pmatrix}\Sigma_{11} & \Sigma_{12}\\ \Sigma_{12}^T & \Sigma_{22}\end{pmatrix}$, where

$$\begin{aligned}\Sigma_{11} &= [\tau_1(D-\rho_1 W)]^{-1} + (\eta_0 I + \eta_1 W)[\tau_2(D-\rho_2 W)]^{-1}(\eta_0 I + \eta_1 W)\\ \Sigma_{12} &= (\eta_0 I + \eta_1 W)[\tau_2(D-\rho_2 W)]^{-1}\\ \Sigma_{22} &= [\tau_2(D-\rho_2 W)]^{-1} .\end{aligned} \tag{10.12}$$

Jin et al. (2005) denote this new model by $GMCAR(\rho_1, \rho_2, \eta_1, \eta_2, \tau_1, \tau_2)$. This bivariate GMCAR model has the same number of parameters as the twofold CAR model in (10.10), and has one more parameter than the $MCAR(\rho_1, \rho_2, \Lambda)$ model in (10.8).

Setting $\rho_1 = \rho_2 = \rho$ and $\eta_1 = 0$ in (10.12), and using a standard result from matrix theory (Harville, 1997, Corollary 8.5.12), produces the separable precision matrix $\Sigma^{-1} = \Lambda \otimes (D - \rho W)$, where $\tau_1 = \Lambda_{11}$, $\tau_2 = \Lambda_{22} - \frac{\Lambda_{12}^2}{\Lambda_{11}}$, and $\eta_0 = -\frac{\Lambda_{12}}{\Lambda_{11}}$. Further assuming $\rho = 1$ produces an improper MIAR (Multivariate Intrinsic Autoregressive) model. If we assume

$\rho_1 \neq \rho_2$ and $\eta_0 = \eta_1 = 0$, then we ignore dependence between the multivariate components, and the model turns out to be equivalent fitting two separate univariate CAR models. Finally, if we instead assume $\rho_1 = \rho_2 = 0$, $\eta_0 \neq 0$, and $\eta_1 = 0$, the model becomes an i.i.d. bivariate normal model.

The MCAR model in (10.6) has $\mathrm{E}(\boldsymbol{\phi}_1 \mid \boldsymbol{\phi}_2) = -\frac{\Lambda_{12}}{\Lambda_{11}}\boldsymbol{\phi}_2$, which reveals that the conditional mean is merely a scale multiple of $\boldsymbol{\phi}_2$. Since $Var(\boldsymbol{\phi}_1|\boldsymbol{\phi}_2) = [\Lambda_{11}(D - \rho_1 W)]^{-1}$, which is free of $\boldsymbol{\phi}_2$, the distribution of the random variable at a particular site in one field is independent of neighbor variables in another field *given* the value of the related variable at the same area. The extended MCAR model (10.8) has

$$\mathrm{E}(\boldsymbol{\phi}_1 \mid \boldsymbol{\phi}_2) = -\frac{\Lambda_{12}}{\Lambda_{11}}(D - \rho_1 W)^{-\frac{1}{2}}(D - \rho_2 W)^{\frac{1}{2}}\boldsymbol{\phi}_2$$

and $\mathrm{Var}(\boldsymbol{\phi}_1|\boldsymbol{\phi}_2)$ identical to that of model (10.6). Therefore, the distribution of the random variable at a particular site in one field is no longer conditionally independent of neighboring variables in another field. However, this dependence is determined implicitly by ρ_1 and ρ_2 and is difficult to interpret.

By contrast, the GMCAR model has $\mathrm{E}(\boldsymbol{\phi}_1 \mid \boldsymbol{\phi}_2) = (\eta_0 I + \eta_1 W)\boldsymbol{\phi}_2$ and $\mathrm{Var}(\boldsymbol{\phi}_1 \mid \boldsymbol{\phi}_2) = [\tau_1(D - \rho_1 W)]^{-1}$. Thus, while the conditional variance remains free of $\boldsymbol{\phi}_2$, the GMCAR allows spatial information (via the W matrix) to enter the conditional mean in an intuitive way, with a free parameter (η_1) to model the weights. That is, the GMCAR models the conditional mean of $\boldsymbol{\phi}_1$ for a given region as a sensible weighted average of the values of $\boldsymbol{\phi}_2$ for that region *and* a neighborhood of that region.

The GMCAR also allows us to incorporate different weighted adjacency matrices in the $MCAR(\rho, \boldsymbol{\Lambda})$ distribution. Suppose, we wish to extend the precision matrix in model (10.6) to

$$\Sigma^{-1} = \begin{pmatrix} (D_1 - \rho W^{(1)})\Lambda_{11} & (D_3 - \rho W^{(3)})\Lambda_{12} \\ (D_3 - \rho W^{(3)})\Lambda_{12} & (D_2 - \rho W^{(2)})\Lambda_{22} \end{pmatrix} , \tag{10.13}$$

where $D_k = Diag\left(\sum_{j=1}^n W_{1j}^{(k)}, \ldots, \sum_{j=1}^n W_{nj}^{(k)}\right)$ and $W^{(k)}$ is the weighted adjacency matrix with ij-element $W_{ij}^{(k)}, k = 1, 2, 3$, and $i, j = 1, \ldots, n$. The conditions for the precision matrix in (10.13) to be positive definite are less obvious. But in our GMCAR case, we obtain

$$\begin{aligned} \boldsymbol{\phi}_1|\boldsymbol{\phi}_2 &\sim N\left((\eta_0 I + \eta_1 W^{(3)})\boldsymbol{\phi}_2\,,\ [\tau_1(D_1 - \rho_1 W^{(1)})]^{-1}\right) , \\ \text{and } \boldsymbol{\phi}_2 &\sim N\left(0,\ [\tau_2(D_2 - \rho_2 W^{(2)})]^{-1}\right) . \end{aligned}$$

The conditions for positive definiteness can be easily seen to be $|\alpha_1| < 1$ and $|\alpha_2| < 1$ using the fact that diagonally dominant matrices are always positive definite.

Since we specify the joint distribution for a multivariate MRF directly through specification of simpler conditional and marginal distributions, an inherent problem with these methods is that their conditional specification imposes a potentially arbitrary order on the variables being modeled, as they lead to different marginal distributions depending upon the conditioning sequence (i.e., whether to model $p(\boldsymbol{\phi}_1|\boldsymbol{\phi}_2)$ and then $p(\boldsymbol{\phi}_2)$, or $p(\boldsymbol{\phi}_2|\boldsymbol{\phi}_1)$ and then $p(\boldsymbol{\phi}_1)$). This problem is somewhat mitigated in certain (e.g., medical and environmental) contexts where a *natural* order is reasonable, but in many disease mapping contexts this is not the case. Although Jin et al. (2005) suggest using model comparison techniques to decide upon the proper modeling order, since all possible permutations of the variables would need to be considered this seems feasible only with relatively few variables. In any case, the principle of choosing among conditioning sequences using model comparison metrics is perhaps not uncontroversial.

We note that, while the previous theory helps to illuminate structure in specifying multivariate CAR models, from a practical point of view, the reader may simply choose

to implement an analogue of coregionalization to create a multivariate dependence model for areal data. In particular, suppose we assume a common proximity specification for each component of the random effects vector, $\boldsymbol{\phi}$. Then, we could write $\boldsymbol{\phi} = A\boldsymbol{\psi}$ where $\boldsymbol{\psi}_j$, the jth component of $\boldsymbol{\psi}$, is a univariate intrinsic CAR with precision parameter τ_j^2 and each of the component CAR models is independent. As above, we can take A to be lower triangular, with prior specifications discussed earlier. The resulting multivariate CAR model is, of course, improper which is fine as a prior for random effects. Moreover, it is easy to work with since we will only fit the model in the space of the independent CAR models, as we do with the coregionalization model fitting for point-referenced data. We could make the prior proper by making each of the components of $\boldsymbol{\psi}$ proper CAR's, should we wish. Jin, Banerjee and Carlin (2007) develop such coregionalized MCAR distributions, which we discuss in greater detail in Section 10.6. Similar modeling, both in the bivariate and trivariate cases, has been presented in Sang and Gelfand (2009) in the context of modeling temperature extremes (see Section 15.2).

10.4 Modeling using the GMCAR distribution

The $GMCAR(\rho_1, \rho_2, \eta_1, \eta_2, \tau_1, \tau_2)$ models are straightforwardly implemented in a Bayesian framework using MCMC methods. Matters are especially simple with Gaussian likelihoods in the first stage. As a specific example, consider the model

$$Y_{ij} \overset{ind}{\sim} N(Z_{ij}\,,\,\sigma^2)\,,\; i = 1, \ldots, n,\; j = 1, 2\,. \tag{10.14}$$

Assume that $Z_{ij} = \beta_j + \phi_{ij}$, where the ϕ_{ij}'s follow the GMCAR distribution in (10.12). This means that the Z_{ij}'s follow the GMCAR distribution in (10.12) but with $\mathrm{E}[Z_{ij}] = \beta_j$, rather than zero. Then, we easily derive conditional distribution for $\mathbf{Z}_1 \mid \mathbf{Z}_2$

$$\mathbf{Z}_1 \mid \mathbf{Z}_2 \;\sim\; N\left(\beta_1\mathbf{1} + (\eta_0 I + \eta_1 W)(\mathbf{Z}_2 - \beta_2\mathbf{1})\,,\; [\tau_1(D - \rho_1 W)]^{-1}\right)\;,$$

and the marginal distribution $\mathbf{Z}_2 \sim N\left(\beta_2\mathbf{1}\,,\,[\tau_2(D - \rho_2 W)]^{-1}\right)$, where $\mathbf{Z}_1 = (Z_{11}, \ldots, Z_{n1})^T$ and $\mathbf{Z}_2 = (Z_{12}, \ldots, Z_{n2})^T$. Therefore, the joint distribution of $\mathbf{Z}^T = (\mathbf{Z}_1^T, \mathbf{Z}_2^T)$ $p(\mathbf{Z} \mid \boldsymbol{\beta}, \boldsymbol{\tau}, \boldsymbol{\alpha}, \boldsymbol{\eta})$ is proportional to

$$\begin{aligned}
&\tau_1^{\frac{n}{2}} |D - \rho_1 W|^{\frac{1}{2}} \exp\{-\frac{\tau_1}{2}[\mathbf{Z}_1 - \beta_1\mathbf{1} - (\eta_0 I + \eta_1 W)(\mathbf{Z}_2 - \beta_2\mathbf{1})]' \\
&\quad \times (D - \rho_1 W)[\mathbf{Z}_1 - \beta_1\mathbf{1} - (\eta_0 I + \eta_1 W)(\mathbf{Z}_2 - \beta_2\mathbf{1})]\} \\
&\quad \times \tau_2^{\frac{n}{2}} |D - \rho_2 W|^{\frac{1}{2}} \exp\left[-\frac{\tau_2}{2}(\mathbf{Z}_2 - \beta_2\mathbf{1})'(D - \rho_2 W)(\mathbf{Z}_2 - \beta_2\mathbf{1})\right]\;,
\end{aligned} \tag{10.15}$$

where $\boldsymbol{\beta} = (\beta_1, \beta_2)$, $\boldsymbol{\tau} = (\tau_1, \tau_2)$, $\boldsymbol{\eta} = (\eta_0, \eta_1)$, and $\boldsymbol{\alpha} = (\rho_1, \rho_2)$.

The joint posterior distribution $p\left(\boldsymbol{\beta}, \sigma^2, \mathbf{Z}, \boldsymbol{\tau}, \boldsymbol{\alpha}, \boldsymbol{\eta} \mid \mathbf{Y}_1, \mathbf{Y}_2\right)$ is proportional to

$$L\left(\mathbf{Y}_1, \mathbf{Y}_2 \mid \mathbf{Z}, \sigma^2\right) p\left(\mathbf{Z} \mid \boldsymbol{\beta}, \boldsymbol{\tau}, \boldsymbol{\alpha}, \boldsymbol{\eta}\right) p(\boldsymbol{\beta}) p(\boldsymbol{\tau}) p(\boldsymbol{\alpha}) p(\boldsymbol{\eta}) p(\sigma^2)\;, \tag{10.16}$$

where $\mathbf{Y}_1 = (Y_{11}, \ldots, Y_{n1})^T$, $\mathbf{Y}_2 = (Y_{12}, \ldots, Y_{n2})^T$, $L\left(\mathbf{Y}_1, \mathbf{Y}_2 \mid \mathbf{Z}, \sigma^2\right)$ is the likelihood

$$\sigma^{-2n} \exp\left\{-\frac{1}{2\sigma^2}[(\mathbf{Y}_1 - \mathbf{Z}_1)'(\mathbf{Y}_1 - \mathbf{Z}_1) + (\mathbf{Y}_2 - \mathbf{Z}_2)'(\mathbf{Y}_2 - \mathbf{Z}_2)]\right\}\;,$$

$p\left(\mathbf{Z} \mid \boldsymbol{\beta}, \boldsymbol{\tau}, \boldsymbol{\alpha}, \boldsymbol{\eta}\right)$ is given by (10.15), and the remaining terms in (10.25) are the prior distributions on $(\boldsymbol{\beta}, \boldsymbol{\tau}, \boldsymbol{\alpha}, \boldsymbol{\eta}, \sigma^2)$.

For the remaining terms, flat priors are chosen for β_1 and β_2, while σ^2 is assigned a vague inverse gamma prior, i.e., a $IG(1, 0.1)$ where we parametrize the $IG(a, b)$ so that $E(\sigma^2) = b/(a-1)$. Next, τ_1 and τ_2 are assigned vague gamma priors, specifically a $G(1, 0.1)$,

which has mean 10 and variance 100. Finally, ρ_1, ρ_2 are given $Unif(0,1)$ priors while η_0 and η_1 are given $N(0,\sigma_1^2)$ and $N(0,\sigma_2^2)$ priors, respectively. Jin et al. (2005) present several simulation studies using these specifications and report robust inference for varying hyperparameter values.

The Gibbs sampler is natural for updating the parameters in this setting because it can take advantage of the conditional specification of the GMCAR model. Each of the full conditional distributions required by the Gibbs sampler must be proportional to (10.25). Furthermore, no matrix inversion is required and only calculations on rather special (e.g., diagonal) n-dimensional matrices are required, regardless of the dimension p ($p = 2$ in our case). To calculate the determinant in (10.15), we have the fact that

$$\begin{aligned} |D - \rho_k W| &= |D^{\frac{1}{2}}(1 - \rho_k D^{-\frac{1}{2}} W D^{-\frac{1}{2}})D^{\frac{1}{2}}| = |D| \prod_{i=1}^{n}(1 - \rho_k \lambda_i) \\ &\propto \prod_{i=1}^{n}(1 - \rho_k \lambda_i)\ ,\ k = 1, 2\ , \end{aligned}$$

where $\lambda_i, i = 1, \ldots n$ are the eigenvalues of the matrix $D^{-\frac{1}{2}} W D^{-\frac{1}{2}}$. The λ_i may be calculated prior to any MCMC iteration. Hence posterior computation for the GMCAR model is simpler and faster than that for existing MCAR models, especially for large areal data sets.

All of the parameters in (10.25) except $\boldsymbol{\eta}$ and $\boldsymbol{\alpha}$ have closed-form full conditionals, and so may be directly updated. For these two remaining parameters, Metropolis-Hastings steps with bivariate Gaussian proposals are convenient (though for $\boldsymbol{\alpha}$, a preliminary logit transformation, having Jacobian $\prod_{k=1}^{2} \rho_k(1 - \rho_k)$, is required). In practice, the ρ_k must be bounded away from 1 (say, by insisting $0 < \rho_k < 0.999$, $k = 1, 2$) to maintain identifiability and hence computational stability.

10.5 Illustration: Fitting conditional GMCAR to Minnesota cancer data

We now use GMCAR distributions as specifications for second-stage random effects in a hierarchical areal data model with a non-Gaussian first stage. Following Jin et al. (2005), we consider modeling the numbers of deaths due to cancers of the lung and esophagus in the years from 1991 to 1998 at the county level in Minnesota. We write the model as

$$Y_{ij} \overset{ind}{\sim} Poisson(E_{ij} e^{Z_{ij}})\ ,\ i = 1, \ldots, 87,\ j = 1, 2\ , \tag{10.17}$$

where Y_{ij} is the observed number of deaths due to cancer j in county i, and E_{ij} is the corresponding expected number of deaths (assumed known). To calculate E_{ij}, we account for each county's age distribution by calculating the expected *age-adjusted* number of deaths due to cancer j in county i as

$$E_{ij} = \sum_{k=1}^{m} \omega_j^k N_i^k\ , i = 1, \ldots, 87,\ j = 1, 2\ ,$$

where $\omega_j^k = (\sum_{i=1}^{87} D_{ij}^k)/(\sum_{i=1}^{87} N_i^k)$ is the age-specific death rate due to cancer j for age group k over all Minnesota counties, D_{ij}^k is the number of deaths in age group k of county i due to cancer j, and N_i^k is the total population at risk in county i, age group k. The GMCAR models can be implemented in `BUGS` (see `www.biostat.umn.edu/~brad/software.html` for the code and the data.).

The county-level maps of the raw standardized mortality ratios (i.e., $\text{SMR}_{ij} = Y_{ij}/E_{ij}$) shown in Figure 10.1 exhibit evidence of correlation both across space and between cancers, motivating use of our proposed GMCAR models. Regarding the selection of the proper order

Figure 10.1 *Maps of raw standard mortality ratios (SMR) of lung and esophagus cancer in Minnesota.*

Figure 10.2 *Exploratory plot to help select modeling order: (a) [lung|esophagus], sample correlation 0.394, regression* $t = 3.956$*; (b) [esophagus|lung], sample correlation 0.193, regression* $t = 1.813$.

in which to model the two cancers, Figure 10.2 gives a helpful data-based exploratory plot. We first obtain crude data-based estimates of the spatial random effects as $\hat{\phi}_{i1} = \log(\text{SMR}_{i1})$ and $\hat{\phi}_{i2} = \log(\text{SMR}_{i2})$. Next, recall the linearity of the conditional GMCAR mean for a given ordering (say, lung given esophagus), i.e.,

$$E(\boldsymbol{\phi}_1|\boldsymbol{\phi}_2) = A\boldsymbol{\phi}_2 = A(\eta_0, \eta_1)\boldsymbol{\phi}_2 = (\eta_0 I + \eta_1 W)\boldsymbol{\phi}_2 \ .$$

This motivates obtaining least-squares estimates $\hat{\eta}_0$ and $\hat{\eta}_1$ by minimizing $(\hat{\boldsymbol{\phi}}_1 - A(\eta_0, \eta_1)\hat{\boldsymbol{\phi}}_2)'(\hat{\boldsymbol{\phi}}_1 - A(\eta_0, \eta_1)\hat{\boldsymbol{\phi}}_2)$ as a function of η_0 and η_1. Finally, we plot $A(\hat{\eta}_0, \hat{\eta}_1)\hat{\boldsymbol{\phi}}_2$ versus $\hat{\boldsymbol{\phi}}_1$, and investigate how well the linearity assumption is supported by the data. Repeating this entire process for the reverse order (here, esophagus given lung) produces a second plot, which may be compared in quality to the first. In our case, Figure 10.2(a) (lung given esophagus) indicates more support for linearity, both in its appearance and in its higher sample correlation and regression t statistic.

Using the likelihood in (10.17), we model the random effects Z_{ij} using the $GMCAR(\rho_1, \rho_2, \eta_1, \eta_2, \tau_1, \tau_2)$ with mean $\boldsymbol{\beta}$. In what follows we compare the GMCAR with other existing MCAR models using DIC. In Table 10.3, Models 1–3 are members of our proposed GMCAR class. Specifically, in Model 1, we have the full model with all six parameters, and the conditioning order of the cancers is [lung | esophagus]. Model 2 assumes $\eta_1 = 0$ and uses the same conditioning order as Model 1. In Model 3, we switch the conditioning order to [esophagus | lung] and return to a full model. To compare the GMCAR to existing MCAR models, we take the $MCAR(\rho_1, \rho_2, \Lambda)$ using the Cholesky method for the U_k as Model 4, the same model but using the spectral decomposition for the U_k as Model 5, and the $2fCAR(\rho_0, \rho_1, \rho_2, \rho_3, \tau_1, \tau_2)$ as Model 6. We choose the prior distributions for each parameter as discussed in Section 10.4, and use Metropolis-Hastings and Gibbs sampling to

	model	$\overline{D}$	p_D	DIC
1	GMCAR (full)	483.4	58.2	541.6
2	GMCAR (reduced; $\eta_1 = 0$)	483.0	63.8	546.8
3	GMCAR (full, reverse order)	480.6	63.3	543.9
4	MCAR (Cholesky decomposition)	483.6	61.3	544.9
5	MCAR (spectral decomposition)	483.8	60.6	544.4
6	2fCAR	482.6	65.1	547.7

Table 10.3 *Model comparison using DIC statistics, Minnesota cancer data analysis.*

update all parameters. We use 5,000 iterations as the pre-convergence burn-in period, and then a further 20,000 iterations as our production run for posterior summarization.

Fit measures $\overline{D}$, effective numbers of parameters p_D, and DIC scores for each model are seen in Table 10.3. Model 1 has the smallest p_D and DIC values, so our $GMCAR(\rho_1, \rho_2, \eta_0, \eta_1, \tau_1, \tau_2)$ full model with the conditioning order [lung | esophagus] emerges as best for this data set. The reduced GMCAR Model 2 does less well, suggesting the need to account for bivariate spatial structure in these data. The two MCAR methods perform similarly to each other and to the reduced GMCAR model, while the 2fCAR model does less well, largely because it does not seem to allow sufficient smoothing of the random effects (larger p_D score). Note that effective degrees of freedom may actually be smaller for apparently more complex models that allow more complicated forms of shrinkage, such as Model 1 in this case. We note that our "focus" parameter is the same for each model (both fixed and random effects are in focus), and the Poisson likelihood is also not changing across models. Also, our priors are all noninformative or quite vague (e.g., uniform priors for all ρ parameters). All of this suggests the DIC comparison in Table 10.3 is fair across models. Moreover, the resulting DIC scores were robust to the moderate changes in the prior distributions.

Regarding estimation of the fixed effects, under Model 1 we obtained point and 95% equal-tail interval estimates of 0.602 and (0.0267, 0.979) for ρ_1, and 0.699 and (0.0802, 0.973) for ρ_2. Recall these are spatial association parameters, but while their values are between 0 and 1 they are not "correlations" in the usual sense; the moderate point estimates and wide confidence intervals suggest a relatively modest degree of spatial association in the random effects. It is also important to remember that in this setup, ρ_2 measures spatial association in the esophagus random effects $\boldsymbol{\phi}_2$, while ρ_1 measures spatial association in the lung random effects $\boldsymbol{\phi}_1$ *given* the esophagus random effects $\boldsymbol{\phi}_2$. Thus the interpretation of the ρ_k would be different for Model 3 (due to the different conditioning order), and much different for Model 4 or 5. Note that for the MCAR model, $E(\boldsymbol{\phi}_1|\boldsymbol{\phi}_2)$ and $E(\boldsymbol{\phi}_2|\boldsymbol{\phi}_1)$ both depend on both ρ_1 and ρ_2. But for the GMCAR, $E(\boldsymbol{\phi}_1|\boldsymbol{\phi}_2)$ is free of both ρ_1 and ρ_2, while of course $E(\boldsymbol{\phi}_2) = 0$. Thus for this model, ρ_1 and ρ_2 unambiguously control only their corresponding variance matrices, and can be set without altering the mean structure.

Turning to τ_1 and τ_2, under Model 1 we obtained 32.65, (16.98, 66.71) and 13.73, (4.73, 38.05) as our point and interval estimates, respectively. Since these parameters measure spatial precision for each disease, they suggest slightly more variability in the esophagus random effects, although again comparison is difficult here since τ_2 is a *marginal* precision for $\boldsymbol{\phi}_2$ while τ_1 is a *conditional* precision for $\boldsymbol{\phi}_1$ given $\boldsymbol{\phi}_2$. Along these lines, Figure 10.3 shows estimated posteriors of the conditional variances $\sigma_1^2 = 1/\tau_1$ for several candidate multivariate spatial models. Panel (a) shows the situation for two separate CAR models, a model that ignores any possibility of connection between the cancers. The remaining panels consider the $MCAR(\rho_1, \rho_2, \boldsymbol{\Lambda})$ model, the reduced $GMCAR(\rho_1, \rho_2, \eta_0, \tau_1, \tau_2)$ model, and the full $GMCAR(\rho_1, \rho_2, \eta_0, \eta_1, \tau_1, \tau_2)$ model. The reduction of uncertainty in $\boldsymbol{\phi}_1$ given $\boldsymbol{\phi}_2$ in these more complex models is a measure of the information content between the cancers, and is readily apparent from the histograms and their empirical means.

Figure 10.3 *Posterior samples of conditional variances* $\sigma_1^2 = 1/\tau_1$ *for various models: (a) two separate CAR models; (b) MCAR model; (c) reduced GMCAR model; (d) full GMCAR model.*

Figure 10.4 *Posterior samples of* η_0 *and* η_1 *using the full GMCAR model with two conditioning orders: (a) estimated posterior for* η_0*, [lung | esophagus]; (b) estimated posterior for* η_1*, [lung | esophagus]; (c) estimated posterior for* η_0*, [esophagus | lung]; (d) estimated posterior for* η_1*, [esophagus | lung].*

DIC's slight preference for Model 1 is consistent with the estimated posteriors of the linking parameters η_0 and η_1 shown in Figure 10.4. The inclusion of 0 within the 95% credible interval for η_1 under the reverse ordering, but not under the natural ordering, is yet further evidence against the former. Note also that the linking parameters η_0 and η_1 have mostly positive support, meaning that the two cancers have positive spatial correlation. This is also evident from the maps of the posterior means of the SMRs for the two cancers under the full model shown in Figure 10.5. Clearly incidence of the two cancers is strongly correlated, with higher fitted ratios extending from the Twin Cities metro area (eastern side, about one third of the way up) to the mining- and tourism-oriented north and northeast, regions where conventional wisdom suggests cigarette smoking may be more common.

Figure 10.5 *Maps of posterior means of standardized mortality ratios (SMR) of lung and esophagus cancer in Minnesota from the full GMCAR model with conditioning order [lung | esophagus].*

10.6 Coregionalized MCAR distributions

As mentioned in Section 10.3, generalizations of GMCAR to settings with a large number of diseases is encumbered by the dependence of the joint distribution on the sequence of ordering in the hierarachy. To obviate this issue, Jin, Banerjee and Carlin (2007) develop an order-free framework for multivariate areal modeling that allows versatile spatial structures, yet is computationally feasible for many outcomes. This approach is based upon an adaptation of the *linear model of coregionalization* (LMC) to areal data.

The essential idea is to develop richer spatial association models using linear transformations of much simpler spatial distributions. The objective is to allow explicit smoothing of cross-covariances while at the same time not being hampered by conditional ordering. The most natural model here would parametrize the cross-covariances themselves as $D - \gamma_{ij}W$, instead of using the U_k's as in (10.6). Unfortunately, except in the separable model with only one smoothing parameter ρ, constructing such dispersion structures is not trivial and leads to identifiability issues on the γ's (see, e.g., Gelfand and Vonatsou, 2003). Kim et al. (2001) resolve these identifiability issues in the bivariate setting using diagonal dominance, but recognize the difficulty in extending this to the multivariate setting. We address this problem using a *linear model of coregionalization* (LMC).

It is worth pointing out that our use of the LMC here is somewhat different from what is usually encountered in geostatistics. In geostatistics we typically transform independent latent effects, which suffices in meeting the primary goal of introducing a different spatial range for each variable. This is akin to introducing different smoothing parameters for each variable and indeed, as we show below in Section 10.6.2, independent latent effects produce the $MCAR(\alpha_1, \ldots, \alpha_p; \Lambda)$ in (10.9).

However, to explicitly smooth the cross-covariances with identifiable parameters, we will relax the independence of latent effects. Still, in our ensuing parametrization, we are able to derive conditions that yield valid joint distributions. To be precise, let $\boldsymbol{\phi} = (\boldsymbol{\phi}_1^T, \ldots, \boldsymbol{\phi}_p^T)^T$ be an $np \times 1$ vector, where each $\boldsymbol{\phi}_j = (\phi_{1j}, \ldots, \phi_{nj})^T$ is $n \times 1$ representing the spatial effects corresponding to disease j. We can write $\boldsymbol{\phi} = (A \otimes I_{n\times n})\mathbf{u}$, where $\mathbf{u} = (\mathbf{u}_1^T, \ldots, \mathbf{u}_p^T)^T$ is $np \times 1$ with each $\mathbf{u}_j$ being an $n \times 1$ areal process. Indeed, a proper distribution for $\mathbf{u}$ ensures a proper distribution for $\boldsymbol{\phi}$ subject only to the non-singularity of A. The flexibility of this approach is apparent: we obtain different multivariate lattice models with rich spatial covariance structures by making different assumptions about the p spatial processes $\mathbf{u}_j$.

10.6.1 Case 1: Independent and identical latent CAR variables

First, we will assume that the random spatial processes $\mathbf{u}_j$, $j = 1, \ldots, p$, are independent and identical. Since each spatial process $\mathbf{u}_j$ is a univariate process over areal units, we might

adopt a CAR structure for each of them, that is

$$\mathbf{u}_j \sim N_n\left(\mathbf{0},\ (D-\alpha W)^{-1}\right),\ j=1,\ldots,p\,. \tag{10.18}$$

Since the $\mathbf{u}_j$'s are independent of each other, the joint distribution of $\mathbf{u}=(\mathbf{u}_1',\ldots,\mathbf{u}_p')'$ is $\mathbf{u}\sim N_{np}\left(\mathbf{0},\ I_{p\times p}\otimes(D-\alpha W)^{-1}\right)$. The joint distribution of $\boldsymbol{\phi}=(A\otimes I_{n\times n})\mathbf{u}$ is

$$\boldsymbol{\phi} \sim N_{np}\left(\mathbf{0},\ \Sigma\otimes(D-\alpha W)^{-1}\right)\,, \tag{10.19}$$

defining $\Sigma=AA^T$. We denote the distribution in (10.19) by $MCAR(\alpha,\ \Sigma)$. Note that the joint distribution of (10.19) is identifiable up to $\Sigma=AA'$, and is independent of the choice of A. Thus, without loss of generality, we can specify the matrix A as the upper-triangular Cholesky decomposition of Σ.

Since $\boldsymbol{\phi}=(A\otimes I_{n\times n})\mathbf{u}$, a valid joint distribution of $\boldsymbol{\phi}$ requires valid joint distributions of the $\mathbf{u}_j$, which happens if and only if $\frac{1}{\xi_{min}}<\alpha<\frac{1}{\xi_{max}}$, where ξ_{min} and ξ_{max} are the minimum and maximum eigenvalues of $D^{-\frac{1}{2}}WD^{-\frac{1}{2}}$. Note if $\alpha=1$ in CAR structure (10.18), which is an ICAR, the joint distribution of $\boldsymbol{\phi}$ in (10.19) becomes the multivariate intrinsic CAR (Gelfand and Vounatsou, 2003).

Currently, `BUGS` offers an implementation of the $MCAR(\alpha=1,\ \Sigma)$ distribution (using its `mv.car` distribution), but not the $MCAR(\alpha,\ \Sigma)$. However through the LMC approach we still can fit the $MCAR(\alpha,\ \Sigma)$ in `BUGS` by writing $\boldsymbol{\phi}=(A\otimes I_{n\times n})\mathbf{u}$ and assigning proper CAR priors (via the `car.proper` distribution) for each $\mathbf{u}_j$, $j=1,\ldots,p$ with a common smoothing parameter α. Regarding the prior on A, note that since $AA'=\Sigma$ and A is the Cholesky decomposition of Σ, there is a one-to-one relationship between the elements of Σ and A. In Section 10.7, we argue that assigning a prior to Σ is computationally preferable.

10.6.2 Case 2: Independent but not identical latent CAR variables

Here, we continue to assume that the $\mathbf{u}_j$ are independent, but relax them being identically distributed. Adopting the CAR structure, we assume

$$\mathbf{u}_j \sim N_n\left(\mathbf{0},\ (D-\alpha_j W)^{-1}\right),\ j=1,\ldots,p\,, \tag{10.20}$$

where α_j is the smoothing parameter for the jth spatial process. Since the $\mathbf{u}_j$'s are independent of each other and $\boldsymbol{\phi}=(A\otimes I_{n\times n})\mathbf{u}$, the joint distribution of $\boldsymbol{\phi}$ is

$$\boldsymbol{\phi} \sim N_{np}\left(\mathbf{0},\ (A\otimes I_{n\times n})\Gamma^{-1}(A\otimes I_{n\times n})^T\right), \tag{10.21}$$

where $\Sigma=AA^T$ and Γ is an $np\times np$ block diagonal matrix with $n\times n$ diagonal entries $\Gamma_j=D-\alpha_j W, j=1,\ldots,p$. We denote the distribution in (10.21) by $MCAR(\alpha_1,\ldots,\alpha_p,\Sigma)$.

It follows from (10.21) that different joint distributions of $\boldsymbol{\phi}$ having different covariance matrices emerge under different linear transformation matrices A. To ensure A is identifiable, we could again specify it to be the upper-triangular Cholesky decomposition of Σ, although this might not be the best choice computationally. Through the LMC approach in this case, the distribution in (10.21) is similar to the $MCAR(\alpha_1,\ldots,\alpha_p,\ \Lambda)$ structure (10.9), developed in Carlin and Banerjee (2003) and Gelfand and Vounatsou (2003). All of these have the same number of parameters, and there is no unique joint distribution for $\boldsymbol{\phi}$ with the $MCAR(\alpha_1,\ldots,\alpha_p,\ \Lambda)$ structure, since there is not a unique R_j matrix such that $R_jR_j^T=R_jPP^TR_j^T=D-\alpha_jW$ (P being an arbitrary orthogonal matrix).

Again, a valid joint distribution in (10.21) requires p valid distributions for $\mathbf{u}_j$, i.e., $\frac{1}{\xi_{min}}<\alpha_j<\frac{1}{\xi_{max}}$, $j=1,\ldots,p$. Through the LMC approach, we can also fit the data

with the $MCAR(\alpha_1, \ldots, \alpha_p, \Sigma)$ prior distribution (10.21) on $\boldsymbol{\phi}$ in `WinBUGS` as in the previous subsection by writing $\boldsymbol{\phi} = (A \otimes I_{n\times n})\mathbf{u}$ and assigning proper CAR priors (via the `car.proper` distribution) with a distinct smoothing parameter α_j for each $\mathbf{u}_j,\ j = 1, \ldots, p$. As mentioned in the preceding section, we assign a prior to $AA^T = \Sigma$ (e.g., an inverse Wishart), and determine A from the one-to-one relationship between the elements of Σ and A; Section 10.7 below provides details.

10.6.3 Case 3: Dependent and not identical latent CAR variables

Finally, in this case we will assume that the random spatial processes $\mathbf{u}_j = (u_{1j}, \ldots, u_{nj})^T,\ j = 1, \ldots, p$ are neither independent nor identically distributed. We now assume that u_{ij} and $u_{i,\, l\neq j}$ are independent given $u_{k\neq i,\, j}$ and $u_{k\neq i,\, l\neq j}$, where $l, j = 1, \ldots, p$ and $i, k = 1, \ldots, n$ implying that latent effects for different diseases in the same region are conditionally independent given those for diseases in the neighboring regions. Based upon the Markov property and similar to the conditional distribution in the univariate case, we specify the ij^{th} conditional distribution as Gaussian with mean

$$E(u_{ij} \mid u_{k\neq i,\, j},\ u_{i,\, l\neq j},\ u_{k\neq i,\, l\neq j}) = b_{jj}\left(\sum_{k\sim i} u_{kj}/m_i\right) + \sum_{l\neq j}\left[b_{jl}\left(\sum_{k\sim i} u_{kl}/m_i\right)\right] ,$$

and conditional variance $Var(u_{ij} \mid u_{k\neq i,\, j},\ u_{i,\, l\neq j},\ u_{k\neq i,\, l\neq j}) \propto 1/m_i$, where b_{jj} denotes the spatial autocorrelation for the random spatial process $\mathbf{u}_j$ while b_{jl} $(l \neq j,\ \ l, j = 1, \ldots, p)$ denotes the cross-spatial correlation between the random spatial process $\mathbf{u}_j$ and $\mathbf{u}_l$. Putting these conditional distributions together reveals the joint distribution of $\mathbf{u} = (\mathbf{u}_1^T, \ldots, \mathbf{u}_p^T)^T$ to be

$$\mathbf{u} \ \sim\ N_{np}\left(\mathbf{0},\ (I_{p\times p} \otimes D - B \otimes W)^{-1}\right) , \tag{10.22}$$

where I is a $p \times p$ identity matrix and B is a $p \times p$ symmetric matrix with the elements b_{jl}, $j, l = 1, \ldots, p$. As long as the dispersion matrix in (10.22) is positive definite, which boils down to $(I_{p\times p} \otimes D - B \otimes W)$ being positive definite, (10.22) is itself a valid model. To assess non-singularity, note $I_{p\times p} \otimes D - B \otimes W = (I_{p\times p} \otimes D)^{\frac{1}{2}}\left(I_{pn\times pn} - B \otimes D^{-\frac{1}{2}} W D^{-\frac{1}{2}}\right)(I_{p\times p} \otimes D)^{\frac{1}{2}}$. Denoting the eigenvalues for $D^{-\frac{1}{2}} W D^{-\frac{1}{2}}$ as $\xi_i,\ i = 1, \ldots, n$, and the eigenvalues for B as $\zeta_j,\ j = 1, \ldots, p$, one finds (see, e.g., Harville, 1997, Theorem 21.11.1) the eigenvalues for $B \otimes (D^{-\frac{1}{2}} W D^{-\frac{1}{2}})$ as $\xi_i \times \zeta_j,\ \ i = 1, \ldots, n,\ j = 1, \ldots, p$. Hence, the conditions for $I_{p\times p} \otimes D - B \otimes W$ being positive definite become $\xi_i \zeta_j < 1$, *i.e.*, $\frac{1}{\xi_{min}} < \zeta_j < \frac{1}{\xi_{max}},\ \ i = 1, \ldots, n,\ j = 1, \ldots, p$, where ξ_{min} and ξ_{max} are the minimum and maximum eigenvalues of $D^{-\frac{1}{2}} W D^{-\frac{1}{2}}$. Thus, $\frac{1}{\xi_{min}} < \zeta_j < 1,\ j = 1, \ldots, p$, ensures the positive definiteness of the matrix $I_{p\times p} \otimes D - B \otimes W$ and, hence, the validity of the distribution of $\mathbf{u}$ given in (10.22). In fact, $\xi_{max} = 1$ and $\xi_{min} < 0$, which makes this formulation easier to work with in practice (e.g., in choosing priors; see Section 10.7) than the alternative parametrization $\frac{1}{\zeta_{min}} < \xi_j < \frac{1}{\zeta_{max}}$.

The model in (10.22) introduces smoothing parameters in the cross-covariance structure through the matrix B, but unlike the $MCAR$ models in Sections 10.6.1 and 10.6.2 does not have the Σ matrix to capture non-spatial variances. To remedy this, we model $\boldsymbol{\phi} = (A \otimes I_{n\times n})\mathbf{u}$ so that the joint distribution for the random effects $\boldsymbol{\phi}$ is

$$\boldsymbol{\phi} \ \sim\ N_{np}\left(\mathbf{0},\ (A \otimes I_{n\times n})\,(I_{p\times p} \otimes D - B \otimes W)^{-1}\,(A \otimes I_{n\times n})^T\right) . \tag{10.23}$$

Since $\boldsymbol{\phi} = (A \otimes I_{n\times n})\mathbf{u}$, it is immediately that the validity of (10.22) ensures a valid joint distribution for (10.23). We denote distribution (10.23) by $MCAR(B, \Sigma)$, where $\Sigma = AA^T$. Again, A identifies with the upper-triangular Cholesky square-root of Σ. Note that with $\Sigma = I$ we recover (10.22), which we henceforth denote as $MCAR(B,\ I)$.

To see the generality of (10.23), we find the joint distribution of $\boldsymbol{\phi}$ reduces to the $MCAR(\alpha_1, \ldots, \alpha_p, \Sigma)$ distribution (10.21) if $b_{jl} = 0$ and $b_{jj} = \alpha_j$, or the $MCAR(\alpha, \Sigma)$ distribution (10.19) if $b_{jl} = 0$ and $b_{jj} = \alpha$, in both cases for $j, l = 1, \ldots, p$. Also note that the distribution in (10.23) is invariant to orthogonal transformations (up to a reparametrization of B) in the following sense: let $T = AP$ with P being a $p \times p$ orthogonal matrix such that $TT^T = APP^TA^T = \Sigma$. Then the covariance matrix in (10.23) can be expressed as $(A \otimes I_{n\times n})\,(I_{p\times p} \otimes D - B \otimes W)^{-1}\,(A \otimes I_{n\times n})^T = (T \otimes I_{n\times n})\,(I_{p\times p} \otimes D - C \otimes W)^{-1}\,(T \otimes I_{n\times n})^T$, where $C = P^TBP$. Without loss of generality, then, we can choose the matrix A as the upper-triangular Cholesky decomposition of Σ.

To understand the features of the $MCAR(B, \Sigma)$ distribution (10.23), we illustrate in the bivariate case ($p = 2$). Define

$$(AA^T)^{-1} = \Sigma^{-1} = \begin{pmatrix} \Lambda_{11} & \Lambda_{12} \\ \Lambda_{12} & \Lambda_{22} \end{pmatrix}$$

and $B = A^T \begin{pmatrix} \gamma_1\Lambda_{11} & \gamma_{12}\Lambda_{12} \\ \gamma_{12}\Lambda_{12} & \gamma_2\Lambda_{22} \end{pmatrix} A$, where $A = \begin{pmatrix} a_{11} & a_{12} \\ 0 & a_{22} \end{pmatrix}$. For convenience, we will denote the entries of B as b_{ij}. Note that the γ's are not identifiable from the matrix Λ and our reparametrization in terms of B must be used to conduct posterior inference on B and Λ (see Section 10.7), from which the cross-covariances may be recovered. The above expression does allow the $MCAR(B, \Sigma)$ distribution (10.23) to be rewritten as

$$\boldsymbol{\phi} \;\sim\; N_{2n}\left(\mathbf{0},\; \begin{pmatrix} (D-\gamma_1 W)\Lambda_{11} & (D-\gamma_{12}W)\Lambda_{12} \\ (D-\gamma_{12}W)\Lambda_{12} & (D-\gamma_2 W)\Lambda_{22} \end{pmatrix}^{-1}\right), \tag{10.24}$$

which is precisely the general dispersion structure we set out to achieve. Jin et al. (2007) provide explicit expressions for the conditional means and variances, which offer further insight into how the parameters in (10.24) affect smoothing.

10.7 Modeling with coregionalized MCAR's

The $MCAR(B, \Sigma)$ model is straightforwardly implemented in a Bayesian framework using MCMC methods. As in Section 10.6.3, we write $\boldsymbol{\phi} = (A \otimes I_{n\times n})\mathbf{u}$, where $\mathbf{u} = (\mathbf{u}_1^T, \mathbf{u}_2^T)$ and $\mathbf{u}_j = (u_{1j}, \ldots, u_{nj})^T$. The joint posterior distribution is $p\left(\boldsymbol{\beta}, \sigma^2, \mathbf{u}, A, B \mid \mathbf{Y}_1, \mathbf{Y}_2\right)$, which is proportional to

$$L\left(\mathbf{Y}_1, \mathbf{Y}_2 \mid \mathbf{u}, \sigma^2, A\right) p\left(\mathbf{u} \mid B\right) p(B) p(\boldsymbol{\beta}) p(A) p(\sigma^2)\;, \tag{10.25}$$

where $\mathbf{Y}_1 = (Y_{11}, \ldots, Y_{n1})^T$ and $\mathbf{Y}_2 = (Y_{12}, \ldots, Y_{n2})^T$, $L\left(\mathbf{Y}_1, \mathbf{Y}_2 \mid \mathbf{u}, \sigma^2, A\right)$ is the data likelihood and $p\left(\mathbf{u} \mid B\right) = N_{np}\left(\mathbf{0},\; (I_{p\times p} \otimes D - B \otimes W)^{-1}\right)$. As mentioned in Section 10.6.3, propriety of this distribution requires the eigenvalues ζ_j of B to satisfy $\frac{1}{\xi_{min}} < \zeta_j < 1$ ($j = 1, \ldots, p$). When p is large, it is hard to determine the intervals over the elements of B that result in $\frac{1}{\xi_{min}} < \zeta_j < 1$, and thus designing priors for B that guarantee this condition is awkward. In principle, one might impose the constraint numerically by assigning a flat prior or a normal prior with a large variance for the elements of B, and then simply check whether the eigenvalues of the corresponding B matrix are in that range during a random-walk Metropolis-Hastings (MH) update. If the resulting eigenvalues are out of range, the values are thrown out since they correspond to prior probability 0; otherwise we perform the standard MH comparison step. In our experience, however, this does not work well, especially when p is large.

Instead, here we outline a different strategy to update the matrix B. Our approach is to represent B using the spectral decomposition, which we write as $B = P\Delta P^T$, where P is the corresponding orthogonal matrix of eigenvectors and Δ is a diagonal matrix of

ordered eigenvalues, $\zeta_1, \ldots, \zeta_p$. We parameterize the $p \times p$ orthogonal matrix P in terms of the $p(p-1)/2$ Givens angles θ_{ij} for $i = 1, \ldots, p-1$ and $j = i+1, \ldots, p$ (Daniels and Kass, 1999). The matrix P is written as the product of $p(p-1)/2$ matrices, each one associated with a Givens angle. Specifically, $P = G_{12}G_{13} \ldots G_{1p} \ldots G_{(p-1)p}$ where i and j are distinct and G_{ij} is the $p \times p$ identity matrix with the ith and jth diagonal elements replaced by $\cos(\theta_{ij})$, and the (i,j)-th and (j,i)-th elements replaced by $\pm \sin(\theta_{ij})$, respectively. Since the Givens angles θ_{ij} are unique with a domain $(-\pi/2, \pi/2)$ and the eigenvalues ζ_j of B are in the range $(\frac{1}{\xi_{min}}, 1)$, we then put a Uniform$(-\pi/2, \pi/2)$ prior on the θ_{ij} and a Uniform$(\frac{1}{\xi_{min}}, 1)$ prior on the ζ_j. To update θ_{ij}'s or ζ_j's using random-walk Metropolis-Hastings steps with Gaussian proposals, we need to transform them to have support equal to the whole real line. A straightforward solution here is to use $g(\theta_{ij}) = \log(\frac{\pi/2+\theta_{ij}}{\pi/2-\theta_{ij}})$, a transformation having Jacobian $\prod_{i=1}^{p-1} \prod_{j=i+1}^{p} (\pi/2 + \theta_{ij})(\pi/2 - \theta_{ij})$. In practice, the ζ_j must be bounded away from 1 (say, by insisting $\frac{1}{\xi_{min}} < \zeta_j < 0.999$, $j = 1, \ldots, p$) to maintain identifiability and hence computational stability. In fact, with our approach it is also easy to calculate the determinant of the precision matrix, that is, $|I_{p\times p} \otimes D - B \otimes W| \propto \prod_{i=1}^{n} \prod_{j=1}^{p} (1 - \xi_i \zeta_j)$, where ξ_i are the eigenvalues of $D^{-\frac{1}{2}} W D^{-\frac{1}{2}}$, which can be calculated prior to any MCMC iteration. For the special case of the $MCAR(\alpha_1, \ldots, \alpha_p, \Sigma)$ models, one could assign each $\alpha_i \sim U(0,1)$, which would be sufficient to ensure a valid model (e.g. Carlin and Banerjee, 2002). However, we also investigated with more informative priors on the α_i's such as the $Beta(2, 18)$ that centers the smoothing parameters closer to 1 and leads to greater smoothing.

With respect to the prior distribution $p(A)$ on the right hand side of (10.25), we can put independent priors on the individual elements of A, such as inverse gamma for the square of the diagonal elements of A and normal for the off-diagonal elements. In practice, we cannot assign non-informative priors here, since then MCMC convergence is poor. In our experience it is easier to assign a vague (i.e., weakly informative) prior on Σ than to put such priors on the elements of A in terms of letting the data drive the inference and obtaining good convergence. Since Σ is a positive definite covariance matrix, the inverse Wishart prior distribution renders itself as a natural choice, that is, $\Sigma^{-1} \sim Wishart\left(\nu, (\nu R)^{-1}\right)$ (see, e.g., Carlin and Louis, 2000, p. 328). Hence, we instead place a prior directly on Σ, and then use the one-to-one relationship between the elements of Σ and the Cholesky factor A. Then, the prior distribution $p(A)$ becomes

$$p(A) \propto |AA^T|^{-\frac{\nu+4}{2}} exp\left\{-\frac{1}{2} tr[\nu D(AA^T)^{-1}]\right\} \left|\frac{\partial \Sigma}{a_{ij}}\right| ,$$

where $\left|\frac{\partial \Sigma}{a_{ij}}\right|$ is the Jacobian $2^p \prod_{i=1}^{p} a_{ii}^{p-i+1}$. For example, when $p = 2$, the Jacobian is $4a_{22}^2 a_{11}$. Rather than updating Σ as a block using a Wishart proposal, updating the elements a_{ij} of A offers better control. These are updated via a random-walk Metropolis, using log-normal proposals for the diagonal elements and normal proposals for the off-diagonal elements. With regard to choosing ν and R in the $Wishart\left(\nu, (\nu R)^{-1}\right)$, since $E(\Sigma^{-1}) = R^{-1}$, if there is no information about the prior mean structure of Σ, a diagonal matrix R can be chosen, with the scale of the diagonal elements being judged using ordinary least squares estimates based on independent models for each response variable. While this leads to a data-dependent prior, typically the Wishart prior lets the data drive the results, leading to robust posterior inference. In this study we adopt $\nu = 2$ (i.e., the smallest value for which this Wishart prior is proper) and $R = Diag(0.1, 0.1)$. Finally, for the remaining terms on the right hand side of (10.25), flat priors are chosen for β_1 and β_2, while σ^2 is assigned a vague inverse gamma prior, i.e., an $IG(1, 0.01)$ parameterized so that $E(\sigma^2) = b/(a-1)$. In this study, $\boldsymbol{\beta}$ and σ^2 have closed-form full conditionals, and so can be directly updated using Gibbs sampling.

10.8 Illustrating coregionalized MCAR models with three cancers from Minnesota

Jin, Banerjee and Carlin (2007) estimate different coregionalized MCAR models methods with a data set consisting of the numbers of deaths due to cancers of the lung, larynx, and esophagus in the years from 1990 to 2000 at the county level in Minnesota. The larynx and esophagus are sites of the upper aerodigestive tract, so they are closely related anatomically. Epidemiological evidence shows a strong and consistent relationship between exposure to alcohol and tobacco and the risk of cancer at these two sites (Baron et al., 1993). Meanwhile, lung cancer is the leading cause of cancer death for both men and women. An estimated 159,260 Americans will die in 2014 from lung cancer, accounting for 27% of all cancer deaths. It has long been established that tobacco, and particularly cigarette smoking, is the major cause of lung cancer. More than 87% of lung cancers are smoking-related (`http://www.lungcancer.org`).

Following Jin et al. (2007), we estimate the model

$$Y_{ij} \overset{ind}{\sim} Po(E_{ij}e^{\mu_{ij}}),\ i = 1, \ldots, n,\ j = 1, 2, 3 \tag{10.26}$$

where Y_{ij} is the observed number of cases of cancer type j (one of three types) in region i, $\log \mu_{ij} = \beta_j + \phi_{ij}$ with the β_j's being cancer-specific intercepts and the ϕ_{ij}'s being spatial random effects that are distributed according to some version of the coregionalized MCAR's we discussed earlier. To calculate the expected counts E_{ij}, we have to take each county's age distribution (over the 18 age groups) into account. To do so, we calculate the expected *age-adjusted* number of deaths due to cancer j in county i as $E_{ij} = \sum_{k=1}^{m} \omega_j^k N_i^k$, $i = 1, \ldots, 87,\ j = 1, 2, 3,\ k = 1, \ldots, 18,$ where $\omega_j^k = (\sum_{i=1}^{87} D_{ij}^k)/(\sum_{i=1}^{87} N_i^k)$ is the age-specific death rate due to cancer j for age group k over all Minnesota counties, D_{ij}^k is the number of deaths in age group k of county i due to cancer j, and N_i^k is the total population at risk in county i, age group k, which we assume to be the same for each type of cancer.

The county-level maps of the raw age-adjusted standardized mortality ratios (i.e., $\text{SMR}_{ij} = Y_{ij}/E_{ij}$) shown in Figure 10.6 exhibit evidence of correlation both across space and among the cancers, motivating use of our proposed multivariate lattice model. Using the likelihood in (10.26), we model the random effects ϕ_{ij} using our proposed $MCAR(B,\ \Sigma)$ model (10.23). In what follows we compare it with other MCAR models, including the $MCAR(\alpha,\ \Sigma)$ and $MCAR(1,\ \Sigma)$ from Section 10.6.1, a "three separate CAR's" model ignoring correlation between cancers, and a trivariate i.i.d. model ignoring correlations of any kind. We also compare one of the $MCAR(\alpha_1, \alpha_2, \alpha_3,\ \Sigma)$ models given in (10.21) of Section 10.6.2 by choosing the matrix A as the upper-triangular Cholesky decomposition of Σ. Note that we do not consider the order-specific GMCAR model (Section 10.3), since with no natural causal order for these three cancers, it is hard to choose among the six possible conditioning orders.

For priors, we follow the guidelines outlined earlier and use the same specifications as in Jin et al. (2007). Since $p = 3$ in this example, we choose the inverse Wishart distribution with $\nu = 3$ and $R = Diag(0.1, 0.1, 0.1)$ for Σ. For a model comparisons using DIC, we retain the same "focus" parameters and likelihood across the models. We used 20,000 pre-convergence burn-in iterations followed by a further 20,000 production iterations for posterior summarization. To see the relative performance of these models, we use DIC. As in the previous section, the deviance is the same for the models we wish to compare since they differ only in their random effect distributions $p(\boldsymbol{\phi}|B, \Sigma)$.

In what follows, Models 1–6 are multivariate lattice models with different assumptions about the smoothing parameters. Model 1 is the full model $MCAR(B,\ \Sigma)$ (with a 3×3 matrix B whose elements are the six smoothing parameters) while Model 2 is the $MCAR(B,\ I)$ model. Model 2 is the $MCAR(\alpha_1, \alpha_2, \alpha_3; \Sigma)$ model (10.21) with a different smoothing

	model	$\overline{D}$	p_D	DIC
1	$MCAR(B, \Sigma)$	138.8	82.5	221.3
2	$MCAR(B,\ I)$	147.6	81.4	229.0
3	$MCAR(\alpha_1, \alpha_2, \alpha_3, \Sigma)$	139.6	86.4	226.0
4	$MCAR(\alpha, \Sigma)$	143.4	81.9	225.3
5	separate CAR	147.6	82.8	230.4
6	trivariate i.i.d.	146.8	91.3	238.1
7	$MCAR(B, \Sigma)$ + trivariate I.I.D	129.6	137.6	267.2
8	$MCAR(B,\ I)$ + trivariate I.I.D	139.5	155.2	294.7
9	$MCAR(\alpha_1, \alpha_2, \alpha_3, \Sigma)$ + trivariate I.I.D	137.4	155.0	292.4
10	$MCAR(\alpha, \Sigma)$ + trivariate I.I.D	138.2	151.0	289.2
11	separate CAR + trivariate i.i.d.	139.2	162.8	302.0

Table 10.4 *Model comparison using DIC statistics, Minnesota cancer data analysis.*

parameter for each cancer. Model 3 assumes a common smoothing parameter α and Model 4 fits the three separate univariate CAR model, while Model 6 is the trivariate i.i.d. model. Fit measures $\overline{D}$, effective numbers of parameters p_D, and DIC scores for each model are seen in Table 10.4. We find that the $MCAR(B, \Sigma)$ model has the smallest $\overline{D}$ and DIC values for this data set. The $MCAR(B, I)$ model again disappoints, excelling over the non-spatial model and the separate CAR models only (very marginally over the latter). The $MCAR(\alpha,\ \Sigma)$ and $MCAR(\alpha_1, \alpha_2, \alpha_3,\ \Sigma)$ models perform slightly worse than the $MCAR(B,\ \Sigma)$ model, suggesting the need for different spatial autocorrelation and cross-spatial correlation parameters for this data set. Note that the effective numbers of parameters p_D in Model 3 is a little larger than in Model 1, even though the latter has three extra parameters. Finally, the MCAR models do better than the separate CAR model or the i.i.d. trivariate model, suggesting that it is worth taking account of the correlations both across counties and among

Figure 10.6 *Maps of raw standardized mortality ratios (SMR) of lung, larynx and esophagus cancer in the years from 1990 to 2000 in Minnesota.*

	Lung median (2.5%, 97.5%)	Larynx median (2.5%, 97.5%)	Esophagus mean (2.5%, 97.5%)
β_1, β_2, β_3	–0.093 (–0.179,–0.006)	–0.128 (–0.316, 0.027)	–0.080 (–0.194, 0.025)
Σ_{11}, Σ_{22}, Σ_{33}	0.048 (0.030, 0.073)	0.173 (0.054, 0.395)	0.107 (0.044, 0.212)
ρ_{12}, ρ_{13}		0.277 (–0.112, 0.643)	0.378 (–0.022, 0.716)
ρ_{23}			0.337 (–0.311, 0.776)
b_{11}, b_{22}, b_{33}	0.442 (–0.302, 0.921)	0.036 (–0.830, 0.857)	0.312 (–0.526, 0.901)
b_{12}, b_{13}		0.323 (–0.156, 0.842)	0.389 (–0.028, 0.837)
b_{23}			0.006 (–0.519, 0.513)

Table 10.5 *Posterior summaries of parameters in $MCAR(B, \Sigma)$ model for Minnesota cancer data.*

Figure 10.7 *Posterior samples of ρ_{12}, ρ_{13} and ρ_{23} in the Minnesota cancer data analysis using the $MCAR(B, \Sigma)$ model: (a) estimated posterior for correlation ρ_{12} between lung and larynx; (b) estimated posterior for correlation ρ_{13} between lung and esophagus; (c) estimated posterior for correlation ρ_{23} between larynx and esophagus.*

cancers. Model 6 exhibits a large p_D score, suggesting it does not seem to allow sufficient smoothing of the random effects. This is what we might have expected, since the spatial correlations are missed by this model.

Models 7–11 are the convolution prior models corresponding to Models 1–5 formed by adding i.i.d. effects (following $N(0, \tau^2)$) to the ϕ_{ij}'s. Here the distinctions between the models are somewhat more pronounced due to the added variability in the models caused by the i.i.d. effects. The relative performances of the models remain the same with the $MCAR(B, \Sigma)$+ i.i.d. model emerging as best. Interestingly, none of the convolution models perform better than their purely spatial counterparts as the improvements in $\bar{D}$ in the former are insignificant compared to the increase in the effective dimensions brought about. This is indicative of the dominance of the spatial effects over the i.i.d. effects whence the convolution models seem to be rendering overparametrized models.

We summarize our results from the $MCAR(B, \Sigma)$, which is Model 1 in Table 10.4. Table 10.5 provides posterior means and associated standard deviations for the parameters $\boldsymbol{\beta}$, Σ and b_{ij} in this model, where b_{ij} is the element of the symmetric matrix B. Instead of

Figure 10.8 *Posterior samples of b_{12}, b_{13} and b_{23} in the Minnesota cancer data analysis using the $MCAR(B, \Sigma)$ model: (a) estimated posterior for b_{12}; (b) estimated posterior for b_{13}; (c) estimated posterior for b_{23}.*

reporting Σ_{12}, Σ_{13} and Σ_{23}, we provide the mean and associated standard deviations for the correlation parameters ρ_{12}, ρ_{13} and ρ_{23}, which are calculated as $\rho_{ij} = \Sigma_{ij}/\sqrt{\Sigma_{ii}\Sigma_{jj}}$. We also plot histograms of the posterior samples ρ_{ij} in Figure 10.7, and histograms of the posterior samples b_{ij} in Figure 10.8.

Table 10.5 and Figure 10.7 reveal correlations between cancers, in particular a strong correlation between lung and esophagus (ρ_{13}). This might explain why the DIC scores for Models 1–4 in Table 10.5 are smaller than that under the separate CAR model. The b_{ij} in Table 10.5 are spatial autocorrelation and cross-spatial correlation parameters for the latent spatial processes $\mathbf{u}_j$, $j = 1, 2, 3$. Figure 10.8 shows most of the b_{12} and b_{13} posterior samples are positive; the means of these two parameters are 0.323 and 0.389, respectively. Consistent with the DIC results in Table 10.4, these suggest it is worth fitting our proposed $MCAR(B, \Sigma)$ model to these data.

Turning to geographical summaries, Figure 10.9 maps the posterior means of the fitted standard mortality ratios (SMR) of lung, larynx and esophagus cancer from our $MCAR(B, \Sigma)$ model. From Figure 10.9, the correlation among the cancers is apparent, with higher fitted ratios extending from the Twin Cities metro area to the north and northeast (an area where previous studies have suggested smoking may be more common). In Figure 10.6, the range of the raw SMRs is seen to be from 0 to 3.3, while in Figure 10.9, the range of the fitted SMRs is from 0.7 to 1.3, due to spatial shrinkage in the random effects.

10.9 Exercises

1. The usual and generalized (but still proper) MCAR models may be constructed using linear transformations of some nonspatially correlated variables. Consider a vector blocked by components, say $\boldsymbol{\phi} = \left(\boldsymbol{\phi}_1^T, \boldsymbol{\phi}_2^T\right)^T$, where each $\boldsymbol{\phi}_i$ is $n \times 1$, n being the number of areal

Figure 10.9 *Maps of posterior means of the fitted standard mortality ratios (SMR) of lung, larynx and esophagus cancer in the years from 1990 to 2000 in Minnesota from* $MCAR(B,\, \Sigma)$ *model.*

units. Suppose we look upon these vectors as arising from linear transformations

$$\boldsymbol{\phi}_1 = A_1\mathbf{v}_1 \text{ and } \boldsymbol{\phi}_2 = A_2\mathbf{v}_2 \ ,$$

where A_1 and A_2 are any $n \times n$ matrices, $\mathbf{v}_1 = (v_{11}, \ldots, v_{1n})^T$ and $\mathbf{v}_2 = (v_{21}, \ldots, v_{2n})^T$ with covariance structure

$$Cov\,(v_{1i}, v_{1j}) = \lambda_{11} I_{[i=j]},\ Cov\,(v_{1i}, v_{2j}) = \lambda_{12} I_{[i=j]},$$
$$\text{and } Cov\,(v_{2i}, v_{2j}) = \lambda_{22} I_{[i=j]}\ ,$$

where $I_{[i=j]} = 1$ if $i = j$ and 0 otherwise. Thus, although $\mathbf{v}_1$ and $\mathbf{v}_2$ are associated, their nature of association is nonspatial in that covariances remain same for every areal unit, and there is no association between variables in different units.

(a) Show that the dispersion matrix $\Sigma_{(\mathbf{v}_1, \mathbf{v}_2)}$ equals $\Lambda \otimes I$, where $\Lambda = (\lambda_{ij})_{i,j=1,2}$.

(b) Show that setting $A_1 = A_2 = A$ yields a separable covariance structure for $\boldsymbol{\phi}$. What choice of A would render a separable MCAR model, analogous to (10.4)?

(c) Show that appropriate (different) choices of A_1 and A_2 yield the generalized MCAR model with covariance matrix given by (10.7).

2. Derive the covariance matrix in (10.12) for the bivariate $GMCAR(\rho_1,\, \rho_2,\, \eta_1,\, \eta_2,\, \tau_1,\, \tau_2)$ model.

3. Show that the covariance matrix in (10.22) can be expressed as:

$$(I_{p\times p} \otimes D)^{\frac{1}{2}} \left(I_{pn\times pn} - B \otimes D^{-\frac{1}{2}} W D^{-\frac{1}{2}}\right) (I_{p\times p} \otimes D)^{\frac{1}{2}}\ .$$

4. Prove that $|I_{p\times p} \otimes D - B \otimes W| \propto \prod_{i=1}^{n} \prod_{j=1}^{p} (1 - \xi_i \zeta_j)$, where ξ_i's are the eigenvalues of $D^{-\frac{1}{2}} W D^{-\frac{1}{2}}$ and ζ_i's are the eigenvalues of B.

Chapter 11

Spatiotemporal modeling

In both theoretical and applied work, spatiotemporal modeling has received dramatically increased attention in the past few years. The reason is easy to see: the proliferation of data sets that are both spatially and temporally indexed, and the attendant need to understand them. For example, in studies of air pollution, we are interested not only in the spatial nature of a pollutant surface, but also in how this surface changes over time. Customarily, ongoing temporal measurements (e.g., hourly, daily, three-day average, etc.) are collected at monitoring sites yielding long-time series of data. Similarly, with climate data we may be interested in spatial patterns of temperature or precipitation at a given time, but also in dynamic patterns in weather. With real estate markets, we might be interested in how the single-family home sales market changes on a quarterly or annual basis. Here an additional wrinkle arises in that we do not observe the *same* locations for each time period; the data are cross-sectional, rather than longitudinal.

Applications with areal unit data are also commonplace. For instance, we may look at annual lung cancer rates by county for a given state over a number of years to judge the effectiveness of a cancer control program. Or we might consider daily asthma hospitalization rates by zip code, over a period of several months.

From a methodological point of view, the introduction of time into spatial modeling brings a substantial increase in the scope of our work, as we must make separate decisions regarding spatial correlation, temporal correlation, and how space and time interact in our data. Such modeling will also carry an obvious associated increase in notational and computational complexity.

As in previous chapters, we make a distinction between the cases where the geographical aspect of the data is at point level versus where it is at areal unit level. Again the former case is typically handled via Gaussian process models, while the latter often uses CAR specifications. A parallel distinction could be drawn for the temporal scale: Is time viewed as continuous (say, over $\Re^+$ or some subinterval thereof) or discrete (hourly, daily, etc.)? In the former case there is a conceptual measurement at each moment t. But in the latter case, we must determine whether each measurement should be interpreted as a block average over some time interval (analogous to block averaging in space), or whether it should be viewed merely as a measurement, e.g., a count attached to an associated time interval (and thus analogous to an areal unit measurement). Relatedly, when time is discretized, are we observing a time series of spatial data, e.g., the same points or areal units in each time period (as would be the case in our climate and pollution examples)? Or are we observing cross-sectional data, where the locations change with time period (as in our real estate setting)? In the case of time series, we could regard the data as a multivariate measurement vector at each location or areal unit. We could then employ multivariate spatial data models as in the previous chapter. With short series, this might be reasonable; with longer series, we would likely want to introduce aspects of usual time series modeling.

The nature and location of missing data is another issue that we have faced before, yet becomes doubly complicated in the spatiotemporal setting. The major goal of our earlier

kriging methods is to impute missing values at locations for which no data have been observed. Now we may encounter time points for which we lack spatial information, locations for which information is lacking for certain (possibly future) time points, or combinations thereof. Some of these combinations will be extrapolations (e.g., predicting future values at locations for which no data have been observed) that are statistically riskier than others (e.g., filling in missing values at locations for which we have data at some times but not others). Here the Bayesian hierarchical approach is particularly useful, since it not only helps organize our thinking about the model, but also fully accounts for all sources of uncertainty, and properly delivers wider confidence intervals for predictions that are "farther" from the observed data (in either space or time).

Our Atlanta data set (Figure 7.2) illustrates the sort of misalignment problem we face in many spatiotemporal settings. Here the number of ozone monitoring stations is small (just 8 or 10), but the amount of data collected from these stations over time (92 summer days for each of three years) is substantial. In this case, under suitable modeling assumptions, we may not only learn about the temporal nature of the data, but also enhance our understanding of the spatial process.

In the next few sections we consider the case of point-level spatial data, so that point-point and point-block realignment can be contemplated as in Section 7.1. We initially focus on relatively simple *separable* forms for the space-time correlation, but also consider more complex forms that do not impose the strong restrictions on space-time interaction that separability implies. We subsequently move on to spatiotemporal modeling for data where the spatial component can only be thought of as areal (block) level.

11.1 General modeling formulation

11.1.1 Preliminary analysis

Before embarking on a general spatiotemporal modeling formulation, consider the case of point-referenced data where time is discretized to customary integer-spaced intervals. We may look at a spatiotemporally indexed datum $Y(\mathbf{s},t)$ in two ways. Writing $Y(\mathbf{s},t) = Y_{\mathbf{s}}(t)$, it is evident that we have a spatially varying time series model. Writing $Y(\mathbf{s},t) = Y_t(\mathbf{s})$, we instead have a temporally varying spatial model.

In fact, with locations $\mathbf{s}_i$, $i = 1,\ldots,n$ and time points $t = 1,\ldots,T$, we can collect the data into Y, an $n \times T$ matrix. Column averages of Y produce a space-averaged time series, while row averages yield a time-averaged spatial realization. In fact, suppose we center each column of Y by the vector of row averages and call the resulting matrix $\widetilde{Y}_{rows}$. Then clearly $\widetilde{Y}_{rows}\mathbf{1}_T = \mathbf{0}$, but also $\frac{1}{T}\widetilde{Y}_{rows}\widetilde{Y}_{rows}^T$ is an $n \times n$ matrix that is the sample spatial covariance matrix. Similarly, suppose we center each row of Y by the vector of column averages and call the resulting matrix $\widetilde{Y}_{cols}$. Now $\mathbf{1}_n^T\widetilde{Y}_{cols} = \mathbf{0}$ and $\frac{1}{n}\widetilde{Y}_{cols}^T\widetilde{Y}_{cols}$ is the $T \times T$ sample autocorrelation matrix.

One could also center Y by the grand mean of the $Y(\mathbf{s},t)$. Indeed, to examine residual spatiotemporal structure, adjusted for the mean, one could fit a suitable OLS regression to the $Y(\mathbf{s},t)$ and examine $\widehat{E}$, the matrix of residuals $\widehat{e}(\mathbf{s},t)$. As above, $\frac{1}{T}\widehat{E}\widehat{E}^T$ is the residual spatial covariance matrix while $\frac{1}{n}\widehat{E}^T\widehat{E}$ is the residual autocorrelation matrix.

We can create the singular value decomposition (Harville, 1997) for any of the foregoing matrices. Using E which would be most natural in practice to consider spatiotemporal structure, we can write

$$E = UDV^T = \sum_{l=1}^{\mathtt{min}(n,T)} d_l \mathbf{u}_l \mathbf{v}_l^T \,, \tag{11.1}$$

where U is an $n \times n$ orthogonal matrix with columns $\mathbf{u}_l$, V is a $T \times T$ orthogonal matrix with columns $\mathbf{v}_l$, and D is an $n \times T$ matrix of the form $\binom{\Delta}{0}$ where Δ is $T \times T$ diagonal

with diagonal entries d_l, $l = 1, \ldots, T$. Without loss of generality, we can assume the d_l's are arranged in decreasing order of their absolute values. Then, $\mathbf{u}_l \mathbf{v}_l^T$ is referred to as the lth *empirical orthogonal function* (EOF) since $\mathbf{u}_l \mathbf{v}_l^T \perp \mathbf{u}_m \mathbf{v}_m^T, l \neq m$ and $(\mathbf{u}_l \mathbf{v}_l^T)^T \mathbf{u}_l \mathbf{v}_l^T = 1$.

Thinking of $\mathbf{u}_l = (u_l(\mathbf{s}_1), \ldots, u_l(\mathbf{s}_n))^T$ and $\mathbf{v}_l = (v_l(1), \ldots, v_l(T))^T$, the expression in (11.1) represents the observed data as a sum of products of spatial and temporal variables, i.e., $E(\mathbf{s}_i, t) = \sum d_l u_l(\mathbf{s}_i) v_l(t)$. Suppose we approximate E by its first EOF, that is, $E \approx d_1 \mathbf{u}_1 \mathbf{v}_1^T$. Then we are saying that $E(\mathbf{s}_i, t) \approx d_1 u_1(\mathbf{s}_i) v_1(t)$, i.e., the spatiotemporal process can be approximated by a product of a spatial process and a temporal process. If the u_1 and v_1 processes are mean 0 (as they would be for modeling residuals) and independent, this implies a *separable* covariance function for $E(\mathbf{s}, t)$ (see (11.18)).[1] Indeed, if the first term in the sum in (11.1) explains much of the residual matrix E, this is often taken as evidence for specifying a separable model. In any event, it does yield a reduction in dimension, introducing $n + T$ variables to represent E, rather than nT. Adding the second EOF yields the approximation $E(\mathbf{s}_i, t) \approx d_1 u_1(\mathbf{s}_i) v_1(t) + d_2 u_2(\mathbf{s}_i) v_2(t)$, a representation involving only $2(n + T)$ variables, and so on.

Note that, if, say, $T < n$,

$$EE^T = UDD^T U^T = U \begin{pmatrix} \Delta^2 & 0 \\ 0 & 0 \end{pmatrix} U^T = \sum_{l=1}^{T} d_l^2 \mathbf{u}_l \mathbf{u}_l^T \ ,$$

clarifying the interpretation of the d_l's. (Of course, $EE^T = V^T D^T D V = V^T \Delta^2 V$ as well.) Altogether, when applicable, EOFs provide an exploratory tool for learning about spatial structure and suggesting models, in the spirit of the tools described in Section 2.3. For full inference, however, we require a full spatiotemporal model specification, the subject to which we now turn.

11.1.2 Model formulation

Modeling for spatiotemporal data can be given a fairly general formulation that naturally extends that of Chapter 6. Consider point-referenced locations and continuous time. Let $Y(\mathbf{s}, t)$ denote the measurement at location $\mathbf{s}$ at time t. Extending (6.1), for continuous data assumed to be roughly normally distributed, we can write the general form

$$Y(\mathbf{s}, t) = \mu(\mathbf{s}, t) + e(\mathbf{s}, t) \ , \tag{11.2}$$

where $\mu(\mathbf{s}, t)$ denotes the mean structure and $e(\mathbf{s}, t)$ denotes the residual. If $\mathbf{x}(\mathbf{s}, t)$ is a vector of covariates associated with $Y(\mathbf{s}, t)$ then we can set $\mu(\mathbf{s}, t) = \mathbf{x}(\mathbf{s}, t)^T \boldsymbol{\beta}(\mathbf{s}, t)$. Note that this form allows spatiotemporally varying coefficients (in the spirit of Section 9.6), which is likely more general than we would want; $\boldsymbol{\beta}(\mathbf{s}, t) = \boldsymbol{\beta}$ is frequently adopted. If t is discretized, $\boldsymbol{\beta}(\mathbf{s}, t) = \boldsymbol{\beta}_t$ might be appropriate if there were enough time points to suggest a temporal change in the coefficient vector. Similarly, setting $\boldsymbol{\beta}(\mathbf{s}, t) = \boldsymbol{\beta}(\mathbf{s})$ yields spatially varying coefficients, again following Section 9.6. Finally, $e(\mathbf{s}, t)$ would typically be rewritten as $w(\mathbf{s}, t) + \epsilon(\mathbf{s}, t)$, where $\epsilon(\mathbf{s}, t)$ is a Gaussian white noise process and $w(\mathbf{s}, t)$ is a mean-zero spatiotemporal process.

We can therefore view (11.2) as a hierarchical model with a conditionally independent first stage given $\{\mu(\mathbf{s}, t)\}$ and $\{w(\mathbf{s}, t)\}$. But then, in the spirit of Section 6.2, we can replace the Gaussian first stage with another first-stage model (say, an exponential family model) and write $Y(\mathbf{s}, t) \sim f(y(\mathbf{s}, t) \mid \mu(\mathbf{s}, t), w(\mathbf{s}, t))$, where

$$f(y(\mathbf{s}, t) \mid \mu(\mathbf{s}, t), w(\mathbf{s}, t)) = h(y(\mathbf{s}, t)) \exp\{\gamma[\eta(\mathbf{s}, t) y(\mathbf{s}, t) - \chi(\eta(\mathbf{s}, t))]\} \ , \tag{11.3}$$

[1] It is routine to see that this will not be the case if the u_1 process and the v_1 process are not independent.

where γ is a positive dispersion parameter. In (11.3), $g(\eta(\mathbf{s},t)) = \mu(\mathbf{s},t) + w(\mathbf{s},t)$ for some link function g.

For areal unit data with discrete time, let Y_{it} denote the measurement for unit i at time period t. (In some cases we might obtain replications at i or t, e.g., the jth cancer case in county i, or the the jth property sold in school district i.) Analogous to (11.2) we can write

$$Y_{it} = \mu_{it} + e_{it} \ . \tag{11.4}$$

Now $\mu_{it} = \mathbf{x}_{it}^T \boldsymbol{\beta}_t$ (or perhaps just $\boldsymbol{\beta}$), and $e_{it} = w_{it} + \epsilon_{it}$ where the ϵ_{it} are unstructured heterogeneity terms and the w_{it} are spatiotemporal random effects, typically associated with a spatiotemporal CAR specification. Choices for this latter part of the model will be presented in Section 11.7.

Since areal unit data are often non-Gaussian (e.g., sparse counts), again we would view (11.4) as a hierarchical model and replace the first stage Gaussian specification with, say, a Poisson model. We could then write $Y_{it} \sim f(y_{it}|\mu_{it}, w_{it})$, where

$$f(y_{it} \mid \mu_{it}, w_{it}) = h(y_{it}) \exp\{\gamma[\eta_{it} y_{it} - \chi(\eta_{it})]\} \ , \tag{11.5}$$

with γ again a dispersion parameter, and $g(\eta_{it}) = \mu_{it} + w_{it}$ for some suitable link function g. With replications, we obtain Y_{ijt} hence $\mathbf{x}_{ijt}, \mu_{ijt}$, and η_{ijt}. Now we can write $g(\eta_{ijt}) = \mu_{ijt} + w_{ijt} + \epsilon_{ijt}$, enabling separation of spatial and heterogeneity effects.

Returning to the point-referenced data model (11.2), spatiotemporal richness is captured by extending $e(\mathbf{s},t)$ beyond $\epsilon(\mathbf{s},t)$, a white noise process, as noted above. As a result, we need forms for $w(\mathbf{s},t)$. Below, α's denote temporal effects and w's denote spatial effects. Following Gelfand, Ecker, Knight, and Sirmans (2004) with t discretized, consider the following forms for $w(\mathbf{s},t)$:

$$w(\mathbf{s},t) = \alpha(t) + w(\mathbf{s}) \ , \tag{11.6}$$

$$w(\mathbf{s},t) = \alpha_{\mathbf{s}}(t) \ , \tag{11.7}$$

$$\text{and } w(\mathbf{s},t) = w_t(\mathbf{s}) \ . \tag{11.8}$$

The given forms avoid specification of space-time interactions. With regard to (11.6), (11.7), and (11.8), the $\epsilon(\mathbf{s},t)$ are i.i.d. $N(0, \sigma_\epsilon^2)$ and independent of the other processes. This pure error is viewed as a residual adjustment to the spatiotemporal explanation. (One could allow $Var(\epsilon(\mathbf{s},t)) = \sigma_\epsilon^{2(t)}$, i.e., an error variance that changes with time. Modification to the details below is straightforward.)

Expression (11.6) provides an additive form in temporal and spatial effects. In fact, we can also introduce a multiplicative form, $\alpha(t)w(\mathbf{s})$ which would, of course, become additive on the log scale. Expression (11.7) provides temporal evolution at each site; temporal effects are nested within sites. Expression (11.8) provides spatial evolution over time; spatial effects are nested within time. Spatiotemporal modeling beyond (11.6), (11.7), and (11.8) (particularly if t is continuous) necessitates the choice of a specification to connect the space and time scales; this is the topic of Section 11.2.

Next, we consider the components in (11.6), (11.7), and (11.8) in more detail. In (11.6), if t were continuous we could model $\alpha(t)$ as a one-dimensional stationary Gaussian process. In particular, for the set of times, $\{t_1, t_2, ..., t_m\}$, $\boldsymbol{\alpha} = (\alpha(t_1), ..., \alpha(t_m))' \sim N(\mathbf{0}, \sigma_\alpha^2 \Sigma(\phi))$ where $(\Sigma(\phi))_{rs} = Corr(\alpha(t_r), \alpha(t_s)) = \rho(\mid t_r - t_s \mid; \phi)$ for ρ a valid one-dimensional correlation function. A typical choice for ρ would be the exponential, $\exp(-\phi \mid t_r - t_s \mid)$ though other forms, analogous to the spatial forms in Table 2.1 are possible.

With t confined to an indexing set, $t = 1, 2, \ldots T$, we can simply view $\alpha(1), ..., \alpha(T)$ as the coefficients associated with a set of time dummy variables. With this assumption for the $\alpha(t)$'s, suppose in (11.6), $w(\mathbf{s})$ is set to zero, $\boldsymbol{\beta}(t)$ is assumed constant over time and $\mathbf{X}(\mathbf{s},t)$

is assumed constant over t. Then, upon differencing, we find models described in the real estate literature, e.g., the seminal model for repeat property sales given in Bailey, Muth, and Nourse (1963). Also within these assumptions but restoring $\boldsymbol{\beta}$ to $\boldsymbol{\beta}(t)$, we obtain the extension of Knight, Dombrow, and Sirmans (1995). In very recent work (Paci et al., 2013) the multiplicative form is used, with differencing, to implement real-time ozone forecasting.

Alternatively, we might set $\alpha(t+1) = \rho\alpha(t) + \eta(t)$ where $\eta(t)$ are i.i.d. $N(0, \sigma_\alpha^2)$. If $\rho < 1$ we have the familiar stationary $AR(1)$ time series, a special case of the continuous time model of the previous paragraph. If $\rho = 1$ the $\alpha(t)$ follow a random walk. With a finite set of times, time-dependent coefficients are handled analogously to the survival analysis setting (see, e.g., Cox and Oakes, 1984, Ch. 8).

The autoregressive and random walk specifications are naturally extended to provide a model for the $\alpha_{\mathbf{s}}(t)$ in (11.7). That is, we assume $\alpha_{\mathbf{s}}(t+1) = \rho\alpha_{\mathbf{s}}(t) + \eta_{\mathbf{s}}(t)$ where again the $\eta_{\mathbf{s}}(t)$ are all i.i.d. Thus, there is no spatial modeling; rather, we imagine independent conceptual time series at each location. With spatial time series we can fit this model. With cross-sectional data, there is no information in the data about ρ so the likelihood can only identify the stationary variance $\sigma_\alpha^2/(1-\rho^2)$ but not σ_α^2 or ρ. The case $\rho < 1$ with $\boldsymbol{\beta}(t)$ constant over time provides the models proposed in Hill, Knight, and Sirmans (1997) and in Hill, Sirmans, and Knight (1999). If $\rho = 1$ with $\boldsymbol{\beta}(t)$ and $\mathbf{X}(\mathbf{s}, t)$ constant over time, upon differencing we obtain the widely used model of Case and Shiller (1989). In application, it will be difficult to learn about the $\alpha_{\mathbf{s}}$ processes with typically one or at most two observations for each $\mathbf{s}$. The $w(\mathbf{s})$ are modeled as a Gaussian process following Section 3.1.

For $w_t(\mathbf{s})$ in (11.8), assuming t restricted to an index set, we can view the $w_t(\mathbf{s})$ as a collection of independent spatial processes. That is, rather than defining a dummy variable at each t, we conceptualize a separate spatial dummy *process* at each t. The components of $\mathbf{w}_t$ correspond to the sites at which measurements were observed in the time interval denoted by t. Thus, we capture the dynamics of location in a very general fashion. In particular, comparison of the respective process parameters reveals the nature of spatial evolution over time.

With a single time dummy variable at each t, assessment of temporal effects would be provided through inference associated with these variables. For example, a plot of the point estimates against time would clarify size and trend for the effects. With distinct spatial processes, how can we see such temporal patterns? A convenient reduction of each spatial process to a univariate random variable is the block average (see Expression (7.1)).

To shed the independence assumption for the $w_t(\mathbf{s})$, we could instead assume that $w_t(\mathbf{s}) = \sum_{j=1}^{t} v_j(\mathbf{s})$ where the $v_j(\mathbf{s})$ are i.i.d. processes, again of one of the foregoing forms. Now, for $t < t^*$, $\mathbf{w}_t$ and $\mathbf{w}_{t^*}$ are not independent but $\mathbf{w}_t$ and $\mathbf{w}_{t^*} - \mathbf{w}_t$ are. This leads us to dynamic spatiotemporal models that are the focus of Section 11.4.

11.1.3 Associated distributional results

We begin by developing the likelihood under model (11.2) using (11.6), (11.7), or (11.8). Assuming t belongs to the set $\{1, 2, ..., T\}$, it is convenient to first obtain the joint distribution for $\mathbf{Y}' = (\mathbf{Y}_1', ..., \mathbf{Y}_T')$ where $\mathbf{Y}_t' = (Y(\mathbf{s}_1, t), ..., Y(\mathbf{s}_n, t))$. That is, each $\mathbf{Y}_t$ is $n \times 1$ and $\mathbf{Y}$ is $Tn \times 1$. This joint distribution will be multivariate normal. Thus, the joint distribution for the observed $Y(\mathbf{s}, t)$ requires only pulling off the appropriate entries from the mean vector and appropriate rows and columns from the covariance matrix. This simplifies the computational bookkeeping, though care is still required.

In the constant $\boldsymbol{\beta}$ case, associate with $\mathbf{Y}_t$ the matrix X_t whose ith row is $\mathbf{X}(\mathbf{s}_i, t)'$. Let $\boldsymbol{\mu}_t = X_t\boldsymbol{\beta}$ and $\boldsymbol{\mu}' = (\mu_1', ..., \mu_T')$. In the time-dependent parameter case we merely set $\boldsymbol{\mu}_t = X_t\boldsymbol{\beta}(t)$.

Under (11.6), let $\boldsymbol{\alpha}' = (\alpha(1), \ldots, \alpha(T))$, $\mathbf{w}' = (\mathbf{w}(\mathbf{s}_1), \ldots, \mathbf{w}(\mathbf{s}_n))$ and $\boldsymbol{\epsilon}' = (\epsilon(\mathbf{s}_1, 1), \epsilon(\mathbf{s}_1, 2), ..., \epsilon(\mathbf{s}_n, T)$. Then,

$$\mathbf{Y} = \boldsymbol{\mu} + \boldsymbol{\alpha} \otimes \mathbf{1}_{n\times 1} + \mathbf{1}_{T\times 1} \otimes \mathbf{w} + \boldsymbol{\epsilon} \tag{11.9}$$

where $\otimes$ denotes the Kronecker product. Hence, given $\boldsymbol{\beta}$ along with the temporal and spatial effects,

$$\mathbf{Y} \mid \boldsymbol{\beta}, \boldsymbol{\alpha}, \mathbf{w}, \sigma_\epsilon^2 \sim N(\boldsymbol{\mu} + \boldsymbol{\alpha} \otimes \mathbf{1}_{n\times 1} + \mathbf{1}_{T\times 1} \otimes \mathbf{w}\ ,\ \sigma_\epsilon^2 I_{Tn\times Tn})\ . \tag{11.10}$$

Let $\mathbf{w} \sim N(\mathbf{0}, \sigma_w^2 H(\delta))$. Suppose the $\alpha(t)$ follow an $AR(1)$ model, so that $\boldsymbol{\alpha} \sim N(\mathbf{0}, \sigma_\alpha^2 A(\rho))$ where $(A(\rho))_{ij} = \rho^{|i-j|}/(1-\rho^2)$. Hence, if $\boldsymbol{\alpha}, \mathbf{w}$ and $\boldsymbol{\epsilon}$ are independent, marginalizing over $\boldsymbol{\alpha}$ and $\mathbf{w}$, i.e., integrating (11.10) with regard to the prior distribution of $\boldsymbol{\alpha}$ and $\mathbf{w}$, we obtain

$$\begin{aligned}&\mathbf{Y} \mid \boldsymbol{\beta}, \sigma_\epsilon^2, \sigma_\alpha^2, \rho, \sigma_w^2, \delta \\ &\sim N\left(\boldsymbol{\mu}\ ,\ \sigma_\alpha^2 A(\rho) \otimes \mathbf{1}_{n\times 1}\mathbf{1}'_{n\times 1} + \sigma_w^2 \mathbf{1}_{T\times 1}\mathbf{1}'_{T\times 1} \otimes H(\delta) + \sigma_\epsilon^2 I_{Tn\times Tn}\right)\ .\end{aligned} \tag{11.11}$$

If the $\alpha(t)$ are coefficients associated with dummy variables (now $\boldsymbol{\beta}$ does not contain an intercept) we only marginalize over $\mathbf{w}$ to obtain

$$\begin{aligned}&\mathbf{Y} \mid \boldsymbol{\beta}, \boldsymbol{\alpha}, \sigma_\epsilon^2, \sigma_w^2, \delta \\ &\quad\sim N\left(\boldsymbol{\mu} + \boldsymbol{\alpha} \otimes \mathbf{1}_{n\times 1}\ ,\ \sigma_w^2 \mathbf{1}_{T\times 1}\mathbf{1}'_{T\times 1} \otimes H(\delta) + \sigma_\epsilon^2 I_{Tn\times Tn}\right)\ .\end{aligned} \tag{11.12}$$

The likelihood resulting from (11.10) arises as a product of independent normal densities by virtue of the conditional independence. This can facilitate model fitting but at the expense of a very high-dimensional posterior distribution. Marginalizing to (11.11) or (11.12) results in a much lower-dimensional posterior. Note, however, that while the distributions in (11.11) and (11.12) can be determined, evaluating the likelihood (joint density) requires evaluation of a high-dimensional quadratic form and determinant calculation.

Turning to (11.7), if $\boldsymbol{\alpha}'(t) = (\alpha_{s_1}(t), ..., \alpha_{s_n}(t))$ and now we also define $\boldsymbol{\alpha}' = (\boldsymbol{\alpha}'(1), \ldots, \boldsymbol{\alpha}'(T))$ with $\boldsymbol{\epsilon}$ as above, then

$$\mathbf{Y} = \boldsymbol{\mu} + \boldsymbol{\alpha} + \boldsymbol{\epsilon}\ .$$

Now

$$\mathbf{Y} \mid \boldsymbol{\beta}, \boldsymbol{\alpha}, \sigma_\epsilon^2 \sim N\left(\boldsymbol{\mu} + \boldsymbol{\alpha}\ ,\ \sigma_\epsilon^2 I_{Tn\times Tn}\right)\ .$$

If the $\alpha_{s_i}(t)$ follow an $AR(1)$ model independently across i, then marginalizing over $\boldsymbol{\alpha}$,

$$\mathbf{Y} \mid \boldsymbol{\beta}, \sigma_\epsilon^2, \sigma_\alpha^2, \rho \sim N(\boldsymbol{\mu}\ ,\ A(\rho) \otimes I_{Tn\times Tn} + \sigma_\epsilon^2 I_{Tn\times Tn}). \tag{11.13}$$

For (11.8), let $\mathbf{w}'_t = (w_t(\mathbf{s}_1), \ldots, w_t(\mathbf{s}_n))$ and $\mathbf{w}' = (\mathbf{w}'_1, \ldots, \mathbf{w}'_T)$. Then with $\boldsymbol{\epsilon}$ as above,

$$\mathbf{Y} = \boldsymbol{\mu} + \mathbf{w} + \boldsymbol{\epsilon} \tag{11.14}$$

and

$$\mathbf{Y} \mid \boldsymbol{\beta}, \mathbf{w}, \sigma_\epsilon^2 \sim N\left(\boldsymbol{\mu} + \mathbf{w}\ ,\ \sigma_\epsilon^2 I_{Tn\times Tn}\right)\ . \tag{11.15}$$

If $\mathbf{w}_t \sim N(\mathbf{0}, \sigma_w^{2(t)} H(\delta^{(t)}))$ independently for $t = 1, \ldots, T$, then, marginalizing over $\mathbf{w}$,

$$\mathbf{Y} \mid \boldsymbol{\beta}, \sigma_\epsilon^2, \boldsymbol{\sigma}_w^2, \boldsymbol{\delta} \sim N(\boldsymbol{\mu}\ ,\ D(\boldsymbol{\sigma}_w^2, \boldsymbol{\delta}) + \sigma_\epsilon^2 I_{Tn\times Tn})\ , \tag{11.16}$$

where $\boldsymbol{\sigma}_w^{2\prime} = (\sigma_w^{2(1)}, ..., \sigma_w^{2(T)})$, $\boldsymbol{\delta}' = (\delta^{(1)}, \ldots, \delta^{(T)})$, and $D(\boldsymbol{\sigma}_w^2, \boldsymbol{\delta})$ is block diagonal with the tth block being $\sigma_w^{2(t)}(H(\delta^{(t)}))$. Because D is block diagonal, likelihood evaluation associated with (11.16) is less of an issue than for (11.11) and (11.12).

We note that with either (11.7) or (11.8), $e(\mathbf{s}, t)$ is comprised of two sources of error that the data cannot directly separate. However, by incorporating a stochastic assumption on the $\alpha_{\mathbf{s}}(t)$ or on the $w_t(\mathbf{s})$, we can learn about the processes that guide the error components, as (11.13) and (11.16) reveal.

11.1.4 Prediction and forecasting

We now turn to forecasting under (11.2) with models (11.6), (11.7), or (11.8). Such forecasting involves prediction at location $\mathbf{s}_0$ and time t_0, i.e., of $Y(\mathbf{s}_0, t_0)$. Here $\mathbf{s}_0$ may correspond to an already observed location, perhaps to a new location. However, typically $t_0 > T$ is of interest. Such prediction requires specification of an associated vector of characteristics $\mathbf{X}(\mathbf{s}_0, t_0)$. Also, prediction for $t_0 > T$ is available in the fixed coefficients case. For the time-varying coefficients case, we would need to specify a temporal model for $\boldsymbol{\beta}(t)$.

In general, within the Bayesian framework, prediction at $(\mathbf{s}_0, t_0)$ follows from the posterior predictive distribution of $f(Y(\mathbf{s}_0, t_0) \mid \mathbf{Y})$ where $\mathbf{Y}$ denotes the observed data vector. Assuming $\mathbf{s}_0$ and t_0 are new, and for illustration, taking the form in (11.6),

$$\begin{aligned} f(Y(\mathbf{s}_0, t_0) \mid \mathbf{Y}) = \textstyle\int f(Y(\mathbf{s}_0, t_0) \mid \boldsymbol{\beta}, \sigma_\epsilon^2, \alpha(t_0), w(\mathbf{s}_0)) \\ \times dF(\boldsymbol{\beta}, \boldsymbol{\alpha}, \mathbf{w}, \sigma_\epsilon^2, \sigma_\alpha^2, \rho, \sigma_w^2, \delta, \alpha(t_0), w(\mathbf{s}_0) \mid \mathbf{Y}) \ . \end{aligned} \tag{11.17}$$

Using (11.17), given a random draw $(\boldsymbol{\beta}^*, \sigma_\epsilon^{2*}, \alpha(t_0)^*, w(\mathbf{s}_0)^*)$ from the posterior $f(\boldsymbol{\beta}, \sigma_\epsilon^2, \alpha(t_0), w(\mathbf{s}_0) \mid \mathbf{Y})$, if we draw $Y^*(\mathbf{s}_0, t_0)$ from $N(X'(\mathbf{s}_0, t_0)\boldsymbol{\beta}^* + \alpha(t_0)^* + w(\mathbf{s}_0)^*, \sigma_e^{2*})$, marginally, $Y^*(\mathbf{s}_0, t_0) \sim f(Y(\mathbf{s}_0, t_0) \mid \mathbf{Y})$.

Using sampling-based model fitting and working with (11.10), we obtain samples $(\boldsymbol{\beta}^*, \sigma_\epsilon^{2*}, \sigma_\alpha^{2*}, \rho^*, \sigma_w^{2*}, \delta^*, \boldsymbol{\alpha}^*, \mathbf{w}^*)$ from the posterior distribution, $p(\boldsymbol{\beta}, \sigma_\epsilon^2, \sigma_\alpha^2, \rho, \sigma_w^2, \delta, \boldsymbol{\alpha}, \mathbf{w} \mid \mathbf{Y})$. But $f(\boldsymbol{\beta}, \sigma_\epsilon^2, \sigma_\alpha^2, \rho, \sigma_w^2, \delta, \boldsymbol{\alpha}, \mathbf{w}, \alpha(t_0), w(\mathbf{s}_0) \mid \mathbf{Y}) = f(\alpha(t_0) \mid \boldsymbol{\alpha}, \sigma_\alpha^2, \rho) \cdot f(w(\mathbf{s}_0) \mid \mathbf{w}, \sigma_w^2, \delta) \cdot f(\boldsymbol{\beta}, \sigma_\epsilon^2, \sigma_\alpha^2, \rho, \sigma_w^2, \delta, \boldsymbol{\alpha}, \mathbf{w} \mid \mathbf{Y})$. If, e.g., $t_0 = T + 1$, and $\alpha(t)$ is modeled as a time series, $f(\alpha(T+1) \mid \boldsymbol{\alpha}, \sigma_\alpha^2, \rho)$ is $N(\rho\alpha(T), \sigma_\alpha^2)$. If the $\alpha(t)$ are coefficients associated with dummy variables, setting $\alpha(T+1) = \alpha(T)$ is, arguably, the best one can do. The joint distribution of $\mathbf{w}$ and $w(\mathbf{s}_0)$ is a multivariate normal from which $f(w(\mathbf{s}_0) \mid \mathbf{w}, \sigma_w^2, \delta)$ is a univariate normal. So if $\alpha(t_0)^* \sim f(\alpha(t_0) \mid \boldsymbol{\alpha}^*, \sigma_\alpha^{2*}, \rho^*)$ and $w(\mathbf{s}_0)^* \sim f(w(\mathbf{s}_0) \mid \mathbf{w}^*, \sigma_w^{2*}, \delta^*)$, along with $\boldsymbol{\beta}^*$ and σ_ϵ^{2*} we obtain a draw from $f(\boldsymbol{\beta}, \sigma_\epsilon^2, \alpha(t_0),\ w(\mathbf{s}_0) \mid \mathbf{Y})$. (If $t_0 \epsilon \{1, 2, ..., T\}$, $\alpha(t_0)$ is a component of $\boldsymbol{\alpha}$, then $\alpha(t_0)^*$ is a component of $\boldsymbol{\alpha}^*$. If $\mathbf{s}_0$ is one of the $\mathbf{s}_1, \mathbf{s}_2, ..., \mathbf{s}_n$, $w^*(\mathbf{s}_0)$ is a component of $\mathbf{w}^*$.) Alternatively, one can work with (11.13). Now, having marginalized over $\boldsymbol{\alpha}$ and $\mathbf{w}$, $Y(\mathbf{s}, t)$ and $\mathbf{Y}$ are no longer independent. They have a multivariate normal distribution from which $f(Y(\mathbf{s}, t) \mid \mathbf{Y}, \boldsymbol{\beta}, \sigma_\epsilon^2, \sigma_\alpha^2, \rho, \sigma_w^2, \delta)$ must be obtained. Note that for multiple predictions, $w(\mathbf{s}_0)$ is replaced by a vector, say, $\mathbf{w}_0$. Now $f(\mathbf{w}_0 | \mathbf{w}, \sigma_w^2, \delta)$ is a multivariate normal distribution. No additional complications arise.

Example 11.1 *(Baton Rouge home sales).* We present a portion of the data analysis developed in Gelfand et al. (2004) for sales of single-family homes drawn from two regions in the city of Baton Rouge, LA. The two areas are known as Sherwood Forest and Highland Road. These regions are approximately the same size and have similar levels of transaction activity; they differ chiefly in the range of neighborhood characteristics and house amenities found within. Sherwood Forest is a large, fairly homogeneous neighborhood located east, southeast of downtown Baton Rouge. Highland Road, on the other hand, is a major thoroughfare connecting downtown with the residential area to the southeast. Rather than being one homogeneous neighborhood, the Highland Road area consists, instead, of heterogeneous subdivisions. Employing two regions makes a local isotropy assumption more comfortable and allows investigation of possibly differing time effects and location dynamics.

For these regions, a subsample of all homes sold only once during the period 1985 through 1995 (single-sale transactions) and a second subsample of homes sold more than once (repeat-sale transactions) were drawn. These two samples can be studied separately to assess whether the population of single-sale houses differs from that of repeat-sale houses. The sample sizes are provided by year in Table 11.1. The location of each property is defined by its latitude and longitude coordinates, rescaled to UTM projection. In addition, a variety of house characteristics, to control for physical differences among the properties, are recorded at the time of sale. We use age, living area, other area (e.g., patios, garages, and

	Highland		Sherwood	
Year	Repeat	Single	Repeat	Single
1985	25	40	32	29
1986	20	35	32	39
1987	27	32	27	37
1988	16	26	20	34
1989	21	25	24	35
1990	42	29	27	37
1991	29	30	25	31
1992	33	38	39	27
1993	24	40	31	40
1994	26	35	20	34
1995	26	35	21	32
Total	289	365	298	375

Table 11.1 *Sample size by region, type of sale, and year.*

	Highland		Sherwood	
Variable	Repeat	Single	Repeat	Single
Age	11.10	12.49	14.21	14.75
	(8.15)	(11.37)	(8.32)	(10.16)
Bathrooms	2.18	2.16	2.05	2.02
	(0.46)	(0.56)	(0.36)	(0.40)
Living area	2265.4	2075.8	1996.0	1941.5
	(642.9)	(718.9)	(566.8)	(616.2)
Other area	815.1	706.0	726.0	670.6
	(337.7)	(363.6)	(258.1)	(289.2)

Table 11.2 *Mean (standard deviation) for house characteristics by region and type of sale.*

carports) and number of bathrooms as covariates in our analysis. Summary statistics for these attributes appear in Table 11.2. We see that the homes in the Highland Road area are somewhat newer and slightly larger than those in the Sherwood area. The greater heterogeneity of the Highland Road homes is borne out by the almost uniformly higher standard deviations for each covariate. In fact, we have more than 20 house characteristics in our data set, but elaborating the mean with additional features provides little improvement in R^2 and introduces multicollinearity problems. So, we confine ourselves to the four explanatory variables above and turn to spatial modeling to explain a portion of the remaining variability. Empirical semivariograms (2.9) offer evidence of spatial association, after adjusting for house characteristics.

We describe the results of fitting the model with mean $\mu(\mathbf{s}) = \mathbf{x}(\mathbf{s})^T\boldsymbol{\beta}$ and the error structure in (11.8). This is also the preferred model using the predictive model choice approach of Gelfand and Ghosh (5.14); we omit details. Fixed coefficients were justified by the shortness of the observation period. Again, an exponential isotropic correlation function was adopted.

Variable	Repeat	Single
Highland region:		
intercept (β_0)	11.63 (11.59, 11.66)	11.45 (11.40, 11.50)
age (β_1)	–0.04 (–0.07, –0.02)	–0.08 (–0.11, –0.06)
bathrooms (β_2)	0.02 (–0.01, 0.04)	0.02 (–0.01, 0.05)
living area (β_3)	0.28 (0.25, 0.31)	0.33 (0.29, 0.37)
other area (β_4)	0.08 (0.06, 0.11)	0.07 (0.04, 0.09)
Sherwood region:		
intercept (β_0)	11.33 (11.30, 11.36)	11.30 (11.27, 11.34)
age (β_1)	–0.06 (–0.07, –0.04)	–0.05 (–0.07, –0.03)
bathrooms (β_2)	0.05 (0.03, 0.07)	0.00 (–0.02, 0.02)
living area (β_3)	0.19 (0.17, 0.21)	0.22 (0.19, 0.24)
other area (β_4)	0.02 (0.01, 0.04)	0.06 (0.04, 0.08)

Table 11.3 *Parameter estimates (median and 95% interval estimates) for house characteristics.*

To complete the Bayesian specification, we adopt rather noninformative priors in order to resemble a likelihood/least squares analysis. In particular, we assume a flat prior on the regression parameter $\boldsymbol{\beta}$ and inverse gamma (a, b) priors for σ_ϵ^2, $\sigma_w^{2(t)}$ and $\delta^{(t)}$, $t = 1, \ldots T$. The shape parameter for these inverse gamma priors was fixed at 2, implying an infinite prior variance. We choose the inverse gamma scale parameter for all $\delta^{(t)}$'s to be equal, i.e., $b_{\delta^{(1)}} = b_{\delta^{(2)}} = \ldots = b_{\delta^{(T)}} = b_\delta$, say, and likewise for $\sigma_w^{2(t)}$. Furthermore, we set $b_{\sigma_\epsilon} = b_{\sigma_w^2}$ reflecting uncertain prior contribution from the nugget to the sill. Finally, the exact values of b_{σ_ϵ}, $b_{\sigma_w^2}$ and b_δ vary between region and type of sale reflecting different prior beliefs about these characteristics.

Inference for the house characteristic coefficients is provided in Table 11.3 (point and 95% interval estimates). Age, living area, and other area are significant in all cases; number of bathrooms is significant only in Sherwood repeat sales. Significance of living area is much stronger in Highland than in Sherwood. The Highland sample is composed of homes from several heterogeneous neighborhoods. As such, living area not only measures differences in house size, but may also serve as a partial proxy for construction quality and for neighborhood location within the sample. The greater homogeneity of homes in Sherwood implies less variability in living area (as seen in Table 11.2) and reduces the importance of these variables in explaining house price.

Turning to the error structure, the parameters of interest for each region are the $\sigma_w^{2(t)}$, the $\delta^{(t)}$, and σ_e^2. The sill at time t is $Var(Y(s,t)) = \sigma_w^{2(t)} + \sigma_e^2$. Figure 11.1 plots the posterior medians of these sills. We see considerable difference in variability over the groups and over time, providing support for distinct spatial models at each t. Variability is highest for Highland single sales, lowest for Sherwood repeats. The additional insight is the effect of time. Variability is generally increasing over time.

We can obtain posterior median and interval estimates for $\sigma_w^{2(t)}/(\sigma_e^2 + \sigma_w^{2(t)})$, the proportion of spatial variance to total. The strength of the spatial story is considerable; 40 to 80% of the variability is spatial.

In Figure 11.2 we provide point and interval estimates for the range. The ranges for the repeat sales are quite similar for the two regions, showing some tendency to increase in the later years of observation. By contrast, the range for the Highland single sales is much different from that for Sherwood. It is typically greater and much more variable. The latter again is a refection of the high variability in the single-sale home prices in Highland. The resulting posteriors are more dispersed.

Figure 11.1 *Posterior median sill by year.*

Figure 11.2 *Posterior median and 95% interval estimates for the range by year for (a) Highland repeat sales, (b) Highland single sales, (c) Sherwood repeat sales, and (d) Sherwood single sales.*

Finally, in Figure 11.3, we present the posterior distribution of the block averages, mentioned at the end of Subsection 11.1.2, for each of the four analyses. Again, these block averages are viewed as analogues of more familiar time dummy variables. Time effects are evident. In all cases, we witness somewhat of a decline in magnitude in the 1980s and an increasing trend in the 1990s.

Figure 11.3 *Posterior median and 95% interval estimates for the block averages by year for (a) Highland repeat sales, (b) Highland single sales, (c) Sherwood repeat sales, and (d) Sherwood single sales.*

11.2 Point-level modeling with continuous time

Suppose now that $\mathbf{s} \in \Re^2$ and $t \in \Re^+$ and we seek to define a spatiotemporal process $Y(\mathbf{s}, t)$. As in Subsection 3.1 we have to provide a joint distribution for an uncountable number of random variables. Again, we do this through arbitrary finite dimensional distributions. Confining ourselves to the Gaussian case, we only need to specify a valid spatiotemporal covariance function. Here, "valid" means that for any set of locations and any set of time points, the covariance matrix for the resulting set of random variables is positive definite. An important point here is that it is not sensible to combine $\mathbf{s}$ and t and propose a valid correlation function on $\Re^3$. This is because distance in space has nothing to do with "distance" on the time scale.

As a result, a stationary spatiotemporal covariance specification is assumed to take the form $cov(Y(\mathbf{s}, t), Y(\mathbf{s}', t')) = c(\mathbf{s} - \mathbf{s}', t - t')$. An isotropic form sets $cov(Y(\mathbf{s}, t), Y(\mathbf{s}', t')) = c(||\mathbf{s} - \mathbf{s}'||, |t - t'|)$. A frequently used choice is the *separable* form

$$cov(Y(\mathbf{s}, t), Y(\mathbf{s}', t')) = \sigma^2 \rho^{(1)}(\mathbf{s} - \mathbf{s}'; \boldsymbol{\phi})\, \rho^{(2)}(t - t'; \boldsymbol{\psi}) \ , \tag{11.18}$$

where $\rho^{(1)}$ is a valid two-dimensional correlation function and $\rho^{(2)}$ is a valid one-dimensional correlation function. Expression (11.18) shows that dependence attenuates in a multiplicative manner across space and time. Forms such as (11.18) have a history in spatiotemporal modeling; see, e.g., Mardia and Goodall (1993) and references therein.

Why is (11.18) valid? For locations $\mathbf{s}_1, \ldots, \mathbf{s}_I$ and times $t_1, \ldots, t_J$, collecting the variables a vector $\mathbf{Y}_s^T = (\mathbf{Y}^T(\mathbf{s}_1), \ldots, \mathbf{Y}^T(\mathbf{s}_I))$ where $\mathbf{Y}(\mathbf{s}_i) = (Y(\mathbf{s}_i, t_1), \ldots, Y(\mathbf{s}_i, t_J))^T$, the covariance matrix of $\mathbf{Y}_s$ is

$$\Sigma_{\mathbf{Y}_s}(\sigma^2, \boldsymbol{\phi}, \boldsymbol{\psi}) = \sigma^2 H_s(\boldsymbol{\phi}) \otimes H_t(\boldsymbol{\psi}) \ , \tag{11.19}$$

where "$\otimes$" again denotes the Kronecker product. In (11.19), $H_s(\boldsymbol{\phi})$ is $I \times I$ with $(H_s(\boldsymbol{\phi}))_{ii'} = \rho^{(1)}(\mathbf{s}_i - \mathbf{s}_i'; \boldsymbol{\theta})$, and $H_t(\boldsymbol{\psi})$ is $J \times J$ with $(H_t(\boldsymbol{\psi}))_{jj'} = \rho^{(2)}(t_j - t_{j'}; \boldsymbol{\psi})$. Expression (11.19) clarifies that $\Sigma_{\mathbf{Y}_s}$ is positive definite, following the argument below (9.11). So, $\mathbf{Y}_s$ will

be IJ-dimensional multivariate normal with, in obvious notation, mean vector $\boldsymbol{\mu}_s(\boldsymbol{\beta})$ and covariance matrix (11.19).

Given a prior for $\boldsymbol{\beta}, \sigma^2, \boldsymbol{\phi}$, and $\boldsymbol{\psi}$, the Bayesian model is completely specified. Simulation-based model fitting can be carried out similarly to the static spatial case by noting the following. The log-likelihood arising from $\mathbf{Y}_s$ is

$$\begin{aligned} &-\tfrac{1}{2}\log\left|\sigma^2 H_s(\boldsymbol{\phi})\otimes H_t(\boldsymbol{\psi})\right| \\ &\qquad -\tfrac{1}{2\sigma^2}(\mathbf{Y}_s-\boldsymbol{\mu}_s(\boldsymbol{\beta}))^T(H_s(\boldsymbol{\phi})\otimes H_t(\boldsymbol{\psi}))^{-1}(\mathbf{Y}_s-\boldsymbol{\mu}_s(\boldsymbol{\beta}))\ . \end{aligned}$$

But in fact $\left|\sigma^2 H_s(\boldsymbol{\phi})\otimes H_t(\boldsymbol{\psi})\right| = (\sigma^2)^{IJ}\left|H_s(\boldsymbol{\phi})\right|^J\left|H_t(\boldsymbol{\psi})\right|^I$ and $(H_s(\boldsymbol{\phi})\otimes H_t(\boldsymbol{\psi}))^{-1} = H_s^{-1}(\boldsymbol{\phi})\otimes H_t^{-1}(\boldsymbol{\psi})$ by properties of Kronecker products. In other words, even though (11.19) is $IJ\times IJ$, we need only the determinant and inverse for an $I\times I$ and a $J\times J$ matrix, expediting likelihood evaluation and hence Gibbs sampling.

With regard to prediction, first consider new locations $\mathbf{s}'_1, ..., \mathbf{s}'_K$ with interest in inference for $Y(\mathbf{s}'_k, t_j)$. As with the observed data, we collect the $Y(\mathbf{s}'_k, t_j)$ into vectors $\mathbf{Y}(\mathbf{s}'_k)$, and the $\mathbf{Y}(\mathbf{s}'_k)$ into a single $KJ\times 1$ vector $\mathbf{Y}_{s'}$. Even though we may not necessarily be interested in every component of $\mathbf{Y}_{s'}$, the simplifying forms that follow suggest that, with regard to programming, it may be easiest to simulate draws from the entire predictive distribution $f(\mathbf{Y}_{s'}\mid\mathbf{Y}_s)$ and then retain only the desired components.

Since $f(\mathbf{Y}_{s'}|\mathbf{Y}_s)$ has a form analogous to (7.3), given posterior samples $(\boldsymbol{\beta}^*_g, \sigma^{2*}_g, \boldsymbol{\phi}^*_g, \boldsymbol{\psi}^*_g)$, we draw $\mathbf{Y}^*_{s',g}$ from $f(\mathbf{Y}_{s'}\mid\mathbf{Y}_s, \boldsymbol{\beta}^*_g, \sigma^{2*}_g, \boldsymbol{\phi}^*_g, \boldsymbol{\psi}^*_g)$, $g=1,\ldots,G$. Analogous to (7.4),

$$f\left(\begin{pmatrix}\mathbf{Y}_s\\ \mathbf{Y}_{s'}\end{pmatrix}\middle|\boldsymbol{\beta},\sigma^2,\boldsymbol{\phi},\boldsymbol{\psi}\right) = N\left(\begin{pmatrix}\boldsymbol{\mu}_s(\boldsymbol{\beta})\\ \boldsymbol{\mu}_{s'}(\boldsymbol{\beta})\end{pmatrix}\ ,\ \Sigma_{\mathbf{Y}_s,\mathbf{Y}_{s'}}\right) \tag{11.20}$$

where

$$\Sigma_{\mathbf{Y}_s,\mathbf{Y}_{s'}} = \sigma^2\begin{pmatrix} H_s(\boldsymbol{\phi})\otimes H_t(\boldsymbol{\psi}) & H_{s,s'}(\boldsymbol{\phi})\otimes H_t(\boldsymbol{\psi}) \\ H^T_{s,s'}(\boldsymbol{\phi})\otimes H_t(\boldsymbol{\psi}) & H_{s'}(\boldsymbol{\phi})\otimes H_t(\boldsymbol{\psi})\end{pmatrix}\ ,$$

with obvious definitions for $H_{s'}(\boldsymbol{\phi})$ and $H_{s,s'}(\boldsymbol{\phi})$. But then the conditional distribution $\mathbf{Y}_{s'}\mid\mathbf{Y}_s,\boldsymbol{\beta},\sigma^2,\boldsymbol{\phi},\boldsymbol{\psi}$ is also normal, with mean

$$\begin{aligned} &\boldsymbol{\mu}_{s'}(\boldsymbol{\beta}) + (H^T_{s,s'}(\boldsymbol{\phi})\otimes H_t(\boldsymbol{\phi}))(H_s(\boldsymbol{\phi})\otimes H_t(\boldsymbol{\psi}))^{-1}(Y_s-\boldsymbol{\mu}_s(\boldsymbol{\beta})) \\ &\quad = \boldsymbol{\mu}_{s'}(\boldsymbol{\beta}) + (H^T_{s,s'}(\boldsymbol{\phi})H_s^{-1}(\boldsymbol{\phi})\otimes I_{J\times J})(\mathbf{Y}_s-\boldsymbol{\mu}_s(\boldsymbol{\beta}))\ , \end{aligned} \tag{11.21}$$

and covariance matrix

$$\begin{aligned} &H_{s'}(\boldsymbol{\phi})\otimes H_t(\boldsymbol{\psi}) \\ &\quad -(H^T_{s,s'}\otimes H_t(\boldsymbol{\psi}))(H_s(\boldsymbol{\phi})\otimes H_t(\boldsymbol{\psi}))^{-1}(H_{s,s'}(\boldsymbol{\phi})\otimes H_t(\boldsymbol{\psi})) \\ &\qquad = (H_{s'}(\boldsymbol{\phi}) - H^T_{s,s'}(\boldsymbol{\phi})H_s^{-1}(\boldsymbol{\phi})H_{s,s}(\boldsymbol{\phi}))\otimes H_t(\boldsymbol{\psi})\ , \end{aligned} \tag{11.22}$$

using standard properties of Kronecker products.

In (11.21), time disappears apart from $\boldsymbol{\mu}_{s'}(\boldsymbol{\beta})$, while in (11.22), time "factors out" of the conditioning. Sampling from this normal distribution usually employs the inverse square root of the conditional covariance matrix, but conveniently, this is

$$(H_{s'}(\boldsymbol{\phi}) - H^T_{s,s'}(\boldsymbol{\phi})H_s^{-1}(\boldsymbol{\phi})H_{s,s'}(\boldsymbol{\phi}))^{-\frac{1}{2}}\otimes H_t^{-\frac{1}{2}}(\boldsymbol{\psi})\ ,$$

so the only work required beyond that in (7.5) is obtaining $H_t^{-\frac{1}{2}}(\boldsymbol{\psi})$, since $H_t^{-1}(\boldsymbol{\psi})$ will already have been obtained in evaluating the likelihood, following the discussion above.

For prediction not for points but for areal units (blocks) $B_1,\ldots,B_K$, we would set $\mathbf{Y}^T(B_k) = (Y(B_k,t_1),\ldots,Y(B_k,t_J))$ and then further set $\mathbf{Y}^T_B = (\mathbf{Y}^T(B_1),\ldots,\mathbf{Y}^T(B_K))$. Analogous to (7.6) we seek to sample $f(\mathbf{Y}_B\mid\mathbf{Y}_s)$, so we require $f(\mathbf{Y}_B\mid$

Figure 11.4 *Observed 1-hour maximum ozone measurement by day, July 1995, 10 Atlanta monitoring sites.*

$\mathbf{Y}_s, \boldsymbol{\beta}, \sigma^2, \boldsymbol{\phi}, \boldsymbol{\psi})$. Analogous to (11.20), this can be derived from the joint distribution $f((\mathbf{Y}_s, \mathbf{Y}_B)^T \mid \boldsymbol{\beta}, \sigma^2, \boldsymbol{\phi}, \boldsymbol{\psi})$, which is

$$N\left(\begin{pmatrix} \boldsymbol{\mu}_s(\boldsymbol{\beta}) \\ \boldsymbol{\mu}_B(\boldsymbol{\beta}) \end{pmatrix}, \sigma^2 \begin{pmatrix} H_s(\boldsymbol{\phi}) \otimes H_t(\boldsymbol{\psi}) & H_{s,B}(\boldsymbol{\phi}) \otimes H_t(\boldsymbol{\psi}) \\ H_{s,B}^T(\boldsymbol{\phi}) \otimes H_t(\boldsymbol{\psi}) & H_B(\boldsymbol{\phi}) \otimes H_t(\boldsymbol{\psi}) \end{pmatrix}\right),$$

with $\boldsymbol{\mu}_B(\boldsymbol{\beta})$, $H_B(\boldsymbol{\phi})$, and $H_{s,B}(\boldsymbol{\phi})$ defined as in Section 7.1.2. Thus the distribution $f(\mathbf{Y}_B|\mathbf{Y}_s, \boldsymbol{\beta}, \sigma^2, \boldsymbol{\phi}, \boldsymbol{\psi})$ is again normal with mean and covariance matrix as given in (11.21) and (11.22), but with $\boldsymbol{\mu}_B(\boldsymbol{\beta})$ replacing $\boldsymbol{\mu}_{s'}(\boldsymbol{\beta})$, $H_B(\boldsymbol{\phi})$ replacing $H_{s'}(\boldsymbol{\phi})$, and $H_{s,B}(\boldsymbol{\phi})$ replacing $H_{s,s'}(\boldsymbol{\phi})$. Using the same Monte Carlo integrations as proposed in Section 7.1.2 leads to sampling the resultant $\widehat{f}(\mathbf{Y}_B \mid \mathbf{Y}_s, \boldsymbol{\beta}, \sigma^2, \boldsymbol{\phi}, \boldsymbol{\psi})$, and the same technical justification applies.

If we started with block data, $Y(B_i, t_j)$, then following (7.11) and (11.19),

$$f(\mathbf{Y}_B \mid \boldsymbol{\beta}, \sigma^2, \boldsymbol{\phi}, \boldsymbol{\psi}) = N(\mu_B(\boldsymbol{\beta}), \sigma^2(H_B(\boldsymbol{\phi}) \otimes H_t(\boldsymbol{\psi})). \tag{11.23}$$

Given (11.23), the path for prediction at new points or at new blocks is clear, following the above and the end of Section 7.1.2; we omit the details.

Note that the association structure in (11.18) allows *forecasting* of the spatial process at time t_{J+1}. This can be done at observed or unobserved points or blocks following the foregoing development. To retain the above simplifying forms, we would first simulate the variables at t_{J+1} associated with observed points or blocks (with no change of support). We would then revise $H_t(\boldsymbol{\phi})$ to be $(J+1) \times (J+1)$ before proceeding as above.

Example 11.2 To illustrate the methods above, we use a spatiotemporal version of the Atlanta ozone data set. As mentioned in Section 7.1, we actually have ozone measurements at the 10 fixed monitoring stations shown in Figure 1.3 over the 92 summer days in 1995. Figure 11.4 shows the daily 1-hour maximum ozone reading for the sites during July of this same year. There are several sharp peaks, but little evidence of a weekly (seven-day) period in the data. The mean structure appears reasonably constant in space, with the ordering

	Spatial only		Spatiotemporal	
	Point	95% Interval	Point	95% Interval
Point A	.125	(.040, .334)	.139	(.111, .169)
Point B	.116	(.031, .393)	.131	(.098, .169)
Zip 30317 (east-central)	.130	(.055, .270)	.138	(.121, .155)
Zip 30344 (south-central)	.123	(.055, .270)	.135	(.112, .161)
Zip 30350 (north)	.112	(.040, .283)	.109	(.084, .140)

Table 11.4 *Posterior medians and 95% equal-tail credible intervals for ozone levels at two points, and for average ozone levels over three blocks (zip codes), purely spatial model versus spatiotemporal model, Atlanta ozone data for July 15, 1995.*

of the site measurements changing dramatically for different days. Moreover, with only 10 "design points" in the metro area, any spatial trend surface we fit would be quite speculative over much of the study region (e.g., the northwest and southwest metro; see Figure 1.3). The temporal evolution of the series is not inconsistent with a constant mean autoregressive error model; indeed, the lag 1 sample autocorrelation varies between .27 and .73 over the 10 sites, strongly suggesting the need for a model accounting for both spatial and temporal correlations.

We thus fit our spatiotemporal model with mean $\boldsymbol{\mu}(\mathbf{s}, t; \boldsymbol{\beta}) = \mu$, but with spatial and temporal correlation functions $\rho^{(1)}(\mathbf{s}_i - \mathbf{s}_{i'}; \phi) = e^{-\phi\|\mathbf{s}_i - \mathbf{s}_{i'}\|}$ and $\rho^{(2)}(t_j - t_{j'}; \psi) = \psi^{|j-j'|}$. Hence our model has four parameters: we use a flat prior for μ, an $IG(3, 0.5)$ prior for σ^2, a $G(0.003, 100)$ prior for ϕ, and a $U(0, 1)$ prior for ψ (thus eliminating the implausible possibility of *negative* autocorrelation in our data, but favoring no positive value over any other). To facilitate our Gibbs-Metropolis approach, we transform to $\theta = \log \phi$ and $\lambda = \log(\psi/(1-\psi))$, and subsequently use Gaussian proposals on these transformed parameters.

Running 3 parallel chains of 10,000 iterations each, sample traces (not shown) again indicate virtually immediate convergence of our algorithm. Posterior medians and 95% equal-tail credible intervals for the four parameters are as follows: for μ, 0.068 and (0.057, 0.080); for σ^2, 0.11 and (0.08, 0.17); for ϕ, 0.06 and (0.03, 0.08); and for ψ, 0.42 and (0.31, 0.52). The rather large value of ψ confirms the strong temporal autocorrelation suspected in the daily ozone readings.

Comparison of the posteriors for σ^2 and ϕ with those obtained for the static spatial model in Example 7.1 is not sensible, since these parameters have different meanings in the two models. Instead, we make this comparison in the context of point-point and point-block prediction. Table 11.4 provides posterior predictive summaries for the ozone concentrations for July 15, 1995, at points A and B (see Figure 1.3), as well as for the block averages over three selected Atlanta city zips: 30317, an east-central city zip very near to two monitoring sites; 30344, the south-central zip containing the points A and B; and 30350, the northernmost city zip. Results are shown for both the spatiotemporal model of this subsection and for the static spatial model previously fit in Example 7.1. Note that all the posterior medians are a bit higher under the spatiotemporal model, except for that for the northern zip, which remains low. Also note the significant increase in precision afforded by this model, which makes use of the data from all 31 days in July 1995, instead of only that from July 15. Figure 11.5 shows the estimated posteriors giving rise to the first and last rows in Table 11.4 (i.e., corresponding to the the July 15, 1995, ozone levels at point A and the block average over the northernmost city zip, 30350). The Bayesian approach's ability to reflect differing amounts of predictive uncertainty for the two models is clearly evident.

Finally, Figure 11.6 plots the posterior medians and upper and lower .025 quantiles produced by the spatiotemporal model by day for the ozone concentration at point A, as well as those for the block average in zip 30350. Note that the overall temporal pattern

Figure 11.5 *Posterior predictive distributions for ozone concentration at point A and the block average over zip 30350, purely spatial model versus spatiotemporal model, Atlanta ozone data for July 15, 1995.*

is quite similar to that for the data shown in Figure 11.4. Since point A is rather nearer to several data observation points, the confidence bands associated with it are often a bit narrower than those for the northern zip, but this pattern is not perfectly consistent over time. Also note that the relative positions of the bands for July 15 are consistent with the data pattern for this day seen in Figure 1.3, when downtown ozone exposures were higher than those in the northern metro. Finally, the day-to-day variability in the predicted series is substantially larger than the predictive variability associated with any given day.

11.3 Nonseparable spatiotemporal models ⋆

The separable form for the spatiotemporal covariance function in (11.18) is convenient for computation and offers attractive interpretation. However, its form limits the nature of space-time interaction. Additive forms, arising from $w(\mathbf{s},t) = w(\mathbf{s}) + \alpha(t)$ with $w(\mathbf{s})$ and $\alpha(t)$ independent may be even more unsatisfying.

A simple way to extend (11.18) is through *mixing*. For instance, suppose $w(\mathbf{s},t) = w_1(\mathbf{s},t) + w_2(\mathbf{s},t)$ with w_1 and w_2 independent processes, each with a separable spatiotemporal covariance function, say $c_\ell(\mathbf{s}-\mathbf{s}', t-t) = \sigma_\ell^2 \rho_\ell^{(1)}(\mathbf{s}-\mathbf{s}')\rho_\ell^{(2)}(t-t')$, $\ell = 1,2$. Then the covariance function for $w(\mathbf{s},t)$ is evidently the sum and is not separable. Building covariance functions in this way is easy to interpret but yields an explosion of parameters with finite mixing. Continuous parametric mixing, e.g.,

$$c(\mathbf{s}-\mathbf{s}', t-t) = \sigma^2 \int \rho^{(1)}(\mathbf{s}-\mathbf{s}', \boldsymbol{\phi})\rho^{(2)}(t-t', \boldsymbol{\psi})\, G_{\boldsymbol{\gamma}}(d\boldsymbol{\phi}, d\boldsymbol{\psi}) \ , \tag{11.24}$$

yields a function that depends only on σ^2 and $\boldsymbol{\gamma}$. Extensions of these ideas are developed

Figure 11.6 *Posterior medians and upper and lower .025 quantiles for the predicted 1-hour maximum ozone concentration by day, July 1995; solid lines, point A; dotted lines, block average over zip 30350 (northernmost Atlanta city zip).*

in De Iaco et al. (2002). However, these forms have not received much attention in the literature to date.

Cressie and Huang (1999) introduce a flexible class of nonseparable stationary covariance functions that allow for space-time interaction. However, they work in the spectral domain and require that $c(\mathbf{s}-\mathbf{s}', t-t')$ can be computed explicitly, i.e., the Fourier inversion can be obtained in closed-form. Unfortunately this occurs only in very special cases. Recent work by Gneiting (2002) adopts a similar approach but obtains very general classes of valid space-time models that do not rely on closed form Fourier inversions. One simple example is the class $c(\mathbf{s}-\mathbf{s}', t-t') = \sigma^2(|t-t'|+1)^{-1}\exp(-||\mathbf{s}-\mathbf{s}'||(|t-t'|+1)^{-\beta/2})$. Here, β is a space-time interaction parameter; $\beta = 0$ provides a separable specification.

Stein (2005) also works in the spectral domain, providing a class of spectral densities whose resulting spatiotemporal covariance function is nonseparable with flexible analytic behavior. These spectral densities extend the Matérn form; see (3.5) or the discussion below on Equation (A.2) in Appendix A. In particular, the spectral density is

$$\widehat{c}(\mathbf{w}, v) \propto [c_1(\alpha_1^2 + ||\mathbf{w}||^2)^{\alpha_1} + c_2(\alpha_2 + v^2)^{\alpha_2}]^{-v} .$$

Unfortunately, the associated covariance function cannot be computed explicitly; fast Fourier transforms (see Appendix Section A.1) offer the best computational prospects. Also, unlike Gneiting's class, separability does not arise as a special or limiting case. For further discussion of the above, see Chapter 23 of Gelfand et al. (2010). We also mention related work using "blurring" discussed in Brown, Kåresen, Roberts, and Tonellato (2000).

11.4 Dynamic spatiotemporal models ⋆

In this section we follow the approach taken in Banerjee, Gamerman, and Gelfand, (2003), viewing the data as arising from a time series of spatial processes. In particular, we work in

the setting of dynamic models (West and Harrison, 1997), describing the temporal evolution in a latent space. We achieve a class of dynamic models for spatiotemporal data.

Here, there is a growing literature. Non-Bayesian approaches include Huang and Cressie (1996), Wikle and Cressie (1999), and Mardia et al. (1998). Bayesian approaches include Tonellato (1997), Sanso and Guenni (1999), Stroud et al. (2001), and Huerta et al. (2003). The paper by Stroud et al. (2001) is attractive in being applicable to any data set that is continuous in space and discrete in time and allows straightforward computation using Kalman filtering.

11.4.1 Brief review of dynamic linear models

Dynamic linear models, often referred to as state-space models in the time-series literature, offer a versatile framework for fitting several time-varying models (West and Harrison, 1997). We briefly outline the general dynamic linear modeling framework. Thus, let $\mathbf{Y}_t$ be a $m \times 1$ vector of observables at time t. $\mathbf{Y}_t$ is related to a $p \times 1$ vector, $\boldsymbol{\theta}_t$, called the state vector, through a *measurement equation*. In general, the elements of $\boldsymbol{\theta}_t$ are not observable, but are generated by a first-order Markovian process, resulting in a *transition equation*. Therefore, we can describe the above framework as

$$\begin{aligned} \mathbf{Y}_t &= F_t \boldsymbol{\theta}_t + \boldsymbol{\epsilon}_t,\ \boldsymbol{\epsilon}_t \sim N\left(\mathbf{0}, \Sigma_t^{\epsilon}\right). \\ \boldsymbol{\theta}_t &= G_t \boldsymbol{\theta}_{t-1} + \boldsymbol{\eta}_t,\ \boldsymbol{\eta}_t \sim N\left(\mathbf{0}, \Sigma_t^{\eta}\right)\ , \end{aligned}$$

where F_t and G_t are $m \times p$ and $p \times p$ matrices, respectively. The first equation is the measurement equation, where $\boldsymbol{\epsilon}_t$ is a $m \times 1$ vector of serially uncorrelated Gaussian variables with mean $\mathbf{0}$ and an $m \times m$ covariance matrix, Σ_t^{ϵ}. The second equation is the transition equation with $\boldsymbol{\eta}_t$ being a $p \times 1$ vector of serially uncorrelated zero-centered Gaussian disturbances and Σ_t^{η} the corresponding $p \times p$ covariance matrix. Note that under (11.25), the association structure can be computed explicitly across time, e.g., $Cov\left(\boldsymbol{\theta}_t, \boldsymbol{\theta}_{t-1}\right) = G_t Var\left(\boldsymbol{\theta}_{t-1}\right)$ and $Cov\left(\mathbf{Y}_t, \mathbf{Y}_{t-1}\right) = F_t G_t Var\left(\boldsymbol{\theta}_{t-1}\right) F_t^T$.

F_t (in the measurement equation) and G_t (in the transition equation) are referred to as *system matrices* that may change over time. F_t and G_t may involve unknown parameters but, given the parameters, temporal evolution is in a predetermined manner. The matrix F_t is usually specified by the design of the problem at hand, while G_t is specified through modeling assumptions; for example, $G_t = I_p$, the $p \times p$ identity matrix would provide a random walk for $\boldsymbol{\theta}_t$. Regardless, the system is linear, and for any time point t, $\mathbf{Y}_t$ can be expressed as a linear combination of the present $\boldsymbol{\epsilon}_t$ and the present and past $\boldsymbol{\eta}_t$'s.

11.4.2 Formulation for spatiotemporal models

In this section we adapt the above dynamic modeling framework to univariate spatiotemporal models with spatially varying coefficients. For this we consider a collection of sites $S = \{\mathbf{s}_1, ..., \mathbf{s}_{N_s}\}$, and time-points $T = \{t_1, ..., t_{N_t}\}$, yielding observations $Y(\mathbf{s}, t)$, and covariate vectors $\mathbf{x}(\mathbf{s}, t)$, for every $(\mathbf{s}, t) \in S \times T$.

The response, $Y(\mathbf{s}, t)$, is first modeled through a measurement equation, which incorporates the measurement error, $\epsilon(\mathbf{s}, t)$, as serially and spatially uncorrelated zero-centered Gaussian disturbances. The transition equation now involves the regression parameters (slopes) of the covariates. The slope vector, say $\tilde{\boldsymbol{\beta}}(\mathbf{s}, t)$, is decomposed into a purely temporal component, $\boldsymbol{\beta}_t$, and a spatiotemporal component, $\boldsymbol{\beta}(\mathbf{s}, t)$. Both these are generated through transition equations, capturing their Markovian dependence in time. While the transition equation of the purely temporal component is as in usual state-space modeling, the spatiotemporal component is generated by a multivariate Gaussian spatial process.

Thus, we may write the spatiotemporal modeling framework as

$$\begin{aligned} Y(\mathbf{s},t) &= \mu(\mathbf{s},t)+\epsilon(\mathbf{s},t)\,;\ \epsilon(\mathbf{s},t)\overset{ind}{\sim} N\left(0,\sigma^2\right), && (11.25)\\ \mu(\mathbf{s},t) &= \mathbf{x}^T(\mathbf{s},t)\,\tilde{\boldsymbol{\beta}}(\mathbf{s},t), \\ \tilde{\boldsymbol{\beta}}(\mathbf{s},t) &= \boldsymbol{\beta}_t+\boldsymbol{\beta}(\mathbf{s},t), && (11.26)\\ \boldsymbol{\beta}_t &= \boldsymbol{\beta}_{t-1}+\boldsymbol{\eta}_t,\ \boldsymbol{\eta}_t\overset{ind}{\sim} N_p\left(\mathbf{0},\Sigma_{\boldsymbol{\eta}}\right), \\ \text{and } \boldsymbol{\beta}(\mathbf{s},t) &= \boldsymbol{\beta}(\mathbf{s},t-1)+\mathbf{w}(\mathbf{s},t)\ . \end{aligned}$$

In (11.26), we introduce a linear model of coregionalization (Section 9.5) for $\mathbf{w}(\mathbf{s},t)$, i.e., $\mathbf{w}(\mathbf{s},t)=A\mathbf{v}(\mathbf{s},t)$, with $\mathbf{v}(\mathbf{s},t)=(v_1(\mathbf{s},t),...,v_p(\mathbf{s},t))^T$, yielding $\Sigma_w=AA^T$. The $v_l(\mathbf{s},t)$ are serially independent replications of a Gaussian process with unit variance and correlation function $\rho_l(\cdot\,;\phi_l)$, henceforth denoted by $GP(0,\rho_l(\cdot\,;\phi_l))$, for $l=1,\ldots,p$ and independent across l. In the current context, we assume that A does not depend upon $(\mathbf{s},t)$. Nevertheless, this still allows flexible modeling for the spatial covariance structure, as we discuss below.

Moreover, allowing a spatially varying coefficient $\boldsymbol{\beta}(\mathbf{s},t)$ to be associated with $\mathbf{x}(\mathbf{s},t)$ provides an arbitrarily rich explanatory relationship for the x's with regard to the Y's (see Section 9.6 in this regard). By comparison, in Stroud et al. (2001), at a given t, a locally weighted mixture of linear regressions is proposed and only the purely temporal component of $\tilde{\boldsymbol{\beta}}(\mathbf{s},t)$ is used. Such a specification requires both number of basis functions and number of mixture components.

Returning to our specification, note that if $v_l(\cdot,t)\overset{ind}{\sim} GP(0,\rho(\cdot;\phi))$, we have the intrinsic or separable model for $w(\mathbf{s},t)$. Allowing different correlation functions and decay parameters for the $v_l(\mathbf{s},t)$, i.e., $v_l(\cdot,t)\overset{ind}{\sim} GP(0,\rho_l(\cdot;\phi_l))$ yields the linear model of coregionalization (Section 9.5).

Following Section 11.4.1, we can compute the general association structure for the Y's under (11.25) and (11.26). For instance, we have the result that

$$\operatorname{Cov}(Y(\mathbf{s},t),Y(\mathbf{s}',t-1))=\mathbf{x}^T(\mathbf{s},t)\,\Sigma_{\tilde{\boldsymbol{\beta}}(\mathbf{s},t),\tilde{\boldsymbol{\beta}}(\mathbf{s}',t-1)}\mathbf{x}(\mathbf{s},t-1)\ ,$$

where $\Sigma_{\tilde{\boldsymbol{\beta}}(\mathbf{s},t),\tilde{\boldsymbol{\beta}}(\mathbf{s}',t-1)}=(t-1)\left(\Sigma_{\boldsymbol{\eta}}+\sum_{l=1}^p\rho_l(\mathbf{s}-\mathbf{s}';\phi_l)\,\mathbf{a}_l\mathbf{a}_l^T\right)$. Furthermore,

$$\operatorname{Var}(Y(\mathbf{s},t))=\mathbf{x}^T(\mathbf{s},t)\,t\left[\Sigma_{\boldsymbol{\eta}}+AA^T\right]\mathbf{x}(\mathbf{s},t)$$

with the result that $\operatorname{Corr}(Y(\mathbf{s},t),Y(\mathbf{s}',t-1))=O(1)$ as $t\to\infty$.

A Bayesian hierarchical model for (11.25) and (11.26) may be completed by prior specifications such as

$$\begin{aligned} \boldsymbol{\beta}_0 &\sim N(\mathbf{m}_0,C_0)\ \text{ and } \boldsymbol{\beta}(\cdot,0)\equiv 0. && (11.27)\\ \Sigma_{\boldsymbol{\eta}} &\sim IW\left(a_{\boldsymbol{\eta}},B_\eta\right),\ \Sigma_{\mathbf{w}}\sim IW(a_{\mathbf{w}},B_{\mathbf{w}})\ \text{ and } \sigma_\epsilon^2\sim IG(a_\epsilon,b_\epsilon), \\ \mathbf{m}_0 &\sim N(\mathbf{0},\Sigma_0)\,;\ \Sigma_0=10^5\times I_p\ , \end{aligned}$$

where B_η and $B_{\mathbf{w}}$ are $p\times p$ precision (hyperparameter) matrices for the inverted Wishart distribution.

Consider now data, in the form $(Y(\mathbf{s}_i,t_j))$ with $i=1,2,...,N_s$ and $j=1,2,...,N_t$. Let us collect, for each time point, the observations on all the sites. That is, we form, $\mathbf{Y}_t=(Y(\mathbf{s}_1,t),...,Y(\mathbf{s}_{N_s},t))^T$ and the $N_s\times N_sp$ block diagonal matrix $F_t=\left(\mathbf{x}^T(\mathbf{s}_1,t),\mathbf{x}^T(\mathbf{s}_2,t),...,\mathbf{x}^T(\mathbf{s}_N,t)\right)$ for $t=t_1,...,t_{N_t}$. ^{T}. Analogously we form the $N_sp\times 1$ vector $\boldsymbol{\theta}_t=\mathbf{1}_{N_s}\otimes\boldsymbol{\beta}_t+\boldsymbol{\beta}_t^*$, where $\boldsymbol{\beta}_t^*=(\boldsymbol{\beta}(\mathbf{s}_1,t),\ldots,\boldsymbol{\beta}(\mathbf{s}_{N_s},t))^T$, $\boldsymbol{\beta}_t=\boldsymbol{\beta}_{t-1}+\boldsymbol{\eta}_t$, $\boldsymbol{\eta}_t\overset{ind}{\sim}$

$N_p\left(\mathbf{0}, \Sigma_{\boldsymbol{\eta}}\right)$; and, with $\mathbf{w}_t = \left(\mathbf{w}^T\left(\mathbf{s}_1, t\right), \ldots, \mathbf{w}^T\left(\mathbf{s}_{N_s}, t\right)\right)^T$,

$$\boldsymbol{\beta}_t^* = \boldsymbol{\beta}_{t-1}^* + \mathbf{w}_t,\ \mathbf{w}_t \overset{ind}{\sim} N\left(\mathbf{0}, \sum_{l=1}^{p}\left(R_l\left(\phi_l\right) \otimes \Sigma_{\mathbf{w},l}\right)\right) ,$$

where $[R_l(\phi_l)]_{ij} = \rho_l(\mathbf{s}_i - \mathbf{s}_j; \phi_l)$ is the correlation matrix for $v_l(\cdot, t)$. We then write the data equation for a dynamic spatial model as

$$\mathbf{Y}_t = F_t\boldsymbol{\theta}_t + \boldsymbol{\epsilon}_t;\ t = 1, \ldots, N_t;\ \boldsymbol{\epsilon}_t \sim N\left(\mathbf{0}, \sigma_\epsilon^2 I_{N_s}\right) .$$

With the prior specifications in (11.27), we can design a Gibbs sampler with Gaussian full conditionals for the temporal coefficients $\{\boldsymbol{\beta}_t\}$, the spatiotemporal coefficients $\{\boldsymbol{\beta}_t^*\}$, inverted Wishart for $\Sigma_{\boldsymbol{\eta}}$, and Metropolis steps for ϕ and the elements of $\Sigma_{\mathbf{w},l}$. Updating of $\Sigma_{\mathbf{w}} = \sum_{l=1}^{p}\Sigma_{\mathbf{w},l}$ is most efficiently done by reparametrizing the model in terms of the matrix square root of $\Sigma_{\mathbf{w}}$, say, A, and updating the elements of the lower triangular matrix A. To be precise, consider the full conditional distribution,

$$\begin{aligned} f\left(\Sigma_{\mathbf{w}}|\boldsymbol{\gamma}, \phi_1, \phi_2\right) \propto\ & f\left(\Sigma_{\mathbf{w}}|a_\gamma, B_\gamma\right) \frac{1}{\left|\sum_{l=1}^{p} R_l(\phi_l)\otimes\Sigma_{\mathbf{w},l}\right|} \\ & \times \exp\left(-\tfrac{1}{2}\boldsymbol{\beta}^{*T}\left(J^{-1} \otimes \left(\textstyle\sum_{l=1}^{p} R_l\left(\phi_l\right) \otimes \Sigma_{\mathbf{w},l}\right)^{-1}\right)\boldsymbol{\beta}^*\right) . \end{aligned}$$

The one-to-one relationship between elements of $\Sigma_{\mathbf{w}}$ and the Cholesky square root A is well known (see, e.g., Harville, 1997, p. 235). So, we reparametrize the above full conditional as

$$\begin{aligned} f\left(A|\boldsymbol{\gamma}, \phi_1, \phi_2\right) \propto\ & f\left(h\left(A\right)|a_\gamma, B_\gamma\right) \left|\frac{\partial h}{\partial a_{ij}}\right| \frac{1}{\left|\sum_{l=1}^{p} R_l(\phi_l)\otimes\left(\mathbf{a}_l\mathbf{a}_l^T\right)\right|} \\ & \times \exp\left(-\tfrac{1}{2}\boldsymbol{\beta}^{*T}\left(J^{-1} \otimes \left(\textstyle\sum_{l=1}^{p} R_l\left(\phi_l\right) \otimes \left(\mathbf{a}_l\mathbf{a}_l^T\right)\right)^{-1}\right)\boldsymbol{\beta}^*\right) . \end{aligned}$$

Here, h is the function taking the elements of A, say, a_{ij}, to those of the symmetric positive definite matrix $\Sigma_{\mathbf{w}}$. In the 2×2 case we have

$$h\left(a_{11}, a_{21}, a_{22}\right) = \left(a_{11}^2, a_{11}a_{21}, a_{21}^2 + a_{22}^2\right) ,$$

and the Jacobian is $4a_{11}^2a_{22}$. Now, the elements of A are updated with univariate random-walk Metropolis proposals: lognormal or gamma for a_{11} and a_{22}, and normal for a_{21}. Additional computational burden is created, since now the likelihood needs to be computed for each of the three updates, but the chains are much better tuned (by controlling the scale of the univariate proposals) to move around the parameter space, thereby leading to better convergence behavior.

Example 11.3 (*Modeling temperature given precipitation*). Our spatial domain, shown in Figure 11.7 along with elevation contours (in 100-m units), provides a sample of 50 locations (indicated by "+") in the state of Colorado. Each site provides information on monthly maximum temperature, and monthly mean precipitation. We denote the temperature summary in location $\mathbf{s}$ at time t, by $Y\left(\mathbf{s}, t\right)$, and the precipitation by $x\left(\mathbf{s}, t\right)$. Forming a covariate vector $\mathbf{x}^T\left(\mathbf{s}, t\right) = \left(1, x\left(\mathbf{s}, t\right)\right)$, we analyze the data using a coregionalized dynamic model, as outlined in Subsection 11.4.2. As a result, we have an intercept process $\tilde{\beta}_0\left(\mathbf{s}, t\right)$ and a slope process $\tilde{\beta}_1\left(\mathbf{s}, t\right)$, and the two processes are dependent.

Figure 11.8 displays the time-varying intercepts and slopes (coefficient of precipitation). As expected, the intercept is higher in the summer months and lower in the winter months, highest in July, lowest in December. In fact, the gradual increase from January to July, and the subsequent decrease toward December is evident from the plot. Precipitation seems to have a negative impact on temperature, although this seems to be significant only in the months of January, March, May, June, November, and December, i.e., seasonal pattern is retrieved although no such structure is imposed.

Figure 11.7 *Map of the region in Colorado that forms the spatial domain. The data for the illustrations come from 50 locations, marked by "+" signs in this region.*

Table 11.5 displays the credible intervals for elements of the $\Sigma_{\boldsymbol{\eta}}$ matrix. Rows 1 and 2 show the medians and credible intervals for the respective *variances*; while Row 3 shows the *correlation*. The corresponding results for the elements of $\Sigma_{\mathbf{w}}$ are given in Table 11.6. A significant negative correlation is seen between the intercept and the slope processes, justifying our use of dependent processes. Next, in Table 11.7, we provide the measurement error variances for temperature along with the estimates of the spatial correlation parameters for the intercept and slope process. Also presented are the ranges implied by ϕ_1 and ϕ_2 for the marginal intercept process, $w_1(\mathbf{s})$, and the marginal slope process, $w_2(\mathbf{s})$. The first range is computed by solving for the distance d, $\rho_1(\phi_1, d) = 0.05$, while the second range is obtained by solving $\left(a_{21}^2 \exp(-\phi_1 d) + a_{22}^2 \exp(-\phi_2 d)\right) / \left(a_{21}^2 + a_{22}^2\right) = 0.05$. The ranges are presented in units of 100 km with the maximum observed distance between our sites being approximately 742 km.

Finally, Figure 11.9 displays the time-sliced image-contour plots for the slope process; similar figures can be drawn for the intercept process. For both processes, the spatial variation is better captured in the central and western edges of the domain. In Figure 11.9, all the months display broadly similar spatial patterns, with denser contour variations toward the west than the east. However, the spatial pattern does seem to be more pronounced in the months with more extreme weather, namely in the winter months of November through January and the summer months of June through August.

11.4.3 Spatiotemporal data

A natural extension of the modeling of the previous sections is to the case where we have data correlated at spatial locations across time. If, as in Section 11.2, we assume that time

Figure 11.8 *Posterior distributions for the time-varying parameters in the temperature given precipitation example. The top graph corresponds to the intercept, while the lower one is the coefficient of precipitation. Solid lines represent the medians while the dashed lines correspond to the upper and lower credible intervals.*

Σ_η	Median (2.5%, 97.5%)
$\Sigma_\eta\,[1,1]$	0.296 (0.130, 0.621)
$\Sigma_\eta\,[2,2]$	0.786 (0.198, 1.952)
$\Sigma_\eta\,[1,2]\,/\sqrt{\Sigma_\eta\,[1,1]\,\Sigma_\eta\,[2,2]}$	–0.562 (–0.807, –0.137)

Table 11.5 *Estimates of the variances and correlation from Σ_η, dynamic spatiotemporal modeling example.*

is discretized to a finite set of equally spaced points on a scale, we can conceptualize a time series of spatial processes that are observed only at the spatial locations $\mathbf{s}_1, \ldots, \mathbf{s}_n$.

Adopting a general notation that parallels (9.54), let

$$Y(\mathbf{s},t) = \mathbf{X}^T(\mathbf{s},t)\,\tilde{\boldsymbol{\beta}}(\mathbf{s},t) + \epsilon(\mathbf{s},t)\ ,\ \ t = 1,2,\ldots,M\ . \tag{11.28}$$

That is, we introduce spatiotemporally varying intercepts and spatiotemporally varying slopes. Alternatively, if we write $\tilde{\boldsymbol{\beta}}(\mathbf{s},t) = \boldsymbol{\beta}(\mathbf{s},t) + \boldsymbol{\mu}_\beta$, we are partitioning the total error into $p+1$ spatiotemporal intercept pieces including $\epsilon(\mathbf{s},t)$, each with an obvious interpretation. So we continue to assume that $\epsilon(\mathbf{s},t) \overset{iid}{\sim} N\left(0,\tau^2\right)$, but need to specify a model for $\tilde{\boldsymbol{\beta}}(\mathbf{s},t)$. Regardless, (11.28) defines a nonstationary process having moments $E(Y(\mathbf{s},t)) = \mathbf{X}^T(\mathbf{s},t)\tilde{\boldsymbol{\beta}}(s,t)$, $Var(Y(\mathbf{s},t)) = \mathbf{X}^T(\mathbf{s},t)\Sigma_{\tilde{\boldsymbol{\beta}}(,t)}\mathbf{X}(\mathbf{s},t) + \tau^2$, and $Cov(Y(\mathbf{s},t), Y(\mathbf{s}',t')) = \mathbf{X}^T(\mathbf{s},t)\Sigma_{\tilde{\beta}(\mathbf{s},t),\tilde{\beta}(\mathbf{s}',t')}\mathbf{X}(\mathbf{s}',t')$.

$\Sigma_{\mathbf{w}}$	Median (2.5%, 97.5%)
$\Sigma_{\mathbf{w}}[1,1]$	0.017 (0.016, 0.019)
$\Sigma_{\mathbf{w}}[2,2]$	0.026 (0.0065, 0.108)
$\Sigma_{\mathbf{w}}[1,2]/\sqrt{\Sigma_{\mathbf{w}}[1,1]\Sigma_{\mathbf{w}}[2,2]}$	–0.704 (–0.843, –0.545)

Table 11.6 *Estimates of the variances and correlation from $\Sigma_{\mathbf{w}}$, dynamic spatiotemporal modeling example.*

Parameters	Median (2.5%, 97.5%)
σ_ϵ^2	0.134 (0.106, 0.185)
ϕ_1	1.09 (0.58, 2.04)
ϕ_2	0.58 (0.37, 1.97)
Range for intercept process	2.75 (1.47, 5.17)
Range for slope process	4.68 (1.60, 6.21)

Table 11.7 *Nugget effects and spatial correlation parameters, dynamic spatiotemporal modeling example.*

Section 11.4 handled (11.28) using a dynamic model. Here we consider four alternative specifications for $\boldsymbol{\beta}(\mathbf{s},t)$. Paralleling the customary assumption from longitudinal data modeling (where the time series are usually short), we could set

- **Model 1:** $\boldsymbol{\beta}(\mathbf{s},t) = \boldsymbol{\beta}(\mathbf{s})$, where $\boldsymbol{\beta}(\mathbf{s})$ is modeled as in the previous sections. This model can be viewed as a locally linear growth curve model.
- **Model 2:** $\boldsymbol{\beta}(\mathbf{s},t) = \boldsymbol{\beta}(\mathbf{s}) + \boldsymbol{\alpha}(t)$, where $\boldsymbol{\beta}(\mathbf{s})$ is again as in Model 1. In modeling $\boldsymbol{\alpha}(t)$, two possibilities are (i) treat the $\alpha_k(t)$ as time dummy variables, taking this set of pM variables to be *a priori* independent and identically distributed; and (ii) model the $\boldsymbol{\alpha}(t)$ as a random walk or autoregressive process. The components could be assumed independent across k, but for greater generality, we take them to be dependent, using a separable form that replaces $\mathbf{s}$ with t and takes ρ to be a valid correlation function in just one dimension.
- **Model 3:** $\boldsymbol{\beta}(\mathbf{s},t) = \boldsymbol{\beta}^{(t)}(\mathbf{s})$, i.e., we have spatially varying coefficient processes nested within time. This model is an analogue of the nested effects areal unit specification in Waller et al. (1997); see also Gelfand, Ecker et al. (2003). The processes are assumed independent across t (essentially dummy time processes) and permit temporal evolution of the coefficient process. Following Subsection 9.6.2, the process $\boldsymbol{\beta}^{(t)}(\mathbf{s})$ would be mean-zero, second-order stationary Gaussian with cross-covariance specification at time t, $C^{(t)}(\mathbf{s},\mathbf{s}')$ where $\left(C^{(t)}(\mathbf{s},\mathbf{s}')\right)_{lm} = \rho\left(\mathbf{s}-\mathbf{s}';\phi^{(t)}\right)\tau_{lm}^{(t)}$. We have specified Model 3 with a common $\boldsymbol{\mu}_\beta$ across time. This enables some comparability with the other models we have proposed. However, we can increase flexibility by replacing $\boldsymbol{\mu}_\beta$ with $\boldsymbol{\mu}_\beta^{(t)}$.
- **Model 4:** For $\rho^{(1)}$ a valid two-dimensional correlation function, $\rho^{(2)}$ a valid one-dimensional choice, and T positive definite symmetric, $\boldsymbol{\beta}(\mathbf{s},t)$ such that $\Sigma_{[\boldsymbol{\beta}(\mathbf{s},t),\boldsymbol{\beta}(\mathbf{s}',t')]} = \rho^{(1)}(\mathbf{s}-\mathbf{s}';\phi)\,\rho^{(2)}(t-t';\gamma)\,T$. This model proposes a separable covariance specification in space and time, as in Section 11.2. Here $\rho^{(1)}$ obtains spatial association as in earlier subsections that is attenuated across time by $\rho^{(2)}$. The resulting covariance matrix for the full vector $\boldsymbol{\beta}$, blocked by site and time within site has the convenient form $H_2(\gamma)\otimes H_1(\phi)\otimes T$.

In each of the above models we can marginalize over $\boldsymbol{\beta}(\mathbf{s},t)$ as we did earlier in this section. Depending upon the model it may be more computationally convenient to block the data by site or by time. We omit the details and notice only that, with n sites and T

Figure 11.9 *Time-sliced image-contour plots displaying the posterior mean surface of the spatial residuals corresponding to the slope process in the temperature given precipitation model.*

Model	Independent process G	P	D_∞	Dependent process G	P	D_∞
1	88.58	56.15	144.73	54.54	29.11	83.65
2a	77.79	50.65	128.44	47.92	26.95	74.87
2b	74.68	50.38	125.06	43.38	29.10	72.48
3a	59.46	48.55	108.01	43.74	20.63	64.37
3b	57.09	48.41	105.50	42.35	21.04	63.39
4	53.55	52.98	106.53	37.84	26.47	64.31

Table 11.8 *Model choice criteria for various spatiotemporal process models.*

time points, the resulting likelihood will involve the determinant and inverse of an $nT \times nT$ matrix (typically a large matrix; see Appendix Section A.2).

Note that all of the foregoing modeling can be applied to the case of cross-sectional data where the set of observed locations varies with t. This is the case, for instance, with our real estate data. We only observe a selling price at the time of a transaction. With n_t locations in year t, the likelihood for all but Model 3 will involve a $\sum n_t \times \sum n_t$ matrix.

Example 11.4 *(Baton Rouge housing prices (contd.)).* We now turn to the dynamic version of Baton Rouge dataset presented in Example 9.4. From the Baton Rouge database we drew a sample of 120 transactions at distinct spatial locations for the years 1989, 1990, 1991, and

1992. We compare Models 1–4. In particular, we have two versions of Model 2; 2a has the $\boldsymbol{\alpha}(t)$ as four i.i.d. time dummies, while 2b uses the multivariate temporal process model for $\boldsymbol{\alpha}(t)$. We also have two versions of Model 3; 3a has a common $\boldsymbol{\mu}_\beta$ across t, while 3b uses $\boldsymbol{\mu}_\beta^{(t)}$. In all cases the five-dimensional spatially varying coefficient model for $\boldsymbol{\beta}$'s was employed. Table 11.8 shows the results. Model 3, where space is nested within time, turns out to be the best with Model 4 following closely behind. We omit the posterior inference summary for Model 3b, noting only that the overall coefficients $\left(\boldsymbol{\mu}_\beta^{(t)}\right)$ do not change much over time. However, there is some indication that spatial range is changing over time.

11.5 Fitting dynamic spatiotemporal models using `spBayes`

`spBayes` offers a relatively simple, but rather flexible, univariate version of the dynamic models discussed in the preceding section. Suppose, $y_t(\mathbf{s})$ denotes the observation at location $\mathbf{s}$ and time t. We model $y_t(\mathbf{s})$ through a *measurement equation* that provides a regression specification with a space-time varying intercept and serially and spatially uncorrelated zero-centered Gaussian disturbances as measurement error $\epsilon_t(\mathbf{s})$. Next a *transition equation* introduces a $p \times 1$ coefficient vector, say, $\boldsymbol{\beta}_t$, which is a purely temporal component (i.e., time-varying regression parameters), and a spatio-temporal component $u_t(\mathbf{s})$. Both these are generated through transition equations, capturing their Markovian dependence in time. While the transition equation of the purely temporal component is akin to usual state-space modeling, the spatio-temporal component is generated using Gaussian spatial processes. The overall model, for $t = 1, 2, \ldots, N_t$, is written as

$$\begin{aligned} y_t(\mathbf{s}) &= \mathbf{x}_t(\mathbf{s})^\top \boldsymbol{\beta}_t + u_t(\mathbf{s}) + \epsilon_t(\mathbf{s}), \quad \epsilon_t(\mathbf{s}) \overset{ind.}{\sim} N(0, \tau_t^2)\ ; \\ \boldsymbol{\beta}_t &= \boldsymbol{\beta}_{t-1} + \boldsymbol{\eta}_t, \quad \boldsymbol{\eta}_t \overset{i.i.d.}{\sim} N(0, \boldsymbol{\Sigma}_\eta)\ ; \\ u_t(\mathbf{s}) &= u_{t-1}(\mathbf{s}) + w_t(\mathbf{s}), \quad w_t(\mathbf{s}) \overset{ind.}{\sim} GP\left(\mathbf{0}, C_t(\cdot, \boldsymbol{\theta}_t)\right)\ , \end{aligned} \tag{11.29}$$

where the abbreviations *ind.* and *i.i.d* are *independent* and *independent and identically distributed*, respectively. Here $\mathbf{x}_t(\mathbf{s})$ is a $p \times 1$ vector of predictors and $\boldsymbol{\beta}_t$ is a $p \times 1$ vector of coefficients. In addition to an intercept, $\mathbf{x}_t(\mathbf{s})$ can include location specific variables useful for explaining the variability in $y_t(\mathbf{s})$. The $GP(\mathbf{0}, C_t(\cdot, \boldsymbol{\theta}_t))$ denotes a spatial Gaussian process with covariance function $C_t(\cdot; \boldsymbol{\theta}_t)$. We customarily specify $C_t(\mathbf{s}_1, \mathbf{s}_2; \boldsymbol{\theta}_t) = \sigma_t^2 \rho(\mathbf{s}_1, \mathbf{s}_2; \phi_t)$, where $\boldsymbol{\theta}_t = \{\sigma_t^2, \phi_t\}$ and $\rho(\cdot; \phi)$ is a *correlation function* with ϕ controlling the correlation decay and σ_t^2 represents the spatial variance component. We further assume $\boldsymbol{\beta}_0 \sim N(\mathbf{m}_0, \boldsymbol{\Sigma}_0)$ and $u_0(\mathbf{s}) \equiv 0$, which completes the prior specifications leading to a well-identified Bayesian hierarchical model with reasonable dependence structures. In practice, estimation of model parameters are usually very robust to these hyper-prior specifications. Also note that (11.29) reduces to a simple spatial regression model for $t = 1$.

We consider settings where the inferential interest lies in spatial prediction or interpolation over a region for a set of discrete time points. We also assume that the same locations are monitored for each time point resulting in a space-time matrix whose rows index the locations and columns index the time points, i.e., the (i, j)-th element is $y_j(\mathbf{s}_i)$. Our algorithm will accommodate the situation where some cells of the space-time data matrix may have missing observations, as is common in monitoring environmental variables.

The dynamic model (11.29) and a computationally efficient low-rank version using the *predictive process* (see Section 12.4) are implemented in the `spDynLM` function. Here we illustrate the full rank dynamic model using an ozone monitoring dataset that was previously analyzed by Sahu and Bakar (2012). This is a relatively small dataset and does not require dimension reduction.

Figure 11.10 *Open and filled circle symbols indicate the location of 28 ozone monitoring stations across New York State. Filled circle symbols identify those stations that have half of the daily ozone measurements withheld to assess model predictive performance.*

The dataset comprises 28 Environmental Protection Agency monitoring stations that recorded ozone from July 1 to August 31, 2006. The outcome is daily 8-hour maximum average ozone concentrations (parts per billion; O3.8HRMAX), and predictors include maximum temperature (Celsius; cMAXTMP), wind speed (knots; WDSP), and relative humidity (RM). Of the $1,736$ possible observations, i.e., n=28 locations times N_t=62 daily O3.8HRMAX measurements, 114 are missing. In this illustrative analysis we use the predictors cMAXTMP, WDSP, and RM as well as the spatially and temporally structured residuals to predict missing O3.8HRMAX values. To gain a better sense of the dynamic model's predictive performance, we withheld half of the observations from the records of three stations for subsequent validation. Figure 11.10 shows the monitoring station locations and identifies those stations where data were withheld.

The first `spDynLM` function argument is a list of N_t symbolic model statements representing the regression within each time step. This can be easily assembled using the `lapply` function as shown in the code below. Here too, we define the station coordinates as well as starting, tuning, and prior distributions for the model parameters. Exploratory data analysis using time step specific variograms can be helpful for defining starting values and prior support for parameters in $\boldsymbol{\theta}_t$ and τ_t^2. To avoid cluttering the code, we specify the same prior for the ϕ_t's, σ_t^2's, and τ_t^2's. As in the other `spBayes` model functions, one can choose among several popular spatial correlation functions including the exponential, spherical, Gaussian and Matérn. The exponential correlation function is specified in the `spDynLM` call below. Unlike other model functions described in the preceding sections, the `spDynLM` function will accept `NA` $y_t(\mathbf{s})$ values. The sampler will provide posterior predictive samples for these missing values. If the `get.fitted` argument is `TRUE` then these posterior predictive samples are saved along with posterior *fitted* values for locations where the outcomes are observed.

```
> mods <- lapply(paste("O3.8HRMAX.",1:N.t, "~cMAXTMP.",1:N.t,
```

Figure 11.11 *Posterior distribution medians and 95% credible intervals for model intercept and predictors.*

```
+                        "+WDSP.",1:N.t, "+RH.",1:N.t, sep=""),
+                         as.formula)
> p <- 4 ##number of predictors
> coords <- NYOzone.dat[,c("X.UTM","Y.UTM")]/1000
> max.d <- max(iDist(coords))
> starting <- list("beta"=rep(0,N.t*p),
+                  "phi"=rep(3/(0.5*max.d),N.t),
+                  "sigma.sq"=rep(2,N.t), "tau.sq"=rep(1,N.t),
+                  "sigma.eta"=diag(rep(0.01, p)))
> tuning <- list("phi"=rep(2, N.t))
> priors <- list("beta.0.Norm"=list(rep(0,p), diag(100000,p)),
+                "phi.Unif"=list(rep(3/(0.9*max.d), N.t),
+                                rep(3/(0.05*max.d), N.t)),
+                "sigma.sq.IG"=list(rep(2,N.t), rep(25,N.t)),
+                "tau.sq.IG"=list(rep(2,N.t), rep(25,N.t)),
+                "sigma.eta.IW"=list(2, diag(0.001,p)))
> n.samples <- 5000
> m.i <- spDynLM(mods, data=NYOzone.dat,
+                coords=as.matrix(coords), starting=starting,
+                tuning=tuning, priors=priors, get.fitted=TRUE,
+                cov.model="exponential", n.samples=n.samples,
+                n.report=2500)
```

Time series plots of parameters' posterior summary statistics are often useful for exploring the temporal evolution of the parameters. In the case of the regression coefficients, these plots describe the time-varying trend in the outcome and impact of covariates. For example, the sinusoidal pattern in the model intercept, β_0, seen in Figure 11.11, correlates strongly with both cMAXTMP, RM, and to a lesser degree with WDSP. With only a maximum of 28 observations within each time step, there is not much information to inform estimates of $\boldsymbol{\theta}$. As seen in Figure 11.12, this paucity of information is reflected in the imprecise CI's for the ϕ's and small deviations from the priors on σ^2 and τ^2. There are, however, noticeable trends in the variance components over time.

Figure 11.13 shows the observed and predicted values for the three stations used for validation. Here, open circle symbols indicate those observations used for parameter

Figure 11.12 *Posterior distribution medians and 95% credible intervals for* $\boldsymbol{\theta}$ *and* τ^2.

Figure 11.13 *Posterior predicted distribution medians and 95% credible intervals, solid and dashed lines, respectively, for three stations. Open circle symbols indicate those observations use for model parameter estimation and filled circle symbols indicate those observations withheld for validation.*

estimation and filled circles identify holdout observations. The posterior predicted median and 95% CI's are overlaid using solid and dashed lines, respectively. Three of the 36 holdout measurements fell outside of their 95% predicted CI, a ∼92% coverage rate. As noted in Sahu and Bakar (2012), there is a noticeable reduction in ozone levels in the last two weeks in August.

11.6 Geostatistical space-time modeling driven by differential equations

The objective of this section is to consider space-time modeling in the context of stochastic differential equations, in particular, using stochastic diffusion processes. Such processes arise frequently in application. For example, we find various environmental diffusions: for emerging diseases such as avian or H1N1 flu, for the progression of invasive species, and for transformation of the landscape. In Section 8.8.3, we find a diffusion model to describe a space-time point pattern, urban development with regard to single family homes. Here, we consider spread in space and time in the geostatistical setting, with associated uncertainty, with potential explanatory covariates. Our starting point is a deterministic integro-difference

equation or partial differential equation. Many of the ideas in this section arise from the work of Wikle and colleagues. See, e.g., Wikle and Hooten (2006, 2010) and references therein. In general, differential equations that have analytical solutions are too simple to capture what we seek in practice. So, to accommodate more flexible forms, we adopt a strategy which carries out discretization in time. In fact, it may be argued that we should begin with a temporally discretized version, incorporating the features we seek, rather than attempting to frame a particular SDE.

We continue to work within our hierarchical paradigm,

$$[data|process, parameters][process|parameters][parameters]$$

We continue to work within the Bayesian framework, employing structured dependence in space and time. So, model fitting using MCMC will be challenging and we will typically have to resort to dimension reduction techniques (see Chapter 12).

More precisely, our discretization envisions continuous space with discrete time, i.e., $w_t(\mathbf{s})$. Without loss of generality we can take time to be $t \in \{1, 2, ..., T\}$. The customary terminology here refers to $w_t(\mathbf{s})$ as a *dynamical* process. In fact, we simplify to a first order Markov process, i.e., for the finite set of locations $\mathbf{s}_1, \mathbf{s}_2, ..., \mathbf{s}_n$, let $\mathbf{w}_t = (w_t(\mathbf{s}_1), w_t(\mathbf{s}_2), ..., w_t(\mathbf{s}_n))^T$. Then $[\mathbf{w}_t|\mathbf{w}_0, \mathbf{w}_1, \ldots, \mathbf{w}_{t-1}] = [\mathbf{w}_t \mid \mathbf{w}_{t-1}]$. For example, a linear update would be

$$\mathbf{w}_t = H\mathbf{w}_{t-1} + \boldsymbol{\eta}_t \tag{11.30}$$

where $\eta_t(\mathbf{s})$ incorporates spatial structure. In the literature, this is referred to as a vector AR(1) model and H is called the propagator matrix. The modeling challenge is the specification of H.

We can look at several cases. For instance, consider $H = I$. Note that this form will not provide stationary behavior but, more importantly, there is no interaction across space and time and so this choice would not be realistic for most dynamic processes of interest. Next, consider $H = \text{Diag}(h)$ where $\text{Diag}(h)$ has diagonal elements $0 < h_i < 1$. Now, we achieve stationarity but still we have no space-time interaction. A more general form is an integro-difference equation (IDE),

$$w_t(\mathbf{s}) = \int h(\mathbf{s}, \mathbf{r}; \phi) w_{t-1}(\mathbf{r}) d\mathbf{r} + \eta_t(\mathbf{s}). \tag{11.31}$$

We see that (11.31) does enable the dynamics we seek. In particular, in (11.31), h is a "redistribution kernel," providing redistribution in space which determines the rate of *diffusion* and the *advection*. If we require $w > 0$, then we could work with $\log w_t(\mathbf{s}) = \log(\int h(\mathbf{s}, \mathbf{r}; \phi) w_{t-1}(\mathbf{r}) d\mathbf{r}) + \eta_t(\mathbf{s})$. Again, we resort to discretization in order to supply the H matrix. In this regard, we might begin with forms for h in (11.31); would we want a stationary choice, $h(\mathbf{s}, \mathbf{r}; \phi)$ or would a time-dependent choice, $h_t(\mathbf{s}, \mathbf{r}; \phi)$ be more appropriate? Might ϕ depend upon $\mathbf{r}$?

Recall the linear partial differential equation (PDE), $\frac{dw(\mathbf{s},t)}{dt} = h(\mathbf{s})w(\mathbf{s}, t)$. Applying finite differencing yields $w(\mathbf{s}, t + \Delta t) - w(\mathbf{s}, t) = h(\mathbf{s})w(\mathbf{s}, t)\Delta t$, i.e., $w(\mathbf{s}, t + 1) \approx \tilde{h}(\mathbf{s})w(\mathbf{s}, t)$. We see that the linear PDE suffers the same problems as we noted above. There is no space-time interaction; there is no redistribution over space. So, we need more general PDE's which, as noted above, can motivate an IDE, can illuminate the choice of H.

In this regard, it may be useful to note the "forward" vs. "backward" perspective associated with an IDE. In one sense, we can think of (11.31) as moving forward in time, taking us from a current "state," $w_t(\mathbf{s}), \mathbf{s} \in D$ to a new state, $w_{t+1}(\mathbf{s}), \mathbf{s} \in D$. In another sense, we can look backwards, thinking of (11.31) as clarifying how the "state" $w_t(\mathbf{r}), \mathbf{r} \in D$ contributed to give us the current state, $w_{t+1}(\mathbf{s}), \mathbf{s} \in D$. Depending upon the process we are modeling, specification of h may emerge more naturally under one perspective rather

than the other. Also, IDE's can be specified directly without using PDE's. That is, $h(\mathbf{s}, \mathbf{r})$ can be developed using process-based assumptions, e.g., as a sum of a survival/growth term plus a birth/replenishment term as in Ghosh et al. (2012).

Now, consider a diffusion in one dimension. Fick's Law of diffusion (Fick, 1855) asserts that the diffusive flux from *high* concentration to *low* is $-\delta\frac{\partial w(x,t)}{\partial x}$ with δ being the diffusion coefficient. More flexible dynamics arise with a location varying diffusion coefficient $\delta(x)$ supplying a location varying diffusion. The associated diffusion equation is $\partial w/\partial t = -\partial \texttt{flux}/\partial x$, i.e., $\frac{\partial w(x,t)}{\partial t} = \frac{\partial}{\partial x}(\delta(x)\frac{\partial w(x,t)}{\partial x})$ Applying the chain rule, the one-dimensional diffusion equation is

$$\frac{\partial w(x,t)}{\partial t} = \delta^{'}(x)\frac{\partial w(x,t)}{\partial x} + \delta(x)\frac{\partial^2 w(x,t)}{\partial x^2}. \tag{11.32}$$

Moving to two-dimensional space, writing $\mathbf{s} = (x, y)$, the diffusive flux is $-\delta(x, y)\nabla w(x, y, t)$ where $\nabla w(x, y, t)$ is the concentration gradient at time t. Now, the resulting diffusion PDE is

$$\frac{\partial w(x,y,t)}{\partial t} = \frac{\partial}{\partial x}(\delta(x,y)\frac{\partial w(x,y,t)}{\partial x}) + \frac{\partial}{\partial y}(\delta(x,y)\frac{\partial w(x,y,t)}{\partial y}). \tag{11.33}$$

We can complete the chain rule calculation to explicitly obtain the spatial diffusion equation, noting that it will involve the second order partial derivatives, $\frac{\partial^2 w}{\partial x^2}$ and $\frac{\partial^2 w}{\partial y^2}$. Now, suppose we introduce Δt, Δx, Δy and replace ∂'s with finite differences (first forward and second order centered). The resulting expressions are elaborate and messy but are developed in careful detail in Hooten and Wikle (2007). The critical point is that, after the smoke clears, we obtain the propagator matrix H to insert into $\mathbf{w}_{t+\Delta t} = H\mathbf{w}_t$. Again, we would add independent spatial noise at each time point, $\boldsymbol{\eta}_t$. Evidently, we are back to our earlier redistribution form in (11.6).

Hooten and Wikle (2007) illustrate with data from the U.S. Breeding Bird Survey, focusing on the Eurasian collared dove. For the years 1986–2003, the data consist of recorded bird counts by sight (for three minutes) over a collection of routes each roughly of length 40kms with 50 stops per route. The counts are attached to grid boxes i in year t and are given a Poisson specification reflecting the number of visits to the site in a given year and a model for the associated expected counts (intensities, see Chapter 8) per visit. The H matrix is a function of the vector of local diffusion coefficients. Since the number of sites is large, dimension reduction is applied to the vector of w's (see Chapter 12). We do not offer further detail here, encouraging the reader to consul the Hooten and Wikle (2007) paper. Instead we present a different geostatistical example below.

The foregoing dynamics redistribute the existing population spatially over time. However, in many situations it would be the case that there is growth or decline in the population. For instance, with housing stock, while some new homes are built, others are torn down. With species populations, change in population size is *density dependent*; competition may encourage or discourage population growth. Hence, we might attempt to add a growth rate to the model. An illustrative choice, which we employ below, is the logistic differential equation (see, e.g., Kot, 2001),

$$\frac{\partial w(\mathbf{s},t)}{\partial t} = rw(\mathbf{s},t)\left(1 - \frac{w(\mathbf{s},t)}{K}\right). \tag{11.34}$$

Here, r is the growth rate and K is the carrying capacity. In practice, we would imagine a spatially varying growth rate, $r(\mathbf{s})$ and perhaps even a spatially varying capacity, $K(\mathbf{s})$.

Turning to more general structures, suppose $w(\mathbf{s}, t)$ is a mean (second stage) specification for a space-time geostatistical model or GLM (or perhaps an intensity for a space-time point pattern, as in Section 10.3.3). A general deterministic diffusion PDE for $w(\mathbf{s}, t)$ looks like

$$\frac{\partial w(\mathbf{s},t)}{\partial t} = a(w(\mathbf{s},t), v(\mathbf{s},t); \theta) \tag{11.35}$$

where $v(\mathbf{s},t)$ includes other potential variables; in its simplest form, $v(\mathbf{s},t)=t$. Furthermore, we might extend θ to $\theta(\mathbf{s})$ or perhaps to $\theta(\mathbf{s},t)$.

To think about adding uncertainty, ignore location $\mathbf{s}$ for the moment and consider a usual nonlinear differential equation, $d\mu(t)=a(\mu(t),t,\theta)dt$ with $\mu(0)=\mu_0$. A simple way to add stochasticity is to make θ random. However, this imposes the likely unreasonable assumption that the functional form of the equation is *true*. Instead, we might assume $d\mu(t)=a(\mu(t),t,\theta)dt+b(\mu(t),t,\theta)dZ(t)$ where $Z(t)$ is variance 1 Brownian motion over R^1 (Section 3.2) with a and b the "drift" and "volatility," respectively. Now we obtain a *stochastic* differential equation (SDE) in which we would still assume θ to be random. A bit more generality is achieved with the form, $d\mu(t)=a(\mu(t),t,\theta(t))dt$ where $\theta(t)$ is driven by an SDE, $d\theta(t)=g(\theta(t),t,\beta)dt+h(\theta(t),\sigma)dZ(t)$ where, again, $Z(t)$ is variance 1 Brownian motion.

Now, we add space. We write $d\mu(\mathbf{s},t)=a(\mu(\mathbf{s},t),t,\theta(\mathbf{s}))dt$ with $\mu(\mathbf{s},0))=\mu_0(\mathbf{s})$, a partial differential equation (PDE). We add randomness through $\theta(\mathbf{s})$, a process realization, resulting in a stochastic process of differential equations. Again, we don't believe the form of the PDE is true. So, we write

$$d\mu(\mathbf{s},t)=a(\mu(\mathbf{s},t),t,\theta(\mathbf{s}))dt+b(\mu(\mathbf{s},t),t,\theta(\mathbf{s}))dZ(\mathbf{s},t)\ , \tag{11.36}$$

where we need to extend our modeling of Brownian motion to $Z(\mathbf{s},t)$. For a fixed finite set of spatial locations we customarily assume independent Brownian motion at each location, allowing the $\theta(\mathbf{s})$ process to provide the spatial dependence. Again, we can work with the more general form, $d\mu(\mathbf{s},t)=a(\mu(\mathbf{s},t),t,\theta(\mathbf{s},t))dt$ where, simplifying the volatility, $d\theta(\mathbf{s},t)=g(\theta(\mathbf{s},t))dt+bdZ(\mathbf{s},t)$. When $d\theta(\mathbf{s},t)=\alpha(\theta(\mathbf{s},t)-\theta(\mathbf{s}))dt+bdZ(\mathbf{s},t)$ with $\theta(\mathbf{s})$ a process realization as above, we have characterized $\theta(\mathbf{s},t)$ through an infinite dimensional SDE and it can be shown that the associated covariance function of the space time process is separable.

The usual space-time "geostatistics" setting with observations at locations and times in the foregoing context takes the form $Y(\mathbf{s},t)=w(\mathbf{s},t)+\epsilon(\mathbf{s},t)$ where now, $w(\mathbf{s},t)$ is modeled through a stochastic PDE. In particular, the logistic equation now extended to space and time becomes

$$\frac{\partial w(\mathbf{s},t)}{\partial t}=r(\mathbf{s},t)w(\mathbf{s},t)\left(1-\frac{w(\mathbf{s},t)}{K(\mathbf{s})}\right). \tag{11.37}$$

Again, with time discretized to intervals Δt, indexed as $t_j, j=0,1,2,...J$ and locations $\mathbf{s}_i$, our data takes the form $\{Y(\mathbf{s}_i,t_j)\}$ with resulting dynamic model $Y(\mathbf{s}_i,t_j)=w(\mathbf{s}_i,t_j)+\varepsilon(\mathbf{s}_i,t_j)$. Using Euler's approximation yields the difference equation:

$$\Delta w(\mathbf{s},t_j)=r(\mathbf{s},t_j)w(\mathbf{s},t_{j-1})\left[1-\frac{w(\mathbf{s},t_{j-1})}{K(\mathbf{s})}\right]\Delta t, \tag{11.38}$$

$$w(\mathbf{s},t_j)\approx w(\mathbf{s},0)+\sum_{l=1}^{j}\Delta w(\mathbf{s},t). \tag{11.39}$$

For $r(\mathbf{s},t)$, we adopt the model given above for $\theta(\mathbf{s},t)$, i.e., $dr(\mathbf{s},t)=\alpha(r(\mathbf{s},t)-r(\mathbf{s}))dt+bdZ(\mathbf{s},t)$. For the initial positive $w(\mathbf{s},0)$ and $K(\mathbf{s})$ we can use log-Gaussian spatial processes with regression forms for the means:

$$\log w(\mathbf{s},0)=\mu_w(X_w(\mathbf{s}),\beta_w)+\eta_w(\mathbf{s})$$

with $\eta_w(\mathbf{s})\sim \mathrm{GP}\left(0,\sigma_w^2\rho_w(\mathbf{s}-\mathbf{s}';\phi_w)\right)$ and

$$\log K(\mathbf{s})=\mu_K(X_K(\mathbf{s}),\beta_K)+\eta_K(\mathbf{s})$$

Figure 11.14 *Simulated data at four selected locations marked as 1, 2, 3, and 4 are shown as small circles. Hold-out data at four randomly chosen locations for out-of-sample validation is shown in diamond shape and marked as A, B, C and D.*

with $\eta_K(\mathbf{s}) \sim \mathrm{GP}\left(0, \sigma_K^2 \rho_K(\mathbf{s}-\mathbf{s}'; \phi_K)\right)$. In fact, to simplify here, we assume $K(s)$ known and set to 1, yielding interpretation of w on a *percent* scale with K as 100% or full capacity. We add similar modeling for $\mu_r(\mathbf{s})$. Altogether, we have specified a hierarchical model. After specifying priors, we fit with MCMC cumulatively in w over time. We consider the usual prediction questions: (i) interpolating the past at new locations and (ii) forecasting the future at current and new locations

For our specific example, we use a 10×10 study region introducing 44 locations over 30 time periods with four sites retained as holdout for validation (see Figure 11.14). To illustrate with a different choice of covariance function, we adopted the Matérn with smoothness parameter $\nu = 3/2$ for both $w_0(\mathbf{s})$ and $r(\mathbf{s})$. Hence, the resulting space time covariance function for $r(\mathbf{s}, t)$ becomes

$$\sigma_r^2 \exp\left(-\alpha\left|t_{j_1} - t_{j_2}\right|\right)\left(1 + \phi_r\left|\left|\mathbf{s}_{i_1} - \mathbf{s}_{i_2}\right|\right|\right) \exp\left(-\phi_r\left|\mathbf{s}_{i_1} - \mathbf{s}_{i_2}\right|\right) \ .$$

We use the simulated $\mathbf{r}$ and $\mathbf{w}_0$ with the transition equation recursively to obtain $\boldsymbol{\Delta} w_j$ and $\mathbf{w}_j$ for each of the 30 periods. The observed data are sampled as mutually independent given $\mathbf{w}_j$ with the random noise $\boldsymbol{\varepsilon}_j$. The data at four selected locations marked as 1, 2, 3, and 4 in Figure 11.14 are shown as small circles. We leave out the data at four randomly chosen locations (shown in diamond shape and marked as A, B, C and D in Figure 11.14) for spatial prediction and out-of-sample validation for our model.

We fit the foregoing model to the data at the remaining 40 locations (hence a 40×30 spatiotemporal data set). We use very vague priors for the constant means: $\pi(\mu_w) \sim N\left(0, 10^8\right)$ and $\pi(\mu_r) \sim N\left(0, 10^8\right)$. We use conjugate gamma priors for the precision parameters of r and w_0: $\pi\left(1/\sigma_r^2\right) \sim Gamma(1,1)$ and $\pi\left(1/\sigma_w^2\right) \sim Gamma(1,1)$. The positive parameter for the temporal correlation of r also has a vague log-normal prior: $\pi(\alpha) \sim log\text{-}N\left(0, 10^8\right)$. Because the spatial range parameters ϕ_r and ϕ_w are only weakly identified, we only use informative and discrete prior for them. Indeed, we have chosen 20 values (from 0.1 to 2.0) and assume uniform priors over them for both ϕ_r and ϕ_w.

We use the random-walk Metropolis-Hastings algorithm to simulate posterior samples of $\mathbf{r}$ and $\mathbf{w}_0$. We draw the entire vector of $\mathbf{w}_0$ for all 40 locations as a single block in every

Model Parameters	True Value	Posterior Mean	95% Equal-tail Interval
μ_w	-4.2	-4.14	(-4.88, -3.33)
σ_w	1.0	0.91	(0.62, 1.46)
ϕ_w	0.7	0.77	(0.50, 1.20)
σ_ε	0.05	0.049	(0.047, 0.052)
μ_r	0.24	0.24	(0.22, 0.26)
σ_r	0.08	0.088	(0.077, 0.097)
ϕ_r	0.7	0.78	(0.60, 1.10)
α	0.6	0.64	(0.51, 0.98)

Table 11.9 *Parameters and their posterior inference for the simulated example.*

iteration. Because $\mathbf{r}$ is very high-dimensional (a 40×30 matrix concatenated into a vector), we cannot draw the entire matrix of $\mathbf{r}$ as one block and achieve a satisfactory acceptance rate. So, we partition $\mathbf{r}$ into 40 row blocks (location-wise) in every odd-numbered iteration and 30 column blocks (period-wise) in every even numbered iteration. Each block is drawn in one Metropolis step. The posterior samples start to converge after about 30,000 iterations. Given the sampled $\mathbf{r}$ and $\mathbf{w}_0$, the mean parameters μ_r, μ_w and the precision parameters $1/\sigma_r^2$ and $1/\sigma_\Lambda^2$ all have conjugate priors, and therefore their posterior samples are drawn directly. ϕ_r and ϕ_w have discrete priors and therefore are also directly sampled. We use the random-walk Metropolis-Hastings algorithm to draw α.

We obtain 200,000 samples from the algorithm and discard the first 100,000 as burn-in. For the posterior inference, we use 4,000 subsamples from the remaining 100,000 samples, with a thinning equal to 25. The posterior means and 95% equal-tail Bayesian posterior predictive intervals for the model parameters are presented in Table 11.9. Evidently we are recovering the true parameter values very well.

Figure 11.15 displays the posterior mean of the growth curves and 95% Bayesian predictive intervals for the four locations which were used in the fitting (1, 2, 3 and 4), compared with the actual latent growth curve $w(\mathbf{s}, t)$ and observed data. Up to the uncertainty in the model we approximate the actual curves very well. The fitted mean growth curves almost perfectly overlap with the actual simulated growth curves. The empirical coverage of the Bayesian predictive bounds is 93.4%.

Interpolation yields the predictive growth curve for the four hold out locations (A, B, C and D). In Figure 11.15 we display the means of the predicted curves and 95% Bayesian predictive intervals, together with the hold-out data. We can see the spatial prediction captures the patterns of the hold-out data very well. The predicted mean growth curves overlap with the actual simulated growth curves very well except for location D (because location D is rather far from all the observed locations). The empirical coverage of the Bayesian predictive intervals is 95.8%.

Finally, for comparison and in the absence of covariates, we also fit the following customary process realization model with space-time random effects to the simulated data set

$$\mathbf{y}_j = \mu \mathbf{1} + \boldsymbol{\xi}_j + \boldsymbol{\epsilon}_j \; ; \;\; \boldsymbol{\epsilon}_j \sim N\left(0, \sigma_\epsilon^2 I_n\right), \;\; \text{for} \;\; j = 0, 1, \ldots, J \tag{11.40}$$

where the random effects $\boldsymbol{\xi} = [\boldsymbol{\xi}_0, \ldots, \boldsymbol{\xi}_J]$ come from a Gaussian process with a separable spatio-temporal correlation of the form:

$$C_\xi(t - t', s - s') = \sigma_\xi^2 \exp\left(-\alpha_\xi \left|t - t'\right|\right) \left(\phi_\xi \left|s - s'\right|\right)^\nu \kappa_\nu \left(\phi_\xi \left|s - s'\right|\right), \; \nu = \frac{3}{2} \, . \tag{11.41}$$

Comparison of model performance between the SPDE model and this model is done using spatial prediction at the four new locations. In Figure 11.16 we display the means of the predicted curves and 95% Bayesian predictive intervals, together with the hold-out data.

Figure 11.15 *Hold-out space-time geostatistical data at four locations, actual (dashed line) and predicted mean growth curves (solid line) and 95% predictive intervals (dotted line) by our model (11.36) for the simulated data example.*

For the four hold-out sites, the average mean square error for the SPDE model is 1.75×10^{-3} versus 3.34×10^{-3} for the standard model. The average length of the 95% predictive intervals for the SPDE model is 0.29 versus 0.72 for the standard model. It is evident that, when we have an SPDE driving the data, discretized as above, we can learn about it and will do better in terms of prediction than using a standard model.

11.7 Areal unit space-time modeling

We now return to spatiotemporal modeling for areal unit data, following the discussion of Equations (11.4) and (11.5) in Section 11.1. Recall that we have briefly discussed general space-time SAR modeling in Section 4.4.2. Here, we focus on the spatiotemporal disease mapping setting.

11.7.1 Aligned data

In the aligned data case, matters are relatively straightforward. Consider for example the spatiotemporal extension of the standard disease mapping setting described in Section 6.4.1. Here we would have $Y_{i\ell t}$ and $E_{i\ell t}$, the observed and expected disease counts in county i and demographic subgroup ℓ (race, gender, etc.) during time period t (without loss of generality we let t correspond to years in what follows). Again the issue of whether the $E_{i\ell t}$ are internally or externally standardized arises; in the more common former case we would use $n_{i\ell t}$, the number of persons at risk in county i during year t, to compute $E_{i\ell t} = n_{i\ell t}(\sum_{i\ell t} Y_{i\ell t}/\sum_{i\ell t} n_{i\ell t})$. That is, $E_{i\ell t}$ is the number of cases we would expect if the grand disease rate (all regions, subgroups, and years) were in operation throughout. The extension

Figure 11.16 *Hold-out space-time geostatistical data at four locations, actual (dashed line) and predicted mean growth curves (solid line) and 95% predictive intervals (dotted line) by the benchmark model (11.40) for the simulated data example.*

of the basic Section 6.4.1 Poisson regression model is then

$$Y_{i\ell t} \mid \mu_{i\ell t} \overset{ind}{\sim} Po\left(E_{i\ell t}\, e^{\mu_{i\ell t}}\right) ,$$

where $\mu_{i\ell t}$ is the log-relative risk of disease for region i, subgroup ℓ, and year t.

It now remains to specify the main effect and interaction components of $\mu_{i\ell t}$, and corresponding prior distributions. First the main effect for the demographic subgroups can be taken to have ordinary linear regression structure, i.e., $\varepsilon_\ell = \mathbf{x}'_\ell \boldsymbol{\beta}$, with a flat prior for $\boldsymbol{\beta}$. Next, the main effects for time (say, δ_t) can be assigned flat priors (if we wish to view them as fixed effects, i.e., temporal dummy variables), or an $AR(1)$ specification (if we wish them to reflect temporal autocorrelation). In some cases an even simpler structure (say, $\delta_t = \gamma t$) may be appropriate.

Finally, the main effects for space are similar to those assumed in the nontemporal case. Specifically, we might let

$$\psi_i = \mathbf{z}'_i \boldsymbol{\omega} + \theta_i + \phi_i ,$$

where $\boldsymbol{\omega}$ has a flat prior, the θ_i capture *heterogeneity* among the regions via the i.i.d. specification,

$$\theta_i \overset{iid}{\sim} N(0\,,\, 1/\tau) ,$$

and the ϕ_i capture regional *clustering* via the CAR prior,

$$\phi_i \mid \phi_{j \neq i} \sim N(\bar{\phi}_i\,,\, 1/(\lambda m_i)) .$$

As usual, m_i is the number of neighbors of region i, and $\bar{\phi}_i = m_i^{-1} \Sigma_{j \in \partial_i}\, \phi_j$.

Turning to spatiotemporal interactions, suppose for the moment that demographic effects are not affected by region and year. Consider then the *nested* model,

$$\theta_{it} \overset{iid}{\sim} N(0\,,\, 1/\tau_t) \;\text{ and }\; \phi_{it} \sim CAR(\lambda_t) , \tag{11.42}$$

where $\tau_t \overset{iid}{\sim} G(a,b)$ and $\lambda_t \overset{iid}{\sim} G(c,d)$. Provided these hyperpriors are not too informative, this allows "shrinkage" of the year-specific effects toward their grand mean, and in a way that allows the data to determine the amount of shrinkage.

Thus our most general model for $\mu_{i\ell t}$ is

$$\mu_{i\ell t} = \mathbf{x}_\ell'\boldsymbol{\beta} + \delta_t + \mathbf{z}_i'\boldsymbol{\omega} + \theta_{it} + \phi_{it} \ ,$$

with corresponding joint posterior distribution proportional to

$$L(\boldsymbol{\beta}, \delta, \boldsymbol{\omega}, \boldsymbol{\theta}, \boldsymbol{\phi}; \mathbf{y}) p(\delta) p(\boldsymbol{\theta}|\tau) p(\boldsymbol{\phi}|\lambda) p(\tau) p(\lambda) \ .$$

Computation via univariate Metropolis and Gibbs updating steps is relatively straightforward (and readily available in this aligned data setting in the `WinBUGS` language). However, convergence can be rather slow due to the weak identifiability of the joint parameter space. As a possible remedy, consider the the simple space-only case again for a moment. We may transform from $(\boldsymbol{\theta}, \boldsymbol{\phi})$ to $(\boldsymbol{\theta}, \boldsymbol{\eta})$ where $\eta_i = \theta_i + \phi_i$. Then $p(\boldsymbol{\theta}, \boldsymbol{\eta}|\mathbf{y}) \propto L(\boldsymbol{\eta}; \mathbf{y}) p(\boldsymbol{\theta}) p(\boldsymbol{\eta} - \boldsymbol{\theta})$, so that

$$p(\eta_i \mid \eta_{j\neq i}, \boldsymbol{\theta}, \mathbf{y}) \propto L(\eta_i; y_i)\, p(\eta_i - \theta_i \mid \{\eta_j - \theta_j\}_{j\neq i})$$

and

$$p(\theta_i \mid \theta_{j\neq i}, \boldsymbol{\eta}, \mathbf{y}) \propto p(\theta_i)\, p(\eta_i - \theta_i \mid \{\eta_j - \theta_j\}_{j\neq i}) \ .$$

This simple transformation improves matters since each η_i full conditional is now well identified by the data point Y_i, while the weakly identified (indeed, "Bayesianly unidentified") θ_i now emerges in closed form as a normal distribution (since the nonconjugate Poisson likelihood no longer appears).

Example 11.5 The study of the trend of risk for a given disease in space and time may provide important clues in exploring underlying causes of the disease and helping to develop environmental health policy. Waller, Carlin, Xia, and Gelfand (1997) consider the following data set on lung cancer mortality in Ohio. Here Y_{ijkt} is the number of lung cancer deaths in county i during year t for gender j and race k in the state of Ohio. The data are recorded for $J = 2$ genders (male and female, indexed by s_j) and $K = 2$ races (white and nonwhite, indexed by r_k) for each of the $I = 88$ Ohio counties over $T = 21$ years (1968–1988).

We adopt the model,

$$\mu_{ijkt} = s_j\alpha + r_k\beta + s_j r_k \xi + \theta_{it} + \phi_{it} \ , \tag{11.43}$$

where $s_j = 1$ if $j = 2$ (female) and 0 otherwise, and $r_k = 1$ if $k = 2$ (nonwhite) and 0 otherwise. That is, there is one subgroup (white males) for which there is no contribution to the mean structure (11.52). For our prior specification, we select

$$\theta_{it} \overset{ind}{\sim} N\left(0\,,\, \tfrac{1}{\tau_t}\right) \quad \text{and} \quad \phi_{it} \sim CAR(\lambda_t) \ ;$$
$$\alpha, \beta, \xi \sim \text{flat} \ ;$$
$$\tau_t \overset{iid}{\sim} G(1, 100) \quad \text{and} \quad \lambda_t \overset{iid}{\sim} G(1, 7) \ ,$$

where the relative sizes of the hyperparameters in these two gamma distributions were selected following guidance given in Bernardinelli et al. (1995); see also Best et al. (1999) and Eberly and Carlin (2000).

Regarding implementation, five parallel, initially overdispersed MCMC chains were run for 500 iterations. Graphical monitoring of the chains for a representative subset of the parameters, along with sample autocorrelations and Gelman and Rubin (1992) diagnostics, indicated an acceptable degree of convergence by around the 100th iteration.

Demographic subgroup	Contribution to ε_{jk}	Fitted relative risk
White males	0	1
White females	α	0.34
Nonwhite males	β	1.02
Nonwhite females	$\alpha+\beta+\xi$	0.28

Table 11.10 *Fitted relative risks, four sociodemographic subgroups in the Ohio lung cancer data.*

Histograms of the sampled values showed θ_{it} distributions centered near 0 in most cases, but ϕ_{it} distributions typically removed from 0, suggesting that the heterogeneity effects are not really needed in this model. Plots of $E(\tau_t|\mathbf{y})$ and $E(\lambda_t|\mathbf{y})$ versus t suggest increasing clustering and slightly increasing heterogeneity over time. The former might be the result of flight from the cities to suburban "collar counties" over time, while the latter is likely due to the elevated mean levels over time (for the Poisson, the variance increases with the mean).

Fitted relative risks obtained by Waller et al. (1997) for the four main demographic subgroups are shown in Table 11.10. The counterintuitively positive fitted value for nonwhite females may be an artifact of the failure of this analysis to age-standardize the rates prior to modeling (or at least to incorporate age group as another demographic component in the model). To remedy this, consider the following revised and enhanced model, described by Xia and Carlin (1998), where we assume that

$$Y^*_{ijkt} \sim Poisson(E_{ijkt}\exp(\mu_{ijkt}))\ , \tag{11.44}$$

where again Y^*_{ijkt} denotes the observed age-adjusted deaths in county i for sex j, race k, and year t, and E_{ijkt} are the expected death counts. We also incorporate an ecological level smoking behavior covariate into our log-relative risk model, namely,

$$\mu_{ijkt} = \mu + s_j\alpha + r_k\beta + s_jr_k\xi + p_i\rho + \gamma t + \phi_{it}\ , \tag{11.45}$$

where p_i is the true smoking proportion in county i, γ represents the fixed time effect, and the ϕ_{it} capture the random spatial effects over time, wherein clustering effects are nested within time. That is, writing $\boldsymbol{\phi}_t = (\phi_{1t},\ldots,\phi_{It})'$, we let $\boldsymbol{\phi}_t \sim CAR(\lambda_t)$ where $\lambda_t \overset{iid}{\sim} G(c,d)$. We assume that the sociodemographic covariates (sex and race) do not interact with time or space. Following the approach of Bernardinelli, Pascutto et al. (1997), we introduce both sampling error and spatial correlation into the smoking covariate. Let

$$q_i \mid p_i \sim N(p_i, \sigma_q^2),\ i=1,\ldots,I, \text{ and} \tag{11.46}$$

$$\mathbf{p} \sim CAR(\lambda_p) \iff p_i \mid p_{j\neq i} \sim N(\mu_{p_i}, \sigma^2_{p_i}),\ i=1,\ldots,I\ , \tag{11.47}$$

where q_i is the current smoking proportion observed in a sample survey of county i (an imperfect measurement of p_i), $\mu_{p_i} = \sum_{j\neq i} w_{ij}p_j/\sum_{j\neq i} w_{ij}$, and $\sigma^2_{p_i} = (\lambda_p \sum_{j\neq i} w_{ij})^{-1}$. Note that the amount of smoothing in the two CAR priors above may differ, since the smoothing is controlled by different parameters λ_ϕ and λ_p. Like λ_ϕ, λ_p is also assigned a gamma hyperprior, namely, a $G(e,f)$.

We ran 5 independent chains using our Gibbs-Metropolis algorithm for 2200 iterations each; plots suggested discarding the first 200 samples as an adequate burn-in period. We obtained the 95% posterior credible sets [–1.14, –0.98], [0.07, 0.28], and [–0.37, –0.01] for α, β, and ξ, respectively. Note that all 3 fixed effects are significantly different from 0, in contrast to our Table 11.10 results, which failed to uncover a main effect for race. The corresponding point estimates are translated into the fitted relative risks for the four sociodemographic

Demographic subgroup	Contribution to ε_{jk}	Fitted log-relative risk	Fitted relative risk
White males	0	0	1
White females	α	–1.06	0.35
Nonwhite males	β	0.18	1.20
Nonwhite females	$\alpha + \beta + \xi$	–1.07	0.34

Table 11.11 *Fitted relative risks, four sociodemographic subgroups in the Ohio lung cancer data*

subgroups in Table 11.11. Nonwhite males experience the highest risk, followed by white males, with females of both races having much lower risks.

11.7.2 Misalignment across years

In this subsection we develop a spatiotemporal model to accommodate the situation of Figure 11.17, wherein the response variable and the covariate are spatially aligned within any given timepoint, but not across timepoints (due to periodic changes in the regional grid). Assuming that the observed disease count Y_{it} for zip i in year t is conditionally independent of the other zip-level disease counts given the covariate values, we have the model,

$$Y_{it} \mid \mu_{it} \overset{ind}{\sim} Po(E_{it} \exp(\mu_{it})),\ i = 1, \ldots, I_t,\ t = 1, \ldots, T,$$

where the expected count for zip i in year t, E_{it}, is proportional to the population count. In our case, we set $E_{it} = Rn_{it}$, where n_{it} is the population count in zip i at year t and $R = (\sum_{it} Y_{it})/(\sum_{it} n_{it})$, the grand asthma hospitalization rate (i.e., the expected counts assume homogeneity of disease rates across all zips and years). The log-relative risk is modeled as

$$\mu_{it} = x_{it}\beta_t + \delta_t + \theta_{it} + \phi_{it}, \tag{11.48}$$

Figure 11.17 *Traffic density (average vehicles per km of major roadway) in thousands by zip code for 1983, San Diego County.*

where x_{it} is the zip-level exposure covariate (traffic density) depicted for 1983 in Figure 11.17, β_t is the corresponding main effect, δ_t is an overall intercept for year t, and θ_{it} and ϕ_{it} are zip- and year-specific heterogeneity and clustering random effects, analogous to those described in Section 11.7.1. The changes in the zip grid over time cloud the interpretation of these random effects (e.g., a particular region may be indexed by different i in different years), but this does not affect the interpretation of the main effects β_t and δ_t; it is simply the analogue of unbalanced data in a longitudinal setting. In the spatiotemporal case, the distributions on these effects become

$$\boldsymbol{\theta}_t \overset{ind}{\sim} N\left(0, \frac{1}{\tau_t} I\right) \text{ and } \boldsymbol{\phi}_t \overset{ind}{\sim} CAR(\lambda_t) \,, \tag{11.49}$$

where $\boldsymbol{\theta}_t = (\theta_1, \ldots, \theta_{I_t})'$, $\boldsymbol{\phi}_t = (\phi_1, \ldots, \phi_{I_t})'$, and we encourage similarity among these effects across years by assuming $\tau_t \overset{iid}{\sim} G(a, b)$ and $\lambda_t \overset{iid}{\sim} G(c, d)$, where G again denotes the gamma distribution. Placing flat (uniform) priors on the main effects β_t and δ_t completes the model specification. Note that the constraints $\sum_i \phi_{it} = 0$, $t = 1, \ldots, T$ must be added to identify the year effects δ_t, due to the location invariance of the CAR prior.

Example 11.6 Asthma is the most common chronic disease diagnosis for children in the U.S. (National Center for Environmental Health, 1996). A large number of studies have shown a correlation between known products and byproducts of auto exhaust (such as ozone, nitrogen dioxide, and particulate matter) and pediatric asthma ER visits or hospitalizations. Several studies (e.g., Tolbert et al., 2000; Zidek et al., 1998; Best et al., 2000) have used hierarchical Bayesian methods in such investigations. An approach taken by some authors is to use proximity to major roadways (or some more refined measure of closeness to automobile traffic) as an omnibus measure of exposure to various asthma-inducing pollutants. We too adopt this approach and use the phrase "exposure" in what follows, even though in fact our traffic measures are really surrogates for the true exposure.

Our data set arises from San Diego County, CA, the region pictured in Figure 11.17. The city of San Diego is located near the southwestern corner of the map; the map's western boundary is the Pacific Ocean, while Mexico forms its southern boundary. The subregions pictured are the zip codes as defined in 1983; as mentioned earlier this grid changes over time. Specifically, during the course of our eight-year (1983–1990) study period, the zip code boundaries changed four times: in 1984, 1987, 1988, and 1990.

The components of our data set are as follows. First, for a given year, we have the number of discharges from hospitalizations due to asthma for children aged 14 and younger by zip code (California Office of Statewide Health Planning and Development, 1997). The primary diagnosis was asthma based on the International Classification of Diseases, code 493 (U.S. Department of Health and Human Services, 1989). Assuming that patient records accurately report the correct zip code of residence, these data can be thought of as error-free.

Second, we have zip-level population estimates (numbers of residents aged 14 and younger) for each of these years, as computed by Scalf and English (1996). These estimates were obtained in `ARC/INFO` using the following process. First, a land-use covariate was used to assist in a linear interpolation between the 1980 and 1990 U.S. Census figures, to obtain estimates at the census block group level. Digitized hard-copy U.S. Postal Service maps or suitably modified street network files provided by the San Diego Association of Governments (SANDAG) were then used to reallocate these counts to the zip code grid for the year in question. To do this, the GIS first created a subregional grid by intersecting the block group and zip code grids. The block group population totals were allocated to the subregions per a combination of subregional area and population density (the latter again based on the land-use covariate). Finally, these imputed subregional counts were reaggregated to the zip grid. While there are several possible sources of uncertainty in these

calculations, we ignore them in our initial round of modeling, assuming these population counts to be fixed and known.

Finally, for each of the major roads in San Diego County, we have mean yearly traffic counts on each road segment in our map. Here "major" roads are defined by SANDAG to include interstate highways or equivalent, major highways, access or minor highways, and arterial or collector routes. The sum of these numbers within a given zip divided by the total length of its major roads provides an aggregate measure of traffic exposure for the zip. These zip-level *traffic densities* are plotted for 1983 in Figure 11.17; this is the exposure measure we use in the following text.

We set $a = 1$, $b = 10$ (i.e., the τ_t have prior mean and standard deviation both equal to 10) and $c = 0.1$, $d = 10$ (i.e., the λ_t have prior mean 1, standard deviation $\sqrt{10}$). These are fairly vague priors designed to let the data dominate the allocation of excess spatial variability to heterogeneity and clustering. (As mentioned near Equation (6.26), simply setting these two priors equal to each other would not achieve this, since the prior for the θ_{it} is specified *marginally*, while that for the ϕ_{it} is specified *conditionally* given the neighboring ϕ_{jt}.) Our MCMC implementation ran 3 parallel sampling chains for 5000 iterations each, and discarded the first 500 iterations as preconvergence "burn-in."

Plots of the posterior medians and 95% equal-tail Bayesian confidence intervals for β_t (not shown) makes clear that, with the exception of that for 1986, all of the β_t's are significantly greater than 0. Hence, the traffic exposure covariate in Figure 11.17 is positively associated with increased pediatric asthma hospitalization in seven of the eight years of our study. To interpret these posterior summaries, recall that their values are on the *log*-relative risk scale. Thus a zip having a 1983 traffic density of 10,000 cars per km of roadway would have median relative risk $e^{10(.065)} = 1.92$ times higher than a zip with essentially no traffic exposure, with a corresponding 95% confidence interval of $(e^{10(.000)}, e^{10(.120)}) = (1.00, 3.32)$. There also appears to be a slight weakening of the traffic-asthma association over time.

Figure 11.18 provides `ARC/INFO` maps of the crude and fitted asthma rates (per thousand) in each of the zips for 1983. The crude rates are of course given by $r_{it} = Y_{it}/n_{it}$, while the fitted rates are given by $R\exp(\hat{\mu}_{it})$, where R is again the grand asthma rate across all zips and years and $\hat{\mu}_{it}$ is obtained by plugging in the estimated posterior means for the various components in Equation (11.48). The figure clearly shows the characteristic Bayesian shrinkage of the crude rates toward the grand rate. In particular, no zip is now assigned a rate of exactly zero, and the rather high rates in the thinly populated eastern part of the map have been substantially reduced. However, the high observed rates in urban San Diego continue to be high, as the method properly recognizes the much higher sample sizes in these zips. There also appears to be some tendency for clusters of similar crude rates to be preserved, the probable outcome of the CAR portion of our model.

11.7.3 Nested misalignment both within and across years

In this subsection we extend our spatiotemporal model to accommodate the situation of Figure 11.19, wherein the covariate is available on a grid that is a refinement of the grid for which the response variable is available (i.e., nested misalignment within years, as well as misalignment across years). Letting the subscript j index the subregions (which we also refer to as *atoms*) of zip i, our model now becomes

$$Y_{ijt} \mid \mu_{ijt} \sim Po(E_{ijt}\exp(\mu_{ijt})),\ i = 1, \ldots, I_t,\ j = 1, \ldots, J_{it},\ t = 1, \ldots, T,$$

where the expected counts E_{ijt} are now Rn_{ijt}, with the grand rate R as before. The population of atom ijt is not known, and so we determine it by areal interpolation as $n_{ijt} = n_{it}(\text{area of atom } ijt)/(\text{area of zip } it)$. The log-relative risk in atom ijt is then modeled as

$$\mu_{ijt} = x_{ijt}\beta_t + \delta_t + \theta_{it} + \phi_{it}, \tag{11.50}$$

Figure 11.18 *Pediatric asthma hospitalization rate (per thousand children) by zip code for 1983, San Diego County: (a) crude rate, (b) temporally misaligned model fitted rate.*

where x_{ijt} is now the atom-level exposure covariate (depicted for 1983 in Figure 11.19), but β_t, δ_t, θ_{it} and ϕ_{it} are as before. Thus our prior specification is exactly that of the previous subsection; priors for the $\boldsymbol{\theta}_t$ and $\boldsymbol{\phi}_t$ as given in Equation (11.49), exchangeable gamma hyperpriors for the τ_t and λ_t with $a = 1$, $b = 10$, $c = 0.1$, and $d = 10$, and flat priors for the main effects β_t and δ_t.

Since only the zip-level hospitalization totals Y_{it} (and not the atom-level totals Y_{ijt}) are observed, we use the additivity of conditionally independent Poisson distributions to obtain

$$Y_{it} \mid \beta_t, \delta_t, \theta_{it}, \phi_{it} \sim Po\left(\sum_{j=1}^{J_{it}} E_{ijt} \exp(\mu_{ijt})\right), \; i = 1, \ldots, I_t, \; t = 1, \ldots, T. \tag{11.51}$$

Figure 11.19 *Adjusted traffic density (average vehicles per km of major roadway) in thousands by zip code subregion for 1983, San Diego County.*

Using Expression (11.51), we can obtain the full Bayesian model specification for the observed data as

$$\left[\prod_{t=1}^{T}\prod_{i=1}^{I_t} p\left(y_{it}|\beta_t,\delta_t,\theta_{it},\phi_{it}\right)\right]\left[\prod_{t=1}^{T} p(\boldsymbol{\theta}_t|\tau_t)p(\boldsymbol{\phi}_t|\lambda_t)p(\tau_t)p(\lambda_t)\right] \tag{11.52}$$

As in the previous section, only the τ_t and λ_t parameters may be updated via ordinary Gibbs steps, with Metropolis steps required for the rest.

Note that model specification (11.52) makes use of the atom-level covariate values x_{ijt}, but only the zip-level hospitalization counts Y_{it}. Of course, we might well be interested in *imputing* the values of the missing subregional counts Y_{ijt}, whose full conditional distribution is multinomial, namely,

$$(Y_{i1t},\ldots,Y_{iJ_it})\,|\,Y_{it},\beta_t,\delta_t,\theta_{it},\phi_{it} \sim Mult(Y_{it},\{q_{ijt}\})\ , \tag{11.53}$$

$$\text{where}\qquad q_{ijt} = \frac{E_{ijt}e^{\mu_{ijt}}}{\sum_{j=1}^{J_{it}} E_{ijt}e^{\mu_{ijt}}}\ .$$

Since this is a purely predictive calculation, Y_{ijt} values need not be drawn as part of the MCMC sampling order, but instead at the very end, conditional on the post-convergence samples.

Zhu, Carlin, English, and Scalf (2000) use Figure 11.19 to refine the definition of exposure used in Example 11.6 by subdividing each zip into subregions based on whether or not they are closer than 500 m to a major road. This process involves creating "buffers" around each road and subsequently overlaying them in a GIS, and has been previously used in several studies of vehicle emissions. This definition leads to some urban zips becoming "entirely exposed," as they contain no point further than 500 m from a major road; these are roughly the zips with the darkest shading in Figure 11.17 (i.e., those having traffic densities greater than 10,000 cars per year per km of major roadway). Analogously, many zips in the thinly populated eastern part of the county contained at most one major road, suggestive of little

or no traffic exposure. As a result, we (somewhat arbitrarily) defined those zips in the two lightest shadings (i.e., those having traffic densities less than 2,000 cars per year per km of roadway) as being "entirely unexposed." This typically left slightly less than half the zips (47 for the year shown, 1983) in the middle range, having some exposed and some unexposed subregions, as determined by the intersection of the road proximity buffers. These subregions are apparent as the lightly shaded regions in Figure 11.19; the "entirely exposed" regions continue to be those with the darkest shading, while the "entirely unexposed" regions have no shading.

The fitted rates obtained by Zhu et al. (2000) provide a similar overall impression as those in Figure 11.18, except that the newer map is able to show subtle differences within several "partially exposed" regions. These authors also illustrate the interpolation of missing subregional counts Y_{ijt} using Equation (11.53). Analogous to the block-block FMPC imputation in Subsection 7.2, the sampling-based hierarchical Bayesian method produces more realistic estimates of the subregional hospitalization counts, with associated confidence limits emerging as an automatic byproduct.

11.7.4 Nonnested misalignment and regression

In this subsection we consider spatiotemporal *regression* in the misaligned data setting motivated by our Atlanta ozone data set. Recall that the first component of this data set provides ozone measurements X_{itr} at between 8 and 10 fixed monitoring sites i for day t of year r, where $t = 1, \ldots, 92$ (the summer days from June 1 through August 31) and $r = 1, 2, 3$, corresponding to years 1993, 1994, and 1995 For example, Figure 1.3 shows the 8-hour daily maximum ozone measurements (in parts per million) at the 10 monitoring sites for a particular day (July 15, 1995), along with the boundaries of the 162 zip codes in the Atlanta metropolitan area.

A *second* component of this data set (about which we so far have said far less) provides relevant health outcomes, but only at the zip code level. Specifically, for each zip l, day t, and year r, we have the number of pediatric emergency room (ER) visits for asthma, Y_{ltr}, as well as the total number of pediatric ER visits, n_{ltr}. These data come from a historical records-based investigation of pediatric asthma emergency room visits to seven major emergency care centers in the Atlanta metropolitan statistical area during the same three summers. Our main substantive goal is an investigation of the relationship between ozone and pediatric ER visits for asthma in Atlanta, controlling for a range of sociodemographic covariates. Potential covariates (available only as zip-level summaries in our data set) include average age, percent male, percent black, and percent using Medicaid for payment (a crude surrogate for socioeconomic status). Clearly an investigation of the relationship between ozone exposure and pediatric ER visit count cannot be undertaken until the mismatch in the support of the (point-level) predictor and (zip-level) response variables is resolved.

A naive approach would be to average the ozone measurements belonging to a specific zip code, then relate this average ozone measurement to the pediatric asthma ER visit count in this zip. In fact, there are few monitoring sites relative to the number of zip codes; Figure 1.3 shows most of the zip codes contain no sites at all, so that most of the zip-level ER visit count data would be discarded. An alternative would be to aggregate the ER visits over the entire area and model them as a function of the average of the ozone measurements (that is, eliminate the spatial aspect of the data and fit a temporal-only model). Using this idea in a Poisson regression, we obtained a coefficient for ozone of 2.48 with asymptotic standard error 0.71 (i.e., significant positive effect of high ozone on ER visit rates). While this result is generally consistent with our findings, precise comparison is impossible for a number of reasons. First, this approach requires use of data from the entire Atlanta metro area (due to the widely dispersed locations of the monitoring stations), not data from the city only as our approach allows. Second, it does not permit use of available covariates (such

as race and SES) that were spatially but not temporally resolved in our data set. Third, standardizing using expected counts E_i (as in Equation (11.54) below) must be done only over days (not regions), so the effect of including them is now merely to adjust the model's intercept.

We now describe the disease component of our model, and subsequently assemble the full Bayesian hierarchical modeling specification for our spatially misaligned regression. Similar to the model of Subsection 11.7.3, we assume the zip-level asthma ER visit counts, Y_{ltr} for zip l during day t of summer r, follow a Poisson distribution,

$$Y_{ltr} \sim Poisson\left(E_{ltr} \exp(\lambda_{ltr})\right) \ , \tag{11.54}$$

where the E_{ltr} are expected asthma visit counts, determined via internal standardization as $E_{ltr} = n_{ltr}(\sum_{ltr} Y_{ltr} / \sum_{ltr} n_{ltr})$, where n_{ltr} is the total number of pediatric ER visits in zip code l on day t of year r. Thus E_{ltr} is the number of pediatric ER asthma visits we would expect from the given zip and day if the proportion of such visits relative to the total pediatric ER visit rate was homogeneous across all zips, days, and years. Hence λ_{ltr} in (11.54) can be interpreted as a log-relative risk of asthma among those children visiting the ER in group ltr. Our study design is thus a *proportional admissions model* (Breslow and Day, 1987, pp. 153–155).

We do not take n_{ltr} equal to the total number of children *residing* in zip l on day t of year r, since this standardization would implicitly presume a constant usage of the ER for pediatric asthma management across all zips, which seems unlikely (children from more affluent zips are more likely to have the help of family doctors or specialists in managing their asthma, and so would not need to rely on the ER; see Congdon and Best, 2000, for a solution to the related problem of adjusting for patient referral practices). Note however that this in turn means that our disease (pediatric asthma visits) is not particularly "rare" relative to the total (all pediatric visits). As such, our use of the Poisson distribution in (11.54) should not be thought of as an approximation to a binomial distribution for a rare event, but merely as a convenient and sensible model for a discrete variable.

For the log-relative risks in group ltr, we begin with the model,

$$\lambda_{ltr} = \beta_0 + \beta_1 X_{l,t-1,r} + \sum_{c=1}^{C} \alpha_c Z_{cl} + \sum_{d=1}^{D} \delta_d W_{dt} + \theta_l \ . \tag{11.55}$$

Here, β_0 is an intercept term, and β_1 denotes the effect of ozone exposure $X_{l,t-1,r}$ in zip l during day $t-1$ of year r. Note that we model pediatric asthma ER visit counts as a function of the ozone level on the *previous* day, in keeping with the most common practice in the epidemiological literature (see, e.g., Tolbert et al., 2000). This facilitates next-day predictions for pediatric ER visits given the current day's ozone level, with our Bayesian approach permitting full posterior inference (e.g., 95% prediction limits). However, it also means we have only $(J-1) \times 3 = 273$ days worth of usable data in our sample. Also, $\mathbf{Z}_l = (Z_{1l}, \ldots, Z_{Cl})^T$ is a vector of C zip-level (but not time-varying) sociodemographic covariates with corresponding coefficient vector $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_C)^T$, and $\mathbf{W}_t = (W_{1t}, \ldots, W_{Dt})^T$ is a vector of D day-level (but not spatially varying) temporal covariates with corresponding coefficient vector $\boldsymbol{\delta} = (\delta_1, \ldots, \delta_D)^T$. Finally, θ_l is a zip-specific random effect designed to capture extra-Poisson variability in the observed ER visitation rates. These random effects may simply be assumed to be exchangeable draws from a $N(0, 1/\tau)$ distribution (thus modeling overall *heterogeneity*), or may instead be assumed to vary spatially using a conditionally autoregressive (CAR) specification.

Of course, model (11.54)–(11.55) is not fittable as stated, since the zip-level previous-day ozone values $X_{l,t-1,r}$ are not observed. Fortunately, we may use the methods of Section 7.1 to perform the necessary point-block realignment. To connect our Equation (11.55) notation

with that used in Section 7.1, let us write $\mathbf{X}_{B,r} \equiv \{X_{l,t-1,r},\ l = 1, \ldots, L, t = 2, \ldots, J\}$ for the unobserved block-level data from year r, and $\mathbf{X}_{s,r} \equiv \{X_{itr},\ i = 1, \ldots, I, t = 1, \ldots, J\}$ for the observed site-level data from year r. Then, from Equations (11.21) and (11.22) and assuming no missing ozone station data for the moment, we can find the conditional predictive distribution $f(\mathbf{X}_{B,r}|\mathbf{X}_{s,r}, \boldsymbol{\gamma}_r, \sigma_r^2, \boldsymbol{\phi}_r, \boldsymbol{\rho}_r)$ for year r. However, for these data some components of the $\mathbf{X}_{s,r}$ will be missing, and thus replaced with imputed values $\mathbf{X}_{s,r}^{(m)}$, $m = 1, \ldots, M$, for some modest number of imputations M (say, $M = 3$). (In a slight abuse of notation here, we assume that any *observed* component of $\mathbf{X}_{s,r}^{(m)}$ is simply set equal to that observed value for all m.)

Thus, the full Bayesian hierarchical model specification is given by

$$\begin{aligned}&[\textstyle\prod_r \prod_t \prod_l f(Y_{ltr}|\boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{\delta}, \boldsymbol{\theta}, X_{l,t-1,r})]\, p(\boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{\delta}, \boldsymbol{\theta}) \\ &\quad \times \Big[\textstyle\prod_r f(\mathbf{X}_{B,r}|\mathbf{X}_{s,r}^{(m)}, \boldsymbol{\gamma}_r, \sigma_r^2, \boldsymbol{\phi}_r, \boldsymbol{\rho}_r) \\ &\qquad\qquad \times f(\mathbf{X}_{s,r}^{(m)}|\boldsymbol{\gamma}_r, \sigma_r^2, \boldsymbol{\phi}_r, \boldsymbol{\rho}_r) p(\boldsymbol{\gamma}_r, \sigma_r^2, \boldsymbol{\phi}_r, \boldsymbol{\rho}_r)\Big]\ ,\end{aligned} \tag{11.56}$$

where $\boldsymbol{\beta} = (\beta_0, \beta_1)^T$, and $\boldsymbol{\gamma}_r, \sigma_r^2, \boldsymbol{\phi}_r$ and $\boldsymbol{\rho}_r$ are year-specific versions of the parameters in (7.6). Note that there is a posterior distribution for each of the M imputations. Model (11.56) assumes the asthma-ozone relationship does not depend on year; the misalignment parameters are year-specific only to permit year-by-year realignment.

Zhu, Carlin, and Gelfand (2003) offer a reanalysis of the Atlanta ozone and asthma data by fitting a version of model (11.55), namely,

$$\lambda_{ltr} = \beta_0 + \beta_1 X_{l,t-1,r}^{*(m,v)} + \alpha_1 Z_{1l} + \alpha_2 Z_{2l} + \delta_1 W_{1t} + \delta_2 W_{2t} + \delta_3 W_{3t} + \delta_4 W_{4t}\ , \tag{11.57}$$

where $X_{l,t-1,r}^{*(m,v)}$ denotes the (m, v)th imputed value for the zip-level estimate of the 8-hour daily maximum ozone measurement on the previous day $(t-1)$. Our zip-specific covariates are Z_{1l} and Z_{2l}, the percent high socioeconomic status and percent black race of those pediatric asthma ER visitors from zip l, respectively. Of the day-specific covariates, W_{1t} indexes day of summer ($W_{1t} = t \bmod 91$) and $W_{2t} = W_{1t}^2$, while W_{3t} and W_{4t} are indicator variables for days in 1994 and 1995, respectively (so that 1993 is taken as the reference year). We include both linear and quadratic terms for day of summer in order to capture the rough U-shape in pediatric ER asthma visits, with June and August higher than July.

The analysis of Zhu et al. (2003) is only approximate, in that they run *separate* MCMC algorithms on the portions of the model corresponding to the two lines of model (11.56). In the spirit of the multiple imputation approach to the missing (point-level) ozone observations, they also retain $V = 3$ post-convergence draws from each of our $M = 3$ imputed data sets, resulting in $MV = 9$ zip-level approximately imputed ozone vectors $\mathbf{X}_{B,r}^{*(m,v)}$.

The results of this approach are shown in Table 11.12. The posterior median of β_1 (.7860) is positive, as expected. An increase of .02 ppm in 8-hour maximum ozone concentration (a relatively modest increase, as seen from Figure 1.3) thus corresponds to a fitted relative risk of $\exp(.7860 \times .02) \approx 1.016$, or a 1.6% increase in relative risk of a pediatric asthma ER visit. However, the 95% credible set for β_1 does include 0, meaning that this positive association between ozone level and ER visits is not "Bayesianly significant" at the 0.05 level. Using a more naive approach but data from all 162 zips in the Atlanta metro area, Carlin et al. (1999) estimate the above relative risk as 1.026, marginally significant at the .05 level (that is, the lower limit of the 95% credible set for β_1 was precisely 0).

Regarding the demographic variables, the effects of both percent high SES and percent black emerge as significantly different from 0. The relative risk for a zip made entirely of high SES residents would be slightly more than half that of a comparable all-low SES zip, while a zip with a 100% black population would have a relative risk nearly 1.8 times that of a 100% nonblack zip. As for the temporal variables, day of summer is significantly negative

Parameter	Effect	Posterior median	95% Posterior credible set	Fitted relative risk
β_0	intercept	–0.4815	(–0.5761, –0.3813)	—
β_1	ozone	0.7860	(–0.7921, 2.3867)	1.016†
α_1	high SES	–0.5754	(–0.9839, –0.1644)	0.562
α_2	black	0.5682	(0.3093, 0.8243)	1.765
δ_1	day	–0.0131	(–0.0190, –0.0078)	—
δ_2	day^2	0.00017	(0.0001, 0.0002)	—
δ_3	year 1994	0.1352	(0.0081, 0.2478)	1.145
δ_4	year 1995	0.4969	(0.3932, 0.5962)	1.644

Table 11.12 *Fitted relative risks for the parameters of interest in the Atlanta pediatric asthma ER visit data, full model. (†This is the posterior median relative risk predicted to arise from a .02 ppm increase in ozone.)*

and its square is significantly positive, confirming the U-shape of asthma relative risks over a given summer. Both year 1994 and year 1995 show higher relative risk compared with year 1993, with estimated increases in relative risk of about 15% and 64%, respectively.

11.8 Areal-level continuous time modeling

Quick, Banerjee and Carlin (2013) address the less common setting where space is discrete and time is continuous. This can be envisioned in situations where a collection of N_s *spatially associated* functions of time over N_s regions are posited. Put another way, functions arising from neighboring regions are believed to resemble each other. The functional data analysis literature (Ramsay and Silverman, 1997, and references therein) deals almost exclusively with kernel smoothers and roughness-penalty type (spline) models; recent discrete-space, continuous time examples using spline-based methods include the works by MacNab and Gustafson (2007) and Ugarte et al. (2010). Baladandayuthapani et al. (2008) consider spatially correlated functional data modeling for point-referenced data by treating space as continuous. Delicado et al. (2010) provide a review of and point out that spatially associated functional modeling of time has received little attention, especially for regionally aggregated data.

Quick et al. (2013) propose a class of Bayesian space-time models based upon a dynamic MRF that evolves continuously over time. This accommodates spatial processes that are posited to be spatially indexed over a geographical map with a well-defined system of neighbors. Rather than modeling time using simple parametric forms, as is often done in longitudinal contexts, these authors employ a stochastic process, enhancing the model's adaptability to the data.

The benefits of using a continuous-time model over a discrete-time model are pronounced when investigators (e.g. public health officials) seek to understand the local effects of temporal impact at a resolution finer than that at which the data were sampled. For instance, despite collecting data monthly, there may be interest in interpolating over a particular week or even at a given day of that month. Dynamic space-time models that treat time discretely can offer statistically legitimate inference only at the level of the data. In additin, the modeling also allows us to subsequently carry out inference on temporal gradients; that is, the rate of change of the underlying process over time (see Chapter 13 for inference on spatial gradients). Quick et al. (2013) show how such inference can be carried out in fully model-based fashion using exact posterior predictive distributions for the gradients at any arbitrary time point.

11.8.1 Areally referenced temporal processes

Here we provide a brief overview of the approach proposed by Quick et al. (2013). Consider a map of a geographical region comprising N_s regions that are delineated by well-defined boundaries, and let $Y_i(t)$ be the outcome arising from region i at time t. For every region i, we believe that $Y_i(t)$ exists, at least conceptually, at every time point. However, the observations are collected not continuously but at discrete time points, say $\mathcal{T} = \{t_1, t_2, \ldots, t_{N_t}\}$. For simplicity, let us assume that the data comes from the same set of time points in $\mathcal{T}$ for each region. This is not strictly necessary for the ensuing development, but will facilitate the notation.

A spatial random effect model that treats space as continuous and time as discrete assumes that

$$Y_i(t) = \mu_i(t) + Z_i(t) + \epsilon_i(t), \quad \epsilon_i(t) \overset{ind}{\sim} N(0, \tau_i^2) \text{ for } i = 1, 2, \ldots, N_s \,, \tag{11.58}$$

where $\mu_i(t)$ captures large scale variation or trends, for example using a regression model, and $Z_i(t)$ is an underlying areally-referenced stochastic process over time that captures smaller-scale variations in the time scale while also accommodating spatial associations. Each region also has its own variance component, τ_i^2, which captures residual variation not captured by the other components.

The process $Z_i(t)$ specifies the probability distribution of correlated space-time random effects while treating space as discrete and time as continuous. We seek a specification that will allow temporal processes from neighboring regions to be more alike than from non-neighbors. As regards spatial associations, we will respect the discreteness inherent in the aggregated outcome. Rather than model an underlying response surface continuously over the region of interest, we want to treat the $Z_i(t)$'s as functions of time that are smoothed across neighbors.

The neighborhood structure arises from a discrete topology comprising a list of neighbors for each region. This is described using an $N_s \times N_s$ adjacency matrix $W = \{w_{ij}\}$, where $w_{ij} = 0$ if regions i and j are not neighbors and $w_{ij} = c \neq 0$ when regions i and j are neighbors, denoted by $i \sim j$. By convention, the diagonal elements of W are all zero. To account for spatial association in the $Z_i(t)$'s, a temporally evolving MRF for the areal units at any arbitrary time point t specifies the full conditional distribution for $Z_i(t)$ as depending only upon the neighbors of region i,

$$p(Z_i(t) \,|\, \{Z_{j \neq i}(t)\}) \sim N\left(\sum_{j \sim i} \alpha \frac{w_{ij}}{w_{i+}} Z_j(t), \frac{\sigma^2}{w_{i+}}\right),$$

where $w_{i+} = \sum_{j \sim i} w_{ij}$, $\sigma^2 > 0$, and α is a propriety parameter described below. This means that the $N_s \times 1$ vector $\mathbf{Z}(t) = (Z_1(t), Z_2(t), \ldots, Z_{N_s}(t))^T$ follows a multivariate normal distribution with zero mean and a precision matrix $\frac{1}{\sigma^2}(D - \alpha W)$, where D is a diagonal matrix with w_{i+} as its i-th diagonal elements. The precision matrix is invertible as long as $\alpha \in (1/\lambda_{(1)}, 1/\lambda_{(n)})$, where $\lambda_{(1)}$ (which can be shown to be negative) and $\lambda_{(n)}$ (which can be shown to be 1) are the smallest (i.e., most negative) and largest eigenvalues of $D^{-1/2} W D^{-1/2}$, respectively, and this yields a proper distribution for $\mathbf{Z}(t)$ at each timepoint t.

The MRF in (11.59) does not allow temporal dependence; the $\mathbf{Z}(t)$'s are independently and identically distributed as $N\left(\mathbf{0}, \sigma^2 (D - \alpha W)^{-1}\right)$. We could allow time-varying parameters σ_t^2 and α_t so that $\mathbf{Z}(t) \overset{ind}{\sim} N\left(\mathbf{0}, \sigma_t^2 (D - \alpha_t W)^{-1}\right)$ for every t. If time were treated discretely, then we could envision dynamic autoregressive priors for these time-varying parameters, or some transformations thereof. However, there are two reasons why we do not

pursue this further. First, we do not consider time as discrete because that would preclude inference on temporal gradients, which, as we have mentioned, is a major objective here. Second, time-varying hyperparameters, especially the α_t's, in MRF models are usually weakly identified by the data; they permit very little prior-to-posterior learning and often lead to over-parametrized models that impair predictive performance over time.

Quick et al. (2013) prefer to jointly build spatialtemporal associations into the model using a multivariate process specification for $\mathbf{Z}(t)$. A highly flexible and computationally tractable option is to assume that $\mathbf{Z}(t)$ is a zero-centered multivariate Gaussian process, $GP(\mathbf{0}, K_Z(\cdot,\cdot))$, where the matrix-valued covariance function (e.g., "*cross-covariance* matrix function," Cressie, 1993) $K_Z(t,u) = \text{cov}\{\mathbf{Z}(t), \mathbf{Z}(u)\}$ is defined to be the $N_s \times N_s$ matrix with (i,j)-th entry $\text{cov}\{Z_i(t), Z_j(u)\}$ for any $(t,u) \in \Re^+ \times \Re^+$. Thus, for any two positive real numbers t and u, $K_Z(t,u)$ is an $N_s \times N_s$ matrix with (i,j)-th element given by the covariance between $Z_i(t)$ and $Z_j(u)$. These multivariate processes are *stationary* when the covariances are functions of the separation between the time-points, in which case we write $K_Z(t,u) = K_Z(\Delta)$, and *fully symmetric* when $K_Z(t,u) = K_Z(|\Delta|)$, where $\Delta = t - u$.

To ensure valid joint distributions for process realizations, we use a constructive approach similar to that used in *linear models of coregionalization* (LMC) and, more generally, belonging to the class of multivariate latent process models. We assume that $\mathbf{Z}(t)$ arises as a (possibly temporally-varying) linear transformation $\mathbf{Z}(t) = A(t)\mathbf{v}(t)$ of a simpler process $\mathbf{v}(t) = (v_1(t), v_2(t), \ldots, v_{N_s}(t))^T$, where the $v_i(t)$'s are univariate temporal processes, independent of each other, and with unit variances. This differs from the conventional LMC approach based on *spatial* processes, which treats space as continuous. The matrix-valued covariance function for $\mathbf{v}(t)$, say, $K_{\mathbf{v}}(t,u)$, thus has a simple diagonal form and $K_Z(t,u) = A(t)K_{\mathbf{v}}(t,u)A(u)^T$. The dispersion matrix for $\mathbf{Z}$ is $\Sigma_Z = \mathcal{A}\Sigma_{\mathbf{v}}\mathcal{A}^T$, where $\mathcal{A}$ is a block-diagonal matrix with $A(t_j)$'s as blocks, and $\Sigma_{\mathbf{v}}$ is the dispersion matrix constructed from $K_{\mathbf{v}}(t,u)$. Constructing simple valid matrix-valued covariance functions for $\mathbf{v}(t)$ automatically ensures valid probability models for $\mathbf{Z}(t)$. Also note that for $t = u$, $K_{\mathbf{v}}(t,t)$ is the identity matrix so that $K_Z(t,t) = A(t)A(t)^T$ and $A(t)$ is a square-root (e.g. obtained from the triangular Cholesky factorization) of the matrix-valued covariance function at time t.

The above framework subsumes several simpler and more intuitive specifications. One particular specification that we pursue here assumes that each $v_i(t)$ follows a stationary Gaussian Process $GP(0, \rho(\cdot,\cdot;\boldsymbol{\phi}))$, where $\rho(\cdot,\cdot;\phi)$ is a positive definite correlation function parametrized by $\boldsymbol{\phi}$ (e.g. Stein, 1999), so that $\text{cov}(v_i(t), v_i(u)) = \rho(t,u;\phi)$ for every $i = 1, 2, \ldots, N_s$ for all non-negative real numbers t and u. Since the $v_i(t)$ are independent across i, $\text{cov}\{v_i(t), v_j(u)\} = 0$ for $i \neq j$.

The matrix-valued covariance function for $\mathbf{Z}(t)$ becomes $K_Z(t,u) = \rho(t,u;\boldsymbol{\phi})A(t)A(u)^T$. If we further assume that $A(t) = A$ is constant over time, then the process $\mathbf{Z}(t)$ is stationary if and only if $\mathbf{v}(t)$ is stationary. Further, we obtain a *separable* specification, so that $K_Z(t,u) = \rho(t,u;\boldsymbol{\phi})AA^T$. Letting A be some square-root (e.g., Cholesky) of the $N_s \times N_s$ dispersion matrix $\sigma^2(D - \alpha W)^{-1}$ and $R(\boldsymbol{\phi})$ be the $N_t \times N_t$ temporal correlation matrix having (i,j)-th element $\rho(t_i, t_j;\boldsymbol{\phi})$ yields

$$K_Z(t,u) = \sigma^2\rho(t,u;\boldsymbol{\phi})(D - \alpha W)^{-1} \text{ and } \Sigma_Z = R(\boldsymbol{\phi}) \otimes \sigma^2(D - \alpha W)^{-1} . \tag{11.59}$$

It is straightforward to show that the marginal distribution from this constructive approach for each $\mathbf{Z}(t_i)$ is $N\left(\mathbf{0}, \sigma^2(D - \alpha W)^{-1}\right)$, the same marginal distribution as the temporally independent MRF specification in (11.59). Therefore, our constructive approach ensures a valid space-time process, where associations in space are modeled discretely using a MRF, and those in time through a continuous Gaussian process.

This separable specification is easily interpretable because it factorizes the dispersion into a spatial association component (areal) and a temporal component. Another significant practical advantage is its computational feasibility. Estimating more general space-time

models usually entails matrix factorizations with $O(N_s^3 N_t^3)$ computational complexity. The separable specification allows us to reduce this complexity substantially by avoiding factorizations of $N_s N_t \times N_s N_t$ matrices. One could design algorithms to work with matrices whose dimension is the smaller of N_s and N_t, thereby accruing massive computational gains. More general models using this approach are introduced and discussed in the online supplement (Quick et al., 2013), but since they do not offer anything new in terms of temporal gradients, we do not pursue them further.

11.8.2 Hierarchical modeling

Following Quick et al. (2013), we build a hierarchical modeling framework using the likelihood from our spatial random effects model in (11.58) and the distributions emerging from the temporal Gaussian process discussed in Section 11.8.1. The mean $\mu_i(t)$ in (11.58) is often indexed by a parameter vector $\boldsymbol{\beta}$, for example a linear regression with regressors indexed by space and time so that $\mu_i(t;\boldsymbol{\beta}) = \mathbf{x}_i(t)^T\boldsymbol{\beta}$.

The posterior distribution is

$$\begin{aligned} p(\boldsymbol{\theta}, \mathbf{Z} \mid \mathbf{Y}) \quad \propto \quad & p(\boldsymbol{\phi}) \times IG(\sigma^2 \mid a_\sigma, b_\sigma) \times \left(\prod_{i=1}^{M} IG(\tau_i^2 \mid a_\tau, b_\tau) \right) \\ & \times N(\boldsymbol{\beta} \mid \mu_\beta, \Sigma_\beta) \times Beta(\alpha \mid a_\alpha, b_\alpha) \\ & \times N(\mathbf{Z} \mid \mathbf{0}, R(\phi) \otimes \sigma^2 (D - \alpha W)^{-1}) \\ & \times \prod_{j=1}^{N_t} \prod_{i=1}^{N_s} N(Y_i(t_j) \mid \mathbf{x}_i(t_j)^T\boldsymbol{\beta} + Z_i(t_j), \tau_i^2) \ , \end{aligned} \tag{11.60}$$

where $\boldsymbol{\theta} = \{\boldsymbol{\phi}, \alpha, \sigma^2, \boldsymbol{\beta}, \tau_1^2, \tau_2^2, \ldots, \tau_{N_s}^2\}$ and $\mathbf{Y}$ is the vector of observed outcomes defined analogous to $\mathbf{Z}$. The parametrizations for the standard densities are as in Carlin and Louis (2008). We assume all the other hyperparameters in (11.60) are known.

Recall the separable matrix-valued covariance function in (11.59). The correlation function $\rho(\cdot;\boldsymbol{\phi})$ determines process smoothness and we choose it to be a fully symmetric Matérn correlation function. Markov chain Monte Carlo (MCMC) can be used to evaluate the joint posterior in (11.60), using Metropolis steps for updating $\boldsymbol{\phi}$ and Gibbs steps for all other parameters; details are available in the supplemental article (Quick et al., 2013). Sampling-based Bayesian inference seamlessly delivers inference on the residual spatial effects. Specifically, if t_0 is an arbitrary unobserved timepoint, then, for any region i, we sample from the posterior predictive distribution

$$p(Z_i(t_0) \mid \mathbf{Y}) = \int p(Z_i(t_0) \mid \mathbf{Z}, \boldsymbol{\theta})\, p(\boldsymbol{\theta}, \mathbf{Z} \mid \mathbf{Y}) d\boldsymbol{\theta}\, d\mathbf{Z} \ .$$

This is achieved using *composition sampling*: for each sampled value of $\{\boldsymbol{\theta}, \mathbf{Z}\}$, we draw $Z_i(t_0)$, one for one, from $p(Z_i(t_0) \mid \mathbf{Z}, \boldsymbol{\theta})$, which is Gaussian. Also, our sampler easily adapts to situations where $Y_i(t)$ is missing (or not monitored) for some of the time points in region i. We simply treat such variables as missing values and update them, from their associated full conditional distributions, which of course are $N(\mathbf{x}_i(t)^T\boldsymbol{\beta} + Z_i(t), \tau_i^2)$. We assume that all predictors in $\mathbf{x}_i(t)$ will be available in the space-time data matrix, so this temporal interpolation step for missing outcomes is straightforward and inexpensive.

Model checking is facilitated by simulating *independent* replicates for each observed outcome: for each region i and observed timepoint t_j, we sample from $p(Y_{rep,i}(t_j) \mid \mathbf{Y})$, which is equal to

$$\int N(Y_{rep,i}(t_j) \mid \mathbf{x}_i(t_j)^T\boldsymbol{\beta} + Z_i(t_j), \tau_i^2)\, p(\boldsymbol{\beta}, Z_i(t_j), \tau_i^2 \mid \mathbf{Y}) d\boldsymbol{\beta} dZ_i(t_j) d\tau_i^2 \ ,$$

	p_D	DIC*
Simple Linear Regression	79	9,894
Random Intercept and Slope	165	4,347
CAR Model	117	7,302
Areally Referenced Gaussian Process	5,256	0

Table 11.13 *Comparisons between our areally referenced Gaussian process model and the three alternatives. p_D is a measure of model complexity, as it represents the effective number of parameters. Smaller values of DIC indicate a better trade-off between in sample model fit and model complexity. DIC* is standardized relative to the areally referenced Gaussian Process model.*

where $p(\boldsymbol{\beta}, Z_i(t_j), \tau_i^2 \mid \mathbf{Y})$ is the marginal posterior distribution of the unknowns in the likelihood. Sampling from the posterior predictive distribution is straightforward, again, using composition sampling.

Example 11.7 Quick et al. (2013) analyze a dataset consisting of monthly asthma hospitalization rates in the 58 counties of California over an 18-year period. As such, $N_t = 12 * 18 = 216$, and we can simply set $t_j = j = 1, 2, \ldots, N_t$. The covariates in this model include population density, ozone level, the percent of the county under 18, and percent black. Population-based covariates are calculated for each county using the 2000 U.S. Census, so they do not vary temporally. In order to accommodate seasonality in the data, monthly fixed effects are included, using January as a baseline. Thus, after accounting for the monthly fixed effects and the four covariates of interest, $\mathbf{x}_i(t)$ is a 16×1 vector.

We compare the model in (11.29) with three alternative models using the DIC criterion (Spiegelhalter et al. 2002). These models are all still of the form

$$Y_i(t) = \mathbf{x}_i(t)'\boldsymbol{\beta} + Z_i(t) + \epsilon_i(t), \quad \epsilon_i(t) \overset{ind}{\sim} N(0, \tau_i^2) \text{ for } i = 1, 2, \ldots, N_s \,, \tag{11.61}$$

but with different $Z_i(t)$. Our first model is a simple linear regression model which ignores both the spatial and the temporal autocorrelation, i.e., $Z_i(t) = 0 \; \forall \; i, t$. The second model allows for a random intercept and random temporal slope, but ignores the spatial nature of the data, i.e., here $Z_i(t) = \alpha_{0i} + \alpha_{1i}t$, where $\alpha_{ki} \overset{iid}{\sim} N(0, \sigma_k^2)$, for $k = 0, 1$. In this model, to preserve model identifiability, we must remove the global intercept from our design matrix, $\mathbf{x}_i(t)$. Our third model builds upon the second, but introduces spatial autocorrelation by letting $\boldsymbol{\alpha}_k = (\alpha_{k1}, \ldots, \alpha_{kN_s})' \sim CAR(\sigma_k^2), k = 0, 1$. The results of the model comparison can be seen in Table 11.13, which indicates that our Gaussian process model has the lowest DIC value, and is thus the preferred model and the only one we consider henceforth. The surprisingly large p_D for the areally referenced Gaussian process model arises due to the very large size of the dataset (58 counties $\times$ 216 timepoints).

The estimates for our model parameters can be seen in Table 11.14. The coefficients for the monthly covariates indicate decreased hospitalization rates in the summer months, a trend which is consistent with previous findings. The coefficients for population density, percent under 18, and percent black are all significantly positive, also as expected. There is a large range of values for the county-specific residual variance parameters, τ_i^2. Perhaps not surprisingly, the magnitude of these terms seems to be negatively correlated with the population of the given counties, demonstrating the effect a (relatively) small denominator can have when computing and modeling rates. The strong spatial story seen in the maps is reflected by the size of σ^2 compared to the majority of the τ_i^2. There is also relatively strong temporal correlation, with $\phi = 0.9$ corresponding to $\rho(t_i, t_j; \phi) \geq 0.4$ for $|t_j - t_i|$ less than 2 months.

Maps of the yearly (averaged across month) spatiotemporal random effects can be seen in Figure 11.20. Since here we are dealing with the *residual* curve after accounting for a number

Parameter	Median (95% CI)	Parameter	Median (95% CI)
β_0 (Intercept)	9.17 (8.93, 9.42)	β_{11} (August)	-3.58 (-4.02, -3.13)
β_1 (Pop Den)	0.60 (0.49, 0.70)	β_{12} (September)	-1.96 (-2.37, -1.54)
β_3 (% Under 18)	1.24 (1.15, 1.34)	β_{13} (October)	-1.36 (-1.73, -1.00)
β_4 (% Black)	1.12 (1.01, 1.24)	β_{14} (November)	-0.71 (-1.02, -0.42)
β_5 (February)	-0.25 (-0.46, -0.04)	β_{15} (December)	0.63 (0.41, 0.86)
β_6 (March)	-0.21 (-0.48, 0.07)	ϕ	0.90 (0.84, 0.97)
β_7 (April)	-1.47 (-1.81, -1.12)	α	0.77 (0.71, 0.80)
β_8 (May)	-1.17 (-1.53, -0.8)	σ^2	21.52 (20.18, 23.06)
β_9 (June)	-2.79 (-3.21, -2.4)	$\bar{\tau}_{\cdot}^2$	3.32 (0.18, 213.16)
β_{10} (July)	-3.78 (-4.21, -3.37)		

Table 11.14 *Parameter estimates for asthma hospitalization data, where estimates for $\bar{\tau}_{\cdot}^2$ represent the median (95% CI) of the $\tau_i^2, i = 1, \ldots, N_s = 58$.*

Figure 11.20 *Spatial random effects for asthma hospitalization data, by year*

of mostly non-time-varying covariates, it comes as no surprise that the spatiotemporal random effects capture most of the variability in the model, including the striking decrease in yearly hospitalization rates over the study period. It also appears that our model is providing a better fit to the data in the years surrounding 2000, perhaps indicating that we could improve our fit by allowing our demographic covariates to vary temporally. Our model also appears to be performing well in the central counties, where asthma hospitalization rates remained relatively stable for much of the study period.

11.9 Exercises

1. Suppose $Var(\epsilon(\mathbf{s}, t))$ in (11.6), (11.7), and (11.8) is revised to $\sigma_\epsilon^{2(t)}$.
   (a) Revise expressions (11.11), (11.13), and (11.16), respectively.
   (b) How would these changes affect simulation-based model fitting?
2. The data `www.biostat.umn.edu/~brad/data/ColoradoS-T.dat` contain the maximum monthly temperatures (in tenths of a degree Celcius) for 50 locations over 12 months in 1997. The elevation at each of the 50 sites is also given.
   (a) Treating month as the discrete time unit, temperature as the dependendent variable, and elevation as a covariate, fit the additive space-time model (11.6) to this data. Provide posterior estimates of the important model parameters, and draw image-contour plots for each month.

   (*Hint:* Modify the `WinBUGS` code in Example 6.1 to fit a simple, nested spatiotemporal model. That is, use either the "direct" approach or the `spatial.exp` command to build an exponential kriging model for the data for a given month t with a range

parameter ϕ_t, and then assume these parameters are in turn i.i.d. from (say) a $U(0, 10)$ distribution.)

(b) Compare a few sensible models (changing the prior for the ϕ_t, including/excluding the covariate, etc.) using the DIC tool in `WinBUGS`. How does DIC seem to perform in this setting?

(c) Repeat part (a) assuming the error structures (11.7) and (11.8). Can these models still be fit in `WinBUGS`, or must you now resort to your own `C, Fortran`, or `R` code?

3. Suppose $Y(\mathbf{s}_i, t_j)$, $i = 1, \ldots, n$, $j = 1, \ldots, m$ arise from a mean-zero stationary spatiotemporal process. Let $a_{ii'} = \sum_{j=1}^{m} Y(\mathbf{s}_i, t_j)Y(\mathbf{s}_{i'}, t_j)/m$, let $b_{jj'} = \sum_{i=1}^{n} Y(\mathbf{s}_i, t_j)Y(\mathbf{s}_i, t_{j'})/n$, and let $c_{ii',jj'} = Y(\mathbf{s}_i, t_j)Y(\mathbf{s}_{i'}, t_{j'})$.

(a) Obtain $E(a_{ii'})$, $E(b_{jj'})$, and $E(c_{ii',jj'})$.

(b) Argue that if we plot $c_{ii',jj'}$ versus $a_{ii'} \cdot b_{jj'}$, under a separable covariance structure, we can expect the plotted points to roughly lie along a straight line. (As a result, we might call this a *separability plot.*) What is the slope of this theoretical line?

(c) Create a separability plot for the data in Exercise 2. Was the separability assumption there justified?

4. Consider again the data and model of Example 11.5, the former located at `www.biostat.umn.edu/~brad/data2.html`. Fit the Poisson spatiotemporal disease mapping model (11.44), but where we discard the smoking covariate, and also reverse the gender scores ($s_j = 1$ if male, 0 if female) so that the log-relative risk (11.45) is reparametrized as

$$\mu_{ijkt} = \mu + s_j\alpha + r_k\beta + s_j r_k(\xi - \alpha - \beta) + \gamma t + \phi_{it} \ .$$

Under this model, β now unequivocally captures the difference in log-relative risk between white and nonwhite females.

(a) Use either `WinBUGS` or your own `R, C++`, or `Fortran` code to find point and 95% interval estimates of β. Is there any real difference between the two female groups?

(b) Use either the mapping tool within `WinBUGS` or your own `ArcView` or other GIS code to map the fitted median nonwhite female lung cancer death rates per 1000 population for the years 1968, 1978, and 1988. Interpret your results. Is a temporal trend apparent?

5. In the following, let C_1 be a valid two-dimensional isotropic covariance function and let C_2 be a valid one-dimensional isotropic covariance function. Let $C_A(\mathbf{s}, t) = C_1(\mathbf{s}) + C_2(t)$ and $C_M(\mathbf{s}, t) = C_1(\mathbf{s})C_2(t)$. C_A is referred to as an *additive* (or *linear*) space-time covariance function, while C_M is referred to as a *multiplicative* space-time covariance function.

(a) Why are C_A and C_M valid?

(b) Comment on the behavior of C_A and C_M as $||\mathbf{s} - \mathbf{s}', t - t'|| \to 0$ (local limit), and as $||\mathbf{s} - \mathbf{s}', t - t'|| \to \infty$ (global limit).

6. Suppose we observe a constant mean space-time process, $Y(\mathbf{s}, t)$ at equally spaced time points over a regular lattice. How might we obtain simple sample estimates for the covariance functions, $C(\mathbf{s}, t)$, $C(\mathbf{s})$, and $C(t)$? How might we use these to do some exploratory data analysis with regard to separability of the covariance function?

7. Suppose a simple dynamic model $Y_t(\mathbf{s}) = \gamma Y_{t-1}(\mathbf{s}) + \eta_t(\mathbf{s})$ where $|\gamma| < 1$ and the $\eta_t(\mathbf{s})$ are independent and identically distributed stationary mean 0 Gaussian processes for $\mathbf{s} \in D$. Show that the $Y_t(\mathbf{s})$ has a separable covariance function in space and time. Show that this is not the case if γ depends upon $\mathbf{s}$. More generally, consider the dynamical process with integro-difference equation $Y_t(\mathbf{s}) = \int h(\mathbf{s} - \mathbf{s}')Y_{t-1}(\mathbf{s}')d\mathbf{s}' + \eta_t(\mathbf{s})$ with $\eta(\mathbf{s})$ as above. Show that the process is not separable unless $h(\mathbf{0}) = \gamma \neq 0$ and $h(\mathbf{u}) = 0$ for almost all $\mathbf{u} \neq 0$.

Chapter 12

Modeling large spatial and spatiotemporal datasets

12.1 Introduction

Implementing Gibbs sampling or other MCMC algorithms requires repeated evaluation of various full conditional density functions. In the case of hierarchical models built from random effects using Gaussian processes, this requires repeated evaluation of the likelihood and/or joint or conditional densities arising under the Gaussian process; see Section A.2. In particular, such computation requires evaluation of quadratic forms involving the inverse of covariance matrix and also the determinant of that matrix. Strictly speaking, we do not have to obtain the inverse in order to compute the quadratic form. Letting $\mathbf{z}^T A^{-1}\mathbf{z}$ denote a general object of this sort, if we obtain $A^{\frac{1}{2}}$ (e.g. a triangular Cholesky factorization) and solve $\mathbf{z} = A^{\frac{1}{2}}\mathbf{v}$ for $\mathbf{v}$, then $\mathbf{v}^T\mathbf{v} = \mathbf{z}^T A^{-1}\mathbf{z}$. Still, with large n, computation associated with resulting $n \times n$ matrices can be unstable, and repeated computation (as for simulation-based model fitting) can be very slow, perhaps infeasible. After all, spatial covariance matrices, in general, are dense and the Cholesky factorization requires about $O(n^3/3)$ flops. We refer to this situation informally as *"the big n problem."*

Extension to multivariate models with, say, p measurements at a location leads to $np \times np$ matrices (see Section 9.5). Extension to spatiotemporal models (say, spatial time series at T time points) leads to $nT \times nT$ matrices (see Section 11.2). Of course, there may be modeling strategies that will simplify to, say, T $n \times n$ matrices or an $n \times n$ and a $T \times T$ matrix, but the problem will still persist if n is large. The objective of this section is thus to review approaches for handling spatial process models in this case. Our emphasis is on the predictive process, a straightforward, off-the-shelf approach which we are quite familiar with and which can handle many challenging settings. In fact, it is already incorporated into the `spBayes` software.

Broadly speaking, the approaches for tackling the big n problem can be classified as those that seek approximations to the exact likelihood and those that develop models which can handle fitting with large values of n. We first describe the approximate likelihood approaches and then turn to the arguably, richer option of devising models for large spatial datasets.

12.2 Approximate likelihood approaches

12.2.1 Spectral methods

A rich, and theoretically attractive, option is to work in the spectral domain (as advocated by Stein, 1999a, and Fuentes, 2002a). The idea is to transform to the space of frequencies, develop a periodogram (an estimate of the spectral density), and utilize the Whittle likelihood (Whittle, 1954; Guyon, 1995) in the spectral domain as an approximation to the data likelihood in the original space. The Whittle likelihood requires no matrix inversion

so, as a result, computation is very rapid. In principle, inversion back to the original space is straightforward.

The practical concerns here are the following. First, there is discretization to implement a fast Fourier transform (see Section A.1). Then, there is a certain arbitrariness to the development of a periodogram. Empirical experience is employed to suggest how many low frequencies should be discarded. Also, there is concern regarding the performance of the Whittle likelihood as an approximation to the exact likelihood. Some empirical investigation we have attempted suggests that this approximation is reasonably well centered, but does a less than satisfactory job in the tails (thus leading to poor estimation of model variances). Lastly, with non-Gaussian first stages, we will be doing all of this with random spatial effects that are never observed, making the implementation impossible. In summary, use of the spectral domain with regard to handling large n is limited in its application, and requires considerable familiarity with spectral analysis (discussed briefly in Subsection 3.1.2).

12.2.2 Lattice and conditional independence methods

Though Gaussian Markov random fields have received a great deal of recent attention for modeling areal unit data, they were originally introduced for points on a regular lattice. In fact, using inverse distance to create a proximity matrix, we can immediately supply a joint spatial distribution for variables at an arbitrary set of locations. As in Section 4.2, this joint distribution will be defined through its full conditional distribution. The joint density is recaptured using Brook's Lemma (4.7). The inverse of the covariance matrix is directly available, and the joint distribution can be made proper through the inclusion of an autocorrelation parameter. Other than the need to sample a large number of full conditional distributions, there is no big n problem. Indeed, many practitioners immediately adopt Gaussian Markov random field models as the spatial specification due to the computational convenience.

The disadvantages arising with the use of Gaussian Markov random fields should by now be familiar. First, and perhaps most importantly, we do not model association directly, which precludes the specification of models exhibiting desired correlation behavior. The joint distribution of the variables at two locations depends not only on their joint distribution given the rest of the variables, but also on the joint distribution of the rest of the variables. In fact, the relationship between entries in the inverse covariance matrix and the actual covariance matrix is very complex and highly nonlinear. Besag and Kooperberg (1995) showed, using a fairly small n that entries in the covariance matrix resulting from a Gaussian Markov random field specification need not behave as desired. They need not be positive nor decay with distance. With large n, the implicit transformation from inverse covariance matrix to covariance matrix is even more ill-behaved (Conlon and Waller, 1999; Wall, 2003).

In addition, with a Gaussian Markov random field there is no notion of a stochastic process, i.e., a collection of variables at all locations in the region of interest with joint distributions determined through finite dimensional distributions. In particular, we cannot write down the distribution of the variable at a selected location in the region. Rather, the best we can do is determine a conditional distribution for this variable given the variables at some prespecified number of and set of locations. Also, introduction of nonspatial error is confusing. The conditional variance in the Gaussian Markov random field cannot be aligned in magnitude with the marginal variance associated with a white noise process. Also, as we clarified in 4.3, Markov random field models preclude valid interpolation diminishing their utility as approximations .

Some authors have proposed approximating a Gaussian process with a Gaussian Markov random field. More precisely, a given set of spatial locations $s_1, ..., s_n$ along with a choice of correlation function yields an $n \times n$ covariance matrix Σ_1. How might we specify a Gaussian Markov random field with full rank inverse matrix Σ_2^{-1} such that $\Sigma_2 \approx \Sigma_1$? That is, unlike

the previous paragraph where we start with a Gaussian Markov random field, here we start with the Gaussian spatial process.

A natural metric in this setting is Kullback-Liebler distance (see Besag and Kooperberg, 1995). If $f_1 \sim N(0, \Sigma_1)$ and $f_2 \sim N(0, \Sigma_2)$, the Kullback-Leibler distance of f_2 from f_1 is

$$KL(f_1, f_2) = \int f_1 \log(f_1/f_2) = -\frac{1}{2}\log\left|\Sigma_2^{-1}\Sigma_1\right| + \frac{1}{2}tr(\Sigma_2^{-1}\Sigma_1 - I). \tag{12.1}$$

Hence, we only need Σ_1 and Σ_2^{-1} to compute (12.1). Using an algorithm originally proposed by Dempster (1972), Besag and Kooperberg provide approximation based upon making (12.1) small. Rue and Tjelmeland (2002) note that this approach does not well approximate the correlation function of the Gaussian process. In particular, it will not do well when spatial association decays slowly. Rue and Tjelmeland propose a "matched correlation" criterion that accommodates both local and global behavior.

12.2.3 INLA

Laplace approximation provided an early approach to handling challenging Bayesian computation (Kass, Tierney and Kadane, 1989; 1991). It was particularly successful in the context of so-called conditionally independent hierarchical models, models with conditionally independent first stage specifications and with exchangeable parameters at the second stage. Its application diminished with the arrival of MCMC model fitting but, recently, it has enjoyed a rejuvenation through the popular Integrated Nested Laplace Approximation (INLA) package (Rue, Martino, Chopin (2009). INLA handles spatial analysis through Markov random field approximation on regular grids. The software runs very rapidly and the Laplace approximation under the hood is very well done. However, overall inference is limited and application to more challenging multi-level models may be difficult.

The previous section offered some insight into how we might develop an approximation to a Gaussian process using a Gaussian Markov random field (GMRF). INLA adopts a very attractive choice through the use of the stochastic partial differential equation approach (SPDE) (Lindgren, Rue, and Lindstrom, 2011). This approach offers an explicit link between a Gaussian process and a GMRF. The SPDE approach extends work of Besag (1981), who proposed to approximate a Gaussian process when $\nu \to 0$ in the Matérn correlation function. This approximation imagines a regular two-dimensional lattice where the number of sites tends to infinity with local conditional distributions having $E(Y_{ij}|Y_{-ij}) = \frac{1}{a}(Y_{i-1,j}+Y_{i+1,j}+Y_{i,j-1}+Y_{i,j+1})$ and $\mathrm{Var}(Y_{ij}|Y_{-ij}) = \frac{1}{a}$ for $|a| > 4$. In the precision matrix, for site (i, j), we have the value a at (i, j) and the values -1 at each of the four immediate N, E, S, and W neighbors of (i, j).

In Lindgren et al. (2011) it is noted that a Gaussian process $X(\mathbf{s})$ with the Matérn covariance is a solution to the linear fractional stochastic partial differential equation (SPDE) $\kappa^2 - \Delta)^{\alpha/2}X(\mathbf{s}) = W(\mathbf{s})$ where $W(\mathbf{s})$ is a white noise process, $\alpha = \nu + d/2$ with $\kappa > 0$, $\nu > 0$ (the usual Matérn smoothness parameter) and, in the spatial setting, $d = 2$. When $\nu = 1$ we achieve an extension of Besag, adding 8 more neighbors, expanding from 4 to 12 neighbors, providing local entries in the precision matrix as follows: $4 + a^2$ at $i, j)$, $-2a$ at each of the immediate N, E, S, and W neighbors, 2 at the NW, NE, SE and SW neighbors and 1 at the two-away N, E, S, and W neighbors. Extension to $\nu = 2$ adds 12 more neighbors yielding a total of 24.

Intuition suggests that if we have larger ν in the Matérn correlation function, i.e., a smoother process realization, we need more non-zero neighbors sites in the GMRF representation. If the spatial locations are on an irregular grid, it is necessary to use a second result in Lindgren et al. (2011) which overlays a regular grid and employs the finite element method for an interpolation of the locations of observations to the nearest grid point. A basis function representation with random Gaussian weights provides the interpolation.

12.2.4 Approximate likelihood

Evidently, the big n challenge arises because we have a high-dimensional joint distribution in the likelihood. So, a natural approximation would be to attempt some sort of pseudo-likelihood approximation, replacing the joint density $f(y(\mathbf{s}_1), y(\mathbf{s}_2), ..., y(\mathbf{s}_n)$ with a product approximation, using a conditional density for $Y(\mathbf{s}_i)$ given a subset of the remaining y's. This idea dates at least to Vecchia (1988) who proposed this as an early spatial analysis computational trick. However, the approach suffers many problems. First, it is not formally defined. Second, it will typically be sequence dependent, though there is no natural ordering of the spatial locations. Most troubling is the arbitrariness in the number of and choice of "neighbors." Moreover, perhaps counter-intuitively, we can not merely select locations close to $\mathbf{s}_i$. Stein, Chi and Welty (2004) pointed out that we need locations at larger distances from each of the $\mathbf{s}_i$ (as we would have with the full data likelihood) in order to learn about the spatial decay in dependence for the process. So, altogether, we do not see such approximations as useful approach.

For hierarchical models with a non-Gaussian first stage, the foregoing forms the basis for coarse-fine coupling as in Higdon, Lee, and Holloman (2003). The idea here is, with a non-Gaussian first stage, if spatial random effects (say, $\theta(\mathbf{s}_1), \ldots, \theta(\mathbf{s}_n)$) are introduced at the second stage, then, as in Subsection 6.2, the set of $\theta(\mathbf{s}_i)$ will have to be updated at each iteration of a Gibbs sampling algorithm.

Suppose n is large and that a "fine" chain does such updating. This chain will proceed very slowly. Suppose, concurrently, we run a "coarse" chain using a much smaller subset n' of the $\mathbf{s}_i$'s. The coarse chain will update very rapidly. Since the process for $\theta(\cdot)$ is the same in both chains it will be the case that the coarse one will explore the posterior more rapidly. However, we need realizations from the fine chain to fit the model using all of the data.

The coupling idea is to let both the fine and coarse chains run, and after a specified number of updates of the fine chain (and many more updates of the coarse chain, of course) we attempt a "swap," i.e., we propose to swap the current value of the fine chain with that of the coarse chain. The swap attempt ensures that the equilibrium distributions for both chains are not compromised (see Higdon, Lee, and Holloman, 2003). For instance, given the values of the θ's for the fine iteration, we might just use the subset of θ's at the locations for the coarse chain. Given the values of the θ's for the coarse chain, we might do an appropriate kriging to obtain the θ's for the fine chain.

With regard to specifying the coarse chain, one could employ a subsample of the sampled locations. Subsampling can be formalized into a model-fitting approach following the ideas of Pardo-Igúzquiza and Dowd (1997). Specifically, for observations $Y(\mathbf{s}_i)$ arising from a Gaussian process with parameters $\boldsymbol{\theta}$, they propose replacing the joint density of $\mathbf{Y} = (Y(\mathbf{s}_1), \ldots, Y(\mathbf{s}_n))^T$, $f(\mathbf{y}|\boldsymbol{\theta})$, by

$$\prod_{i=1}^{n} f(y(\mathbf{s}_i) \mid y(\mathbf{s}_j),\ \mathbf{s}_j \in \partial \mathbf{s}_i)\ , \tag{12.2}$$

where $\partial \mathbf{s}_i$ defines some neighborhood of $\mathbf{s}_i$. For instance, it might be all $\mathbf{s}_j$ within some specified distance of $\mathbf{s}_i$, or perhaps the m $\mathbf{s}_j$'s closest to $\mathbf{s}_i$ for some integer m. Pardo-Igúzquiza and Dowd (1997) suggest the latter, propose $m = 10$ to 15, and check for stability of the inference about $\boldsymbol{\theta}$. Formal argument for approximating $f(\mathbf{y}|\boldsymbol{\theta})$ by (12.2) is essentially from Vecchia (1988) above. In light of the above, we view this approach as purely an algorithm for large datasets rather than a recommendable modeling approach.

12.2.5 Variational Bayes algorithm for spatial models

Variational methods have their origins in the 18th century with the work of Euler, Lagrange, and others on the calculus of variations. Here, we define a functional as a mapping that

takes a function as input instead of a variable and returns the value of the functional as the output. An example would be the entropy $H(p) = -\int p(y) \ln p(y)\, dy$, which takes a probability density function $p(y)$ as the input and returns the quantity value. In particular, entropy is a functional which we work with below.

Many problems can be expressed in terms of an optimization problem in which the quantity being optimized is a functional. The solution is obtained by exploring all possible functions to find the one that maximizes (or minimizes) the functional. Usually no closed form solution can be found. Therefore, variational methods naturally focus on approximations to the optimal solutions. In the case of applications to probabilistic inference, the mean field approximation is used.

Consider how variational optimization can be applied to the Bayes inference problem. Let $\mathbf{y}$ denote the observed variables and $\boldsymbol{\theta}$ denote the unobserved parameters. We assume a prior distribution $p(\boldsymbol{\theta})$ for parameter $\boldsymbol{\theta}$. Then the marginal likelihood $p(\mathbf{y}) = \int p(\mathbf{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})\, d\boldsymbol{\theta}$ can be bounded below using any distribution over the parameter $\boldsymbol{\theta}$. To see how, let $q(\boldsymbol{\theta})$ be any probability density function on $\boldsymbol{\theta}$. Then,

$$\log p(\mathbf{y}) = \log \frac{p(\mathbf{y}, \boldsymbol{\theta})}{p(\boldsymbol{\theta}|\mathbf{y})} = \log p(\mathbf{y}, \boldsymbol{\theta}) - \log p(\boldsymbol{\theta}|\mathbf{y}) = \log \frac{p(\mathbf{y}, \boldsymbol{\theta})}{q(\boldsymbol{\theta})} + \log \frac{q(\boldsymbol{\theta})}{p(\boldsymbol{\theta}|\mathbf{y})} \ .$$

Multiplying both sides by $q(\boldsymbol{\theta})$ and integrating with respect to $\boldsymbol{\theta}$, we obtain

$$\begin{aligned} \log p(\mathbf{y}) &= \int q(\boldsymbol{\theta}) \log \frac{p(\mathbf{y}, \boldsymbol{\theta})}{q(\boldsymbol{\theta})}\, d\boldsymbol{\theta} + \int q(\boldsymbol{\theta}) \log \frac{q(\boldsymbol{\theta})}{p(\boldsymbol{\theta}|\mathbf{y})}\, d\boldsymbol{\theta} \\ &= \mathcal{L}(q) + KL(q, p) \geq \mathcal{L}(q) \ , \end{aligned}$$

where $\mathcal{L}(q)$ is a function of $\mathbf{y}$ and $KL(q, p)$ is the Kullback-Liebler (KL) distance from $q(\boldsymbol{\theta})$ to $p(\boldsymbol{\theta} \mid \mathbf{y})$. Since $KL(q, p)$ satisfies Gibb's inequality (MacKay, 2003) it is always nonnegative, hence $\mathcal{L}(q)$ is a lower bound for the log marginal likelihood. Thus to find a $q(\boldsymbol{\theta})$ that approximates $p(\boldsymbol{\theta}|\mathbf{y})$ well, we can either maximize $\mathcal{L}(q)$ or minimize $KL(q, p)$. Let $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_m)$ and $\mathcal{Q} = \{q(\boldsymbol{\theta}) : q(\boldsymbol{\theta}) = \prod_{i=1}^m q_i(\boldsymbol{\theta}_i)\}$, where each $\boldsymbol{\theta}_i$ can be scalar or vector. Then $\mathcal{L}(q)$ for $q(\boldsymbol{\theta}) \in \mathcal{Q}$ can be written as:

$$\mathcal{L}(q) = \int \prod_{i=1}^m q_i(\boldsymbol{\theta}_i) \log p(\mathbf{y}, \boldsymbol{\theta})\, d\boldsymbol{\theta} - \int q(\boldsymbol{\theta}) \log q(\boldsymbol{\theta})\, d\boldsymbol{\theta}.$$

Using variational calculus we can now show that the optimal $q_i^*(\boldsymbol{\theta}_i)$, which maximizes $\mathcal{L}(q)$, is given by $\log q_i^*(\boldsymbol{\theta}_i) = \mathrm{E}_{j \neq i}[\log p(\mathbf{y}, \boldsymbol{\theta})]$+constant, where $\mathrm{E}_{j \neq i}[\log p(\mathbf{y}, \boldsymbol{\theta})]$ is the expectation of $\log p(\mathbf{y}, \boldsymbol{\theta})$ over $\prod_{j \neq i} q_j(\boldsymbol{\theta}_j)$. Then, it can be shown that

$$q_i^*(\boldsymbol{\theta}_i) = \frac{\exp\{\mathrm{E}_{j \neq i}[\log p(\mathbf{y}, \boldsymbol{\theta})]\}}{\int \exp\{\mathrm{E}_{j \neq i}[\log p(\mathbf{y}, \boldsymbol{\theta})]\}\, d\boldsymbol{\theta}_i} \ . \tag{12.3}$$

Equation (12.3) represents a set of consistent conditions for the maximum of the lower bound subject to the factorization constraint. However, it does not represent an explicit solution because the right hand side of (12.3) depends on the expectation computed with respect to the other parameters $\boldsymbol{\theta}_j$. So we must initialize the distribution of all the $\boldsymbol{\theta}_j$ and then cycle through them iteratively. Each parameter's distribution is updated in turn with a revised function given by (12.3) and evaluated using the current estimate of the distribution function for all other parameters. Convergence is guaranteed because the bound is convex with respect to each of the factors $q_i(\boldsymbol{\theta}_i)$ (Attias, 2000). Ren, Banerjee, Finley and Hodges (2011) develop VB methods for univariate and multivariate (Bayesian LMC) hierarchical spatial models.

12.2.6 Covariance tapering

A wide class of likelihood approximations rely upon sparse approximations. One such approach comes from tapered processes. Recall that we discussed covariance tapering in Section 3.4 in the context of using product form for constructing valid covariance functions. Returning to that theme, tapered processes offer an alternative means to dimension reduction, by producing sparse spatial covariance matrices (e.g., Furrer et al., 2006; Kaufman et al., 2009; Du et al., 2010). The underlying idea is to use a compactly supported covariance function (Wendland, 1995; Gneiting, 2002) as a *tapering kernel* $C_\nu(\mathbf{s}_1, \mathbf{s}_2)$, which is a positive-definite function satisfying

$$C_\nu(\mathbf{s}_1, \mathbf{s}_2) = 0 \quad \text{if} \quad ||\mathbf{s}_1 - \mathbf{s}_2|| > \nu \,, \tag{12.4}$$

where ν is the distance beyond which the covariance becomes zero. Tapering introduces a sparse structure for the spatial covariance matrix from the Gaussian process model. Let T_ν be the $n \times n$ matrix with (i, j)-th element $C_\nu(\mathbf{s}_i, \mathbf{s}_j)$. Clearly the matrix T will have zero entries for any pair of locations separated by more than ν units and, therefore, is sparse. There are choices aplenty for the tapering kernel, but the more widely used kernels use the Wendland family of tapered covariance functions (Wendland, 1995; Furrer et al., 2006). One particularly popular choice is given by

$$C_\nu(\mathbf{s}_1, \mathbf{s}_2) = \left(1 - \frac{h}{\nu}\right)_+^4 \left(1 + 4\frac{h}{\nu}\right) \,, \tag{12.5}$$

where $h = ||\mathbf{s}_1 - \mathbf{s}_2||$. Note that ν is typically not estimated, but fixed to achieve the desired degree of sparsity in the spatial covariance matrix (Kaufman et al., 2009). In a Bayesian context, we can estimate ν using some prior distribution, but such priors will need to be strongly informative for ν to be identified, which may not be straightforward.

Tapering a covariance function $C(\mathbf{s}_1, \mathbf{s}_2; \boldsymbol{\theta}_1)$ yields

$$C_{tap}(\mathbf{s}_1, \mathbf{s}_2; \boldsymbol{\theta}_1) = C_\nu(\mathbf{s}_1, \mathbf{s}_2) C(\mathbf{s}_1, \mathbf{s}_2, \boldsymbol{\theta}_1) \,.$$

Tapered covariances have been used effectively for analyzing large spatial datasets (e.g., Furrer et al., 2009) as its process realizations yield a sparse dispersion matrix $C(\boldsymbol{\theta}_1) \odot T_\nu$, where $\odot$ is the elementwise matrix product (or the Hadamard product). A standard property of the Hadamard product ensures that $C(\boldsymbol{\theta}_1) \odot T_\nu$ will be positive definite because $C(\boldsymbol{\theta}_1)$ and T_ν are. Since T_ν is sparse, so is $C(\boldsymbol{\theta}_1) \odot T_\nu$; therefore, sparse matrix algorithms can be employed to estimate tapered spatial process models.

12.3 Models for large spatial data: low rank models

A popular way of dealing with large spatial datasets is to devise models that bring about dimension reduction. The essential idea is to replace the spatial process $w(\mathbf{s})$ with $\tilde{w}(\mathbf{s})$, where the latter is a dimension-reducing process. How can we construct such dimension-reducing processes? A reduced rank *low rank* or *reduced rank* specification is typically based upon a representation in terms of the realizations of some latent process over a smaller set of coordinates called *knots*. To be precise,

$$\tilde{w}(\mathbf{s}) = \sum_{j=1}^{m} l(\mathbf{s}, \mathbf{s}_j^*) Z(\mathbf{s}_j^*), \tag{12.6}$$

where $Z(\mathbf{s})$ is a well defined process. The surface/process realization for $\tilde{w}(\mathbf{s})$ is completely determined by the function $l(\cdot, \cdot)$ and the set of variables, $\{Z(\mathbf{s}_j^*), j = 1, 2, ..., m\}$. The

collection of $\mathbf{s}_j^*$'s are the knots. For a collection of locations, with associated vector denoted by $\tilde{\mathbf{w}} = (\tilde{w}(\mathbf{s}_1), \tilde{w}(\mathbf{s}_2), ..., \tilde{w}(\mathbf{s}_n))^T$, we write

$$\tilde{\mathbf{w}} = \mathbf{L}\mathbf{z}^*, \tag{12.7}$$

where $\mathbf{L}$ is the $n \times m$ matrix with (i, j)-th element $l(\mathbf{s}_i, \mathbf{s}_j^*)$ and $\mathbf{z}^*$ is the $m \times 1$ vector with entries $Z(\mathbf{s}_j^*)$ with $m < n$.

Equation (12.7) immediately reveals dimension reduction: despite there being n $\tilde{w}(\mathbf{s}_i)$'s, we will only have to work with m $Z(\mathbf{s}_j^*)$'s. Since we anticipate $m << n$, the consequential dimension reduction is evident and, since we will write the model in terms of the Z's (with the $\tilde{w}$'s being deterministic from the Z's, given $l(\cdot, \cdot)$), the associated matrices we work with will be $m \times m$. Evidently, $\tilde{w}(\mathbf{s})$ as defined in (12.6) spans only an m-dimensional space; we create an uncountable number of variables through a finite number of variables. When $n > m$, the joint distribution of $\tilde{\mathbf{w}}$ is singular. However, we do create a valid stochastic process. In particular, the valid covariance function is

$$\text{cov}(\tilde{w}(\mathbf{s}), \tilde{w}(\mathbf{s}')) = \mathbf{l}(\mathbf{s})^T \Sigma_{\mathbf{Z}^*} \mathbf{l}(\mathbf{s}') \tag{12.8}$$

where $\mathbf{l}(\mathbf{s})$ is the $m \times 1$ vector with entries $l(\mathbf{s}, \mathbf{s}_j^*)$. From (12.8), we see that, even if $l(\cdot, \cdot)$ is stationary, i.e., of the form $l(\cdot - \cdot)$, the induced covariance function is not. Also, if the Z's are Gaussian, then $\tilde{w}(\mathbf{s})$ is a Gaussian process.

How do we view the Z's? Are they a collection of w's from a process of interest, whence the $\tilde{w}$'s provide an approximation to enable computational tractability? Or are they merely a specification to provide a spatial process model? Also, are the $\mathbf{s}_j^*$ a subset of the observed $\mathbf{s}$'s or chosen otherwise? We conclude here with a few more words about the first three questions. The last question takes us to the design problem.

The most prevalent specification for the Z's is i.i.d. normal with mean 0 and variance σ^2, i.e., $Z(\mathbf{s})$ is a white noise process, whence (12.8) simplifies to $\sigma^2 \mathbf{l}(\mathbf{s})^T \mathbf{l}(\mathbf{s}')$. This form appears in Barry and Ver Hoef (1996) and in a series of papers by Higdon and collaborators (e.g., Higdon et al., 1998; Higdon, 2002), the former calling it a "moving average" model, the latter, "kernel convolution." In particular, a natural choice for l is a kernel function, say, $K(\mathbf{s} - \mathbf{s}')$ which puts more weight on $\mathbf{s}'$ near $\mathbf{s}$. The kernel would have parameters (which induces a parametric covariance function) and might be spatially varying (Higdon, 2002; Paciorek and Schervish, 2006). The reduced rank form can be viewed as a discretization of a process specification of the form $\tilde{w}(\mathbf{s}) = \int_{R^2} K(\mathbf{s} - \mathbf{s}')Z(\mathbf{s}')d\mathbf{s}'$ (see Xia and Gelfand, 2006, for discussion regarding this discrete approximation). Gaussian kernels are frequently used though they lead to Gaussian covariance functions which, typically, yield process realizations too smooth to be satisfactory in practice (see Section 3.5 as well as Stein, 1999; Paciorek and Schervish, 2006). Moreover, the scope of processes that can be obtained through kernel convolution is limited; for instance, the widely used exponential covariance function does not arise from kernel convolution.

A different approach to specification for the Z's is to endow them with a stochastic process model having a selected covariance function. Again, from (12.8), this will impart a covariance function to the $\tilde{w}$'s. Reversing the perspective, if we have a particular covariance function that we wish for the $\tilde{w}(\mathbf{s})$, what covariance function shall we choose for the Z's? We argue in Section 12.4 that, in some sense, the *predictive process* provides an *optimal* choice.

12.3.1 Kernel-based dimension reduction

Recall the idea of kernel convolution (see Subsection 3.2.2) where we represent the process $Y(s)$ by

$$Y(\mathbf{s}) = \int k(\mathbf{s} - \mathbf{s}')z(\mathbf{s}')d\mathbf{s}' \,, \tag{12.9}$$

where k is a kernel function (which might be parametric, and might be spatially varying) and $z(\mathbf{s})$ is a stationary spatial process (which might be white noise, that is, $\int_A z(\mathbf{s})d\mathbf{s} \sim N(0, \sigma^2 A)$ and $cov(\int_A z(\mathbf{s})d\mathbf{s}, \int_B z(\mathbf{s})d\mathbf{s}) = \sigma^2|A \cap B|$). A finite version of (12.9) yields

$$Y(\mathbf{s}) = \sum_{j=1}^{J} k(\mathbf{s} - \mathbf{s}_j^*)z(\mathbf{s}_j^*) \ . \tag{12.10}$$

Expression (12.10) shows that given k, every variable in the region is expressible as a linear combination of the set $\{z(\mathbf{s}),\ j = 1, ..., J\}$. Hence, no matter how large n is, working with the z's, we never have to handle more than a $J \times J$ matrix. The richness associated with the class in (12.9) suggests reasonably good richness associated with (12.10). Versions of (12.10) to accommodate multivariate processes and spatiotemporal processes can be readily envisioned.

Concerns regarding the use of (12.10) involve two issues. First, how does one determine the number of and choice of the $\mathbf{s}_j^*$'s? How sensitive will inference be to these choices? Also, the joint distribution of $\{Y(\mathbf{s}_i),\ i = 1, ..., n\}$ will be singular for $n > J$. While this does not mean that $Y(\mathbf{s}_i)$ and $Y(\mathbf{s}_i')$ are perfectly associated, it does mean that specifying $Y(\cdot)$ at J distinct locations determines the value of the process at all other locations. As a result, such modeling may be more attractive for spatial random effects than for the data itself.

A variant of this strategy is a conditioning idea. Suppose we partition the region of interest into M subregions so that we have the total of n points partitioned into n_m in subregion m with $\sum_{m=1}^{M} n_m = n$. Suppose we assume that $Y(\mathbf{s})$ and $Y(\mathbf{s}')$ are conditionally independent given $\mathbf{s}$ lies in subregion m and $\mathbf{s}'$ lies in subregion m'. However, suppose we assign random effects $\gamma(\mathbf{s}_1^*), ..., \gamma(\mathbf{s}_M^*)$ with $\gamma(\mathbf{s}_m^*)$ assigned to subregion m. Suppose the $\mathbf{s}_M^*$'s are "centers" of the subregions (using an appropriate definition) and that the $\gamma(\mathbf{s}_M^*)$ follows a spatial process that we can envision as a *hyper*spatial process. There are obviously many ways to build such multilevel spatial structures, achieving a variety of spatial association behaviors. We do not elaborate here but note that we will now have $n_m \times n_m$ matrices with an $M \times M$ matrix rather than a single $n \times n$.

12.3.2 The Karhunen-Loéve representation of Gaussian processes

Reduced rank approaches approximate the parent process $w(\mathbf{s})$ by a process $\tilde{w}(\mathbf{s})$ that lies in a fixed, finite-dimensional space. In seeking such approximations, one can consider the Karhunen-Love theorem (named after Kari Karhunen and Michel Loéve) that represents a stochastic process as an infinite linear combination of orthogonal functions, analogous to a Fourier series representation of a function on a bounded interval. In the case of a zero-centered spatial process $w_D = \{w(\mathbf{s}) : \mathbf{s} \in D\}$ with covariance function $C(\mathbf{s}_1, \mathbf{s}_2)$, where D is a compact subset of $\Re^d$, the Karhunen-Loéve expansion can be written as

$$w(\mathbf{s}) = \sum_{i=1}^{\infty} \sqrt{\lambda_i}\phi_i(\mathbf{s})Z_i \ , \tag{12.11}$$

where Z_i's are a sequence of independent and identically distributed $N(0, 1)$ random variables, λ_i's are the (positive) eigenvalues, often arranged in non-increasing order $\lambda_1 \geq \lambda_2 \geq \cdots$, of the symmetric positive definite function $C(\mathbf{s}_1, \mathbf{s}_2)$, and the $\phi_i(\mathbf{s})$'s are the corresponding eigenfunctions. The $\phi_i(\mathbf{s})$'s are continuous real-valued functions on D, which are mutually orthogonal in the L^2 space of functions over D. They form a set of "basis" functions to represent $w(\mathbf{s})$ with Z_i being the random coefficients with respect to this basis and λ_i's providing a scale adjustment to the random coefficients. Using standard inversion techniques, we have $Z_i = \dfrac{1}{\sqrt{\lambda_i}} \displaystyle\int_D w(\mathbf{s})\phi_i(\mathbf{s})d\mathbf{s}$, $i = 1, 2, \ldots$. The eigen-pairs $\{\lambda_i, \phi_i(\mathbf{s})\}$ satisfy

the integral equation

$$\int_D C(\mathbf{s},\mathbf{u})\phi_i(\mathbf{u})d\mathbf{u} = \lambda_i\phi_i(\mathbf{s}) \ . \tag{12.12}$$

Note that $\sqrt{\lambda_i}$'s are often referred to as the *singular values* of the process $w(\mathbf{s})$. The underlying theme in reduced rank approaches is that only the leading terms in the K-L expansion capture the main feature of the process, so the remaining terms can be dropped from the expansion in (12.11) to yield a reasonable reduced rank approximation of the process. Assuming that $\lambda_i \approx 0$ for $i = m+1, m+2, \ldots$, we retain only the first m terms in (12.11) to arrive at a rank-m approximation:

$$w(\mathbf{s}) \approx \tilde{w}(\mathbf{s}) = \sum_{i=1}^{m} \sqrt{\lambda_i}\phi_i(\mathbf{s})Z_i \ . \tag{12.13}$$

The covariance function for the rank-m process $\tilde{w}(\mathbf{s})$ is given by $\tilde{C}(\mathbf{s}_1,\mathbf{s}_2) = \boldsymbol{\phi}(\mathbf{s}_1)^T\Lambda\boldsymbol{\phi}(\mathbf{s}_2)$, where $\boldsymbol{\phi}(\mathbf{s}) = (\phi_1(\mathbf{s}),\phi_2(\mathbf{s}),\ldots,\phi_m(\mathbf{s}))^T$ and Λ is an $m\times m$ diagonal matrix with λ_i as its i-th diagonal entry. Note that irrespective of how many locations we have, the rank of the matrix Λ will always remain fixed at m. Therefore, the process $\tilde{w}(\mathbf{s})$ is a *degenerate* Gaussian process whose (partial) realizations yield singular (rank-deficient) normal distributions over sets with more than m locations.

Curiously, the reduced-rank representation in (12.13) does not explicitly depend upon knots as does the representation (12.6). After all, knots, or any subset of locations, do not arise in constructing (12.13). Instead, (12.13) truncates a basis expansion based upon the magnitude of the eigenvalues of the parent covariance function. However, the predictive process (Banerjee et al., 2008), which is our topic of discussion in the next section, emerges as a special case of the reduced-rank representation in (12.13). To see how, consider implementing (12.13) in practice. This will entail computing the m eigen-pairs $\{\lambda_i, \phi_i(\mathbf{s})\}$ by solving (12.12). One way to solve this is to discretize (12.12) using an approximate linear system. The discretized system will use the "knots," say, $\mathcal{S}^* = \{\mathbf{s}_1^*, \mathbf{s}_2^*, \ldots, \mathbf{s}_m^*\}$, as the arguments in the integrand in (12.12), so that

$$\frac{1}{m}\sum_{i=1}^{m} C(\mathbf{s},\mathbf{s}_i^*)\phi_i(\mathbf{s}_i^*) = \lambda_i\phi_i(\mathbf{s}) \Rightarrow \mathbf{c}^T(\mathbf{s})\boldsymbol{\phi}_i^* = m\lambda_i\phi_i(\mathbf{s}) \ , \quad i = 1,2,\ldots,m, \tag{12.14}$$

where $\mathbf{c}(\mathbf{s}) = (C(\mathbf{s},\mathbf{s}_1^*), C(\mathbf{s},\mathbf{s}_2^*),\ldots,C(\mathbf{s},\mathbf{s}_m))^T$ is the $m\times 1$ vector with $C(\mathbf{s},\mathbf{s}_i)'$ as the i-th element and $\boldsymbol{\phi}_i^* = (\phi_i(\mathbf{s}_1^*),\phi_i(\mathbf{s}_2^*),\ldots,\phi_i(\mathbf{s}_m^*))^T$. Furthermore, substituting $\mathbf{s}_i^*$ for $\mathbf{s}$ in (12.14) leads to the following $m\times m$ non-singular system:

$$\mathbf{c}^T(\mathbf{s}_i^*)\boldsymbol{\phi}_i^* = m\lambda_i\phi_i(\mathbf{s}_i^*) \ , \quad i = 1,2,\ldots,m \Rightarrow \mathbf{C}^*\boldsymbol{\phi}_i^* = m\lambda_i\boldsymbol{\phi}_i^* \ , \tag{12.15}$$

where $\mathbf{C}^*$ is the $m\times m$ matrix with $C(\mathbf{s}_i^*,\mathbf{s}_j^*)$ as its (i,j)-th element. This implies that $(m\lambda_i, \boldsymbol{\phi}_i^*)$'s are eigen-pairs for the full-rank $m\times m$ matrix $\mathbf{C}^*$.

Using (12.14) we write $\sqrt{\lambda_i}\phi_i(\mathbf{s}) = \dfrac{1}{m\sqrt{\lambda_i}}\mathbf{c}(\mathbf{s})'\boldsymbol{\phi}_i^*$ and using (12.15) we we can write $\boldsymbol{\phi}_i^* = m\lambda_i\mathbf{C}^{*-1}\boldsymbol{\phi}_i^*$. Substituting these in (12.13), we can write

$$\begin{aligned}
\tilde{w}(\mathbf{s}) &= \sum_{i=1}^{m}\sqrt{\lambda_i}\phi_i(\mathbf{s})Z_i = \sum_{i=1}^{m}\frac{1}{m\sqrt{\lambda_i}}\mathbf{c}(\mathbf{s})'\boldsymbol{\phi}_i^*Z_i \\
&= = \sum_{i=1}^{m}\frac{1}{m\sqrt{\lambda_i}}(m\lambda_i)\mathbf{c}^T(\mathbf{s})\mathbf{C}^{*-1}\boldsymbol{\phi}_i^*Z_i\sum_{i=1}^{m}\sqrt{\lambda_i}\mathbf{c}^T(\mathbf{s})\mathbf{C}^{*-1}\boldsymbol{\phi}_i^*Z_i \\
&= \mathbf{c}^T(\mathbf{s})\mathbf{C}^{*-1}\sum_{i=1}^{m}\sqrt{\lambda_i}\boldsymbol{\phi}_i^*Z_i \approx \mathbf{c}^T(\mathbf{s})\mathbf{C}^{*-1}\mathbf{w}^* \ ,
\end{aligned} \tag{12.16}$$

where $\mathbf{w}^* = (w(\mathbf{s}_1^*), w(\mathbf{s}_2^*), \ldots, w(\mathbf{s}_m^*))^T$ and the last approximation follows from the low-rank Karhunen-Loéve representation $w(\mathbf{s}_i^*) \approx \sum_{i=1}^{m} \sqrt{\lambda_i}\phi(\mathbf{s}_i^*)Z_i$. The final expression in (12.16) is precisely the predictive process of Banerjee et al. (2008); see Section 12.4. In fact, the predictive process offers a "closed-form" expression for (12.13) by circumventing the difficult problem of computing functional eigen-pairs.

As a last comment here, once we start down the path of basis function representations for processes, we can similarly consider basis representations for surfaces. By now, this is a standard literature where we can flexibly represent surfaces using, say, spline bases or wavelet bases. A general version would take the form $g(\mathbf{s}) = \sum_{l=1}^{L} a_l f_l(\mathbf{s})$ where we have chosen L basis functions (typically orthonormal), the f_l's, and we would estimate the coefficients, the a_l's, in order to fit a surface. We would face the challenge of specifying choice of and number of functions. Typically, these choices are made using knots, e.g., we use local cubic functions. More commonly, these functions are over one-dimensional space rather than two-dimensional space. In this setting, if the coefficients are random, the surface is random. If the coefficients are, say, i.i.d. normal, we have an analogue of the kernel-based dimension reduction. Throughout this book we prefer to work with random surfaces captured through realizations of stochastic processes over $\Re^2$, rather than through linear combination of functions. A useful point in this regard is the following. Basis representations of functions provide an explicit function to evaluate at any location of interest while process representations require interpolation to infer about arbitrary locations; a process realization is not a function. However, process realizations are truly nonparametric while basis representations, though sometimes described as "nonparametric" are, in fact, parametric, based upon a finite set of coefficients.

12.4 Predictive process models

12.4.1 The predictive process

Banerjee, Gelfand, Finley and Sang (2008) propose a class of models based upon the idea of a spatial predictive process (motivated from kriging ideas). Consider a set of "knots" $\mathcal{S}^* = \{\mathbf{s}_1^*, \ldots, \mathbf{s}_m^*\}$, which may but need not be a subset of the entire collection of observed locations in $\mathcal{S} = \{\mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_n\}$. All we require is that m be much smaller than n. Assume that $w(\mathbf{s}) \sim GP(0, C(\cdot;\boldsymbol{\theta}))$ and let $\mathbf{w}^*$ be a realization of $w(\mathbf{s})$ over $\mathcal{S}^*$. That is, $\mathbf{w}^*$ is $m \times 1$ with entries $w(\mathbf{s}_i^*)$ and $\mathbf{w}^* \sim MVN(\mathbf{0}, C^*(\boldsymbol{\theta}))$, where $C^*(\boldsymbol{\theta})$ is the associated $m \times m$ covariance matrix with entries $C(\mathbf{s}_i^*, \mathbf{s}_j^*; \boldsymbol{\theta})$].

The spatial interpolant (that leads to "kriging") at a site $\mathbf{s}_0$ is given by

$$\tilde{w}(\mathbf{s}_0) = E[w(\mathbf{s}_0)|\mathbf{w}^*] = \mathbf{c}^T(\mathbf{s}_0;\boldsymbol{\theta})C^{*-1}(\boldsymbol{\theta})\mathbf{w}^* \ , \tag{12.17}$$

where $\mathbf{c}(\mathbf{s}_0;\boldsymbol{\theta})$ is $m \times 1$ with entries $C(\mathbf{s}_0, \mathbf{s}_j^*; \boldsymbol{\theta})$. This single site interpolator, in fact, defines a spatial process $\tilde{w}(\mathbf{s}) \sim GP(0, \tilde{C}(\cdot))$ with covariance function,

$$\tilde{C}(\mathbf{s}, \mathbf{s}'; \boldsymbol{\theta}) = \mathbf{c}^T(\mathbf{s};\boldsymbol{\theta})C^{*-1}(\boldsymbol{\theta})\mathbf{c}(\mathbf{s}', \boldsymbol{\theta}), \tag{12.18}$$

where $\mathbf{c}(\mathbf{s};\boldsymbol{\theta}) = [C(\mathbf{s}, \mathbf{s}_j^*; \boldsymbol{\theta})]_{j=1}^m$. We refer to $\tilde{w}(\mathbf{s})$ as the *predictive process* derived from the *parent process* $w(\mathbf{s})$. The realizations of $\tilde{w}(\mathbf{s})$ are precisely the kriged predictions conditional upon a realization of $w(\mathbf{s})$ over $\mathcal{S}^*$. The process is completely specified given the covariance function of the parent process and $\mathcal{S}^*$. So, to be precise, we should write $\tilde{w}_{\mathcal{S}^*}(\mathbf{s})$, but we suppress this implicit dependence. From (12.18), this process is nonstationary regardless of whether $w(\mathbf{s})$ is. The connection with (12.6) is clear: we take $Z(\mathbf{s})$ to be the parent process and $\mathbf{l}(\mathbf{s}) = \mathbf{c}(\mathbf{s};\boldsymbol{\theta})^T C^{*-1}(\boldsymbol{\theta})$.

Therefore, every spatial process induces a predictive process model (in fact, arbitrarily many of them). The latter models project process realizations of the former to a lower-dimensional subspace, thereby reducing the computational burden. For example, consider the customary spatial regression model

$$Y(\mathbf{s}) = \mathbf{x}^T(\mathbf{s})\boldsymbol{\beta} + w(\mathbf{s}) + \epsilon(\mathbf{s}). \tag{12.19}$$

Replacing $w(\mathbf{s})$ in (12.19) with $\tilde{w}(\mathbf{s})$, we obtain the predictive process model,

$$Y(\mathbf{s}) = \mathbf{x}^T(\mathbf{s})\boldsymbol{\beta} + \tilde{w}(\mathbf{s}) + \epsilon(\mathbf{s}). \tag{12.20}$$

Since $\tilde{w}(\mathbf{s}) = \mathbf{c}^T(\mathbf{s})C^{*-1}(\boldsymbol{\theta})\mathbf{w}^*$, $\tilde{w}(\mathbf{s})$ is a spatially varying linear transformation of $\mathbf{w}^*$. The dimension reduction is seen immediately. In fitting the model in (12.20), the n random effects $\{w(\mathbf{s}_i), i = 1, 2, ..., n\}$ are replaced with only the m random effects in $\mathbf{w}^*$; we can work with an m-dimensional joint distribution involving only $m \times m$ matrices. Evidently, the model in (12.20) is different from that in (12.19). Hence, though we introduce the same set of parameters in both models, they will not be identical in both models.

Knot-based linear combinations such as $\sum_{i=1}^m a_i(\mathbf{s})w(\mathbf{s}_i^*)$ resemble other process approximation approaches. For instance, motivated by an integral representation of (certain) stationary processes as a kernel convolutions of Brownian motion on $\Re^2$, Higdon (2001) proposes a finite approximation to the parent process of the form $\sum_{i=1}^m a_i(\mathbf{s};\boldsymbol{\theta})u_i$ where u_i's are i.i.d. $N(0,1)$ and $a_i(\mathbf{s};\boldsymbol{\theta}) = k(\mathbf{s},\mathbf{s}_i^*;\boldsymbol{\theta})$ with $k(\cdot;\boldsymbol{\theta})$ being a Gaussian *kernel* function. Evidently, Gaussian kernels only capture Gaussian processes with Gaussian covariance functions (see Paciorek and Schervish, 2006). Xia and Gelfand (2005) suggest extensions to capture more general classes of stationary Gaussian processes by aligning kernels with covariance functions. However, the class of stationary Gaussian process models admitting a kernel representation is limited.

In Higdon's (2001) representation, the $k(\mathbf{s},\mathbf{s}_i^*;\boldsymbol{\theta})$ can be spatially varying, as in Higdon et al. (1999). The u_i can be replaced with realizations from a stationary Gaussian process on $\Re^2$, say, $\mathbf{w}^*$. Then, the original realizations are projected onto an m-dimensional subspace generated by the columns of the $n \times m$ matrix K with entries $k(\mathbf{s}_i,\mathbf{s}_j^*;\boldsymbol{\theta})$, where $\tilde{\mathbf{w}} = K\mathbf{w}^*$. Alternatively, one could project as $\tilde{\mathbf{w}} = Z\mathbf{u}$, where $\mathbf{u} \sim N(\mathbf{0}, I)$, and Z is $n \times m$ with i-th row $\mathbf{c}^T(\mathbf{s}_i;\boldsymbol{\theta})C^{*-1/2}(\boldsymbol{\theta})$ to yield the same joint distribution as the predictive process model in (12.20). This approach has been used in "low-rank kriging" methods (Kamman and Wand, 2003). More general low-rank spline models are also discussed in Ruppert et al. (2003, Ch 13) and Lin et al. (2000).

We regard the predictive process as a competing model specification with, computational advantages, but induced by an underlying full rank process. In fact, these models are, in some sense, *optimal* projections as we clarify in the next subsection. Also, $\tilde{w}(\mathbf{s})$ does not arise as a discretization of an integral representation of a process and we only require a valid covariance function to induce it.

Recall the discussion regarding fixed rank kriging (see also Section 3.2 and Cressie and Johannesson, 2007). Again, letting $\mathbf{g}(\mathbf{s})$ be a $k \times 1$ vector of specified basis functions on $\Re^2$, the proposed covariance function is $C(\mathbf{s},\mathbf{s}') = \mathbf{g}(\mathbf{s})^T K\mathbf{g}(\mathbf{s}')$ with K an unknown positive definite $k \times k$ matrix that is estimated from the data using a method of moments approach. Such an approach may be challenging for the hierarchical models we envision here. We will be providing spatial modelling with random effects at the second stage of the specification. We have no "data" to provide an empirical covariance function. In contrast, we focus upon fitting hierarchical models (including, but not limited to, kriging models) to large spatial datasets. Depending upon where the spatial process $w(\mathbf{s})$ arises in the hierarchy it may be completely unobserved, unlike in classical kriging where the process is typically partially observed, whence empirical data-based estimators of the spatial dispersion matrix will be unavailable. Instead of regarding induced covariance structures of the predictive process as

an approximation to an empirical gold standard, we consider them as models that are, in some sense, *optimal* projections as we clarify in the next subsection. Also, $\tilde{w}(\mathbf{s})$ does not arise as a discretization of an integral representation of a process and we only require a valid covariance function to induce it.

Lastly, we can draw a connection to recent work in spatial dynamic factor analysis (see, e.g., Lopes, Salazar and Gamerman, 2006, and references therein), where K is viewed as an $n \times m$ matrix of factor loadings. Neither K nor $\mathbf{w}^*$ are known but replication over time in the form of a dynamic model is introduced to enable the data to separate them and to infer about them. In our case, the entries in K are "known" given the covariance function C.

12.4.2 Properties of the predictive process

The predictive process $\tilde{w}(\mathbf{s})$ is, in some sense, an *optimal* projection process as we clarify below. First, note that $\tilde{w}(\mathbf{s}_0)$ is an orthogonal projection of $w(\mathbf{s}_0)$ on to a particular linear subspace (e.g., Stein, 1999). Let $\mathcal{H}_{m+1}$ be the Hilbert space generated by $w(\mathbf{s}_0)$ and the m random variables in $\mathbf{w}^*$ (with $\mathcal{H}_m$ denoting the space generated by the latter); hence, $\mathcal{H}_{m+1}$ comprises all linear combinations of these $m+1$ zero-centered, finite variance random variables along with their mean square limit points. If we seek the element in $\tilde{w}(\mathbf{s}_0) \in \mathcal{H}_m$ closest to $w(\mathbf{s}_0)$ in terms of the inner product norm induced by $E[w(\mathbf{s})w(\mathbf{s}')]$, we obtain the linear system $E[(w(\mathbf{s}_0) - \tilde{w}(\mathbf{s}_0))w(\mathbf{s}_j^*)] = 0,\ j = 1, \ldots, m$ with the unique solution $\tilde{w}(\mathbf{s}_0) = \mathbf{c}^T(\mathbf{s}_0)C^{*-1}(\boldsymbol{\theta})\mathbf{w}^*$. Being a conditional expectation, it immediately follows that $\tilde{w}(\mathbf{s}_0)$ minimizes $E[w(\mathbf{s}_0) - f(\mathbf{w}^*)|\mathbf{w}^*]$ over all real-valued functions $f(\mathbf{w}^*)$. In this sense, the predictive process is the best approximation for the parent process.

Also, $\tilde{w}(\mathbf{s}_0)$ *deterministically* interpolates $w(\mathbf{s})$ over $\mathcal{S}^*$. Indeed, if $\mathbf{s}_0 = \mathbf{s}_j^* \in \mathcal{S}^*$ we have

$$\tilde{w}(\mathbf{s}_j^*) = \mathbf{c}^T(\mathbf{s}_j^*; \boldsymbol{\theta})C^{*-1}(\boldsymbol{\theta})\mathbf{w}^* = w(\mathbf{s}_j^*) \tag{12.21}$$

since $\mathbf{e}_j^T C^*(\boldsymbol{\theta}) = \mathbf{c}^T(\mathbf{s}_j^*, \boldsymbol{\theta})$, where $\mathbf{e}_j$ denotes the vector with 1 in the j-th position and 0 elsewhere. So, (12.21) shows that $E[\tilde{w}(\mathbf{s}_j^*)|\mathbf{w}^*] = w(\mathbf{s}_j^*)$ and $\mathrm{Var}(\tilde{w}(\mathbf{s}_j^*)|\mathbf{w}^*) = 0$ (a property of "kriging"). At the other extreme, suppose $\mathcal{S}$ and $\mathcal{S}^*$ are disjoint. Then $\mathbf{w}|\mathbf{w}^* \sim MVN(c^T C^{*-1}\mathbf{w}^*, C - c^T C^{*-1} c)$ where c is the $m \times n$ matrix whose columns are the $\mathbf{c}(\mathbf{s}_i)$ and C is the $n \times n$ covariance matrix of $\mathbf{w}$. We can write $\mathbf{w} = \tilde{\mathbf{w}} + (\mathbf{w} - \tilde{\mathbf{w}})$ and the choice of $\mathbf{w}^*$ determines the (conditional) variability in the second term on the right side, i.e., how close $\Sigma_{\mathbf{w}}$ is to $\Sigma_{\tilde{\mathbf{w}}}$. It also reveals that there will be less variability in the predictive process than in the parent process as n variables are determined by $m < n$ random variables.

Kullback-Leibler based justification for $\tilde{w}(\mathbf{s}^*)$ is discussed in Csató (2002) and Seeger et al. (2003). The former offers a general theory for Kullback-Leibler projections and proposes sequential algorithms for computations. As a simpler and more direct argument, let us assume that $\mathcal{S}^*$ and $\mathcal{S}$ are disjoint and let $\mathbf{w}_a = (\mathbf{w}^*, \mathbf{w})^T$ be the $(m+n) \times 1$ vector of realizations over $\mathcal{S}^* \cup \mathcal{S}$. In (12.19), assuming all other model parameters fixed, the posterior distribution for $p(\mathbf{w}_a \mid \mathbf{Y})$ is proportional to $p(\mathbf{w}_a)p(\mathbf{Y}|\mathbf{w})$ since $p(\mathbf{Y} \mid \mathbf{w}_a) = p(\mathbf{Y} \mid \mathbf{w})$. The corresponding posterior in (12.20) replaces $p(\mathbf{Y} \mid \mathbf{w})$ with a density $q(\mathbf{Y} \mid \mathbf{w}^*)$. Letting $\mathcal{Q}$ be the class of all probability densities satisfying $q(\mathbf{Y} \mid \mathbf{w}_a) = q(\mathbf{Y} \mid \mathbf{w}^*)$, suppose we seek the density $q \in \mathcal{Q}$ that minimizes the reverse Kullback-Leibler divergence $KL(q, p) = \int q \log(q/p)$. Banerjee et al. (2008) argue that $KL(q(\mathbf{w}_a \mid \mathbf{Y}), p(\mathbf{w}_a \mid \mathbf{Y}))$ is minimized when $q(\mathbf{Y} \mid \mathbf{w}^*) \propto \exp\left(E_{\mathbf{w}|\mathbf{w}^*}[\log p(\mathbf{Y}|\mathbf{w}_a)]\right)$. Subsequent calculations from standard multivariate normal theory reveal this to be the Gaussian likelihood corresponding to the predictive process model.

Turning briefly to local smoothness of $\tilde{w}(\mathbf{s})$, often useful in modeling spatial gradients, note that $\tilde{w}(\mathbf{s})$ depends upon $\mathbf{s}$ only though $\mathbf{c}(\mathbf{s})$, hence $C(\mathbf{s}, \mathbf{s}')$, which is a *deterministic*

function of $\mathbf{s}$. Thus smoothness of the predictive process amounts to investigating the analyticity of the parent covariance function. In fact, most stationary covariance functions (e.g., the Matérn family) admit infinite (multivariate) Taylor expansions in D outside a neighborhood of $\mathbf{0}$ leading to infinite differentiability in the almost sure sense everywhere in the domain, except possibly for $\mathbf{s} \in \mathcal{S}^*$. Mean square differentiability of the predictive process, defined in terms of Taylor expansions in the L^2 metric, follows from a similar argument. For further details on process smoothness, see Chapter 13.

Turning to the smoothness properties of the process $\tilde{w}(\mathbf{s})$, let $C(\mathbf{s};\boldsymbol{\theta})$ be a stationary covariance function for the parent process $w(\mathbf{s})$ such that $C(\mathbf{s};\boldsymbol{\theta})$ is infinitely differentiable whenever $\|\mathbf{s}\| > 0$. This assumption is commonly satisfied by most covariance functions (e.g. the Matérn function; see Stein, 1999). Then, $\tilde{w}(\mathbf{s})$ inherits its smoothness properties from the elements of $\mathbf{c}^T(\mathbf{s})$, i.e., from the C's. Hence it is infinitely differentiable in the almost sure sense everywhere in the domain, except possibly for $\mathbf{s} \in \mathcal{S}^*$. Mean square differentiability is also immediate as $C(\mathbf{s};\boldsymbol{\theta})$ admits infinite Taylor expansions for every $\mathbf{s} \neq \mathbf{0}$ and since $\mathbf{w}^*$ is Gaussian with a finite variance. Consequently, for, say, the Matérn correlation family, even with $\nu < 1$ (e.g., with the exponential correlation function) so that the parent process $w(\mathbf{s})$ is *not* mean square differentiable, the predictive process still is.

12.4.3 Biases in low-rank models and the bias-adjusted modified predictive process

Irrespective of their precise specifications, low-rank models tend to underestimate uncertainty (since they are driven by a finite number of random variables), hence, overestimate the residual variance. In different words, this arises from systemic over-smoothing or model under-specification by the low-rank model when compared to the parent model. In fact, this becomes especially transparent from writing the parent likelihood and low-rank likelihood as mixed linear models. To elucidate, suppose, without much loss of generality, that $\mathcal{S} \cap \mathcal{S}^* = \mathcal{S}^*$ with the first m locations in $\mathcal{S}$ acting as the knots. Note that the Gaussian likelihood with the parent process in (12.19) can be written as $N(\mathbf{y} \mid X\boldsymbol{\beta} + Z(\boldsymbol{\theta})\mathbf{u}, \tau^2 I)$, where $Z(\boldsymbol{\theta})$ is the $n \times n$ lower-triangular Cholesky square-root of $C(\boldsymbol{\theta})$ and $\mathbf{u} = (u_1, u_2, \ldots, u_n)^T$ is now an $n \times 1$ vector such that $u_i \overset{iid}{\sim} N(0,1)$. Writing $Z(\boldsymbol{\theta}) = [Z_1(\boldsymbol{\theta}) : Z_2(\boldsymbol{\theta})]$, a low-rank model would work with the likelihood $N(\mathbf{y} \mid X\boldsymbol{\beta} + Z_1(\boldsymbol{\theta})\mathbf{u}_1, \tau^2 I)$, where $\mathbf{u}_1$ is an $m \times 1$ vector whose components are independently and identically distributed $N(0,1)$ variables. Dimension reduction occurs because estimating the low-rank likelihood requires $m \times m$ (instead of $n \times n$) matrix decompositions. The parent and low-rank likelihoods can now be written as

$$\begin{aligned} &\textbf{Parent likelihood:} && \mathbf{y} = X\boldsymbol{\beta} + Z_1(\boldsymbol{\theta})\mathbf{u}_1 + Z_2(\boldsymbol{\theta})\mathbf{u}_2 + \boldsymbol{\epsilon}_1 \\ &\textbf{Low rank likelihood:} && \mathbf{y} = X\boldsymbol{\beta} + Z_1(\boldsymbol{\theta})\mathbf{u}_1 + \boldsymbol{\epsilon}_2 \ , \end{aligned}$$

where $\boldsymbol{\epsilon}_i \sim N(0, \tau_i^2 I)$ for $i = 1, 2$. For fixed $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$, the basis functions forming the columns of $\mathbf{Z}_2(\boldsymbol{\theta})$ in the parent likelihood are absorbed into the residual error in the low rank likelihood, leading to an upward bias in the estimate of the nugget. Put another way, being smoother than the parent process, the low rank process tends to have lower variance which, in turn, inflates the residual variability often manifested as an overestimation of τ^2.

Let us be a bit more precise. Let $P_Z = Z(Z^T Z)^{-1} Z^T$ be the orthogonal projection matrix (or "hat" matrix) into the column space of Z. Customary linear model calculations reveal that the magnitude of the residual vector from the parent model is given by $\mathbf{y}^T(I - P_Z)\mathbf{y}$, while that from the low-rank model is given by $\mathbf{y}^T(I - P_{Z_1})\mathbf{y}$. Using the fact that $P_Z = P_{Z_1} + P_{[(I-P_{Z_1})Z_2]}$ (see exercises), we find the excess residual variability in the low-rank likelihood is summarized by

$$(\mathbf{y} - X\boldsymbol{\beta})^T P_{[(I-P_{Z_1})Z_2]}(\mathbf{y} - X\boldsymbol{\beta}) \ .$$

Although this excess residual variability can be quantified as above, it is less clear how the low-rank spatial likelihood could be modified to compensate for this oversmoothing without adding significantly to the computational burden. Matters are complicated by the fact that expressions for the excess variability will involve the unknown process parameters $\boldsymbol{\theta}$, which must be estimated.

In fact, not all low-rank models lead to a straightforward quantification for this bias. For instance, low-rank models based upon kernel convolutions (Higdon, 2002) approximate $w(\mathbf{s})$ with $w_{KC}(\mathbf{s}) = \sum_{j=1}^{n^*} k(\mathbf{s}-\mathbf{s}_j^*, \boldsymbol{\theta}_1)u_j$, where $k(\cdot, \boldsymbol{\theta}_1)$ is some kernel function and $u_j \overset{iid}{\sim} N(0,1)$, assumed to arise from a Brownian motion $U(\mathbf{v})$ on $\Re^2$. So $w(\mathbf{s}) - w_{KC}(\mathbf{s})$ is

$$\int k(\mathbf{s}-\mathbf{v}, \boldsymbol{\theta})dU(\mathbf{v}) - \sum_{j=1}^{n^*} k(\mathbf{s}-\mathbf{s}_j^*, \boldsymbol{\theta})u_j \approx \sum_{j=n^*+1}^{\infty} k(\mathbf{s}-\mathbf{s}_j^*, \boldsymbol{\theta})u_j \ , \tag{12.22}$$

which does not, in general, render a closed form and may be difficult to compute accurately. Furthermore, Gaussian kernels only capture Gaussian processes with Gaussian covariance functions (see Paciorek and Schervish, 2006); beyond this class, there is no theoretical assurance that $\mathrm{var}\{w(\mathbf{s})\} - \mathrm{var}\{w_{KC}(\mathbf{s})\} = C(\mathbf{s}, \mathbf{s}; \boldsymbol{\theta}) - \sum_{j=1}^{n^*} k^2(\mathbf{s}-\mathbf{s}_j^*, \boldsymbol{\theta})$ will be positive. In fact, the predictive process, being a conditional expectation, orthogonally decomposes the spatial variance so that $\mathrm{var}\{w(\mathbf{s}) - \tilde{w}(\mathbf{s})\} = \mathrm{var}\{w(\mathbf{s})\} - \mathrm{var}\{\tilde{w}(\mathbf{s})\}$; such orthogonal decompositions do not arise naturally in kernel convolution approximations.

The predictive process has a distinct advantage here. Being a conditional expectation, it is a projection onto a Hilbert space (e.g., Banerjee et al., 2008) and, unlike other existing low-rank methods, renders a closed form for the residual process arising from subtracting the predictive process from the parent process. In fact, the following inequality holds for any fixed $\mathcal{S}^*$ and for any spatial process $w(\mathbf{s})$:

$$\mathrm{var}\{w(\mathbf{s})\} = \mathrm{var}\{\mathrm{E}[w(\mathbf{s}) \mid \mathbf{w}^*]\} + \mathrm{E}\{\mathrm{var}[w(\mathbf{s}) \mid \mathbf{w}^*]\} \geq \mathrm{var}\{\mathrm{E}[w(\mathbf{s}) \mid \mathbf{w}^*]\} \ , \tag{12.23}$$

which implies that $\mathrm{var}\{w(\mathbf{s})\} \geq \mathrm{var}\{\tilde{w}(\mathbf{s})\}$. If the process is Gaussian then standard multivariate normal calculations yields the following closed form for this difference at any arbitrary location $\mathbf{s}$,

$$\delta^2(\mathbf{s}) = \mathrm{E}\{\mathrm{var}[w(\mathbf{s}) \mid \mathbf{w}^*]\} = C(\mathbf{s}, \mathbf{s}; \boldsymbol{\theta}_1) - \mathbf{c}(\mathbf{s}; \boldsymbol{\theta}_1)'\mathbf{C}^*(\boldsymbol{\theta}_1)^{-1}\mathbf{c}(\mathbf{s}, \boldsymbol{\theta}_1) \ . \tag{12.24}$$

One simple remedy for the bias in the predictive process model (Finley et al., 2009b; Banerjee et al., 2010) is to use the process

$$\tilde{w}_\epsilon(\mathbf{s}) = \tilde{w}(\mathbf{s}) + \tilde{\epsilon}(\mathbf{s}) \ , \tag{12.25}$$

where $\tilde{\epsilon}(\mathbf{s}) \overset{iid}{\sim} N(0, \delta^2(\mathbf{s}))$ and $\tilde{\epsilon}(\mathbf{s})$ is independent of $\tilde{w}(\mathbf{s})$. We call this the *modified predictive process*. Now, the variance of $\tilde{w}_\epsilon(\mathbf{s})$ equals that of the parent process $w(\mathbf{s})$ and the remedy is computationally efficient — adding an independent space-varying nugget does not incur substantial computational expense. To summarize, we do not recommend the use of *just* a reduced/low rank model. To improve performance, it is necessary to approximate the residual process and, in this regard, the predictive process is especially attractive since the residual process is available explicitly

We present a brief simulation example revealing the benefit of the modified predictive process. We generate 2000 locations within a $[0, 100] \times [0, 100]$ square and then generate the dependent variable from model (12.19) with an intercept as the regressor, an exponential covariance function with range parameter $\phi = 0.06$ (i.e., such that the spatial correlation is $\sim$0.05 at 50 distance units), scale $\sigma^2 = 1$ for the spatial process, and with nugget variance $\tau^2 = 1$. We then fit the predictive process and modified predictive process models using a

	μ	σ^2	τ^2	RMSPE
True	1	1	1	
$m = 49$				
PP	1.37 (0.29,2.61)	1.37 (0.65,2.37)	1.18 (1.07,1.23)	1.21
MPP	1.36 (0.51,2.39)	1.04 (0.52,1.92)	0.94 (0.68.1,14)	1.20
$m = 144$				
PP	1.36 (0.52,2.32)	1.39 (0.76,2.44)	1.09 (0.96, 1.24)	1.17
MPP	1.33 (0.50,2.24)	1.14 (0.64,1.78)	0.93 (0.76,1.22)	1.17
$m = 900$				
PP	1.31 (0.23, 2.55)	1.12 (0.85,1.58)	0.99 (0.85,1.16)	1.17
MPP	1.31 (0.23,2.63)	1.04 (0.76,1.49)	0.98 (0.87,1.21)	1.17

Table 12.1 *Parameter estimates for the predictive process (PP) and modified predictive process (MPP) models in the univariate simulation.*

holding out set of randomly selected sites, along with a separate set of regular lattices for the knots ($m = 49$, 144 and 900). Table 12.1 shows the posterior estimates and the square roots of MSPE based on the prediction for the hold-out data set. The overestimation of τ^2 by the unmodified predictive process is apparent and we also see how the modified predictive process is able to adjust for the τ^2. Not surprisingly, the RMSPE is essentially the same under either process model.

The remedy for the bias, as presented above, is effective in most practical settings and also leads to improved predictive performance by compensating for oversmoothing. However, it may be less effective in capturing small-scale spatial variation as it may not account for the spatial dependence in the residual process. Sang et al. (2011) and Sang and Huang (2012) propose to *taper* the bias adjustment process $\tilde{\epsilon}(\mathbf{s})$. This yields $\tilde{w}_{\tilde{\epsilon}_2}(\mathbf{s}) = \tilde{w}(\mathbf{s}) + \tilde{\epsilon}_2(\mathbf{s})$, where $\tilde{\epsilon}_2(\mathbf{s})$ is now a Gaussian process with covariance function

$$C_{\tilde{\epsilon}_2}(\mathbf{s}_1, \mathbf{s}_2, \boldsymbol{\theta}) = C_{\tilde{\epsilon}}(\mathbf{s}_1, \mathbf{s}_2, \boldsymbol{\theta}) C_{\nu}(\mathbf{s}_1, \mathbf{s}_2).$$

While existing remedies only adjust for the variability in the residual process, the tapered approach, on the other hand, accounts for the residual spatial association. Since the residual process mainly captures the small scale dependence and the tapering has little impact on such dependence other than introducing sparsity, the resulting error of the new approximation is expected to be small. Sang and Huang (2012) refer to this approach as *full-scale approximation* because of its capability of providing high quality approximations at both the small and large spatial scales. In this sense, the proposed tapering approach will enrich the approximation to the underlying parent process.

Very recent work of Katzfuss (2013) has extended the tapering ideas in Sang et al. (2012) with regard to predictive process modeling. He specifies spatial error using two terms, combining a low rank component to capture both medium-to-long-range dependence with a tapered residual component to capture local dependence. Nonstationary Matérn covariance functions are adopted to provide increased flexibility.

12.4.4 Selection of knots

For a given set of observations, Finley et al. (2009) proposed a knot selection strategy designed to improve the induced predictive process as an approximation to the parent process. For a selected set of knots, $\tilde{w}(\mathbf{s}) = \mathrm{E}[w(\mathbf{s}) \mid \mathbf{w}^*]$ is considered as an approximation to the parent process. Given $\boldsymbol{\theta}_1$, the associated predictive variance of $w(\mathbf{s})$ conditional on $\mathbf{w}^*$, denoted $V_{\boldsymbol{\theta}_1}(\mathbf{s}, \mathcal{S}^*)$, is

$$\mathrm{var}[w(\mathbf{s}) \mid \mathbf{w}(\cdot), \mathcal{S}^*, \boldsymbol{\theta}_1] = \mathbf{C}(\mathbf{s}, \mathbf{s}; \boldsymbol{\theta}_1) - \mathbf{c}(\mathbf{s}, \boldsymbol{\theta}_1)^T \mathbf{C}^{*-1}(\boldsymbol{\theta}_1) \mathbf{c}(\mathbf{s}, \boldsymbol{\theta}_1) \ , \tag{12.26}$$

which measures how well we approximate $w(\mathbf{s})$ by the predictive process $\tilde{w}(\mathbf{s})$. This measure is in the spirit of work by Zidek and colleagues (Le and Zidek, 1992; Zidek et al., 2000), i.e., measuring knot *value* in terms of conditional variance. There, the best knot selection maximizes conditional variance given the previously selected knots. Here, we measure the effectiveness of a given collection of selected knots through small conditional variance.

In particular, the knot selection criterion is then defined as a function of $V_{\boldsymbol{\theta}_1}(\mathbf{s}, \mathcal{S}^*)$. One commonly used criterion is:

$$V_{\boldsymbol{\theta}_1}(\mathcal{S}^*) = \int_D V_{\boldsymbol{\theta}_1}(\mathbf{s}, \mathcal{S}^*) g(\mathbf{s}) d\mathbf{s} = \int_D \mathrm{var}[w(\mathbf{s}) \mid \mathbf{w}^*, \boldsymbol{\theta}_1] g(\mathbf{s}) d\mathbf{s} \tag{12.27}$$

where $g(\mathbf{s})$, integrable over D, is the weight assigned to location $\mathbf{s}$ (Zidek et al., 2000; Diggle and Lophaven, 2006). Here, we only consider the simple case for which $g(\mathbf{s}) \equiv 1$. $V_{\boldsymbol{\theta}_1}(\mathcal{S}^*)$ can be regarded as a spatially averaged predictive variance. The integral in (12.27) is analytically intractable and discrete approximations such as numerical quadrature or Monte Carlo integration will be required. We use the discrete approximation which computes the spatially averaged prediction variance over all the observed locations,

$$V_{\boldsymbol{\theta}_1}(\mathcal{S}^*) \approx \frac{\sum_{i=1}^n \mathrm{var}[w(\mathbf{s}_i) \mid \mathbf{w}^*, \boldsymbol{\theta}_1]}{n} \tag{12.28}$$

We ultimately reduce the problem of knot performance to the minimization of a design criterion, which is the function $V_{\boldsymbol{\theta}_1}(\mathcal{S}^*)$.

The following facts are easily verified for $V_{\boldsymbol{\theta}_1}(\mathcal{S}^*)$ defined in (12.27):

- $V_{\boldsymbol{\theta}_1}(\{\mathcal{S}^*, \mathbf{s}_0\}) - V_{\boldsymbol{\theta}_1}(\mathcal{S}^*) < 0$ for a new site $\mathbf{s}_0$;
- $V_{\boldsymbol{\theta}_1}(\{\mathcal{S}^*, \mathbf{s}_0\}) - V_{\boldsymbol{\theta}_1}(\mathcal{S}^*) \longrightarrow 0$ when $\|\mathbf{s}_0 - \mathbf{s}_i^*\| \longrightarrow 0$, where $\mathbf{s}_i^*$ is any member of $\mathcal{S}^*$, and
- $V_{\boldsymbol{\theta}_1}(\mathcal{S}) = 0$, where $\mathcal{S} = \{\mathbf{s}_1, \ldots, \mathbf{s}_n\}$ are the original observed locations.

The variance-covariance matrix under the parent process model is $\Sigma_{\mathbf{Y}} = \mathbf{C}(\boldsymbol{\theta}_1) + \tau^2\mathbf{I}$, while that from the corresponding predictive process is given by $\tilde{\Sigma}_{\mathbf{Y}} = \mathcal{C}(\boldsymbol{\theta}_1)^T \mathbf{C}^{*-1}(\boldsymbol{\theta}_1)\mathcal{C}(\boldsymbol{\theta}_1) + \tau^2\mathbf{I}$. The Frobenius norm between $\Sigma_{\mathbf{Y}}$ and $\tilde{\Sigma}_{\mathbf{Y}}$ is $\|\Sigma_{\mathbf{Y}} - \tilde{\Sigma}_{\mathbf{Y}}\|_F \equiv \mathrm{tr}\left([\mathbf{C}(\boldsymbol{\theta}_1) - \mathcal{C}(\boldsymbol{\theta}_1)^T \mathbf{C}^{*-1}(\boldsymbol{\theta}_1)\mathcal{C}(\boldsymbol{\theta}_1)]^2\right)$. Since $\mathbf{C}(\boldsymbol{\theta}_1) - \mathcal{C}(\boldsymbol{\theta}_1)^T \mathbf{C}^{*-1}(\boldsymbol{\theta}_1)\mathcal{C}(\boldsymbol{\theta}_1)$ is positive definite, the norm $\|\Sigma_{\mathbf{Y}} - \tilde{\Sigma}_{\mathbf{Y}}\|_F \equiv \sum \lambda_i^2$, where λ_i is the i-th eigenvalue of $\Sigma_{\mathbf{Y}} - \tilde{\Sigma}_{\mathbf{Y}}$. Also, the averaged predictive variance is given by $\bar{V} = \dfrac{1}{n}\mathrm{tr}(\Sigma_{\mathbf{Y}} - \tilde{\Sigma}_{\mathbf{Y}}) = \dfrac{1}{n}\sum \lambda_i$.

Note that, even after discretization, we cannot evaluate $V_{\boldsymbol{\theta}_1}(\mathcal{S}^*)$ since it depends upon the unknown $\boldsymbol{\theta}_1$. Available options to accommodate this include obtaining parameter estimates by using a subset of the original data or more fully Bayesian strategies which place a prior on $\boldsymbol{\theta}_1$ and then minimizes $E_{\boldsymbol{\theta}_1}(V_{\boldsymbol{\theta}_1}(\mathcal{S}^*))$ (see, Diggle and Lophaven, 2006). In fact, we might naturally use the same prior as we would use to fit the model.

Regardless of which of these strategies we adopt, how shall we proceed to find a good $\mathcal{S}^*$? Suppose the values of the parameters and the knot size m are given. The following sequential search algorithm finds an approximately optimal design:

- Initialization: As in all cases where the domain is continuous, for implementation of an optimal design, we need to reduce the possible sampling locations to a finite set. Natural choices include a fine grid set, the observed set of locations or the union of these two sets.
- Specify an initial set of locations of size $m_0 << m$ as starting points for knot selection; possible choices include a coarse grid, or a subset of the observed locations, chosen randomly or deterministically.

- At step $t+1$,
  - For each sample point $\mathbf{s}_i$ in the allowable sample set, evaluate $V(\{\mathcal{S}^{*(t)}, \mathbf{s}_i\})$.
  - Remove the sample point with maximum decrease in $\bar{V}$ from the allowable sample set and add it to the knot set.
- Repeat the above procedure until we obtain m knots.

The sequential evaluation of $\bar{V}$ is achieved using a very efficient routine incorporating block matrix computation. Utilizing this, we have successfully implemented the sequential algorithm in a simulation study presented in Section 12.4.5.

We remark that the sequential algorithm does not necessarily achieve the global optimization solution. Alternative computational approaches are available to us in finding approximately optimal designs such as stochastic search and block selection (see Xia et al., 2006). As to the choice of m, the obvious answer is "as large as possible." Evidently, this is governed by computational cost and sensitivity to choice. So, for the former, we will have to implement the analysis over different choices of m to consider run time. For the latter, we look for stability of predictive inference as m increases. We measure this by the value of minimized $\bar{V}$ under different choices of m. Unlike more formal sampling design contexts, our goal here is to achieve "good" knot selection to enable model fitting. We find coarse progression in m to be adequate. Finally, we can implement the analysis in two steps by combining this knot selection procedure with the modified predictive process in the obvious way: (1) choose a set of knots to minimize the averaged predictive variances; (2) then use the modified process in the model fitting. For further discussion of this design problem, see Gelfand et al. (2012). Here we present some examples.

12.4.5 A simulation example using the two step analysis

We generated 1,000 data points in a $[0, 100] \times [0, 100]$ square and then generated the dependent variable from model (12.19) with an intercept $\mu = 1$ as a regressor, an exponential covariance function with range parameter $\phi = 0.06$ (i.e., an effective range of $\sim$50 units), scale $\sigma = 1$ for the spatial process, and with nugget variance $\tau^2 = 1$. We illustrate a comparison among three design strategies, including regular grids, sequential search over all the observed locations and sequential search over a fine regular lattice. In Figure 12.1, we plot the averaged predictive variances under each strategy. The sequential search algorithm is clearly better than choosing a regular grid as knots. For instance, with 180 sites selected, sequential search over the observed locations yielded an averaged predictive variance approximately 0.15. For the regular grids, roughly 150 *additional* sites are needed to achieve the same level of performance.

12.4.6 Non-Gaussian first stage models

There are two typical non-Gaussian first stage settings: (i) binary response at locations modeled using logit or probit regression and (ii) count data at locations modeled using Poisson regression. Diggle, Tawn, and Moyeed (1998) unify the use of generalized linear models in spatial data contexts. See also Section 6.2 as well as Lin et al. (2000) and Kammann and Wand (2003). Essentially we replace (12.19) with the assumption that $E[Y(\mathbf{s})]$ is linear on a transformed scale, i.e., $\eta(\mathbf{s}) \equiv g(E(Y(\mathbf{s}))) = \mathbf{x}^T(\mathbf{s})\boldsymbol{\beta} + w(\mathbf{s})$ where $g(\cdot)$ is a suitable link function. With the Gaussian first stage, we can marginalize over the w's to achieve the covariance matrix $\tau^2 I + \mathbf{c}^T C^{*-1}\mathbf{c}$. Though this matrix is $n \times n$, using the Sherman-Woodbury result (Harville, 1997; also see Section 12.5), inversion only requires C^{*-1}. With, say, a binary or Poisson first stage, such marginalization is precluded; we have to update the w's in running our Gibbs sampler. Using the predictive process, we only have to update the the $m \times 1$ vector $\mathbf{w}^*$.

Figure 12.1 *Averaged prediction variance (V) versus number of knots (m). Solid dots denote results for regular grids; dash-dot line denotes results for the sequential search over the observed data locations (starting with 49 randomly chosen sites from the observed locations), and solid line denotes results for the sequential search over a* 60×60 *regular grid (starting with a* 7×7 *regular grid).*

A bit more clarification may be useful. As described in the previous paragraph, the resulting model would take the form

$$\prod_i p(Y(\mathbf{s}_i)|\boldsymbol{\beta}, \mathbf{w}^*, \phi)p(\mathbf{w}^*|\sigma^2, \phi)p(\boldsymbol{\beta}, \phi, \sigma^2)$$

Though $\mathbf{w}^*$ is only $m \times 1$, making draws of this vector from its full conditional distribution will require Metropolis-Hastings updates.

12.4.7 Spatiotemporal versions

There are various spatiotemporal contexts in which predictive processes can be introduced to render computation feasible. We illustrate three of them here. First, we generalize (12.19) to

$$Y(\mathbf{s}, t) = \mathbf{x}(\mathbf{s}, t)\boldsymbol{\beta} + w(\mathbf{s}, t) + \epsilon(\mathbf{s}, t) \tag{12.29}$$

for $\mathbf{s} \in D$ and $t \in [0, T]$. In (12.29), the ϵ's are again, pure error terms, the $\mathbf{x}(\mathbf{s}, t)$ are local space-time covariate vectors, and $\boldsymbol{\beta}$ is a coefficient vector, here assumed constant over space and time but can be spatially and/or temporally varying (see Section 3). We have replaced the spatial random effects, $w(\mathbf{s})$, with space-time random effects, $w(\mathbf{s}, t)$ that come from a Gaussian process with covariance function $cov(w(\mathbf{s}, t), w(\mathbf{s}', t')) \equiv C(\mathbf{s}, \mathbf{s}'; t, t')$. There has been recent discussion regarding valid nonseparable space-time covariance functions; see Section 11.3. Now, we assume data $Y(\mathbf{s}_i, t_i), i = 1, 2, ...n$, where n can be very large because there are many distinct locations, times or both. In any event, the predictive process will be defined analogous to that above - $\tilde{w}(\mathbf{s}, t) = c(\mathbf{s}, t)^T C^{*-1}\mathbf{w}^*$ where now $\mathbf{w}^*$ is an $m \times 1$ vector associated with m knots over $D \times [0, T]$ having covariance matrix C^* and $c(\mathbf{s}, t)$ is the vector of covariances of $w(\mathbf{s}, t)$ with the entries in $\mathbf{w}^*$. The spatiotemporal predictive process model, $\tilde{w}(\mathbf{s}, t)$ will enjoy the same properties as $\tilde{w}(s)$. Now, the issue of knot selection arises over $D \times [0, T]$. Knot selection over D follows the ideas above. Knot selection over $[0, T]$ is

much easier due to the ordering in one dimension. For example, with annual data, we might use monthly knots, with monthly data we might use weekly knots.

Next, suppose we discretize time to, say, $t = 1, 2, ...T$. Now, we would write the response as $Y_t(\mathbf{s})$ and the random effects as $w_t(\mathbf{s})$. Dynamic evolution of $w_t(\mathbf{s})$ is natural, leading to a spatial dynamic model as discussed in, e.g., Gelfand, Banerjee, and Gamerman (2005). In one scenario the data may arise as a time series of spatial processes, i.e., there is a conceptual time series at each location $\mathbf{s} \in D$. Alternatively, it may arise as cross-sectional data, i.e., there is a set of locations associated with each time point and these can differ from time point to time point. In the latter case, we can anticipate an explosion of locations as time goes on. Use of predictive process modeling, defined through a dynamic sequence of $\mathbf{w}_t^*$'s sharing the same knots enables us to handle this. A detailed explanation can be found in Finley, Banerjee and Gelfand (2012).

It is also not uncommon to find space-time datasets with a very large number of distinct time points, possibly with different time points observed at different locations (e.g., real estate transactions). Predictive processes can be used to improve the applicability of a class of dynamic space-time models proposed by Gelfand et al. (2005) by alleviating a computational bottleneck without sacrificing model flexibility and with minimal loss of information. Finley et al. (2012) focused on the common setting where space is considered continuous but time is taken to be discrete. Here, data is viewed as arising from a time series of spatial processes. Some examples of data that fit this description include: U.S. Environmental Protection Agency's Air Quality System which reports pollutants' mean, minimum, and maximum at 8- and 24-hour intervals; climate model outputs of weather variables generated on hourly or daily intervals, and remotely sensed landuse/landcover change recorded at annual or decadal time steps.

Finally, predictive processes offer an alternative to the dimension-reduction approach to space-time Kalman filtering presented by Wikle and Cressie (1999). With time discretized, they envision evolution through a discretized integro-differential equation with a spatially structured noise. That is, $w_t(\mathbf{s}) = \int h_{\mathbf{s}}(\mathbf{u}) w_{t-1}(\mathbf{u}) d\mathbf{u} + \eta_t(\mathbf{s})$ with $h_{\mathbf{s}}$ a location interaction function and η a *spatially-colored* error process. $w_t(\mathbf{s})$ is decomposed as $\sum_{k=1}^{K} \phi_k(\mathbf{s}) a_{kt}$ where the $\phi_k(\mathbf{s})$'s are deterministic orthonormal basis functions and the a's are mean 0 time series. Then, each $h_{\mathbf{s}}$ has a basis representation in the ϕ's, i.e., $h_{\mathbf{s}}(\mathbf{u}) = \sum_{l=1}^{\infty} b_l(\mathbf{s}) \phi_l(\mathbf{u})$ where the b's are unknown. A dynamic model for the $k \times 1$ vector a_t driven by a linear transformation of the spatial noise process $\eta(\mathbf{s})$ results. Instead of the above decomposition for $w_t(\mathbf{s})$, we would introduce a predictive process model using $\mathbf{w}_t^*$. We replace the projection onto an arbitrary basis with a projection based upon a desired covariance specification.

12.4.8 Multivariate predictive process models

The predictive process immediately extends to multivariate Gaussian process settings. For a $q \times 1$ multivariate Gaussian parent process, $\mathbf{w}(\mathbf{s})$, the corresponding predictive process is

$$\tilde{\mathbf{w}}(\mathbf{s}) = \mathrm{Cov}(\mathbf{w}(\mathbf{s}), \mathbf{w}^*)\mathrm{Var}^{-1}(\mathbf{w}^*)\mathbf{w}^* = \mathcal{C}^T(\mathbf{s}; \boldsymbol{\theta})\mathcal{C}^{*-1}(\boldsymbol{\theta})\mathbf{w}^* , \tag{12.30}$$

where $\mathcal{C}^T(\mathbf{s}; \boldsymbol{\theta}) = [\Gamma_{\mathbf{w}}(\mathbf{s}, \mathbf{s}_1^*; \boldsymbol{\theta}), \ldots, \Gamma_{\mathbf{w}}(\mathbf{s}, \mathbf{s}_m^*; \boldsymbol{\theta})]$ is $q \times mq$, $\Gamma_{\mathbf{w}}(\mathbf{s}, \mathbf{s}')$ is the *cross-covariance* matrix (see Chapter 9), and $\mathcal{C}^*(\boldsymbol{\theta}) = [\Gamma_{\mathbf{w}}(\mathbf{s}_i^*, \mathbf{s}_j^*; \boldsymbol{\theta})]_{i,j=1}^m$ is the $mq \times mq$ covariance matrix of $\mathbf{w}^* = (\mathbf{w}(\mathbf{s}_1^*)^T, \mathbf{w}(\mathbf{s}_2^*)^T, \ldots, \mathbf{w}(\mathbf{s}_m)^T)^T$. Equation (12.30) shows $\tilde{\mathbf{w}}(\mathbf{s})$ is a zero mean $q \times 1$ predictive process with cross-covariance matrix $\Gamma_{\tilde{\mathbf{w}}}(\mathbf{s}, \mathbf{s}') = \mathcal{C}^T(\mathbf{s}; \boldsymbol{\theta})\mathcal{C}^{*-1}(\boldsymbol{\theta})\mathcal{C}(\mathbf{s}'; \boldsymbol{\theta})$. This is especially important for the applications we consider, where each location $\mathbf{s}$ yields observations on q dependent variables given by a $q \times 1$ vector $\mathbf{Y}(\mathbf{s}) = [Y_l(\mathbf{s})]_{l=1}^q$. For each $Y_l(\mathbf{s})$, we also observe a $p_l \times 1$ vector of regressors $\mathbf{x}_l(\mathbf{s})$. Thus, for each location we have q univariate spatial regression equations which can be combined into the following multivariate regression model:

$$\mathbf{Y}(\mathbf{s}) = X^T(\mathbf{s})\boldsymbol{\beta} + \mathbf{w}(\mathbf{s}) + \boldsymbol{\epsilon}(\mathbf{s}) , \tag{12.31}$$

where $X^T(\mathbf{s})$ is a $q \times p$ matrix ($p = \sum_{l=1}^{q} p_l$) having a block diagonal structure with its l-th diagonal being the $1 \times p_l$ vector $\mathbf{x}_l^T(\mathbf{s})$. Note that $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_p)^T$ is a $p \times 1$ vector of regression coefficients with $\boldsymbol{\beta}_l$ being the $p_l \times 1$ vector of regression coefficients corresponding to $\mathbf{x}_l^T(\mathbf{s})$. Likelihood evaluation from (12.34) involves $nq \times nq$ matrices which can be reduced to $mq \times mq$ matrices by simply replacing $\mathbf{w}(\mathbf{s})$ in (12.34) by $\tilde{\mathbf{w}}(\mathbf{s})$.

Further computational gains in computing $\mathcal{C}^{*-1}(\boldsymbol{\theta})$ can be achieved by adopting the linear model of coregionalization, which specifies $\Gamma_{\mathbf{w}}(\mathbf{s}, \mathbf{s}') = A(\mathbf{s})\mathrm{Diag}[\rho_l(\mathbf{s}, \mathbf{s}'; \boldsymbol{\theta})]_{l=1}^{q} A^T(\mathbf{s}')$, where each $\rho_l(\mathbf{s}, \mathbf{s}'; \boldsymbol{\phi})$ is a univariate correlation function satisfying $\rho_l(\mathbf{s}, \mathbf{s}'; \boldsymbol{\phi}) \to 1$ as $\mathbf{s} \to \mathbf{s}'$. Note that $\Gamma_{\mathbf{w}}(\mathbf{s}, \mathbf{s}) = A(\mathbf{s})A^T(\mathbf{s})$, hence $A(\mathbf{s}) = \Gamma_{\mathbf{w}}^{1/2}(\mathbf{s}, \mathbf{s})$ can be taken as any square-root of $\Gamma_{\mathbf{w}}(\mathbf{s}, \mathbf{s})$. Often we assume $A(\mathbf{s}) = A$ and assign an inverse-Wishart prior on AA^T with A a computationally efficient square-root (e.g., Cholesky or spectral). It now easily follows that $\mathcal{C}^*(\boldsymbol{\theta}) = (I_m \otimes A)\Sigma^*(\boldsymbol{\theta})(I_m \otimes A^T)$, where $\Sigma^*(\boldsymbol{\theta})$ is an $mq \times mq$ matrix partitioned into $q \times q$ blocks, whose (i, j)-th block is the diagonal matrix $\mathrm{Diag}[\rho_l(\mathbf{s}_i^*, \mathbf{s}_j^*; \boldsymbol{\theta})]_{l=1}^{q}$. This yields a sparse structure and can be computed efficiently using specialized sparse matrix algorithms. Alternatively, we can write Σ^* as an orthogonally transformed matrix of $m \times m$ block diagonal matrix, $P^T[\oplus_{l=1}^{q}[\rho_l(\mathbf{s}_i^*, \mathbf{s}_j^*; \theta_l)]_{i,j=1}^{m}]P$, where $\oplus$ is the block diagonal operator and P is a permutation (hence orthogonal) matrix. Since $P^{-1} = P^T$, we need to invert q $m \times m$ symmetric correlation matrices rather than a single $qm \times qm$ matrix. Constructing the $nq \times mq$ matrix $\tilde{\Sigma}(\boldsymbol{\theta}) = [\mathrm{Diag}[\rho_l(\mathbf{s}_i, \mathbf{s}_j^*; \boldsymbol{\theta})]_{l=1}^{q}]_{i,j=1}^{n,m}$, we further have

$$\mathrm{Var}(\tilde{\mathbf{w}}) = \mathcal{C}^T(\boldsymbol{\theta})\mathcal{C}^{*-1}(\boldsymbol{\theta})\mathcal{C}(\boldsymbol{\theta}) = (I_n \otimes A)\tilde{\Sigma}(\boldsymbol{\theta})\Sigma^{*-1}(\boldsymbol{\theta})\tilde{\Sigma}^T(\boldsymbol{\theta})(I_m \otimes A^T), \tag{12.32}$$

where the Kronecker structures and sparse matrices render easier computations. For multivariate spatial processes, Banerjee et al. (2010) show that

$$\begin{aligned} \mathrm{Var}\{\mathbf{w}(\mathbf{s})\} - \mathrm{Var}\{\tilde{\mathbf{w}}(\mathbf{s})\} &= C_{\mathbf{w}}(\mathbf{s}, \mathbf{s}) - \mathcal{C}^T(\mathbf{s}; \boldsymbol{\theta})\Sigma^{*-1}(\boldsymbol{\theta})\mathcal{C}(\mathbf{s}, \boldsymbol{\theta}) \\ &= \mathrm{var}\{\mathbf{w}(\mathbf{s}) \mid \mathbf{w}^*\} \succeq 0, \end{aligned} \tag{12.33}$$

where $\mathrm{Var}\{\cdot\}$ denotes the variance-covariance matrix and $\succeq 0$ indicates non-negative definiteness. Equality holds only when $\mathbf{s} \in \mathcal{S}^*$, whereupon the predictive process coincides with the parent process.

Equation (12.33) is analogous to (12.23). The bias-adjustment for the multivariate predictive process is analogous to the univariate case. More precisely, we introduce $\tilde{\mathbf{w}}_{\tilde{\epsilon}}(\mathbf{s}) = \tilde{\mathbf{w}}(\mathbf{s}) + \tilde{\boldsymbol{\epsilon}}(\mathbf{s})$, where $\tilde{\boldsymbol{\epsilon}}(\mathbf{s}) \sim N(\mathbf{0}, \Gamma_{\mathbf{w}}(\mathbf{s}, \mathbf{s}) - \mathcal{C}^T(\mathbf{s}, \boldsymbol{\theta})\mathcal{C}^{*-1}(\boldsymbol{\theta})\mathcal{C}(\mathbf{s}, \boldsymbol{\theta}))$. Notice that $\tilde{\boldsymbol{\epsilon}}(\mathbf{s})$ acts as a nonstationary adjustment to the "residual process" (i.e., the difference between the parent process and the low-rank process) eliminating the systematic bias in variance components.

12.5 Modeling with the predictive process

We outline the implementation details for estimating a modified predictive process version of a multivariate spatial regression model

$$\mathbf{Y}(\mathbf{s}) = X^T(\mathbf{s})\boldsymbol{\beta} + \mathbf{w}(\mathbf{s}) + \boldsymbol{\epsilon}(\mathbf{s}), \tag{12.34}$$

where $X^T(\mathbf{s})$ is a $q \times p$ matrix ($p = \sum_{l=1}^{q} p_l$) having a block diagonal structure with its l-th diagonal being the $1 \times p_l$ vector $\mathbf{x}_l^T(\mathbf{s})$. Note that $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_p)^T$ is a $p \times 1$ vector of regression coefficients with $\boldsymbol{\beta}_l$ being the $p_l \times 1$ vector of regression coefficients corresponding to $\mathbf{x}_l^T(\mathbf{s})$. Likelihood evaluation from (12.34) that involves $nq \times nq$ matrices can be reduced to $mq \times mq$ matrices by simply replacing $\mathbf{w}(\mathbf{s})$ in (12.34) by $\tilde{\mathbf{w}}(\mathbf{s})$.

The modified predictive process model derived from (12.34) has the data likelihood

$$\mathbf{Y} = X\boldsymbol{\beta} + \mathcal{C}^T(\boldsymbol{\theta})\mathcal{C}^{*-1}(\boldsymbol{\theta})\mathbf{w}^* + \tilde{\boldsymbol{\epsilon}} + \boldsymbol{\epsilon};\ \tilde{\boldsymbol{\epsilon}} \sim N(\mathbf{0}, \Sigma_{\tilde{\epsilon}}),\ \boldsymbol{\epsilon} \sim N(\mathbf{0}, I_q \otimes \Psi), \tag{12.35}$$

where $\mathbf{Y} = [\mathbf{Y}(\mathbf{s}_i)]_{i=1}^n$ is the $nq \times 1$ response vector, $X = [X^T(\mathbf{s}_i)]_{i=1}^n$ is the $nq \times p$ matrix of regressors, $\boldsymbol{\beta}$ is the $p \times 1$ vector of regression coefficients and $\mathcal{C}^T(\boldsymbol{\theta}) = [\Gamma_{\mathbf{w}}(\mathbf{s}_i, \mathbf{s}_j^*; \boldsymbol{\theta})]_{i,j=1}^{n,m}$ is $nq \times mq$. In addition, $\Sigma_{\tilde{\epsilon}} = \mathrm{Diag}[\Gamma_{\mathbf{w}}(\mathbf{s}_i, \mathbf{s}_i) - \mathcal{C}^T(\mathbf{s}_i, \boldsymbol{\theta})\mathcal{C}^{*-1}\mathcal{C}(\mathbf{s}_i, \boldsymbol{\theta})]_{i=1}^n$.

Given priors, model fitting employs a Gibbs sampler with Metropolis-Hastings steps using the marginalized likelihood

$$\mathbf{Y} \mid \boldsymbol{\beta}, \boldsymbol{\theta} \sim N(X\boldsymbol{\beta}, \mathcal{C}^T(\boldsymbol{\theta})\mathcal{C}^{*-1}(\boldsymbol{\theta})\mathcal{C}(\boldsymbol{\theta}) + \Sigma_{\tilde{\epsilon}+\epsilon}(\boldsymbol{\theta})) \ ,$$

where $\Sigma_{\tilde{\epsilon}+\epsilon}(\boldsymbol{\theta}) = \mathrm{Diag}[\Psi + \Gamma_{\mathbf{w}}(\mathbf{s}_i, \mathbf{s}_i) - \mathcal{C}^T(\mathbf{s}_i, \boldsymbol{\theta})\mathcal{C}^{*-1}\mathcal{C}(\mathbf{s}_i, \boldsymbol{\theta})]_{i=1}^n$. Computing this marginalized likelihood now requires the inverse and determinant of $\mathcal{C}^T(\boldsymbol{\theta})\mathcal{C}^{*-1}(\boldsymbol{\theta})\mathcal{C}(\boldsymbol{\theta}) + \Sigma_{\tilde{\epsilon}+\epsilon}(\boldsymbol{\theta})$. The inverse is computed using the Sherman-Woodbury-Morrison formula,

$$\Sigma_{\tilde{\epsilon}+\epsilon}^{-1}(\boldsymbol{\theta}) - \Sigma_{\tilde{\epsilon}+\epsilon}^{-1}(\boldsymbol{\theta})\mathcal{C}^T(\boldsymbol{\theta}) \left[\mathcal{C}^*(\boldsymbol{\theta}) + \mathcal{C}(\boldsymbol{\theta})\Sigma_{\tilde{\epsilon}+\epsilon}^{-1}(\boldsymbol{\theta})\mathcal{C}^T(\boldsymbol{\theta})\right]^{-1} \mathcal{C}(\boldsymbol{\theta})\Sigma_{\tilde{\epsilon}+\epsilon}^{-1}(\boldsymbol{\theta}) \ , \tag{12.36}$$

requiring $mq \times mq$ inversions instead of $nq \times nq$ inversions, while the determinant is computed as

$$|\Sigma_{\tilde{\epsilon}+\epsilon}(\boldsymbol{\theta})||\mathcal{C}^*(\boldsymbol{\theta}) + \mathcal{C}(\boldsymbol{\theta})\Sigma_{\tilde{\epsilon}+\epsilon}(\boldsymbol{\theta})\mathcal{C}^T(\boldsymbol{\theta})|/|\mathcal{C}^*(\boldsymbol{\theta})| \ . \tag{12.37}$$

In particular, with coregionalized models, $\mathcal{C}^T(\boldsymbol{\theta})\mathcal{C}^{*-1}(\boldsymbol{\theta})\mathcal{C}(\boldsymbol{\theta})$ can be expressed as in (12.32), while $\Sigma_{\tilde{\epsilon}}(\boldsymbol{\theta})$ is given by

$$\mathrm{Diag}[AA^T - (\mathbf{1}_m^T \otimes A)[\oplus_{j=1}^m \Gamma(\mathbf{s}_i, \mathbf{s}_j^*; \boldsymbol{\theta})]\Sigma^{*-1}(\boldsymbol{\theta})[\oplus_{j=1}^m \Gamma^T(\mathbf{s}_i, \mathbf{s}_j^*; \boldsymbol{\theta})](\mathbf{1}_m \otimes A^T)].$$

To complete the hierarchical specifications, customarily we assign a flat prior for $\boldsymbol{\beta}$, while Ψ could be assigned an inverse-Wishart prior. More commonly, independence of pure error for the different responses at each site is adopted, yielding a diagonal $\Psi = \mathrm{Diag}(\tau_i^2)_{i=1}^q$ with $\tau_i^2 \sim IG(a_i, b_i)$. Also, we model AA^T with an inverse Wishart prior. The priors for $\boldsymbol{\theta}$ will be assigned using customary specifications. Recall that, with the Matérn, the spatial decay parameters are generally weakly identifiable so reasonably informative priors are needed for satisfactory MCMC behavior. Priors for the decay parameters are set relative to the size of $\mathcal{D}$, e.g., prior means that imply the spatial ranges to be a chosen fraction of the maximum distance. The smoothness parameter ν is typically assigned a prior support of $(0, 2)$ as the data can rarely inform about smoothness of higher orders.

We obtain L samples, say $\{\Omega^{(l)}\}_{l=1}^L$, from

$$p(\Omega \mid Data) \propto p(\boldsymbol{\beta})p(A)p(\boldsymbol{\theta})p(\mathbf{Y} \mid \boldsymbol{\beta}, A, \boldsymbol{\theta}, \Psi) \ , \tag{12.38}$$

where $\Omega = \{\boldsymbol{\beta}, A, \boldsymbol{\theta}, \Psi\}$. In each iteration, we update $\boldsymbol{\beta}$ by drawing from a $N(B\mathbf{b}, B)$ distribution, where $B = [\Sigma_{\boldsymbol{\beta}}^{-1} + (X^T\mathcal{C}^T(\boldsymbol{\theta})\mathcal{C}^{*-1}(\boldsymbol{\theta})\mathcal{C}(\boldsymbol{\theta}) + \Sigma_{\tilde{\epsilon}+\epsilon})^{-1}X]^{-1}$ and $\mathbf{b} = X^T(\mathcal{C}^T(\boldsymbol{\theta})\mathcal{C}^{*-1}(\boldsymbol{\theta})\mathcal{C}(\boldsymbol{\theta}) + \Sigma_{\tilde{\epsilon}+\epsilon})^{-1}\mathbf{Y}$. The remaining parameters are updated using Metropolis steps, possibly with block-updates (e.g., all the parameters in Ψ in one block and those in A in another). Typically, random walk Metropolis with (multivariate) normal proposals is adopted; since all parameters with positive support are converted to their logarithms, some Jacobian computation is needed. For instance, while we assign an inverted Wishart prior to AA^T, in the Metropolis update we update A, which requires transforming the prior by the Jacobian $2^k \prod_{i=1}^k a_{ii}^{k-i+1}$. Uniform priors on the spatial decay parameters will require a Hastings step due to the asymmetry in the priors.

Once the posterior samples from $P(\Omega \mid Data)$, $\{\Omega^{(l)}\}_{l=1}^L$, have been obtained, posterior samples from $P(\mathbf{w}^* \mid Data)$ are drawn by sampling $\mathbf{w}^{*(l)}$ for each $\Omega^{(l)}$ from $P(\mathbf{w}^* \mid \Omega^{(l)}, Data)$. This composition sampling is routine because $P(\mathbf{w}^* \mid \Omega, Data)$ is Gaussian (follows from (12.35) with mean $B\mathbf{b}$ and covariance matrix B, where

$$\begin{aligned} \mathbf{b} &= \mathcal{C}^{*-1}(\boldsymbol{\theta})\mathcal{C}(\boldsymbol{\theta})\Sigma_{\tilde{\epsilon}+\epsilon}^{-1}(\mathbf{Y} - X\boldsymbol{\beta}) \text{ and} \\ B &= (\mathcal{C}^{*-1}(\boldsymbol{\theta}) + \mathcal{C}^{*-1}(\boldsymbol{\theta})\mathcal{C}(\boldsymbol{\theta})\Sigma_{\tilde{\epsilon}+\epsilon}^{-1}\mathcal{C}^T(\boldsymbol{\theta})\mathcal{C}^{*-1}(\boldsymbol{\theta}))^{-1} \text{ respectively.} \end{aligned}$$

In some instances (e.g., prediction) it may be desirable to recover $\tilde{\boldsymbol{\epsilon}}$, in which case we again use composition sampling to draw $\tilde{\boldsymbol{\epsilon}}^{(l)}$ from the distribution $N(B\mathbf{b}, B)$, where $\mathbf{b} = (I_n \otimes \Psi^{-1})(\mathbf{Y} - X\boldsymbol{\beta} - \mathcal{C}^T(\boldsymbol{\theta})\mathcal{C}^{*-1}(\boldsymbol{\theta})\mathbf{w}^*)$ and $B = (\Sigma_{\tilde{\epsilon}}^{-1} + (I_n \otimes \Psi^{-1}))^{-1}$. Once $\mathbf{w}^*$ and $\tilde{\boldsymbol{\epsilon}}$ are recovered, prediction is carried out by drawing $\mathbf{Y}^{(l)}(\mathbf{s}_0)$, for each $l = 1, \ldots, L$ from a $q \times 1$ multivariate normal distribution with mean $X^T(\mathbf{s}_0)\boldsymbol{\beta}^{(l)} + \mathcal{C}^T(\boldsymbol{\theta}^{(l)})\mathcal{C}^{*-1}(\boldsymbol{\theta}^{(l)})\mathbf{w}^{*(l)} + \tilde{\boldsymbol{\epsilon}}^{(l)}$ and variance $\Psi^{(l)}$.

Example 12.1 Forest biomass prediction and mapping. Spatial modelling of forest biomass and other variables related to measurements of current carbon stocks and flux have recently attracted much attention for quantifying the current and future ecological and economic viability of forest landscapes. Interest often lies in detecting how biomass changes across the landscape (as a continuous surface) by forest tree species. We consider point-referenced biomass (log-transformed) data observed at 437 forest inventory plots across the USDA Forest Service Bartlett Experimental Forest (BEF) in Bartlett, New Hampshire. Each location yields measurements of metric tons of above-ground biomass per hectare for American beech (BE), eastern hemlock (EH), red maple (RM), sugar maple (SM), and yellow birch (YB) and five covariates: TC1, TC2, and TC3 tasseled cap components derived from a spring date of mid-resolution Landsat 7 ETM+ satellite imagery from the National Land Cover Database and elevation (ELEV) and slope (SLOPE) derived from a digital elevation model data (see `http://seamless.usgs.gov` for metadata).

Figure 12.2 offers interpolated surfaces of the response variables. Covariates were measured on a 30×30 m pixel grid and are available for every location across the BEF. Interest lies in producing pixel-level prediction of biomass by species across large geographic areas. Because data layers such as these serve as input variables to subsequent forest carbon estimation models, it is crucial that each layer also provides a pixel-level measure of uncertainty in prediction.

Figure 12.2 *Interpolation surfaces of log-transformed metric tons of biomass per hectare by species measured on forest inventory plots across the BEF. Response variables ordered BE, EH top row and RM, SM, YB bottom row. The set of 437 forest inventory plots is represented as points in the top left panel.*

Obtaining predictions from a predictive process model could substantially reduce the time necessary to estimate the posterior predictive distributions over a large array of pixels. A similar analysis was conducted by Finley et al. (2008); however, due to computational limitations they were only able to fit models using half of the available data and pixel-level prediction was still infeasible.

Here we considered sub-models of (12.34) including the non-spatial and spatial non-separable models with the modified predictive process and three knot intensities of 51, 126, and 206. For all models Ψ and Γ_w are considered full $q \times q$ cross-covariance matrices where $q = 5$. Predictive process knots were located on a uniform grid within the BEF. We judge the performance of these models based on prediction of a holdout set of 37 inventory plots, and visual similarity between the predicted and observed response surfaces.

We assigned a flat prior to each of the 30 β parameters (i.e., $p = \sum_{l=1}^{5} p_l = 30$ with each p_l including an intercept, TC1, TC2, TC3, ELEV, and SLOPE) . The cross-covariance matrices Ψ and Γ_w each receive an inverse-Wishart, $IW(df, S)$, with the degrees of freedom set to $q + 1 = 6$. Again, diagonal elements in the IW hyperprior scale matrix for Ψ and Γ_w were taken from univariate semi-variograms fit to the residuals of the non-spatial multivariate model. The decay parameter ϕ in the Matérn correlation function spatial follows a $U(0.002, 0.06)$ which corresponds to an effective spatial range between 50 and 1,500 m. Again, the smoothness parameter, ν, was fixed at 0.5, which reduces to the exponential correlation function. For each model, we ran three initially over-dispersed chains for $35,000$ iterations. Unlike in the simulation analysis, substantial effort was required to select tuning values that achieved acceptable Metropolis acceptance rates. Ultimately, we resorted to univariate updates of elements in $\Psi^{1/2}$ and $\Gamma_w^{1/2}$ to gain the control necessary to maintain an acceptance of approximately 20%. Convergence diagnostics revealed 5,000 iterations to be sufficient for initial burn-in so the remaining 30,000 samples from each chain were used for posterior inference. The 206 knot model required approximately 2 hours to complete the MCMC sampling with the 106 and 51 knot models requiring substantially less time to collect the specified number of samples.

For the three knot intensities, there was negligible difference among the $\boldsymbol{\beta}$ parameter estimates. The estimated diagonal elements of Ψ and Γ_w for the three models were also nearly identical. Further, all of the 95% credible intervals for the off-diagonal elements in Ψ and Γ_w overlapped between the 126 and 206 knot models; however, the 206 knot model had several more significant off-diagonal elements (i.e., indicated by a credible interval that does not include zero). For the 51 knot model, off-diagonal elements of Γ_w were general closer to zero and the corresponding elements in Ψ were significantly different than zero, suggesting that the coarseness of this knot grid could not capture the residual spatial process.

Table 12.2 presents the parameter estimates of Γ_w, Ψ, and $\boldsymbol{\phi}$ for the 126 knot model. For brevity we have omitted $\boldsymbol{\beta}$ estimates but note that 15 were significant at the 0.05 level. Those significant off-diagonal elements of Γ_w in Table 12.2 are apparent in the interpolated surface of $\tilde{\mathbf{w}}$ depicted in Figure 12.3, where positive residual spatial correlation can be seen between BE and YB and between EH and RM.

Turning to prediction, it appears that the covariates and spatial proximity of observed inventory plots explain a significant portion of the variation in the response variables, perhaps leading to overfitting. We note that for our 37 holdout plots the 95% prediction intervals are quite broad yielding a 100% empirical coverage for all three knot intensities. Finally, comparing the surface of pixel-level prediction for $1,000$ randomly selected pixels (Figure 12.4) to the observed (Figure 12.2) we see that the model can capture landscape-level variation in biomass and spatial patterns in biomass by species.

Parameter	50% (2.5%, 97.5%)	Parameter	50% (2.5%, 97.5%)
$\Gamma_{w;1,1}$	1.97 (1.93, 2.02)	$\Psi_{1,1}$	1.95 (1.92, 1.98)
$\Gamma_{w;1,2}$	0.0044 (-0.0029, 0.019)	$\Psi_{1,2}$	-0.01 (-0.031, -0.0002)
$\Gamma_{w;1,3}$	-0.014 (-0.034, -0.004)	$\Psi_{1,3}$	-0.0069 (-0.018, 0.001)
$\Gamma_{w;1,4}$	0.011 (-0.0004, 0.027)	$\Psi_{1,4}$	0.01 (-0.0026, 0.019)
$\Gamma_{w;1,5}$	0.012 (0.0009, 0.018)	$\Psi_{1,5}$	-0.0048 (-0.022, 0.013)
$\Gamma_{w;2,2}$	1.96 (1.89, 2.00)	$\Psi_{2,2}$	1.92 (1.88, 1.97)
$\Gamma_{w;2,3}$	0.017 (0.0043, 0.032)	$\Psi_{2,3}$	0.0081 (-0.0001, 0.015)
$\Gamma_{w;2,4}$	0.0032 (-0.01, 0.013)	$\Psi_{2,4}$	-0.0048 (-0.012, 0.0019)
$\Gamma_{w;2,5}$	0.0031 (-0.0058, 0.041)	$\Psi_{2,5}$	0.011 (0.0042, 0.038)
$\Gamma_{w;3,3}$	1.98 (1.9, 2.01)	$\Psi_{3,3}$	1.97 (1.95, 1.98)
$\Gamma_{w;3,4}$	-0.0058 (-0.015, 0.012)	$\Psi_{3,4}$	-0.013 (-0.045, -0.0002)
$\Gamma_{w;3,5}$	0.016 (-0.0017, 0.029)	$\Psi_{3,5}$	0.0018 (-0.0089, 0.016)
$\Gamma_{w;4,4}$	2.03 (1.99, 2.065)	$\Psi_{4,4}$	1.94 (1.90, 1.98)
$\Gamma_{w;4,5}$	0.0064 (-0.0091, 0.016)	$\Psi_{4,5}$	0.0044 (-0.003, 0.012)
$\Gamma_{w;5,5}$	1.91 (1.84, 2.026)	$\Psi_{5,5}$	1.96 (1.93, 1.98)
ϕ_{w_1}	0.0056 (0.0033, 0.01)	Range_{w_1}	536.75 (296.06, 903.66)
ϕ_{w_2}	0.0048 (0.0037, 0.0144)	Range_{w_2}	624.72 (208.76, 806.32)
ϕ_{w_3}	0.0028 (0.0021, 0.0053)	Range_{w_3}	1085.68 (563.5, 1453.63)
ϕ_{w_4}	0.0051 (0.0035, 0.0085)	Range_{w_4}	586.02 (350.93, 846.24)
ϕ_{w_5}	0.0059 (0.0032, 0.0102)	Range_{w_5}	506.06 (293.25, 934.23)

Table 12.2 *BEF biomass parameter estimates for the 126 knot modified predictive process model. Subscripts 1-6 correspond to BE, EH, RM, SM, and YB species.*

Figure 12.3 *Interpolated surfaces of the 126 knot model's median* $\tilde{\mathbf{w}}$ *at each inventory plot. Top left panel shows forest inventory plots (small points) under the 126 knots (large points). The order of response variables in the subsequent panels corresponds to Figure 12.2.*

12.6 Fitting a predictive process model in `spBayes`

We again make use of the Bartlett Experimental Forest (BEF) dataset. This dataset holds 1991 and 2002 forest inventory data for 437 plots. Variables include species specific basal area and total tree biomass; inventory plot coordinates; slope; elevation; and tasseled cap brightness (TC1), greenness (TC2), and wetness (TC3) components from spring, summer, and fall 2002 Landsat images. Total tree biomass is the sum of bole, branches, and foliage biomass. These quantities are strongly associated, suggesting that we could build a richer model by explicitly accounting for spatial association among the $q = 3$ response variables. Our objective is to obtain an estimate, with an associated measure of uncertainty, of bole,

Figure 12.4 *Interpolated surfaces of the 126 knot model's median predicted response value over a random subset of 1,000 pixels in the BEF. Top left panel shows the subset of prediction pixels (small points) under the 126 knots (large points). The order of response variables in the subsequent panels corresponds to Figure 12.2.*

branches, and foliage biomass as a continuous surface over the domain. We begin by removing non-forest inventory plots, converting biomass measurements from kilograms per hectare to the log of metric tons per hectare, and taking a look at plot locations across the forest (Figure 12.5).

```
> library(spBayes)
> data(BEF.dat)
> BEF.dat <- BEF.dat[BEF.dat$ALLBIO02_KGH>0,]
> bio <- BEF.dat$ALLBIO02_KGH*0.001;
> log.bio <- as.matrix(log(0.001*BEF.dat[,
+                              c("BOLE02_KGH",
+                                "BRANCH02_KGH",
+                                "FOLIAGE02_KGH")]))
> colnames(log.bio) <- c("log.bole.mt", "log.branch.mt",
+                        "log.foliage.mt")
> coords <- as.matrix(BEF.dat[,c("XUTM","YUTM")])
> plot(coords, pch=19, cex=0.5, xlab="Easting (m)",
+      ylab="Northing (m)")
> cov(log.bio)
               log.bole.mt log.branch.mt log.foliage.mt
log.bole.mt     0.12661062     0.1214054     0.09020484
log.branch.mt   0.12140543     0.1501818     0.08910730
log.foliage.mt  0.09020484     0.0891073     0.12114077
```

Given a well distributed sample array, we can either place knots on a grid or use one of several clustering algorithms illustrated in the code block below and plotted in Figure 12.6.

```
> m <- 50
> km.knots <- kmeans(coords, m)$centers
> cl.knots <- clara(coords, m)$medoids
> cd.knots <- cover.design(coords, nd=m)$design
> plot(coords, pch=19, cex=0.5, xlab="Easting (m)",
```

Figure 12.5 *Forest inventory plot locations across the BEF.*

```
+      ylab="Northing (m)")
> points(km.knots, pch=5, cex=1, col="blue")
> points(cl.knots, pch=6, cex=1, col="green")
> points(cd.knots, pch=7, cex=1, col="red")
> legend("bottomright", cex=1, pch=c(19,5,6,7),
+        bty="n", col=c("black","blue","green","red"),
+        legend=c("observations","kmeans","clara",
+                 "cover.design"))
```

Given the knots, we can now fit the predictive process model. In the code block below we call `spMvLM` and choose to use the k-means based knots and the bias-adjusted predictive process (i.e., `modified.pp=TRUE`).

```
> q <- 3
> n.samples <- 20000
> n.ltr <- q*(q+1)/2
> model <- list(log.bio[,"log.bole.mt"] ~ ELEV + SLOPE +
+               SUM_02_TC1 + SUM_02_TC2 + SUM_02_TC3,
+               log.bio[,"log.branch.mt"]~ ELEV + SLOPE +
+               SUM_02_TC1 + SUM_02_TC2 + SUM_02_TC3,
+               log.bio[,"log.foliage.mt"] ~ ELEV + SLOPE +
+               SUM_02_TC1 + SUM_02_TC2 + SUM_02_TC3)
> bef.spMvLM <- spMvLM(model, coords=coords, knots=km.knots,
```

Figure 12.6 *Options for choosing knot locations.*

```
+           data=BEF.dat, starting=list("phi"=rep(3/500,q),
+           "A"=rep(0.05,n.ltr), "Psi"=rep(0.05,q)),
+           tuning=list("phi"=rep(0.3,q), "A"=rep(0.0001,n.ltr),
+           "Psi"=rep(0.001,q)),
+           priors=list("phi.Unif"=list(rep(3/2500,q), rep(3/100,q)),
+           modified.pp=TRUE, "K.IW"=list(q+1, diag(0.1,q)),
+           "Psi.IG"=list(rep(2,q), rep(0.05,q))), cov.model="exponential",
+           n.samples=n.samples, verbose=TRUE, n.report=1000)
```

The above code fits a predictive process version of a multivariate spatial regression model to a dataset with 415 locations and 3 jointly modeled outcomes. We derive the modified predictive process model from a linear model of coregionalization (LMC), with an exponential spatial covariance function for each outcome, using 50 knots. We run 20,000 iterations, recover the posterior samples and obtain summaries for inference.

```
> burn.in <- floor(0.75*n.samples)
> bef.spMvLM <- spRecover(bef.spMvLM, start=burn.in)
> round(summary(mcmc(cbind(bef.spMvLM$p.beta.recover.samples,
+      bef.spMvLM$p.theta.recover.samples)))$quantiles[,c(1,3,5)],3)
```

Next we construct plots of the mean of the fitted values' in the posterior distribution.

```
> fitted <- spPredict(bef.spMvLM, start=burn.in, thin=10,
+           pred.covars=bef.spMvLM$X, pred.coords=bef.spMvLM$coords)
```

Figure 12.7 *Mean of the fitted values' posterior distribution.*

```
> y.hat <- rowMeans(fitted$p.y.predictive.samples)
> bole <- y.hat[seq(1,length(y.hat),q)]
> branch <- y.hat[seq(2,length(y.hat),q)]
> foliage <- y.hat[seq(3,length(y.hat),q)]
> res <- 100
> par(mfrow=c(2,2))
> surf <- mba.surf(cbind(coords,bole), no.X=res, no.Y=res,
+                  extend=FALSE)$xyz.est
> image.plot(surf, main="Bole fitted values")
> points(coords)
> surf <- mba.surf(cbind(coords,branch), no.X=res, no.Y=res,
+                  extend=FALSE)$xyz.est
> image.plot(surf, main="Branch fitted values")
> points(coords)
> surf <- mba.surf(cbind(coords,foliage), no.X=res, no.Y=res,
+                  extend=FALSE)$xyz.est
> image.plot(surf, main="Foliage fitted values")
> points(coords)
```

The resulting images are depicted in Figure 12.7.

Given the samples from the parameters' posterior distribution we can now turn to prediction. Using the `spMvLM` object and predictor variables from *new* locations, the function `spPredict` allows us to sample from the posterior predictive distribution of every pixel across the BEF. We are only interested in predictions within the BEF; however, the predictor variable grid extends well beyond the BEF bounds. Therefore, we would like to *clip* the predictor grid to the BEF bounding polygon. The code block below makes use of the `readShapePoly` function from the `maptools` package and `readGDAL` function from the `rgdal` package to read the bounding polygon and predictor variable grid stack, respectively. We then construct the prediction design matrix for the entire grid extent. Then we extract the coordinates of the BEF bounding polygon vertices and use the `pointsInPoly` `spBayes` function to obtain the desired subset of the prediction design matrix and associated prediction coordinates (i.e., pixel centroids). Finally, the `spPredict` function is called and posterior predictive samples are stored in `bef.bio.pred`.

```
> x.range <- range(coords[,1])
> y.range <- range(coords[,2])
> BEF.shp <- readShapePoly("BEF-data/BEF_bound.shp")
> BEF.poly <- as.matrix(BEF.shp@polygons[[1]]@Polygons[[1]]@coords)
> BEF.grids <- readGDAL("BEF-data/dem_slope_lolosptc_clip_60.img")
```

We now perform posterior predictive computations to predict at 3142 locations using the code block below.

```
> pred.covars <- cbind(BEF.grids[["band1"]],BEF.grids[["band2"]],
+    BEF.grids[["band3"]],BEF.grids[["band4"]], BEF.grids[["band5"]])
> pred.covars <- cbind(rep(1, nrow(pred.covars)), pred.covars)
> pred.coords <- SpatialPoints(BEF.grids)@coords
> pred.covars <- pred.covars[pointsInPoly(BEF.poly, pred.coords),]
> pred.coords <- pred.coords[pointsInPoly(BEF.poly, pred.coords),]
> pred.X <- mkMvX(list(pred.covars, pred.covars, pred.covars))
> bef.bio.pred <- spPredict(bef.spMvLM, start=burn.in, thin=5,
+                           pred.coords=pred.coords, pred.covars=pred.X)
```

With access to each pixel's posterior predictive distribution we can map any summary statistics of interest. In Figure 12.8 we compare the log metric tons of biomass interpolated over the observed plots to that of the pixel-level prediction. To do so, we first extract the posterior predictive means and standard deviations of the three outcomes from the posterior predictive samples stored in `bef.bio.pred`.

```
> y.pred.mu <- apply(bef.bio.pred$p.y.predictive.samples, 1, mean)
> y.pred.sd <- apply(bef.bio.pred$p.y.predictive.samples, 1, sd)
> bole.pred.mu <- y.pred.mu[seq(1,length(y.pred.mu),q)]
> branch.pred.mu <- y.pred.mu[seq(2,length(y.pred.mu),q)]
> foliage.pred.mu <- y.pred.mu[seq(3,length(y.pred.mu),q)]
> bole.pred.sd <- y.pred.sd[seq(1,length(y.pred.sd),q)]
> branch.pred.sd <- y.pred.sd[seq(2,length(y.pred.sd),q)]
> foliage.pred.sd <- y.pred.sd[seq(3,length(y.pred.sd),q)]
```

In the above code, the objects `y.pred.mu` and `y.pred.sd` are created to store all the posterior predictive means and standard devations, respectively, across the locations for the $q = 3$ outcomes. The posterior predictive means and standard deviations for each outcome are then extracted by sampling from `y.pred.mu` and `y.pred.sd` at intervals of length q.

For predictions over a grid of points, which is useful to obtain interpolated surfaces of sufficient resolution, it is convenient to construct a data frame consisting of the coordinates where we predict the outcomes and the posterior predictive means and standard deviations for each aoutcome.

Figure 12.8 *Mean of the posterior predictive distributions.*

```
> bio.pred.grid <- as.data.frame(list(x=pred.coords[,1],
+                 y=pred.coords[,2], bole.mu=bole.pred.mu,
+                 branch.mu=branch.pred.mu, foliage.mu=foliage.pred.mu,
+                 bole.sd=bole.pred.sd, branch.sd=branch.pred.sd,
+                 foliage.sd=foliage.pred.sd))
```

```
> coordinates(bio.pred.grid) <- c("x", "y") # to SpatialPointsDataFrame
> gridded(bio.pred.grid) <- TRUE            # to SpatialGridDataFrame
```

Finally, we plot the images at a specified resolution using the MBA package.

```
> toImage <- function(x){as.image.SpatialGridDataFrame(x)}
> res <- 100
> par(mfrow=c(3,2))
> surf <- mba.surf(cbind(coords,log.bio[,"log.bole.mt"]),
+                        no.X=res, no.Y=res, extend=FALSE)$xyz.est
> z.lim <- range(surf[["z"]], na.rm=TRUE)
> image.plot(surf, xaxs = "r", yaxs = "r", main="Observed bole biomass")
> image.plot(toImage(bio.pred.grid["bole.mu"]), zlim=z.lim, xaxs = "r",
+                    yaxs = "r", main="Predicted bole biomass")
> surf <- mba.surf(cbind(coords,log.bio[,"log.branch.mt"]),
+                  no.X=res, no.Y=res,extend=FALSE)$xyz.est
> z.lim <- range(surf[["z"]], na.rm=TRUE)
> image.plot(surf, xaxs = "r", yaxs = "r", main="Observed branch biomass")
> image.plot(toImage(bio.pred.grid["branch.mu"]), zlim=z.lim,
+            xaxs = "r", yaxs = "r", main="Predicted branch biomass")
> surf <- mba.surf(cbind(coords,log.bio[,"log.foliage.mt"]),
+                  no.X=res, no.Y=res, extend=FALSE)$xyz.est
> z.lim <- range(surf[["z"]], na.rm=TRUE)
> image.plot(surf, xaxs = "r", yaxs = "r", main="Observed foliage biomass")
> image.plot(toImage(bio.pred.grid["foliage.mu"]), zlim=z.lim,
+            xaxs = "r", yaxs = "r", main="Predicted foliage biomass")
```

12.7 Exercises

1. Consider the Sherman-Woodbury-Morrison formula (assuming the inverses exist and the matrices have conformable dimensions):

$$(A + BDC)^{-1} = A^{-1} - A^{-1}BMCA^{-1} \ , \quad \text{where} \quad M = (D^{-1} + CA^{-1}B)^{-1} \ .$$

   Derive this in the following two ways:

   (a) Multiply the right hand side by $(A + BDC)$ and show that it equals to the identity matrix.

   (b) Consider the following block of linear equations:

$$\begin{pmatrix} A & -B \\ C & D^{-1} \end{pmatrix} \begin{pmatrix} X_1 \\ X_2 \end{pmatrix} = \begin{pmatrix} I \\ O \end{pmatrix} \tag{12.39}$$

   Eliminate X_2 from the first block of equations in (12.39) to show that $X_1 = (A + BDC)^{-1}$. Now eliminate X_1 from the second block of equations, solve for X_2 and then substitute this solution in the first system to solve for X_1. Show that this gives $X_1 = A^{-1} - A^{-1}BMCA^{-1}$, where $M = (D^{-1} + CA^{-1}B)^{-1}$. These two different expressions for X_1 establishes the Sherman-Woodbury-Morrison identity.

2. Consider the Bayesian hierarchical model:

$$N(\mathbf{x} \mid \mathbf{0}, D) \times N(\mathbf{y} \mid B\mathbf{x}, A) \ , \tag{12.40}$$

   where $\mathbf{y}$ is $n \times 1$, $\mathbf{x}$ is $p \times 1$, A and D are positive definite matrices of sizes $n \times n$ and $p \times p$ respectively, and B is $n \times p$. By computing the marginal covariance matrix $\text{Var}(\mathbf{y})$ in two ways, show that:

$$(A + BDB')^{-1} = A^{-1} - A^{-1}B(D^{-1} + B'A^{-1}B)^{-1}B'A^{-1} \ .$$

3. Let $Z = [Z_1 : Z_2]$ be a matrix of full column rank and let $P_Z = Z(Z'Z)^{-1}Z'$. Prove that $P_Z = P_{Z_1} + P_W$, where $W = (I - P_{Z_1})Z_2$.
4. Use the `spBayes` package in `R` to fit a predictive process model with 10, 20 and 30 knots to the `ColoradoLMC.dat` dataset available on `www.biostat.umn.edu/~brad/data2.html`. First fit a univariate predictive process model with temperature as the dependent variable and precipitation as the explanatory variable. Next fit a bivariate predictive process model with temperature and precipitation having their means.
5. Write a Variational Bayes algorithm to estimate a full geostatistical (parent) and a predictive process model for the BEF data in `spBayes`.

Chapter 13

Spatial gradients and wombling

13.1 Introduction

Much of this text has focused upon spatial process models that presume, for a region of study $\mathcal{D}$, a collection of random variables $\{Y(\mathbf{s}) : \mathbf{s} \in \mathcal{D}\}$, which can be viewed as a randomly realized surface over the region. In practice, this surface is only observed at a finite set of locations and inferential interest typically resides in estimation of the process parameters as well as in spatial interpolation of the process at the unobserved locations.

Once such an interpolated surface has been obtained, investigation of rapid change on the surface may be of interest. In such contexts, interest often lies in the rate of change of a spatial surface at a given location in a given direction. Such slopes or gradients are of interest in so-called digital terrain models for exploring surface roughness. They would also be of interest in meteorology to recognize temperature or rainfall gradients or in environmental monitoring to understand pollution gradients. Such rates of change may also be sought for unobservable or latent spatial processes. For instance, in understanding real estate markets supplying house prices for single family homes, spatial modeling of residuals provides adjustment to reflect desirability of location, controlling for the characteristics for the home and property. Directional gradients at a given location illuminate potential investment decision-making.

Evidently, inferring about rapid change is often equivalent to investigating smoothness of process realizations. Scale is clearly of crucial importance in capturing roughness. For instance, in terrain modelling, at low resolution, roughness recognizes global features such as hills and valleys. At high resolution, we would be identifying more local features. Indeed, to characterize local rates of change without having to specify a scale we can conceptualize infinitesimal gradients. Under suitable conditions, we can formally define directional derivatives and the existence of directional derivative processes (Banerjee and Gelfand, 2002).

Two points are critical here. First, it is evident that the finitely sampled data cannot visually inform about the smoothness of such realizations. Rather, such smoothness is captured in the specification of the process and hence would be motivated by mechanistic considerations associated with the process yielding the data. Second, the choice of the process covariance function, i.e., the function which provides the covariance between $Y(\mathbf{s})$ and $Y(\mathbf{s}')$ for arbitrary pairs of locations $\mathbf{s}$ and $\mathbf{s}'$ in $\mathcal{D}$ characterizes the smoothness of process realizations. For example, Kent (1989) pursues the notion of almost sure (a.s.) continuity while Stein (1999a) follows the path of mean square continuity (and more generally, mean square differentiability). Banerjee and Gelfand (2002) clarify and extend these ideas in various ways.

Such local assessments of spatial surfaces are not restricted to points, but are often desired for curves and boundaries. For instance, environmental scientists are interested in ascertaining whether natural boundaries (e.g., mountains, forest edges etc.) represent a zone of rapid change in weather; ecologists are interested in determining curves that delineate differing zones of species abundance; while public health officials want to identify change in health care delivery across municipal boundaries, counties or states. The above objectives

require the notion of gradients and, in particular, assigning gradients to curves (*curvilinear gradients*) in order to identify curves that track a path through the region where the surface is rapidly changing (Banerjee and Gelfand, 2006). Such boundaries are commonly referred to as difference boundaries or *wombling boundaries*, named after Womble (1951), who discussed their importance in understanding scientific phenomena (also see Fagan, Fortin and Soykan, 2003).

As a concept, wombling is useful because it attempts to quantify spatial information in objects such as curves and paths which is not easy to model using regressors. It is similar to image analysis in that it also seeks to capture lurking "spatial effects" on curves. However, unlike images, where edges and lines represent discontinuities or breaks, wombling boundaries capture rapid surface change; cutting across a wombling boundary should tend to reveal a steep drop in elevation or, equivalently, a sharp gradient. Evidently, gradients are central to wombling and the spatial surfaces under investigation must be sufficiently smooth. This precludes methods such as wavelets that have been employed in detecting image discontinuities, such as ridges and cliffs (e.g., Csillag and Kabos, 2002), but do not admit gradients.

Visual assessment of the surface over D often proceeds from contour and image plots of the surface fitted from the data using surface interpolators. Surface representation and contouring methods range from tensor-product interpolators for gridded data (e.g., Cohen, Riesenfeld and Elber, 2003) to more elaborate adaptive control-lattice or tessellation based interpolators for scattered data (e.g., Akima, 1996; Lee, Wolberg and Shin, 1997). Mitas and Mitasova (1999) provide a review of several such methods available in GIS software (e.g., GRASS: `http://grass.itc.it/`). These methods are often fast and simple to implement and produce contour maps that reveal topographic features. However, they do not account for association and uncertainty in the data. Contrary to being competitive with statistical methods, they play a complementary role creating descriptive plots from the raw data in the pre-modelling stage and providing visual displays of estimated response or residual surfaces in the post-modelling stage. It is worth pointing out that while contours often provide an idea about the local topography, they are not the same as wombling boundaries. Contour lines connect points with the same spatial elevation and may or may not track large gradients, so they may or may not correspond to wombling boundaries.

Existing wombling methods for point-referenced data concentrate upon finding points that have large gradients and attempt to connect them in an algorithmic fashion, which then defines a boundary. Such algorithms have been employed widely in computational ecology, anthropology and geography. For example, Barbujuani et al. (1990, 1997) used wombling on red blood cell markers to identify genetic boundaries in Eurasian human populations by different processes restricting gene flow; Bocquet-Appel and Bacro (1994) investigated genetic, morphometric and physiologic boundaries; Fortin (1994, 1997) delineated boundaries related to specific vegetation zones; Fortin and Drapeau (1995) applied wombling on real environmental data; and Jacquez and Greiling (2003) estimated boundaries for breast, lung, and colorectal cancer rates in males and females in Nassau, Suffolk, and Queens counties in New York. This last application is somewhat different from the others in that the data were areally referenced with counties. Unlike image pixels, these counties are not regularly spaced but still have a well-defined neighborhood structure (a topological graph) and the image analysis methods can be applied directly. The gradient is not explicitly modelled; boundary effects are looked upon as edge effects and modelled using Markov random field specifications. A Bayesian framework for areal boundary analysis has been provided by Lu and Carlin (2005).

Banerjee and Gelfand (2006) formulated a Bayesian framework for point-referenced curvilinear gradients or boundary analysis, a conceptually harder problem due to the lack of definitive candidate boundaries. Spatial process models help in estimating not only response surfaces, but residual surfaces after covariate and systematic trends have been accounted

for. Depending upon the scientific application, boundary analysis may be desirable on either. Algorithmic methods treat statistical estimates of the surface as "data" and apply interpolation-based wombling to obtain boundaries. Although such methods produce useful descriptive surface plots, they preclude formal statistical inference. Indeed, boundary assessment using such reconstructed surfaces will suffer from inaccurate estimation of uncertainty.

In the next section, we present a formal development of inference for directional finite difference processes and directional derivative processes. We then move from points to curves and assign a meaningful gradient to a curve. For a point, if the gradient in a particular direction is large (positive or negative) then the surface is rapidly increasing or decreasing in that direction. For a curve, if the gradients in the direction orthogonal to the curve tend to be large then the curve tracks a path through the region where the surface is rapidly changing. We obtain complete distribution theory results under the assumptions of a stationary Gaussian process model either for the data or for spatial random effects. We present inference under a Bayesian framework which, in this setting, offers several advantages.

13.2 Process smoothness revisited ⋆

We confine ourselves to smoothness properties of a univariate spatial process, say, $\{Y(\mathbf{s}), \mathbf{s} \in \Re^d\}$; for a discussion of multivariate processes, see Banerjee and Gelfand (2003). In our investigation of smoothness properties we look at two types of continuity, continuity in the L_2 sense and continuity in the sense of process realizations. Unless otherwise noted, we assume the processes to have 0 mean and finite second-order moments.

Definition 13.1 A process $\{Y(\mathbf{s}), \mathbf{s} \in \Re^d\}$ is L_2 continuous at $\mathbf{s}_0$ if and only if $\lim_{\mathbf{s}\to\mathbf{s}_0} E[Y(\mathbf{s}) - Y(\mathbf{s}_0)]^2 = 0$. Continuity in the L_2 sense is also referred to as *mean square continuity*, and will be denoted by $Y(\mathbf{s}) \xrightarrow{L_2} Y(\mathbf{s}_0)$.

Definition 13.2 A process $\{Y(\mathbf{s}), \mathbf{s} \in \Re^d\}$ is *almost surely continuous* at $\mathbf{s}_0$ if $Y(\mathbf{s}) \longrightarrow Y(\mathbf{s}_0)$ *a.s.* as $\mathbf{s} \longrightarrow \mathbf{s}_0$. If the process is almost surely continuous for every $\mathbf{s}_0 \in \Re^d$ then the process is said to have continuous realizations.

In general, one form of continuity does not imply the other since one form of convergence does not imply the other. However, if $Y(\mathbf{s})$ is a bounded process then a.s. continuity implies L_2 continuity. Of course, each implies that $Y(\mathbf{s}) \xrightarrow{P} Y(\mathbf{s}_0)$.

Example 13.1 Almost sure continuity does not imply mean square continuity. To see this, let $t \in [0,1]$ with $\omega \sim U(0,1)$ and define

$$Y(t;\omega) = \begin{cases} \left(t - \frac{1}{2}\right)^{-1} I_{\left(\frac{1}{2},t\right)}(\omega) & \text{if } t \in \left(\frac{1}{2},1\right] \\ 0 & \text{if } t \in \left[0,\frac{1}{2}\right] \end{cases}.$$

Then $Y(t;\omega) \longrightarrow 0$ *a.s.* as $t \longrightarrow \frac{1}{2}$. But $E[Y^2(t;\omega)] \longrightarrow \infty$ as $t \longrightarrow \frac{1}{2}$ if $t \in \left(\frac{1}{2},1\right]$ and $E\left[Y^2(t;\omega)\right] = 0$ if $t \in \left[0,\frac{1}{2}\right]$. Thus the process does not converge in L_2 although it does so almost surely. ■

Example 13.2 Mean square continuity does not imply almost sure continuity. To see this, construct a process over $t \in \Re^+$ defined through $\omega \sim U(0,1)$ as follows. Let $Y\left(\frac{1}{t};\omega\right) = 0$, if t is not a positive integer, $Y(1;\omega) = I_{\left(0,\frac{1}{2}\right)}(\omega)$, $Y\left(\frac{1}{2};\omega\right) = I_{\left(\frac{1}{2},1\right)}(\omega)$, $Y\left(\frac{1}{3};\omega\right) = I_{\left(0,\frac{1}{3}\right)}(\omega)$, $Y\left(\frac{1}{4};\omega\right) = I_{\left(\frac{1}{3},\frac{2}{3}\right)}(\omega)$, $Y\left(\frac{1}{5};\omega\right) = I_{\left(\frac{2}{3},1\right)}(\omega)$, and so on. That is, we construct the process as a sequence of moving indicators on successively finer arithmetic divisions of the unit interval. We see here that $E\left[Y^2\left(\frac{1}{t};\omega\right)\right] \longrightarrow 0$ as $t \longrightarrow 0$, so that $Y\left(\frac{1}{t};\omega\right) \xrightarrow{L_2} 0$. However the process is not continuous almost surely since $Y\left(\frac{1}{t};\omega\right)$ is equal to one infinitely often. ■

The above definitions apply to any stochastic process (possibly nonstationary). Cramér and Leadbetter (1967) and Hoel, Port, and Stone (1972) outline conditions on the covariance function for mean square continuity for processes on the real line. For a process on $\Re^d$, we denote the covariance function, as usual, by $C(\mathbf{s}, \mathbf{s}') = cov(Y(\mathbf{s}), Y(\mathbf{s}'))$, so that the definition of mean square continuity is equivalent to $\lim_{\mathbf{s}' \to \mathbf{s}} [C(\mathbf{s}', \mathbf{s}') - 2C(\mathbf{s}', \mathbf{s}) + C(\mathbf{s}, \mathbf{s})] = 0$. It follows that continuity in $\mathbf{s}$ and $\mathbf{s}'$ serve as sufficient conditions for mean square continuity. For a (weakly) stationary process, mean square continuity is equivalent to the covariance function $C(\mathbf{s})$ being continuous at $\mathbf{0}$. This follows easily since $E[Y(\mathbf{s}') - Y(\mathbf{s})]^2 = 2(C(\mathbf{0}) - C(\mathbf{s}' - \mathbf{s}))$ for a weakly stationary process and enables a simple practical check for mean square continuity.

Kent (1989) investigates continuous process realizations through a Taylor expansion of the covariance function. Let $\{Y(\mathbf{s}), \mathbf{s} \in \Re^d\}$ be a real-valued stationary spatial process on $\Re^d$. Kent proves that if $C(\mathbf{s})$ is d-times continuously differentiable and $C_d(\mathbf{s}) = C(\mathbf{s}) - P_d(\mathbf{s})$, where $P_d(\mathbf{s})$ is the Taylor polynomial of degree d for $C(\mathbf{s})$ about $\mathbf{0}$, satisfies the condition,

$$|C_d(\mathbf{s})| = O\left(||\mathbf{s}||^{d+\beta}\right)$$

for some $\beta > 0$, then there exists a version of the spatial process $\{Y(\mathbf{s}), \mathbf{s} \in \Re^d\}$ with continuous realizations. If $C(\mathbf{s})$ is d-times continuously differentiable then it is of course continuous at $\mathbf{0}$ and so, from the previous paragraph, the process is mean square continuous.

Let us suppose that $f : L_2 \longrightarrow \Re^1$ (L_2 is the usual Hilbert space of random variables induced by the L_2 metric) is a continuous function. Let $\{Y(\mathbf{s}), \mathbf{s} \in \Re^d\}$ be a process that is continuous almost surely. Then the process $Z(\mathbf{s}) = f(Y(\mathbf{s}))$ is almost surely continuous, being the composition of two continuous functions. The validity of this statement is direct and does not require checking Kent's conditions. However, the process $Z(\mathbf{s})$ need not be stationary even if $Y(\mathbf{s})$ is. Moreover, the existence of the covariance function $C(\mathbf{s}, \mathbf{s}') = E[f(Y(\mathbf{s}))f(Y(\mathbf{s}'))]$, via the Cauchy-Schwartz inequality, requires $Ef^2(Y(\mathbf{s})) < \infty$.

While almost sure continuity of the new process $Z(\mathbf{s})$ follows routinely, the mean square continuity of $Z(\mathbf{s})$ is not immediate. However, from the remark below Definition 13.2, if $f : \Re^1 \longrightarrow \Re^1$ is a continuous function that is bounded and $Y(\mathbf{s})$ is a process that is continuous almost surely, then the process $Y(\mathbf{s}) = f(Z(\mathbf{s}))$ (a process on $\Re^d$) is mean square continuous.

More generally suppose f is a continuous function that is Lipschitz of order 1, and $\{Y(\mathbf{s}), \mathbf{s} \in \Re^d\}$ is a process which is mean square continuous. Then the process $Z(\mathbf{s}) = f(Y(\mathbf{s}))$ is mean square continuous. To see this, note that since f is Lipschitz of order 1 we have $|f(Y(\mathbf{s}+\mathbf{h})) - f(Y(\mathbf{s}))| \leq K|Y(\mathbf{s}+\mathbf{h}) - Y(\mathbf{s})|$ for some constant K. It follows that $E[f(Y(\mathbf{s}+\mathbf{h})) - f(Y(\mathbf{s}))]^2 \leq K^2 E[Y(\mathbf{s}+\mathbf{h}) - Y(\mathbf{s})]^2$, and the mean square continuity of $Z(\mathbf{s})$ follows directly from the mean square continuity of $Y(\mathbf{s})$.

We next formalize the notion of a mean square differentiable process. Our definition is motivated by the analogous definition of total differentiability of a function of $\Re^d$ in a nonstochastic setting. In particular, $Y(\mathbf{s})$ is mean square differentiable at $\mathbf{s}_0$ if there exists a vector $\boldsymbol{\nabla}_Y(\mathbf{s}_0)$, such that, for any scalar h and any unit vector $\mathbf{u}$,

$$Y(\mathbf{s}_0 + h\mathbf{u}) = Y(\mathbf{s}_0) + h\mathbf{u}^T \boldsymbol{\nabla}_Y(\mathbf{s}_0) + r(\mathbf{s}_0, h\mathbf{u}) \ , \tag{13.1}$$

where $r(\mathbf{s}_0, h\mathbf{u}) \to 0$ in the L_2 sense as $h \to 0$. That is, we require for any unit vector $\mathbf{u}$,

$$\lim_{h \to 0} E\left(\frac{Y(\mathbf{s}_0 + h\mathbf{u}) - Y(\mathbf{s}_0) - h\mathbf{u}^T \boldsymbol{\nabla}_Y(\mathbf{s}_0)}{h}\right)^2 = 0 \ . \tag{13.2}$$

The first-order linearity condition for the process is required to ensure that mean square differentiable processes are mean square continuous. A counterexample when this condition does not hold is the following.

Example 13.3 Let $Z \sim N(0,1)$ and consider the process $\{Y(\mathbf{s}) : \mathbf{s} = (s_1, s_2) \in \Re^2\}$ defined as follows:

$$Y(\mathbf{s}) = \begin{cases} \frac{s_1 s_2^2}{s_1^2 + s_2^4} Z & \text{if } \mathbf{s} \neq \mathbf{0}, \\ 0 & \text{if } \mathbf{s} = \mathbf{0}. \end{cases}$$

Then, the finite difference process $Y_{\mathbf{u},h}(\mathbf{0})$ is

$$Y_{\mathbf{u},h}(\mathbf{0}) = \frac{Y(\mathbf{0} + h\mathbf{u}) - Y(\mathbf{0})}{h} = \frac{Y(h\mathbf{u})}{h} = \frac{u_1 u_2^2}{u_1^2 + h^2 u_2^4}$$

for any $\mathbf{u} = (u_1, u_2)^T$. Therefore, $D_{\mathbf{u}} Y(\mathbf{0}) = \lim_{h \to 0} Y_{h\mathbf{u}}(\mathbf{0}) = (u_2^2/u_1)Z$ for every $\mathbf{u}$ with $u_1 \neq 0$ and $D_{\mathbf{u}} Y(\mathbf{0}) = 0$ for any direction $\mathbf{u}$ with $u_1 = 0$, where the limits are taken in the L^2 sense. However, the above process is not mean square continuous at $\mathbf{0}$ as can be seen by considering the path $\{(s_2^2, s_2) : s_2 \in \Re\}$, along which $\mathrm{E}\{Y(\mathbf{s}) - Y(\mathbf{0})\}^2 = 1/4$. ■

The above example shows that, even though the directional derivatives exist in all directions at $\mathbf{0}$, the process is not mean-square differentiable because there does not exist the required linear function of $\mathbf{u}$. On the other hand, if $Y(\mathbf{s})$ is a mean square differentiable process on $\Re^d$, then $Y(\mathbf{s})$ is mean square continuous as well. That is, any direction $\mathbf{u}$ can be taken to be of the form $h\mathbf{v}$ where $\mathbf{v}$ is the unit vector giving the direction of $\mathbf{u}$ and h is a scalar denoting the magnitude of $\mathbf{u}$. The point here is the assumed existence of the surface, $\boldsymbol{\nabla}_Y(\mathbf{s})$ under a mean square differentiable process. And, it is important to note that this surface is random; $\boldsymbol{\nabla}_Y(\mathbf{s})$ is not a function.

13.3 Directional finite difference and derivative processes

The focus of this subsection is to formally address the problem of the rate of change of a spatial surface at a given point in a given direction. As noted in the introduction to this chapter, such slopes or gradients are of interest in so-called digital terrain models for exploring surface roughness. They would also arise in meteorology to recognize temperature or rainfall gradients or in environmental monitoring to understand pollution gradients. With spatial computer models, where the process generating the $Y(\mathbf{s})$ is essentially a black box and realizations are costly to obtain, inference regarding local rates of change becomes important. The application we study here considers rates of change for unobservable or latent spatial processes. For instance, in understanding real estate markets, i.e., house prices for single-family homes, spatial modeling of residuals provides adjustment to reflect desirability of location, controlling for the characteristics of the home and property. Suppose we consider the rate of change of the residual surface in a given direction at, say, the central business district. Transportation costs to the central business district vary with direction. Increased costs are expected to reduce the price of housing. Since transportation cost information is not included in the mean, directional gradients to the residual surface can clarify this issue.

Spatial gradients are customarily defined as finite differences (see, e.g., Greenwood, 1984, and Meyer, Ericksson, and Maggio, 2001). Evidently the scale of resolution will affect the nature of the resulting gradient (as we illustrate in Example 13.2). To characterize local rates of change without having to specify a scale, infinitesimal gradients may be considered. Ultimately, the nature of the data collection and the scientific questions of interest would determine preference for an infinitesimal or a finite gradient. For the former, gradients (derivatives) are quantities of basic importance in geometry and physics. Researchers in the physical sciences (e.g., geophysics, meteorology, oceanography) often formulate relationships in terms of gradients. For the latter, differences, viewed as discrete approximations to gradients, may initially seem less attractive. However, in applications involving spatial data, scale is usually a critical question (e.g., in environmental, ecological, or demographic settings). Infinitesimal local rates of change may be of less interest than finite differences at the scale of a map of interpoint distances.

Following the discussion of Section 13.2, with $\mathbf{u}$ a unit vector, let

$$Y_{\mathbf{u},h}(\mathbf{s}) = \frac{Y(\mathbf{s}+h\mathbf{u}) - Y(\mathbf{s})}{h} \tag{13.3}$$

be the finite difference at $\mathbf{s}$ in direction $\mathbf{u}$ at scale h. Clearly, for a fixed $\mathbf{u}$ and h, $Y_{\mathbf{u},h}(\mathbf{s})$ is a well-defined process on $\Re^d$, which we refer to as the finite difference process at scale h in direction $\mathbf{u}$.

Next, let $D_{\mathbf{u}}Y(\mathbf{s}) = \lim_{h\to 0} Y_{\mathbf{u},h}(\mathbf{s})$ if the limit exists. We see that if $Y(\mathbf{s})$ is a mean square differentiable process in $\Re^d$, i.e., (13.2) holds for every $\mathbf{s}_0$ in $\Re^d$, then for each $\mathbf{u}$,

$$\begin{aligned} D_{\mathbf{u}}Y(\mathbf{s}) &= \lim_{h\to 0} \frac{Y(\mathbf{s}+h\mathbf{u}) - Y(\mathbf{s})}{h} \\ &= \lim_{h\to 0} \frac{h\mathbf{u}^T\boldsymbol{\nabla}_Y(\mathbf{s}) + r(\mathbf{s},h\mathbf{u})}{h} = \mathbf{u}^T\boldsymbol{\nabla}_Y(\mathbf{s}) \ . \end{aligned}$$

So $D_{\mathbf{u}}Y(\mathbf{s})$ is a well-defined process on $\Re^d$, which we refer to as the directional derivative process in the direction $\mathbf{u}$.

Note that if the unit vectors $\mathbf{e}_1, \mathbf{e}_2, ..., \mathbf{e}_d$ form an orthonormal basis set for $\Re^d$, any unit vector $\mathbf{u}$ in $\Re^d$ can be written as $\mathbf{u} = \sum_{i=1}^d w_i\mathbf{e}_i$ with $w_i = \mathbf{u}^T\mathbf{e}_i$ and $\sum_{i=1}^d w_i^2 = 1$. But then,

$$D_{\mathbf{u}}Y(\mathbf{s}) = \mathbf{u}^T\boldsymbol{\nabla}_Y(\mathbf{s}) = \sum_{i=1}^d w_i\mathbf{e}_i^T\boldsymbol{\nabla}_Y(\mathbf{s}) = \sum_{i=1}^d w_i D_{\mathbf{e}_i}Y(\mathbf{s}) \ . \tag{13.4}$$

Hence, to study directional derivative processes in arbitrary directions we need only work with a basis set of directional derivative processes. Also from (13.4) it is clear that $D_{-\mathbf{u}}Y(\mathbf{s}) = -D_{\mathbf{u}}Y(\mathbf{s})$. Applying the Cauchy-Schwarz inequality to (13.4), for every unit vector $\mathbf{u}$, $D_{\mathbf{u}}^2Y(\mathbf{s}) \le \sum_{i=1}^d D_{\mathbf{e}_i}^2Y(\mathbf{s})$. Hence, $\sum_{i=1}^d D_{\mathbf{e}_i}^2Y(\mathbf{s})$ is the maximum over all directions of $D_{\mathbf{u}}^2Y(\mathbf{s})$. At location $\mathbf{s}$, this maximum is achieved in the direction $\mathbf{u} = \boldsymbol{\nabla}_Y(\mathbf{s}) / ||\boldsymbol{\nabla}_Y(\mathbf{s})||$, and the maximizing value is $||\boldsymbol{\nabla}_Y(\mathbf{s})||$. In the following text we work with the customary orthonormal basis defined by the coordinate axes so that $\mathbf{e}_i$ is a $d \times 1$ vector with all 0's except for a 1 in the ith row. In fact, with this basis, $\boldsymbol{\nabla}_Y(\mathbf{s}) = (D_{\mathbf{e}_1}Y(\mathbf{s}), ..., D_{\mathbf{e}_d}Y(\mathbf{s}))^T$. The result in (13.4) is a limiting result as $h \to 0$. From (13.3), the presence of h shows that to study finite difference processes at scale h in arbitrary directions we have no reduction to a basis set.

A useful comment is that the directional gradients for $g(Z(\mathbf{s}))$, with g continuous and monotonic increasing, are immediately $D_{\mathbf{u}}g(Z(\mathbf{s})) = g'(Z(\mathbf{s}))D_{\mathbf{u}}Z(\mathbf{s})$, by use of the chain rule. This enables us to study gradient behavior of mean surfaces under link functions.

Formally, finite difference processes require less assumption for their existence. To compute differences we need not worry about a *degree* of smoothness for the realized spatial surface. However, issues of numerical stability can arise if h is too small. Also, with directional derivatives in, say, two-dimensional space, following the discussion below (13.4), we only need work with north and east directional derivatives processes in order to study directional derivatives in arbitrary directions.

13.4 Distribution theory for finite differences and directional gradients

If $E(Y(\mathbf{s})) = 0$ for all $\mathbf{s} \in \Re^d$ then $E(Y_{\mathbf{u},h}(\mathbf{s})) = 0$ and $E(D_{\mathbf{u}}Y(\mathbf{s})) = 0$. Let $C_{\mathbf{u}}^{(h)}(\mathbf{s},\mathbf{s}')$ and $C_{\mathbf{u}}(\mathbf{s},\mathbf{s}')$ denote the covariance functions associated with the process $Y_{\mathbf{u},h}(\mathbf{s})$ and $D_{\mathbf{u}}Y(\mathbf{s})$, respectively. If $\boldsymbol{\Delta} = \mathbf{s} - \mathbf{s}'$ and $Y(\mathbf{s})$ is (weakly) stationary we immediately have

$$C_{\mathbf{u}}^{(h)}(\mathbf{s},\mathbf{s}') = \frac{(2C(\boldsymbol{\Delta}) - C(\boldsymbol{\Delta}+h\mathbf{u}) - C(\boldsymbol{\Delta}-h\mathbf{u}))}{h^2} \ , \tag{13.5}$$

whence $Var\left(Y_{\mathbf{u},h}\left(\mathbf{s}\right)\right)=2\left(C\left(\mathbf{0}\right)-C\left(h\mathbf{u}\right)\right)/h^{2}$. If $Y\left(\mathbf{s}\right)$ is isotropic and we replace $C(\mathbf{s},\mathbf{s}')$ by $\widetilde{C}(||\mathbf{s}-\mathbf{s}'||)$, we obtain

$$C_{\mathbf{u}}^{(h)}\left(\mathbf{s},\mathbf{s}'\right)=\frac{\left(2\widetilde{C}\left(||\mathbf{\Delta}||\right)-\widetilde{C}\left(||\mathbf{\Delta}+h\mathbf{u}||\right)-\widetilde{C}\left(||\mathbf{\Delta}-h\mathbf{u}||\right)\right)}{h^{2}}\,. \tag{13.6}$$

Expression (13.6) shows that even if $Y\left(\mathbf{s}\right)$ is isotropic, $Y_{\mathbf{u},h}\left(\mathbf{s}\right)$ is only stationary. Also $Var\left(Y_{\mathbf{u},h}\left(\mathbf{s}\right)\right)=2\left(\widetilde{C}\left(0\right)-\widetilde{C}\left(h\right)\right)/h^{2}=\gamma\left(h\right)/h^{2}$ where $\gamma\left(h\right)$ is the familiar variogram of the $Y\left(\mathbf{s}\right)$ process (Subsection 2.1.2).

Similarly, if $Y\left(\mathbf{s}\right)$ is stationary we may show that if all second-order partial and mixed derivatives of C exist and are continuous, the limit of (13.5) as $h\to 0$ is

$$C_{\mathbf{u}}\left(\mathbf{s},\mathbf{s}'\right)=-\mathbf{u}^{T}\Omega\left(\mathbf{\Delta}\right)\mathbf{u}\,, \tag{13.7}$$

where $\left(\Omega\left(\mathbf{\Delta}\right)\right)_{ij}=\partial^{2}C\left(\mathbf{\Delta}\right)/\partial\mathbf{\Delta}_{i}\partial\mathbf{\Delta}_{j}$. By construction, (13.7) is a valid covariance function on $\Re^{d}$ for any $\mathbf{u}$. Also, $Var\left(D_{\mathbf{u}}Y\left(\mathbf{s}\right)\right)=-\mathbf{u}^{T}\Omega\left(0\right)\mathbf{u}$. If $Y\left(\mathbf{s}\right)$ is isotropic, using standard chain rule calculations, we obtain

$$C_{\mathbf{u}}\left(\mathbf{s},\mathbf{s}'\right)=-\left\{\left(1-\frac{(\mathbf{u}^{T}\mathbf{\Delta})^{2}}{||\mathbf{\Delta}||^{2}}\right)\frac{\widetilde{C}'\left(||\mathbf{\Delta}||\right)}{||\mathbf{\Delta}||}+\frac{\left(\mathbf{u}^{T}\mathbf{\Delta}\right)^{2}}{||\mathbf{\Delta}||^{2}}\widetilde{C}''\left(||\mathbf{\Delta}||\right)\right\}\,. \tag{13.8}$$

Again, if $Y\left(\mathbf{s}\right)$ is isotropic, $D_{\mathbf{u}}Y\left(\mathbf{s}\right)$ is only stationary. In addition, we have $Var\left(D_{\mathbf{u}}Y\left(\mathbf{s}\right)\right)=-\widetilde{C}''\left(0\right)$ which also shows that, provided $\widetilde{C}$ is twice differentiable at 0, $\lim_{h\to 0}\gamma\left(h\right)/h^{2}=-\widetilde{C}''\left(0\right)$, i.e., $\gamma\left(h\right)=O\left(h^{2}\right)$ for h small.

For $Y\left(\mathbf{s}\right)$ stationary we can also calculate

$$cov\left(Y\left(\mathbf{s}\right),Y_{\mathbf{u},h}\left(\mathbf{s}'\right)\right)=\left(C\left(\mathbf{\Delta}-h\mathbf{u}\right)-C\left(\mathbf{\Delta}\right)\right)/h\,,$$

from which $cov\left(Y\left(\mathbf{s}\right),Y_{\mathbf{u},h}\left(\mathbf{s}\right)\right)=\left(C\left(h\mathbf{u}\right)-C\left(\mathbf{0}\right)\right)/h$. But then,

$$\begin{aligned} cov\left(Y\left(\mathbf{s}\right),D_{\mathbf{u}}Y\left(\mathbf{s}'\right)\right) &= \lim_{h\to 0}\left(C\left(\mathbf{\Delta}-h\mathbf{u}\right)-C\left(\mathbf{\Delta}\right)\right)/h \\ &= -D_{\mathbf{u}}C\left(\mathbf{\Delta}\right)\;=\;D_{\mathbf{u}}C\left(-\mathbf{\Delta}\right)\,, \end{aligned}$$

since $C\left(\mathbf{\Delta}\right)=C\left(-\mathbf{\Delta}\right)$. In particular, we have that $cov\left(Y\left(\mathbf{s}\right),D_{\mathbf{u}}Y\left(\mathbf{s}\right)\right)=\lim_{h\to 0}\left(C\left(h\mathbf{u}\right)-C\left(\mathbf{0}\right)\right)/h=D_{\mathbf{u}}C\left(\mathbf{0}\right)$. The existence of the directional derivative process ensures the existence of $D_{\mathbf{u}}C\left(\mathbf{0}\right)$. Moreover, since $C\left(h\mathbf{u}\right)=C\left(-h\mathbf{u}\right)$, $C\left(h\mathbf{u}\right)$ (viewed as a function of h) is even, so $D_{\mathbf{u}}C\left(\mathbf{0}\right)=0$. Thus, $Y\left(\mathbf{s}\right)$ and $D_{\mathbf{u}}Y\left(\mathbf{s}\right)$ are uncorrelated. Intuitively, this is sensible. The level of the process at a particular location is uncorrelated with the directional derivative in any direction at that location. This is not true for directional differences. Also, in general, $cov\left(Y\left(\mathbf{s}\right),D_{\mathbf{u}}Y\left(\mathbf{s}'\right)\right)$ will not be 0.

Under isotropy,

$$cov\left(Y\left(\mathbf{s}\right),Y_{\mathbf{u},h}\left(\mathbf{s}'\right)\right)=\frac{\widetilde{C}\left(||\mathbf{\Delta}-h\mathbf{u}||\right)-\widetilde{C}\left(||\mathbf{\Delta}||\right)}{h}\,.$$

Now $cov\left(Y\left(\mathbf{s}\right),Y_{\mathbf{u},h}\left(\mathbf{s}\right)\right)=\left(\widetilde{C}\left(h\right)-\widetilde{C}\left(0\right)\right)/h=\gamma\left(h\right)/2h$, so this means $cov(Y(\mathbf{s}),D_{\mathbf{u}}Y(\mathbf{s}))=\widetilde{C}'\left(0\right)=\lim_{h\to 0}\gamma\left(h\right)/2h=0$ since, as above, if $\widetilde{C}''\left(0\right)$ exists, $\gamma\left(h\right)=O\left(h^{2}\right)$.

Suppose we consider the bivariate process $\mathbf{Z}_{\mathbf{u}}^{(h)}\left(\mathbf{s}\right)=\left(Y\left(\mathbf{s}\right),Y_{\mathbf{u},h}\left(\mathbf{s}\right)\right)^{T}$. It is clear that this process has mean 0 and, if $Y\left(\mathbf{s}\right)$ is stationary, cross-covariance matrix $V_{\mathbf{u},h}\left(\mathbf{\Delta}\right)$ given by

$$\begin{pmatrix} C(\mathbf{\Delta}) & \frac{C(\mathbf{\Delta}-h\mathbf{u})-C(\mathbf{\Delta})}{h} \\ \frac{C(\mathbf{\Delta}+h\mathbf{u})-C(\mathbf{\Delta})}{h} & \frac{2C(\mathbf{\Delta})-C(\mathbf{\Delta}+h\mathbf{u})-C(\mathbf{\Delta}-h\mathbf{u})}{h^{2}} \end{pmatrix}\,. \tag{13.9}$$

Since $\mathbf{Z}_{\mathbf{u}}^{(h)}(\mathbf{s})$ arises by linear transformation of $Y(\mathbf{s})$, (13.9) is a valid cross-covariance matrix in $\Re^d$. But since this is true for every h, letting $h \to 0$,

$$V_{\mathbf{u}}(\mathbf{\Delta}) = \begin{pmatrix} C(\mathbf{\Delta}) & -D_{\mathbf{u}}C(\mathbf{\Delta}) \\ D_{\mathbf{u}}C(\mathbf{\Delta}) & -\mathbf{u}^T\Omega(\mathbf{\Delta})\mathbf{u} \end{pmatrix} \tag{13.10}$$

is a valid cross-covariance matrix in $\Re^d$. In fact, $V_{\mathbf{u}}$ is the cross-covariance matrix for the bivariate process $\mathbf{Z}_{\mathbf{u}}(s) = \begin{pmatrix} Y(\mathbf{s}) \\ D_{\mathbf{u}}Y(\mathbf{s}) \end{pmatrix}$.

Suppose we now assume that $Y(\mathbf{s})$ is a stationary Gaussian process. Then, it is clear, again by linearity, that $Y_{\mathbf{u},h}(\mathbf{s})$ and, in fact, $\mathbf{Z}_{\mathbf{u}}^h(\mathbf{s})$ are both stationary Gaussian processes. But then, by a standard limiting moment generating function argument, $\mathbf{Z}_{\mathbf{u}}(\mathbf{s})$ is a stationary bivariate Gaussian process and thus $D_{\mathbf{u}}Y(\mathbf{s})$ is a stationary univariate Gaussian process. As an aside, we note that for a given $\mathbf{s}$, $D_{\frac{\nabla Y(\mathbf{s})}{||\nabla Y(\mathbf{s})||}}Y(\mathbf{s})$ is not normally distributed, and in fact the set $\{D_{\frac{\nabla Y(\mathbf{s})}{||\nabla Y(\mathbf{s})||}}Y(\mathbf{s}) : \mathbf{s} \in \Re^d\}$ is not a spatial process.

Extension to a pair of directions with associated unit vectors $\mathbf{u}_1$ and $\mathbf{u}_2$ results in a trivariate Gaussian process $\mathbf{Z}(\mathbf{s}) = (Y(\mathbf{s}), D_{\mathbf{u}_1}Y(\mathbf{s}), D_{\mathbf{u}_2}Y(\mathbf{s}))^T$ with associated cross-covariance matrix $V_{\mathbf{Z}}(\mathbf{\Delta})$ given by

$$\begin{pmatrix} C(\mathbf{\Delta}) & -(\mathbf{\nabla}C(\mathbf{\Delta}))^T \\ \mathbf{\nabla}C(\mathbf{\Delta}) & -\Omega(\mathbf{\Delta}) \end{pmatrix}. \tag{13.11}$$

At $\mathbf{\Delta} = 0$, (13.11) becomes a diagonal matrix.

We conclude this subsection with a useful example. Recall the power exponential family of isotropic covariance functions of the previous subsection, $\widetilde{C}(||\mathbf{\Delta}||) = \alpha \exp(-\phi ||\mathbf{\Delta}||^{\nu})$, $0 < \nu \leq 2$. It is apparent that $\widetilde{C}''(0)$ exists only for $\nu = 2$. The Gaussian covariance function is the only member of the class for which directional derivative processes can be defined. However, as we have noted in Subsection 3.1.4, the Gaussian covariance function produces process realizations that are too smooth to be attractive for practical modeling.

Turning to the Matérn class, $\widetilde{C}(||\mathbf{\Delta}||) = \alpha(\phi ||\mathbf{\Delta}||)^{\nu} K_{\nu}(\phi ||\mathbf{\Delta}||)$, ν is a smoothness parameter controlling the extent of mean square differentiability of process realizations (Stein, 1999a). At $\nu = 3/2$, $\widetilde{C}(||\mathbf{\Delta}||)$ takes the closed form $\widetilde{C}(||\mathbf{\Delta}||) = \sigma^2(1+\phi ||\mathbf{\Delta}||)\exp(-\phi ||\mathbf{\Delta}||)$ where σ^2 is the process variance. This function is exactly twice differentiable at 0. We have a (once but not twice) mean square differentiable process, which therefore does not suffer the excessive smoothness implicit with the Gaussian covariance function.

For this choice one can show that $\mathbf{\nabla}\widetilde{C}(||\mathbf{\Delta}||) = -\sigma^2\phi^2 \exp(-\phi ||\mathbf{\Delta}||)\mathbf{\Delta}$, that $\left(H_{\widetilde{C}}(||\mathbf{\Delta}||)\right)_{ii} = -\sigma^2\phi^2 \exp(-\phi ||\mathbf{\Delta}||)\left(1 - \phi\Delta_i^2/||\mathbf{\Delta}||\right)$, and also that $\left(H_{\widetilde{C}}(||\mathbf{\Delta}||)\right)_{ij} = \sigma^2\phi^2 \exp(-\phi ||\mathbf{\Delta}||)\Delta_i\Delta_j/||\mathbf{\Delta}||$. In particular, $V_{\mathbf{u}}(\mathbf{0}) = \sigma^2 \text{BlockDiag}(1, \phi^2 I)$.

13.5 Directional derivative processes in modeling

We work in $d = 2$-dimensional space and can envision the following types of modeling settings in which directional derivative processes would be of interest. For $Y(\mathbf{s})$ purely spatial with constant mean, we would seek $D_{\mathbf{u}}Y(\mathbf{s})$. In the customary formulation $Y(\mathbf{s}) = \mu(\mathbf{s}) + w(\mathbf{s}) + \epsilon(\mathbf{s})$ we would instead want $D_{\mathbf{u}}w(\mathbf{s})$. In the case of a spatially varying coefficient model $Y(\mathbf{s}) = \beta_0(\mathbf{s}) + \beta_1(\mathbf{s})X(\mathbf{s}) + \epsilon(\mathbf{s})$ such as in Section 9.6, we would examine $D_{\mathbf{u}}\beta_0(\mathbf{s})$, $D_{\mathbf{u}}\beta_1(\mathbf{s})$, and $D_{\mathbf{u}}EY(\mathbf{s})$ with $EY(\mathbf{s}) = \beta_0(\mathbf{s}) + \beta_1(\mathbf{s})X(\mathbf{s})$.

Consider the constant mean purely spatial process for illustration, where we have $Y(\mathbf{s})$ a stationary process with mean μ and covariance function $C(\mathbf{\Delta}) = \sigma^2\rho(\mathbf{\Delta})$ where ρ is a valid two-dimensional correlation function. For illustration we work with the general Matérn class parametrized by ϕ and ν, constraining $\nu > 1$ to ensure the (mean square) existence of

the directional derivative processes. Letting $\boldsymbol{\theta} = \left(\mu, \sigma^2, \phi, \nu\right)$, for locations $\mathbf{s}_1, \mathbf{s}_2, ..., \mathbf{s}_n$, the likelihood $L\left(\boldsymbol{\theta}; \mathbf{Y}\right)$ is proportional to

$$\left(\sigma^2\right)^{-n/2} \left|R\left(\phi, \nu\right)\right|^{-1/2} \exp\left\{-\frac{1}{2\sigma^2}\left(\mathbf{Y} - \mu\mathbf{1}\right)^T R^{-1}\left(\phi, \nu\right)\left(\mathbf{Y} - \mu\mathbf{1}\right)\right\} . \tag{13.12}$$

In (13.12), $\mathbf{Y}^T = \left(Y(\mathbf{s}_1), \ldots, Y(\mathbf{s}_n)\right)^T$ and $\left(R\left(\phi, \nu\right)\right)_{ij} = \rho\left(\mathbf{s}_i - \mathbf{s}_j; \phi, \nu\right)$.

We have discussed prior specification for such a model in Chapter 6. Customary choices include a vague normal (perhaps flat) prior for μ; we would deal with the identifiability issue for σ^2 and ϕ through a centered inverse gamma on the former and a weakly informative prior for the latter. (Below, we work with an inverse gamma and a gamma prior, respectively). With regard to the prior on ν, Banerjee, Gelfand and Sirmans (2003) follow the suggestion of Stein (1999a) and others, who observe that, in practice, it will be very difficult to distinguish $\nu = 2$ from $\nu > 2$ and so, adopt a $U\left(1, 2\right)$ prior. The algorithms discussed in Chapter 6, and available in the `spBayes` library, can be used for fitting the Bayesian model. We can also use slice sampling (Appendix Section A.2; also see Agarwal and Gelfand, 2005) as an easily programmable alternative.

We note that all gradient analysis is a post-model fitting activity, employing posterior samples of the model parameters to obtain samples from posterior predictive distributions. In particular, a contour or a grey-scale plot of the posterior mean surface is of primary interest in providing a smoothed display of spatial pattern and of areas where the process is elevated or depressed. To handle finite differences at scale h, in the sequel we work with the vector of eight compass directions, $N, NE, E, \ldots$. At $\mathbf{s}_i$, we denote this vector by $\mathbf{Y}_h\left(\mathbf{s}_i\right)$ and let $\mathbf{Y}_h = \left\{\mathbf{Y}_h\left(\mathbf{s}_i\right), i = 1, 2, ..., n\right\}$. With directional derivatives we only need $\mathbf{D}\left(\mathbf{s}_i\right)^T = \left(D_{(1,0)}Y(\mathbf{s}_i), D_{(0,1)}Y(\mathbf{s}_i)\right)$ and let $\mathbf{D} = \left\{\mathbf{D}\left(\mathbf{s}_i\right), i = 1, 2, ..., n\right\}$. We seek samples from the predictive distribution $f\left(\mathbf{Y}_h|\mathbf{Y}\right)$ and $f\left(\mathbf{D}|\mathbf{Y}\right)$. In $\mathbf{Y}_h\left(\mathbf{s}_i\right)$, $Y\left(\mathbf{s}_i\right)$ is observed, hence fixed in the predictive distribution. So we can replace $Y_{u,h}\left(\mathbf{s}_i\right)$ with $Y\left(\mathbf{s}_i + h\mathbf{u}\right)$; posterior predictive samples of $Y\left(\mathbf{s}_i + h\mathbf{u}\right)$ are immediately converted to posterior predictive samples of $Y_{u,h}\left(\mathbf{s}_i\right)$ by linear transformation. Hence, the directional finite differences problem is merely a large Bayesian kriging problem requiring spatial prediction at the set of $8n$ locations $\left\{Y\left(\mathbf{s}_i + h\mathbf{u}_r\right), i = 1, 2, ..., n; r = 1, 2, ..., 8\right\}$. Denoting this set by $\widetilde{\mathbf{Y}}_h$, we require samples from $f\left(\widetilde{\mathbf{Y}}_h|\mathbf{Y}\right)$. From the relationship, $f\left(\widetilde{\mathbf{Y}}_h|\mathbf{Y}\right) = \int f\left(\widetilde{\mathbf{Y}}_h|\mathbf{Y}, \boldsymbol{\theta}\right) f\left(\boldsymbol{\theta}|\mathbf{Y}\right) d\boldsymbol{\theta}$ this can be done one for one with the $\boldsymbol{\theta}_l^*$'s by drawing $\widetilde{\mathbf{Y}}_{h,l}^*$ from the multivariate normal distribution $f\left(\widetilde{\mathbf{Y}}_h|\mathbf{Y}, \boldsymbol{\theta}_l^*\right)$, as detailed in Section 6.1. Similarly $f\left(\mathbf{D}|\mathbf{Y}\right) = \int f\left(\mathbf{D}|\mathbf{Y}, \boldsymbol{\theta}\right) f\left(\boldsymbol{\theta}|\mathbf{Y}\right) d\boldsymbol{\theta}$. The cross-covariance function in (13.11) allows us to immediately write down the joint multivariate normal distribution of $\mathbf{Y}$ and $\mathbf{D}$ given $\boldsymbol{\theta}$ and thus, at $\boldsymbol{\theta}_l^*$, the conditional multivariate normal distribution $f\left(\mathbf{D}|\mathbf{Y}, \boldsymbol{\theta}_l^*\right)$. At a specified new location $\mathbf{s}_0$, with finite directional differences we need to add spatial prediction at the nine new locations, $Y\left(\mathbf{s}_0\right)$ and $Y\left(\mathbf{s}_0 + h\mathbf{u}_r\right)$, $r = 1, 2, ..., 8$. With directional derivatives, we again can use (13.9) to obtain the joint distribution of $\mathbf{Y}$, $\mathbf{D}$, $Y\left(\mathbf{s}_0\right)$ and $\mathbf{D}\left(\mathbf{s}_0\right)$ given θ and thus the conditional distribution $f\left(\mathbf{D}, Y\left(\mathbf{s}_0\right), \mathbf{D}\left(\mathbf{s}_0\right)|\mathbf{Y}, \boldsymbol{\theta}\right)$.

Turning to the random spatial effects model we now assume that

$$Y\left(\mathbf{s}\right) = \mathbf{x}^T\left(\mathbf{s}\right)\boldsymbol{\beta} + w\left(\mathbf{s}\right) + \epsilon\left(\mathbf{s}\right) . \tag{13.13}$$

In (13.13), $\mathbf{x}\left(\mathbf{s}\right)$ is a vector of location characteristics, $w\left(\mathbf{s}\right)$ is a mean 0 stationary Gaussian spatial process with parameters σ^2, ϕ, and ν as above, and $\epsilon\left(\mathbf{s}\right)$ is a Gaussian white noise process with variance τ^2, intended to capture measurement error or microscale variability. Such a model is appropriate for the real estate example mentioned in Section 13.3, where $Y\left(\mathbf{s}\right)$ is the log selling price and $\mathbf{x}\left(\mathbf{s}\right)$ denotes associated house and property characteristics. Here $w\left(\mathbf{s}\right)$ measures the spatial adjustment to log selling price at location $\mathbf{s}$ reflecting

relative desirability of the location. $\epsilon(\mathbf{s})$ is needed to capture microscale variability. Here such variability arises because two identical houses arbitrarily close to each other need not sell for essentially the same price due to unobserved differences in buyers, sellers, and brokers across transactions.

For locations $\mathbf{s}_1, \mathbf{s}_2, ..., \mathbf{s}_n$, with $\boldsymbol{\theta} = \left(\boldsymbol{\beta}, \tau^2, \sigma^2, \phi, \nu\right)$ the model in (13.13) produces a marginal likelihood $L(\boldsymbol{\theta}; Y)$ (integrating over $\{w(\mathbf{s}_i)\}$) proportional to

$$|\Sigma(\boldsymbol{\gamma})|^{-1/2} \exp\left\{-\frac{1}{2}(\mathbf{Y} - X\boldsymbol{\beta})^T \Sigma(\boldsymbol{\gamma})^{-1} (\mathbf{Y} - X\boldsymbol{\beta})\right\} ,$$

where $\boldsymbol{\gamma} = \{\tau^2, \sigma^2, \phi, \nu\}$ and $\Sigma(\boldsymbol{\gamma}) = \sigma^2 R(\phi, \nu) + \tau^2 I$. Priors for $\boldsymbol{\theta}$ can be prescribed as for (13.12). Again, the procedures discussed in Chapter 6, or even slice sampling, provides an efficient fitting mechanism and is available through the `spLM` function in the `spBayes` library.

Further inference with regard to (13.13) focuses on the spatial process itself. That is, we would be interested in the posterior spatial effect surface and in rates of change associated with this surface. The former is usually handled with samples of the set of $w(\mathbf{s}_i)$ given $\mathbf{Y}$ along with, perhaps, a grey-scaling or contouring routine. The latter would likely be examined at new locations. For instance in the real estate example, spatial gradients would be of interest at the central business district or at other externalities such as major roadway intersections, shopping malls, airports, or waste disposal sites but not likely at the locations of the individual houses.

As below (13.12) with $\mathbf{w}^T = (w(\mathbf{s}_1), \ldots, w(\mathbf{s}_n))$, we sample $f(\mathbf{w}|Y)$ one for one with the $\boldsymbol{\theta}_l^*$'s using $f(\mathbf{w}|\mathbf{Y}) = \int f(\mathbf{w}|\mathbf{Y}, \boldsymbol{\theta}) f(\boldsymbol{\theta}|\mathbf{Y}) d\boldsymbol{\theta}$, as described in Section 6.1. But also, given $\boldsymbol{\theta}$, the joint distribution of $\mathbf{w}$ and $\mathbf{V}(\mathbf{s}_0)$ where $\mathbf{V}(\mathbf{s}_0)$ is either $\mathbf{w}_h(\mathbf{s}_0)$ or $\mathbf{D}(\mathbf{s}_0)$ is multivariate normal. For instance, with $D(\mathbf{s}_0)$, the joint normal distribution can be obtained using (13.11) and as a result so can the conditional normal distribution $f(\mathbf{V}(\mathbf{s}_0)|\mathbf{w}, \boldsymbol{\theta})$. Lastly, since

$$f(\mathbf{V}(\mathbf{s}_0)|\mathbf{Y}) = \int f(\mathbf{V}(\mathbf{s}_0)|\mathbf{w}, \boldsymbol{\theta}) f(\mathbf{w}|\boldsymbol{\theta}, \mathbf{Y}) f(\boldsymbol{\theta}|\mathbf{Y}) d\boldsymbol{\theta} d\mathbf{w} ,$$

we can also obtain samples from $f(\mathbf{V}(\mathbf{s}_0)|\mathbf{Y})$ one for one with the $\boldsymbol{\theta}_l^*$'s.

13.6 Illustration: Inference for differences and gradients

A simulation example (also see Banerjee, Gelfand and Sirmans, 2003) is provided to illustrate inference on finite differences and directional gradients. We generate data from a Gaussian random field with constant mean μ and a covariance structure specified through the Matèrn ($\nu = 3/2$) covariance function, $\sigma^2(1 + \phi d)\exp(-\phi d)$. This will yield a realized field that will be mean-square differentiable exactly once. The field is observed on a randomly sampled set of points within a 10x10 square. We set $\mu = 0$, $\sigma^2 = 1.0$ and $\phi = 1.05$. In the subsequent illustration our data consists of $n = 100$ observations at the randomly selected sites shown in the left panel of Figure 13.1. The maximum observed distance in our generated field is approximately 13.25 units. The value of $\phi = 1.05$ provides an effective isotropic range of about 4.5 units. We also perform a Bayesian kriging on the data to develop a predicted field. The right panel of Figure 13.1 shows a grey-scale plot with contour lines displaying the topography of the "kriged" field. We will see below that our predictions of the spatial gradients at selected points are consistent with the topography around those points, as depicted in the right panel of Figure 13.1. Adopting a flat prior for μ, an IG$(2, 0.1)$ (mean = 10, infinite variance) prior for σ^2, a $G(2, 0.1)$ prior (mean=20, variance=200) for ϕ, and a uniform on $(1, 2)$ for ν, we obtain the posterior estimates for our parameters shown in Table 13.1.

Figure 13.1 *Left panel: Location of the 100 sites where the random field has been observed. Right panel: A grey scale plot with contour lines showing the topography of the random field in the simulation example.*

Parameter	50% (2.5%,97.5%)
μ	-0.39 (-0.91,0.10)
σ^2	0.74 (0.50,1.46)
ϕ	1.12 (0.85,1.41)
ν	1.50 (1.24,1.77)

Table 13.1 *Posterior estimates for model parameters.*

We next predict the directional derivatives and directional finite differences for the unit vectors corresponding to angles of 0, 45, 90, 135, 180, 225, 270 and 315 degrees with the horizontal axis in a counter-clockwise direction at the point. For the finite differences we consider $h = 1.0, 0.1$ and 0.01. Recall that $D_{-\mathbf{u}}Y(\mathbf{s}) = -D_{\mathbf{u}}Y(\mathbf{s})$.

Table 13.2 presents the resulting posterior predictive inference for the point $(3.5, 3.5)$ in Figure 13.1. We see that $(3.5, 3.5)$ seems to be in a rather interesting portion of the surface, with many contour lines nearby. It is clear from the contour lines that there is a negative northern gradient (downhill) and a positive southern gradient (uphill) around this point. On the other hand, there does not seem to be any significant EW gradient around that point as seen from the contour lines through that point running EW. This is brought out very clearly in column 1 of Table 13.2. The angles of 0 and 180 degrees which correspond to the EW gradients are not at all significant. The NS gradients are indeed pronounced as seen by the 90 and 270 degree gradients. The directional derivatives along the diagonals also indicate presence of a gradient. There is a significant downhill gradient towards the NE and (therefore) a significant uphill gradient towards the SW. Hence the directional derivative process provides inference consistent with features captured descriptively and visually in Figure 13.1.

For the directional finite differences in columns 2, 3 and 4 of Table 13.2, note, for instance, the difference between column 2 and columns 3 and 4. In the former, none of the directional finite differences are significant. The low resolution (large h) fails to capture

Angle	$D_{\mathbf{u}}Y(\mathbf{s})$ $(h=0)$	$h=1.0$	$h=0.1$	$h=0.01$
0	-0.06 (-1.12,1.09)	0.51 (-0.82,1.81)	-0.08 (-1.23,1.20)	-0.07 (-1.11,1.10)
45	-1.49 (-2.81,-0.34)	-0.01 (-1.29,1.32)	-1.55 (-2.93,-0.56)	-1.53 (-2.89,-0.49)
90	-2.07 (-3.44,-0.66)	-0.46 (-1.71,0.84)	-2.13 (-3.40,-0.70)	-2.11 (-3.41,-0.69)
135	-1.42 (-2.68,-0.23)	-0.43 (-1.69,0.82)	-1.44 (-2.64,-0.23)	-1.43 (-2.70,-0.23)
180	0.06 (-1.09,1.12)	-0.48 (-1.74,0.80)	0.08 (-1.19,1.23)	0.06 (-1.10,1.12)
225	1.49 (0.34,2.81)	0.16 (-1.05,1.41)	1.61 (0.52,3.03)	1.52 (0.48,2.90)
270	2.07 (0.66,3.44)	0.48 (-0.91,1.73)	2.12 (0.68,3.43)	2.10 (0.68,3.42)
315	1.42 (0.23,2.68)	1.12 (-0.09,2.41)	1.44 (0.24,2.68)	1.42 (0.23,2.70)

Table 13.2 *Posterior medians and (2.5%, 97.5%) predictive intervals for directional derivatives and finite differences at point (3.5, 3.5).*

local topographic properties. On the other hand the latter very much resemble column 1. As expected, at high resolution, the directional finite difference process results match those of the directional derivative process. Computational simplicity and stability (difficulties may arise with very small h in the denominator of (2)) encourage the use of the latter (see Banerjee, Gelfand and Sirmans 2003, for details).

13.7 Curvilinear gradients and wombling

We now extend the developments in the previous section to an inferential framework for gradients along curves. The conceptual challenge in moving from points to curves is the construction of a sensible measure to associate with a curve in order to assess whether it can be declared a wombling boundary. In this regard, we can consider open or closed curves. In Subsection 13.7.1 we formally develop the notion of an average gradient to associate with a curve and in Subsection 13.7.2 we are able to offer a formal definition of a wombling boundary.

We use differential geometric notions for parametric boundaries as developed in, e.g., Rudin (1976) or Frankel (2003). Since most spatial modelling is done on domains in $\Re^2$ we restrict our attention to this case, focusing upon a real-valued process $Y(\mathbf{s})$ with the spatial domain as an open subset of $\Re^2$. Thus, we offer an independent development of gradients along planar curves without resorting to geometry on manifolds. For hyper-curves in general $\Re^d$, the theory is more complicated (especially if $d > 3$) and must involve development of calculus on abstract manifolds.

13.7.1 Gradients along curves

Let C be an open curve in $\Re^2$ and suppose it is desired to ascertain whether such a curve is a wombling boundary with regard to $Y(\mathbf{s})$. To do so we seek to associate an average gradient with C. In particular, for each point $\mathbf{s}$ lying on C, we let $D_{n(\mathbf{s})}Y(\mathbf{s})$ be the directional derivative in the direction of the unit normal $n(\mathbf{s})$.[1] We can define the *wombling measure* of the curve either as the total gradient along C,

$$\int_C D_{n(\mathbf{s})}Y(\mathbf{s})\,d\nu = \int_C \langle \nabla Y(\mathbf{s}), n(\mathbf{s})\rangle d\nu, \tag{13.14}$$

or perhaps as the average gradient along C,

$$\frac{1}{\nu(C)}\int_C D_{n(\mathbf{s})}Y(\mathbf{s})\,d\nu = \frac{1}{\nu(C)}\int_C \langle \nabla Y(\mathbf{s}), n(\mathbf{s})\rangle d\nu, \tag{13.15}$$

[1] Again, the rationale for the choice of direction normal to the curve is that, for a curve tracking rapid change in the spatial surface, lines orthogonal to the curve should reveal sharp gradients.

where $\nu(\cdot)$ is an appropriate measure. For (13.14) and (13.15), ambiguity arises with respect to the choice of measure. For example, $\nu(C) = 0$ if we take ν as two-dimensional Lebesgue measure and, indeed, this is true for any ν which is mutually absolutely continuous with respect to Lebesgue measure. Upon reflection, an appropriate choice for ν turns out to be arc-length. This can be made clear by a parametric treatment of the curve C.

In particular, a curve C in $\Re^2$ is a set parametrized by a single parameter $t \in \Re^1$ where $C = \{\mathbf{s}(t) : t \in \mathcal{T}\}$, with $\mathcal{T} \subset \Re^1$. We call $\mathbf{s}(t) = \left(s_{(1)}(t), s_{(2)}(t)\right) \in \Re^2$ the position vector of the curve - $\mathbf{s}(t)$ traces out C as t spans its domain. Then, assuming a differentiable curve with non-vanishing derivative $\mathbf{s}'(t) \neq 0$ (such a curve is often called *regular*), we obtain the (component-wise) derivative $\mathbf{s}'(t)$ as the "velocity" vector, with unit velocity (or tangent) vector $\mathbf{s}'(t) / ||\mathbf{s}'(t)||$. Letting $n(\mathbf{s}(t))$ denote the parametrized unit normal vector to C, again if C is sufficiently smooth, then $\langle \mathbf{s}'(t), n(\mathbf{s}(t))\rangle = 0, a.e. \mathcal{T}$. In $\Re^2$ we see that

$$n(\mathbf{s}(t)) = \frac{\left(s'_{(2)}(t), -s'_{(1)}(t)\right)}{||\mathbf{s}'(t)||}. \tag{13.16}$$

Under the above parametrization (and the regularity assumption) the arc-length measure ν can be defined as

$$\nu(\mathcal{T}) = \int_{\mathcal{T}} ||\mathbf{s}'(t)||\, dt. \tag{13.17}$$

In fact, $||\mathbf{s}'(t)||$ is analogous to the "speed" (the norm of the velocity) at "time" t, so the above integral is interpretable as the distance traversed or, equivalently, the arc-length $\nu(C)$ or $\nu(\mathcal{T})$. In particular, if $\mathcal{T}$ is an interval, say $[t_0, t_1]$, we can write

$$\nu(\mathcal{T}) = \nu_{t_0}(t_1) = \int_{t_0}^{t_1} ||\mathbf{s}'(t)||\, dt.$$

Thus we have $d\nu_{t_0}(t) = ||\mathbf{s}'(t)||\, dt$ and, taking ν as the arc-length measure for C, we have the wombling measures in (13.14) (total gradient) and (13.15) (average gradient), respectively, as

$$\Gamma_{Y(\mathbf{s})}(\mathcal{T}) = \int_C \langle \nabla Y(\mathbf{s}), n(\mathbf{s})\rangle d\nu = \int_{\mathcal{T}} \langle \nabla Y(\mathbf{s}(t)), n(\mathbf{s}(t))\rangle\, ||\mathbf{s}'(t)||\, dt$$
$$\text{and } \bar{\Gamma}_{Y(\mathbf{s})}(\mathcal{T}) = \frac{1}{\nu(\mathcal{T})}\Gamma_{Y(\mathbf{s})}(\mathcal{T}). \tag{13.18}$$

This result is important since we want to take ν as the arc-length measure, but it will be easier to use the parametric representation and work in t space. Also, it is a consequence of the implicit mapping theorem in mathematical analysis (see, e.g., Rudin, 1976) that any other parametrization $\mathbf{s}^*(t)$ of the curve C is related to $\mathbf{s}(t)$ through a differentiable mapping g such that $\mathbf{s}^*(t) = \mathbf{s}(g(t))$. This immediately implies (using (13.18)) that our proposed wombling measure is invariant to the parametrization of C and, as desired, a feature of the curve itself.

For some simple curves the wombling measure can be evaluated quite easily. For instance, when C is a segment of length 1 of the straight line through the point $\mathbf{s}_0$ in the direction $\mathbf{u} = (u_{(1)}, u_{(2)})$, then we have $C = \{\mathbf{s}_0 + t\mathbf{u} : t \in [0,1]\}$. Under this parametrization, $\mathbf{s}'(t)^T = (u_{(1)}, u_{(2)})$, $||\mathbf{s}'(t)|| = 1$, and $\nu_{t_0}(t) = t$. Clearly, $n(\mathbf{s}(t)) = \left(u_{(2)}, -u_{(1)}\right)$, (independent of t), which we write as $\mathbf{u}^\perp$ – the normal direction to $\mathbf{u}$. Therefore $\Gamma_{Y(\mathbf{s})}(\mathcal{T})$ in (13.18) becomes

$$\int_0^1 \langle \nabla Y(\mathbf{s}(t)), n(\mathbf{s}(t))\rangle dt = \int_0^1 D_{\mathbf{u}^\perp} Y(\mathbf{s}(t))\, dt.$$

Another example is when C is the arc of a circle with radius r. For example suppose C is traced out by $\mathbf{s}(t) = (r\cos t, r\sin t)$ as $t \in [0, \pi/4]$. Then,since $||\mathbf{s}'(t)|| = r$, the average

gradient is more easily computed as

$$\begin{aligned}\frac{1}{\nu(C)}\int_0^{\pi/4}\langle\nabla Y(\mathbf{s}(t)),n(\mathbf{s}(t))\rangle r dt &= \frac{4}{r\pi}\int_0^{\pi/4}\langle\nabla Y(\mathbf{s}(t)),n(\mathbf{s}(t))\rangle r dt\\ &= \frac{4}{\pi}\int_0^{\pi/4}\langle\nabla Y(\mathbf{s}(t)),n(\mathbf{s}(t))\rangle dt.\end{aligned}$$

In either case, $n(\mathbf{s}(t))$ is given by (13.16).

Note that while the normal component, $D_{n(\mathbf{s})}Y(\mathbf{s})$, seems to be more appropriate for assessing whether a curve provides a wombling boundary, one may also consider the tangential direction, $\mathbf{u}(t)=\mathbf{s}'(t)/||\mathbf{s}'(t)||$ along a curve C. In this case, the average gradient will be given by

$$\frac{1}{\nu(C)}\int_C\langle\nabla Y(\mathbf{s}(t)),\mathbf{u}(t)\rangle\,||\mathbf{s}'(t)||\,dt=\frac{1}{\nu(C)}\int_C\langle\nabla Y(\mathbf{s}(t)),\mathbf{s}'(t)\rangle dt.$$

In fact, we have

$$\begin{aligned}\int_C\langle\nabla Y(\mathbf{s}(t)),\mathbf{s}'(t)\rangle dt &= \int_{t_0}^{t_1}\langle\nabla Y(\mathbf{s}(t)),\mathbf{s}'(t)\rangle dt\\ &= \int_{\mathbf{s}_0}^{\mathbf{s}_1}\langle\nabla Y(\mathbf{s}),d\mathbf{s}\rangle = Y(\mathbf{s}_1)-Y(\mathbf{s}_0),\end{aligned}$$

where $\mathbf{s}_1=\mathbf{s}(t_1)$ and $\mathbf{s}_0=\mathbf{s}(t_0)$ are the endpoints of C. That is, unsatisfyingly, the average directional gradient in the tangential direction is independent of the path C, depending only upon the endpoints of the curve C. Furthermore, Banerjee and Gelfand (2006) show that for a closed path C, the average gradient in the tangential direction is zero. These considerations motivate us to define a "wombling boundary" (see below) with respect to the direction normal to the curve (i.e., perpendicular to the tangent).

13.7.2 Wombling boundary

With the above formulation in place, we now offer a formal definition of a curvilinear *wombling boundary*:

Definition: A curvilinear wombling boundary is a curve C that reveals a large wombling measure, $\Gamma_{Y(\mathbf{s})}(\mathcal{T})$ or $\bar{\Gamma}_{Y(\mathbf{s})}(\mathcal{T})$ (as given in (13.18)) in the direction normal to the curve.

Were the surface fixed, we would have to set a threshold to determine what "large," say in absolute value, means. Since the surface is a random realization, $\Gamma_{Y(\mathbf{s})}(\mathcal{T})$ and $\bar{\Gamma}_{Y(\mathbf{s})}(\mathcal{T})$ are random. Hence, we declare a curve to be a wombling boundary if, say a, 95% credible set for $\bar{\Gamma}_{Y(\mathbf{s})}(\mathcal{T})$ does not contain 0. It is worth pointing out that while one normal direction (as defined in (13.16)) is used in (13.18), $-n(\mathbf{s}(t))$ would also have been a valid choice. Since $D_{-n(\mathbf{s}(t))}Y(\mathbf{s}(t))=-D_{n(\mathbf{s}(t))}Y(\mathbf{s}(t))$, we note that the wombling measure with respect to one is simply the negative of the other. Thus, in the above definition large positive as well as large negative values of the integral in (13.18) would signify a wombling boundary. Being a local concept, an uphill gradient is equivalent to a downhill gradient across a curve.

We also point out that, being a continuous average (or sum) of the directional gradients, the wombling measure may "cancel" the overall gradient effect. For instance, imagine a curve C that exhibits a large positive gradient in the $n(\mathbf{s})$ direction for the first half of its length and a large negative gradient for the second half, thereby cancelling the total or average gradient effect. A potential remedy is to redefine the wombling measure using absolute gradients, $|D_{n(\mathbf{s})}Y(\mathbf{s})|$, in (13.14) and (13.15). The corresponding development does not entail any substantially new ideas, but would destroy the attractive distribution

theory in Section 13.8 below and make the computation less tractable. In particular, it will make calibration of the resulting measure with regard to significance much more difficult; how do we select a threshold? Moreover, in practice a descriptive contour representation is usually available where sharp gradients will usually reflect themselves and one could instead compute the wombling measure for appropriate sub-curves of C. Though somewhat subjective, identifying such sub-curves is usually unambiguous and leads to robust scientific inference. More fundamentally, in certain applications a signed measure may actually be desirable: one might want to classify a curve as a wombling boundary if it reflects either an overall "large positive" or a "large negative" gradient effect across it. For these reasons, we confine ourselves to working with $D_{n(\mathbf{s})}Y(\mathbf{s})$ and turn to the distribution theory for the wombling measure in the next section.

13.8 Distribution theory for curvilinear gradients

Curvilinear wombling amounts to performing predictive inference for a line integral parametrized over $\mathcal{T}$. Let us suppose that $\mathcal{T}$ is an interval, $[0,T]$, which generates the curve $C = \{\mathbf{s}(t) : t \in [0,T]\}$. For any $t^* \in [0,T]$ let $\nu(t^*)$ denote the arc length of the associated curve C_{t^*}. The line integrals for total gradient and average gradient along C_{t^*} are given by $\Gamma_{Y(\mathbf{s})}(t^*)$ and $\bar{\Gamma}_{Y(\mathbf{s})}(t^*)$ respectively as:

$$\Gamma_{Y(\mathbf{s})}(t^*) = \int_0^{t^*} D_{n(\mathbf{s}(t))}Y(\mathbf{s}(t))\|\mathbf{s}'(t)\|dt \text{ and } \bar{\Gamma}_{Y(\mathbf{s})}(t^*) = \frac{1}{\nu(t^*)}\Gamma_{Y(\mathbf{s})}(t^*). \tag{13.19}$$

We seek to infer about $\Gamma_{Y(\mathbf{s})}(t^*)$ based upon data $Y = (Y(\mathbf{s}_1), \ldots, Y(\mathbf{s}_n))$. Since $D_{n(\mathbf{s}(t))}Y(\mathbf{s}(t)) = \langle \nabla Y(\mathbf{s}(t)), n(\mathbf{s}(t))\rangle$ is a Gaussian process, $\Gamma_{Y(\mathbf{s})}(t^*)$ is a Gaussian process on $[0,T]$, equivalently on the curve C. Note that although $D_{n(\mathbf{s})}Y(\mathbf{s})$ is a process on $\Re^d$, our parametrization of the coordinates by $t \in \mathcal{T} \subseteq \Re^1$ induces a valid process on $\mathcal{T}$. In fact, $\Gamma_{Y(\mathbf{s})}(t^*)$ is a Gaussian process whose mean and covariance functions are

$$\begin{aligned}
\mu_{\Gamma_{Y(\mathbf{s})}}(t^*) &= \int_0^{t^*} D_{n(\mathbf{s}(t))}\mu((\mathbf{s}(t)))\|\mathbf{s}'(t)\|dt \\
K_{\Gamma_{Y(\mathbf{s})}}(t_1^*, t_2^*) &= \int_0^{t_1^*}\int_0^{t_2^*} q_{n(\mathbf{s})}(t_1,t_2)\|\mathbf{s}'(t_1)\|\|\mathbf{s}'(t_2)\|dt_1dt_2 \ ,
\end{aligned}$$

where $q_{n(\mathbf{s})}(t_1,t_2) = n^T(\mathbf{s}(t_1))H_K(\Delta(t_1,t_2))\,n(\mathbf{s}(t_2))$ and $\Delta(t_1,t_2) = \mathbf{s}(t_2) - \mathbf{s}(t_1)$. In particular, $\mathrm{Var}\left(\Gamma_{Y(\mathbf{s})}(t^*)\right) = K_{\Gamma_{Y(\mathbf{s})}}(t^*,t^*)$ is

$$\int_0^{t^*}\int_0^{t^*} n^T(\mathbf{s}(t_1))H_K(\Delta(t_1,t_2))\,n(\mathbf{s}(t_2))\|\mathbf{s}'(t_1)\|\,||\mathbf{s}'(t_2)||\,dt_1dt_2.$$

Evidently, $\Gamma_{Y(\mathbf{s})}(t^*)$ is mean square continuous. But, from the above, note that even if $Y(\mathbf{s})$ is a stationary process, $\Gamma_{Y(\mathbf{s})}(t^*)$ is not. For any $\mathbf{s}_j$ in the domain of Y,

$$\begin{aligned}
\mathrm{Cov}(\Gamma_{Y(\mathbf{s})}(t^*), Y(\mathbf{s}_j)) &= \int_0^{t^*} Cov(D_{n(\mathbf{s}(t))}Y(\mathbf{s}(t)), Y(\mathbf{s}_j))\|\mathbf{s}'(t)\|dt \\
&= \int_0^{t^*} D_{n(\mathbf{s}(t))}K(\Delta_j(t))\|\mathbf{s}'(t)\|dt,
\end{aligned} \tag{13.20}$$

where $\Delta_j(t) = \mathbf{s}(t) - \mathbf{s}_j$. Based upon data $\mathbf{Y} = (Y(\mathbf{s}_1), \ldots, Y(\mathbf{s}_n))$, we seek the predictive distribution $P(\Gamma_{Y(\mathbf{s})}(t^*)|\mathbf{Y})$, but note that $Y(\mathbf{s})$ and $\Gamma_{Y(\mathbf{s})}(t^*)$ are processes on different domains — the former is over a connected region in $\Re^2$, while the latter is on a parametrized

curve, $\mathbf{s}(t)$, indexed by $\mathcal{T}$. Nevertheless, $\Gamma_{Y(\mathbf{s})}(t^*)$ is derived from $Y(\mathbf{s})$ and we have a valid *joint distribution* $\left(\mathbf{Y}, \Gamma_{Y(\mathbf{s})}(t^*)\right)$ for any $t^* \in \mathcal{T}$, given by

$$N_{n+1}\left(\begin{pmatrix} \boldsymbol{\mu} \\ \mu_{\Gamma_{Y(s)}}(t^*) \end{pmatrix}, \begin{pmatrix} \Sigma_Y & \boldsymbol{\gamma}_{\Gamma,\mathbf{Y}}(t^*) \\ \boldsymbol{\gamma}^T_{\Gamma,\mathbf{Y}}(t^*) & K_\Gamma(t^*, t^*) \end{pmatrix}\right). \tag{13.21}$$

Here, $\boldsymbol{\mu} = (\mu(\mathbf{s}_1), \ldots, \mu(\mathbf{s}_n))$ and

$$\gamma^T_{\Gamma,\mathbf{Y}}(t^*) = \left(Cov\left(Y(\mathbf{s}_1), \Gamma_{Y(\mathbf{s})}(t^*)\right), \ldots, Cov\left(Y(\mathbf{s}_n), \Gamma_{Y(\mathbf{s})}(t^*)\right)\right),$$

each component being evaluated from (13.20).

Suppose $\mu(\mathbf{s}; \boldsymbol{\beta})$ and $K(\cdot; \boldsymbol{\eta})$ are indexed by regression parameters $\boldsymbol{\beta}$ and covariance parameters $\boldsymbol{\eta}$, respectively. For now, assume that $\mu(\mathbf{s}; \boldsymbol{\beta})$ is a smooth function in $\mathbf{s}$ (as would be needed to do prediction for $Y(\mathbf{s})$). Using MCMC, these model parameters, $\boldsymbol{\theta} = (\boldsymbol{\beta}, \boldsymbol{\eta})$, are available to us as samples, $\{\boldsymbol{\theta}_l\}$, from their posterior distribution $P(\boldsymbol{\theta}|\mathbf{Y})$. Therefore, $P\left(\Gamma_{Y(\mathbf{s})}(t^*)|\mathbf{Y}\right) = \int P\left(\Gamma_{Y(\mathbf{s})}(t^*)|\mathbf{Y}, \boldsymbol{\theta}\right) P(\boldsymbol{\theta}|\mathbf{Y})\, d\boldsymbol{\theta}$ will be obtained by sampling, for each $\boldsymbol{\theta}_l$, $\Gamma^l_{Y(\mathbf{s})}(t^*)$ from $P\left(\Gamma_{Y(\mathbf{s})}(t^*)|\mathbf{Y}, \boldsymbol{\theta}_l\right)$, which, using (13.21), is normally distributed with mean and variance given by

$$\begin{aligned} &\mu_{\Gamma_{Y(\mathbf{s})}}(t^*; \boldsymbol{\beta}_l) - \gamma^T_{\Gamma,\mathbf{Y}}(t^*; \boldsymbol{\eta}_l)\Sigma_Y^{-1}(\boldsymbol{\eta}_l)(\mathbf{Y} - \boldsymbol{\mu}) \text{ and} \\ &K_{\Gamma_{Y(s)}}(t^*, t^*; \boldsymbol{\eta}_l) - \gamma_{\Gamma,\mathbf{Y}}(t^*; \boldsymbol{\eta}_l)^T \Sigma_Y^{-1}(\boldsymbol{\eta}_l)\gamma_{\Gamma,\mathbf{Y}}(t^*; \boldsymbol{\eta}_l), \text{respectively}. \end{aligned} \tag{13.22}$$

In particular, when $C_{t^*} = \{\mathbf{s}_0 + t\mathbf{u} : t \in [0, t^*]\}$, is a line segment of length t^* joining $\mathbf{s}_0$ and $\mathbf{s}_1 = \mathbf{s}_0 + t^*\mathbf{u}$, we have seen below (13.18) that $\Gamma_{Y(s)}(t^*)$ equals $\int_0^{t^*} D_{\mathbf{u}^\perp} Y(\mathbf{s}(t))\, dt$. Thus, defining $\Delta_{0j} = \mathbf{s}_0 - \mathbf{s}_j$, we have

$$\begin{aligned} \mu_{\Gamma_{Y(\mathbf{s})}}(t^*; \boldsymbol{\beta}) &= \int_0^{t^*} \langle \mathbf{u}^\perp, \nabla\mu(\mathbf{s}(t); \boldsymbol{\beta})\rangle dt; \\ (\gamma_{\Gamma,\mathbf{Y}}(t^*; \boldsymbol{\eta}))_j &= \int_0^{t^*} D_{\mathbf{u}^\perp} K(\Delta_{0j} + t\mathbf{u}; \boldsymbol{\eta})\, dt, \ j = 1, \ldots, n; \\ K_{\Gamma_{Y(s)}}(t^*, t^*; \boldsymbol{\eta}) &= \int_0^{t^*}\int_0^{t^*} -\left(\mathbf{u}^\perp\right)^T H_{K(\boldsymbol{\eta})}(\Delta(t_1, t_2)\, \mathbf{u}^\perp dt_1 dt_2. \end{aligned}$$

These integrals need to be computed for each $\boldsymbol{\theta}_l = (\boldsymbol{\beta}_l, \boldsymbol{\eta}_l)$. Though they may not be analytically tractable (depending upon our choice of $\mu(\cdot)$ and $K(\cdot)$), they are one- or two-dimensional integrals that can be efficiently computed using quadrature. Furthermore, since the $\boldsymbol{\theta}_l$'s will already be available, the quadrature calculations (for each $\boldsymbol{\theta}_l$) can be performed ahead of the predictive inference, perhaps using a separate quadrature program, and the output stored in a file for use in the predictive program. The only needed inputs are $\mathbf{s}_0$, $\mathbf{u}$, and the value of t^*. For a specified line segment, we will know these. In fact, for a general curve C, its representation on a given map is as a polygonal curve. As a result, the total or average gradient for C can be obtained through the $\Gamma_{Y(\mathbf{s})}(t^*)$ associated with the line segments that comprise C.

Specifically, with GIS software we can easily extract (at high resolution) the coordinates along the boundary, thus approximating C by line segments connecting adjacent points. Thus, $C = \cup_{k=1}^M C_k$ where C_k's are virtually disjoint (only one common point at the "join") line segments, and $\nu(C) = \sum_{k=1}^M \nu(C_k)$. If we parametrize each line segment as above and compute the line integral along each C_k by the above steps, the total gradient is the sum of the piece-wise line integrals. To be precise, if $\Gamma_k(\mathbf{s}(t_k^*))$ is the line-integral process on the linear segment C_k, we will obtain predictive samples, $\Gamma_k^{(l)}(\mathbf{s}(t_k^*))$ from each

$P\left(\Gamma_k\left(\mathbf{s}\left(t_k^*\right)\right)|\mathbf{Y}\right)$, $k = 1, \ldots, M$. Inference on the average gradient along C will stem from posterior samples of

$$\frac{1}{\sum_{k=1}^{M} \nu\left(C_k\right)} \sum_{k=1}^{M} \Gamma_k^{(l)}\left(\mathbf{s}\left(t_k^*\right)\right).$$

Thus, with regard to boundary analysis, wombling measure reduces a curve to an average gradient and inference to examination of the posterior of the average gradient.

When $Y(\mathbf{s})$ is a Gaussian process with constant mean, $\mu(\mathbf{s}) = \mu$ and an isotropic correlation function $K(\|\Delta\|; \sigma^2, \phi) = \sigma^2 \exp(-\phi\|\Delta\|^2)$, calculations simplify. We have $\nabla\mu(\mathbf{s}) = 0$, $\mu_\Gamma(t^*) = 0$, and

$$H_K(\Delta) = -2\sigma^2\phi \exp(-\phi\|\Delta\|^2)(I - 2\phi\Delta\Delta^T)\ .$$

Further calculations reveal that $(\gamma_{YZ}(t^*; \sigma^2, \phi))_j$ can be computed as,

$$c(\sigma^2, \phi, \mathbf{u}^\perp, \Delta_{0j})(\Phi(\sqrt{2\phi}(t^* + \langle \mathbf{u}, \Delta_{0j}\rangle)) - \Phi(\sqrt{2\phi}\langle \mathbf{u}, \Delta_{0j}\rangle)), \tag{13.23}$$

where $\Phi()$ is the standard Gaussian cumulative distribution function, and

$$c(\sigma^2, \phi, \mathbf{u}^\perp, \Delta_{0j}) = -2\sigma^2\sqrt{\pi\phi}\langle \mathbf{u}^\perp, \Delta_{0j}\rangle \exp(-\phi|\langle \mathbf{u}^\perp, \Delta_{0j}\rangle|^2).$$

These computations can be performed using the Gaussian cdf function, and quadrature is needed only for $Var(\Gamma_{Y(\mathbf{s})}(t^*))$.

Returning to the model $Y(\mathbf{s}) = \mathbf{x}^T(\mathbf{s})\boldsymbol{\beta} + w(\mathbf{s}) + \epsilon(\mathbf{s})$ with $\mathbf{x}(\mathbf{s})$ a general covariate vector, $w(\mathbf{s}) \sim GP(0, \sigma^2\rho(\cdot, \phi))$ and $\epsilon(\mathbf{s})$ a zero-centered white-noise process with variance τ^2, consider boundary analysis for the residual surface $w(\mathbf{s})$. In fact, boundary analysis on the spatial residual surface is feasible in generalized linear modelling contexts with exponential families, where $w(\mathbf{s})$ may be looked upon as a non-parametric latent structure in the mean of the parent process.

Denoting by $\Gamma_{w(\mathbf{s})}(t)$ and $\bar{\Gamma}_{w(\mathbf{s})}(t)$ as the total and average gradient processes (as defined in (13.18)) for $w(\mathbf{s})$, we seek the posterior distributions $P(\Gamma_{w(\mathbf{s})}(t^*)|\mathbf{Y})$ and $P(\bar{\Gamma}_{w(\mathbf{s})}(t^*)|\mathbf{Y})$. Note that

$$P(\Gamma_{w(\mathbf{s})}(t^*)|\mathbf{Y}) = \int P(\Gamma_{w(\mathbf{s})}(t^*)|\mathbf{w}, \boldsymbol{\theta})P(\mathbf{w}|\boldsymbol{\theta}, \mathbf{Y})P(\boldsymbol{\theta}|\mathbf{Y})d\boldsymbol{\theta}d\mathbf{w}, \tag{13.24}$$

where $\mathbf{w} = (w(s_1), \ldots, w(s_n))$ denotes a realization of the residual process and $\boldsymbol{\theta} = (\boldsymbol{\beta}, \sigma^2, \phi, \tau^2)$. Sampling of this distribution again proceeds in a posterior-predictive fashion using posterior samples of $\boldsymbol{\theta}$, and is expedited in a Gaussian setting since $P(\mathbf{w}|\boldsymbol{\theta}, \mathbf{Y})$ and $P(\Gamma_{w(\mathbf{s})}(t^*)|\mathbf{w}, \boldsymbol{\theta})$ are both Gaussian distributions.

Formal inference for a wombling boundary is done more naturally on the residual surface $w(\mathbf{s})$, i.e., for $\Gamma_{w(\mathbf{s})}(t^*)$ and $\bar{\Gamma}_{w(\mathbf{s})}(t^*)$, because $w(\mathbf{s})$ is the surface containing any non-systematic spatial information on the parent process $Y(\mathbf{s})$. Since $w(\mathbf{s})$ is a zero-mean process, $\mu_{\Gamma_{w(\mathbf{s})}}(t^*; \boldsymbol{\beta}) = 0$, and thus one needs to check for the inclusion of this null value in the resulting 95% credible intervals for $\Gamma_{w(\mathbf{s})}(t^*)$ or, equivalently, for $\bar{\Gamma}_{w(\mathbf{s})}(t^*)$. Again, this clarifies the issue of the normal direction mentioned in Subsection 13.7.2; significance using $n(\mathbf{s}(t))$ is equivalent to significance using $-n(\mathbf{s}(t))$. One only needs to select and maintain a particular orthogonal direction relative to the curve. In accord with our remarks concerning absolute gradients in Subsection 13.7.2, we could compute (13.19) using $|D_{n(\mathbf{s}(t))}Y(\mathbf{s}(t))|$ using a Riemann sum, but would be computationally expensive and would not offer a Gaussian calibration of significance.

13.9 Illustration: Spatial boundaries for invasive plant species

Banerjee and Gelfand (2006) consider data collected from 603 locations in Connecticut with presence/absence and abundance scores for some individual invasive plant species, plus

Parameters	50% (2.5%,97.5%)
Intercept	0.983 (-2.619, 4.482)
Habitat Class (Baseline: Type 1)	
Type 2	-0.660 (-1.044,-0.409)
Type 3	-0.553 (-1.254, 0.751)
Type 4	-0.400 (-0.804,-0.145)
Land use Land Cover Types (Baseline: Level 1)	
Type 2	0.591 (0.094, 1.305)
Type 3	1.434 (0.946, 2.269)
Type 4	1.425 (0.982, 1.974)
Type 5	1.692 (0.934, 2.384)
1970 Category Types (Baseline: Category 1)	
Category 2	-4.394 (-6.169,-3.090)
Category 3	-0.104 (-0.504, 0.226)
Category 4	1.217 (0.864, 1.588)
Category 5	-0.039 (-0.316, 0.154)
Category 6	0.613 (0.123, 1.006)
Canopy Closure	0.337 (0.174, 0.459)
Heavily Managed Points (Baseline: No)	
Yes	-1.545 (-2.027,-0.975)
Log Edge Distance	-1.501 (-1.891,-1.194)
σ^2	8.629 (7.005, 18.401)
ϕ	1.75E-3 (1.14E-3, 3.03E-3)
ν	1.496 (1.102, 1.839)
Range (in meters)	1109.3 (632.8, 1741.7)

Table 13.3 *Parameter estimates for the logistic spatial regression example.*

environmental covariates. The covariates are available only at the sample locations, not on a grid. The response variable $Y(\mathbf{s})$ is a presence-absence binary indicator (0 for absence) for one species *Celastrus orbiculatus* at location $\mathbf{s}$. There are four categorical covariates: habitat class (representing the current state of the habitat) of four different types, land use and land cover (LULC) types (Land use/cover history of the location, e.g., always forest, formerly pasture now forest, etc.) at five levels and a 1970 category number (LULC at one point in the past: 1970, e.g., forest, pasture, residential etc.) with six levels. In addition we have an ordinal covariate, canopy closure percentage (percent of the sky that is blocked by "canopy" of leaves of trees. A location under mature forest would have close to 100% canopy closure while a forest edge would have closer to 25%) with four levels in increasing order, a binary variable for heavily managed points (0 if "no"; "heavy management" implies active landscaping or lawn mowing) and a continuous variable measuring the distance from the forest edge in the logarithm scale. Figure 13.2 is a digital terrain image of the study domain, with the labelled curves indicating forest edges extracted using the GIS software `ArcView` (`http://www.esri.com/`). Ecologists are interested in evaluating spatial gradients along these 10 natural curves and identifying them as wombling boundaries.

We fit a logistic regression model with spatial random effects,

$$\log\left(\frac{P(Y(\mathbf{s})=1)}{P(Y(\mathbf{s})=0)}\right) = \mathbf{x}^T(\mathbf{s})\boldsymbol{\beta} + w(\mathbf{s}),$$

where $\mathbf{x}(\mathbf{s})$ is the vector of covariates observed at location $\mathbf{s}$ and $w(\mathbf{s}) \sim GP(0, \sigma^2\rho(\cdot;\phi,\nu))$ is a Gaussian process with $\rho(\cdot;\phi,\nu)$ as a Matérn correlation function. While $Y(\mathbf{s})$ is a binary surface that does not admit gradients, conducting boundary analysis on $w(\mathbf{s})$ is perfectly

Figure 13.2 *A digital image of the study domain in Connecticut indicating the forest edges as marked curves. These are assessed for significant gradients. Note: Eastings range from 699148 to 708961; Northings range from 4604089 to 4615875 for the picture.*

legitimate. The residual spatial surface reflects unmeasured or unobservable environmental features in the mean surface. Again, attaching curvilinear wombling boundaries to the residual surface tracks rapid change in the departure from the mean surface.

We adopt a completely non-informative flat prior for $\boldsymbol{\beta}$, an inverse gamma $IG(2, 0.001)$ prior for σ^2 and the Matérn correlation function with a gamma prior for the correlation decay parameter, ϕ, specified so that the prior spatial range has a mean of about half of the observed maximum inter-site distance (the maximum distance is 11887 meters based on a UTM projection), and a $U(1, 2)$ prior for the smoothness parameter ν. Again, three parallel MCMC chains were run for 15000 iterations each and 10000 iterations revealed sufficient mixing of the chains, with the remaining 15000 samples (5000×3) being used for posterior analysis.

Table 13.3 presents the posterior estimates of the model parameters. We do not have a statistically significant intercept, but most of the categorical variables reveal significance: Types 2 and 4 for habitat class have significantly different effects from Type 1; all the four types of LULC show significant departure from the baseline Type 1; for the 1970 category number, category 2 shows a significant negative effect, while categories 4 and 6 show significant positive effects compared to category 1. Canopy closure is significantly positive, implying higher presence probabilities of *Celastrus orbiculatus* with higher canopy blockage, while points that are more heavily managed appear to have a significantly lower probability of species presence as does the distance from the nearest forest edge. Posterior summaries of the spatial process parameters are also presented and the effective spatial range is approximated to be around 1109.3 meters approximately. These produce the mean posterior surface of $w(\mathbf{s})$, shown in Figure 13.3 with the 20 endpoints of the forest edges from Figure 13.2 labelled to connect the figures.

Finally, in Table 13.4, we present the formal curvilinear gradient analysis for the 10 forest edges in Figure 13.2. We find that 6 out of the 10 edge curves (with the exception of CD, EF, KL and MN) are formally tested to be wombling boundaries. Our methodology proves

Figure 13.3 *The spatial residual surface from the presence-absence application for the 603 observed locations (not shown) in Connecticut. Also shown are the 20 endpoints for the 10 curves in Figure 13.2 to connect the figures.*

Curve	Average Gradient	Curve	Average Gradient
AB	1.021 (0.912, 1.116)	KL	0.036 (-0.154, 0.202)
CD	0.131 (-0.031, 0.273)	MN	0.005 (-0.021, 0.028)
EF	0.037 (-0.157, 0.207)	OP	0.227 (0.087, 0.349)
GH	1.538 (1.343, 1.707)	QR	0.282 (0.118, 0.424)
IJ	0.586 (0.136, 0.978)	ST	0.070 (0.017, 0.117)

Table 13.4 *Curvilinear gradient assessment for the 10 forest edges labelled in Figure 13.2 for the logistic regression example.*

useful here since some of these edge curves meander along the terrain for substantially long distances. Indeed, while the residual surface in Figure 13.3 reveals a general pattern of spatial variation (higher residuals in the South), it is difficult to make visual assessments on the size (and significance) of average gradients for the longer curves. Furthermore, with non-gridded data as here, the surface interpolators (in this case the Akima (1996) interpolator in `R`) often find it difficult to extrapolate beyond a convex hull of the site locations. Consequently, parts of the curve (e.g., endpoints C, G and (almost) T) lie outside the fitted surface, making local visual assessment on them impossible.

Quickly and reliably identifying forest edges could be useful in determining boundaries between areas of substantial anthropogenic activity and minimally managed forest habitats. Such boundaries are important because locations at which forest blocks have not been invaded by exotic plant species may be subject to significant seed rain from these species. These boundaries thus might form important "front lines" for efforts to monitor or control invasive species.

13.10 Areal wombling

Areal wombling refers to ascertaining boundaries on *areally* referenced data. Such methods are valuable in determining boundaries for data sets that, perhaps due to confidentiality

concerns, are available only in ecological (aggregated) format, or are only collected this way (e.g., delivery of health care or cost information). In this case, since we lack smooth realizations of spatial surfaces, areal wombling cannot employ spatial gradients. Here the gradient is not explicitly modeled; instead, boundary effects are looked upon as edge effects and modeled using Markov random field specifications. Boundaries in areal wombling are just a collection of segments (or *arcs*, in geographic information systems (GIS) parlance) dually indexed by ij, corresponding to the two adjacent regions i and j the segment separates. In the fields of image analysis and pattern recognition, there has been much research in using statistical models for capturing "edge" and "line" effects (see, e.g., Geman and Geman, 1984; Geman and McClure, 1985, 1987; Besag, 1974; Helterbrand, Cressie and Davidson, 1984; see also Cressie, 1993 (Sec 7.4) and references therein). Such models are based upon probability distributions such as Gibbs distributions or Markov random fields (see Chapters 12 and 13 and Rue and Held, 2004) that model pixel intensities as conditional dependencies using the neighborhood structure (see, e.g., Chellappa and Jain, 1993). Modeling objectives include identification of edges based upon distinctly different image-intensities in adjacent pixels.

As seen earlier in this book, in areal models, local spatial dependence between the observed image characteristics is captured by a *neighborhood structure*, where a pixel is independent of the rest given the values of its neighbors. Various neighborhood structures are possible, but all propose stronger statistical dependence between data values from areas that are spatially closer, thus inducing local smoothing. However, in the context of areal wombling this leads to a new problem: when real discontinuities (boundaries) exist between neighboring pixels, MRF models tend to smooth across them, thus blurring the very edges we hope to detect.

Although the boundary analysis problem for public health data resembles the edge-detection problem in image processing, significant differences exist. Unlike image pixels, geographical maps that form the domain of most public health data are not regularly spaced but still have a well-defined neighborhood structure (a topological graph). Furthermore, there are usually far fewer of these areas than the number of pixels that would arise in a typical image restoration problem, so we have far less data. Finally, the areal units (polygons) are often quite different in size, shape, and number of neighbors, leading, for example, to different degrees of smoothing in urban and rural regions, as well as near the external boundary of the study region.

In this section, after a brief review of existing algorithmic techniques, we propose a variety of fully model-based frameworks for areal wombling, using Bayesian hierarchical models with posterior summaries computed using MCMC methods. We explore the suitability of various existing hierarchical and spatial software packages (notably `R` and `WinBUGS`) to the task, and indicate the approaches' superiority over existing non-stochastic alternatives. We also illustrate our methods (as well as the solution of advanced modeling issues such as simultaneous inference) using county-level cancer surveillance and zip code-level hospice utilization data in the state of Minnesota.

13.10.1 Review of existing methods

Areal wombling (also known as *polygonal* wombling) is not as well-developed in the literature as point or raster wombling, but some notable papers exist. Oden et al. (1993) provide a wombling algorithm for multivariate categorical data defined on a lattice. The statistic chosen is the average proportion of category mismatches at each pair of neighboring sites, with significance relative to an independence or particular spatial null distribution judged by a randomization test. Csillag et al. (2001) developed a procedure for characterizing the strength of boundaries examined at neighborhood level. In this method, a topological or a metric distance δ defines a neighborhood of the candidate set of polygons (say, p_i). A weighted local statistic is attached to each p_i. The difference statistic calculated as

the squared difference between any two sets of polygons' local statistic and its quantile measure are used as a relative measure of the distinctiveness of the boundary at the scale of neighborhood size δ. Jacquez and Greiling (2003) estimate boundaries of rapid change for colorectal, lung, and breast cancer incidence in Nassau, Suffolk, and Queens counties in New York.

Classical boundary analysis research often proceeds by selecting a *dissimilarity metric* (say, Euclidean distance) to measure the difference in response between the values at (say) adjacent polygon centroids. An absolute (dissimilarity metrics greater than C) or relative (dissimilarity metrics in the top $k\%$) threshold then determines which borders are considered actual barriers, or parts of the boundary. The relative (top $k\%$) thresholding method for determining boundary elements is easily criticized, since for a given threshold, a fixed number of boundary elements will always be found regardless of whether or not the responses separated by the boundary are statistically different. Jacquez and Maruca (1998) suggest use of both local and global statistics to determine where statistically significant boundary elements are, and a randomization test (with or without spatial constraints) for whether the boundaries for the entire surface are statistically unusual.

These older areal wombling approaches, including the algorithm implemented in the `BoundarySeer` software (`http://www.terraseer.com`), tend to be algorithmic, rather than model-based. That is, they do not involve a probability distribution for the data, and therefore permit statements about the "significance" of a detected boundary only relative to predetermined, often unrealistic null distributions. Naturally, we prefer a hierarchical statistical modeling framework for areal wombling, to permit direct estimation of the probability that two geographic regions are separated by the wombled boundary. Such models can account for spatial and/or temporal association and permit the borrowing of strength across different levels of the model hierarchy.

13.10.2 Simple MRF-based areal wombling

Suppose we have regions $i = 1, \ldots, N$ along with the areal adjacency matrix A, and we have observed a response Y_i (e.g., a disease count or rate) for the ith region. Traditional areal wombling algorithms assign a *boundary likelihood value* (BLV) to each areal boundary using a Euclidean distance metric between neighboring observations. This distance is taken as the dissimilarity metric, calculated for each pair of adjacent regions. Thus, if i and j are neighbors, the BLV associated with the edge (i, j) is

$$D_{ij} = ||Y_i - Y_j|| \ ,$$

where $|| \cdot ||$ is a distance metric. Locations with higher BLV's are more likely to be a part of a difference boundary, since the variable changes rapidly there.

The wombling literature and attendant software further distinguishes between *crisp* and *fuzzy* wombling. In the former, BLV's exceeding specified thresholds are assigned a *boundary membership value* (BMV) of 1, so the wombled boundary is $\{(i, j) : ||Y_i - Y_j|| > c,\ i \text{ adjacent to } j\}$ for some $c > 0$. The resulting edges are called *boundary elements* (BE's). In the latter (fuzzy) case, BMV's can range between 0 and 1 (say, $||Y_i - Y_j|| \,/\, \max_{ij}\{||Y_i - Y_j||\}$) and indicate *partial* membership in the boundary.

For our hierarchical modeling framework, suppose we employ the usual Poisson log-linear form for observed and expected disease counts Y_i and E_i,

$$Y_i \sim Poisson(\mu_i) \text{ where } \log \mu_i = \log E_i + \mathbf{x}_i' \boldsymbol{\beta} + \phi_i \ . \tag{13.25}$$

This model allows a vector of region-specific covariates $\mathbf{x}_i$ (if available), and a random effect vector $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_N)'$ that is given a conditionally autoregressive (CAR) specification. As

seen earlier, the intrinsic CAR, or IAR, form has improper joint distribution, but intuitive conditional distributions of the form

$$\phi_i \mid \boldsymbol{\phi}_{j \neq i} \sim N(\bar{\phi}_i \, , \, 1/(\tau m_i)) \, , \tag{13.26}$$

where N denotes the normal distribution, $\bar{\phi}_i$ is the average of the $\boldsymbol{\phi}_{j \neq i}$ that are adjacent to ϕ_i, and m_i is the number of these adjacencies. Finally, τ is typically set equal to some fixed value, or assigned a distribution itself (usually a relatively vague gamma distribution).

As we have seen, MCMC samples $\mu_i^{(g)}$, $g = 1, \ldots, G$ from the marginal posterior distribution $p(\mu_i|\mathbf{y})$ can be obtained for each i, from which corresponding samples of the (theoretical) standardized morbidity ratio,

$$\eta_i = \frac{\mu_i}{E_i}, \; i = 1, \ldots, N,$$

are immediately obtained. We may then define the BLV for boundary (i, j) as

$$\Delta_{ij} = |\eta_i - \eta_j| \quad \text{for all } i \text{ adjacent to } j \, . \tag{13.27}$$

Crisp and fuzzy wombling boundaries are then based upon the posterior distribution of the BLV's. In the crisp case, we might define ij to be part of the boundary if and only if $E(\Delta_{ij}|\mathbf{y}) > c$ for some constant $c > 0$, or if and only if $P(\Delta_{ij} \geq c \mid \mathbf{y}) > c^*$ for some constant $0 < c^* < 1$.

Model (13.25)–(13.26) can be easily implemented in the `WinBUGS` software package. The case of *multivariate* response variables requires multivariate CAR (MCAR) models; see Chapter 10 above, as well as Ma and Carlin (2007).

Posterior draws $\{\eta_i^{(g)}, g = 1, \ldots, G\}$ and their sample means $\widehat{E}(\eta_i|\mathbf{y}) = \frac{1}{G}\sum_{g=1}^{G} \eta_i^{(g)}$ are easily obtained for our problem. Bayesian areal wombling would naturally obtain posterior draws of the BLV's in (13.27) by simple transformation as $\Delta_{ij}^{(g)} = |\eta_i^{(g)} - \eta_j^{(g)}|$, and then base the boundaries on their empirical distribution. For instance, we might estimate the posterior means as

$$\widehat{E}(\Delta_{ij}|\mathbf{y}) = \frac{1}{G}\sum_{g=1}^{G} \Delta_{ij}^{(g)} = \frac{1}{G}\sum_{g=1}^{G} |\eta_i^{(g)} - \eta_j^{(g)}| \, , \tag{13.28}$$

and take as our wombled boundaries the borders corresponding to the top 20% or 50% of these values.

Turning to fuzzy wombling, suppose we select a cutoff c such that, were we *certain* a particular BLV exceeded c, we would also be certain the corresponding segment was part of the boundary. Since our statistical model (13.25)–(13.26) delivers the full posterior distribution of every Δ_{ij}, we can compute $P(\Delta_{ij} > c \mid \mathbf{y})$, and take this probability as our fuzzy BMV for segment ij. In fact, the availability of the posterior distribution provides another benefit: a way to directly assess the *uncertainty* in our fuzzy BMV's. Our Monte Carlo estimate of $P(\Delta_{ij} > c \mid \mathbf{y})$ is

$$\hat{p}_{ij} \equiv \hat{P}(\Delta_{ij} > c \mid \mathbf{y}) = \frac{\#\Delta_{ij}^{(g)} > c}{G} \, . \tag{13.29}$$

This is nothing but a binomial proportion; were its components independent, basic binomial theory implies an approximate standard error for it would be

$$\widehat{se}(\hat{p}_{ij}) = \sqrt{\frac{\hat{p}_{ij}(1 - \hat{p}_{ij})}{G}} \, . \tag{13.30}$$

Of course, our Gibbs samples Δ_{ij} are *not* independent in general, since they arise from a Markov chain, but we can make them approximately so by subsampling, retaining only every M^{th} sample. Note that this subsampling does *not* remove the *spatial* dependence among the Δ_{ij}, so repeated use of formula (13.30) would not be appropriate if we wanted to make a *joint* probability statement involving more than one of the Δ_{ij} at the same time; Subsection 13.10.4 revisits this "simultaneous inference" issue.

Example 13.4 Minnesota breast cancer late detection data. As an illustration, we consider boundary analysis for a dataset recording the rate of late detection of several cancers collected by the Minnesota Cancer Surveillance System (MCSS), a population-based cancer registry maintained by the Minnesota Department of Health. The MCSS collects information on geographic location and stage at detection for colorectal, prostate, lung and breast cancers. For each county, the late detection rate is defined as the number of regional or distant case detections divided by the total cases observed, for the years 1995 to 1997. Since higher late detection rates are indicative of possibly poorer cancer control in a county, a wombled boundary for this map might help identify barriers separating counties with different cancer control methods. Such a boundary might also motivate more careful study of the counties that separate it, in order to identify previously unknown covariates (population characteristics, dominant employment type, etc.) that explain the difference.

In this example, we consider n_i, the total number of breast cancers occurring in county i, and Y_i, the number of these that were detected late (i.e., at regional or distant stage). To correct for the differing number of detections (i.e., the total population) in each county, we womble not on the Y scale, but on the *standardized late detection ratio* (SLDR) scale, $\eta_i = \mu_i/E_i$, where the expected counts E_i are computed via internal standardization as $E_i = n_i\bar{r}$, where $\bar{r} = \sum_i Y_i / \sum_i n_i$, the statewide late detection rate.

As discussed above, areal wombling is naturally based here on the absolute SLDR differences $\Delta_{ij} = |\eta_i - \eta_j|$; we refer to this as *mean-based* wombling. But one might redefine the BLV's $\Delta_{ij} = |\phi_i - \phi_j|$, resulting in *residual-based* wombling. Comparison of the mean and residual based wombling maps may provide epidemiologists with information regarding barriers that separate regions with different cancer prevention and control patterns. For instance, if segment ij is picked up as a BE in the mean-based map but not in the residual-based map, this suggests that the difference between the two regions in the response counts may be due to the differences in their fixed effects. If on the other hand segment ij is a BE in the residual-based map but not the mean-based map, this may suggest a boundary between region i and j caused by excess spatial heterogeneity, possibly indicating missing spatial covariates.

Our model assumes the mean structure $\log \mu_i = \log E_i + \beta_1(x_i - \overline{x}) + \phi_i$, where x_i is the average annual age-adjusted cancer mortality rate in county i for 1993–1997, and with $\boldsymbol{\phi}$ following the usual CAR prior with fixed adjacency matrix. Regarding prior distributions for this model, we selected a $N(0,1)$ prior for β_1 and a $G(0.1, 0.1)$ prior for τ, choices designed to be vague enough to allow the data to dominate determination of the posterior.

Figure 13.4 shows the resulting crisp wombling maps based on the top 20% of the BLV's. The county shading (blue for low values and red for high values) indicates the posterior mean SLDR η_i or residual ϕ_i, while the boundary segments are shown as dark lines. The residual-based map indicates many boundaries in the southern borders of the state and also in the more thinly-populated north region of the state, an area with large farms and native American reservation land. County 63 (Red Lake, a T-shaped county in the northwest) seems "isolated" from its two neighbors, a finding also noted by Lu and Carlin (2005). The mean map seems to identify the regions with differing fixed effects for mortality.

Note that wombling on the spatial residuals ϕ_i instead of the fitted SLDR's η_i changes not only the scale of the c cutoff (to difference in *log*-relative risk), but the interpretation of the results as well. Specifically, we can borrow an interpretation often mentioned in spatial

Figure 13.4 *Lu and Carlin crisp (top 20%) wombled maps based on SLDR's η_i (left) and residuals ϕ_i (right).*

epidemiology regarding spatially oriented covariates $\mathbf{x}_i$ still missing from model (13.25). Since boundaries based on the ϕ_i separate regions that differ in their unmodeled spatial heterogeneity, a careful comparison of such regions identified by the wombled map should prove the most fruitful in any missing covariate search.

13.10.2.1 Adding covariates

Lu et al. (2007) develop randomly weighted hierarchical areal wombling models that allow the data to help determine the degree and nature of spatial smoothing to perform areal wombling. Specifically, they propose a model that allows the choice of the neighborhood structure to be determined by the value of the process in each region and by variables determining the similarity of two regions. This approach is natural for Bayesian areal wombling since detecting boundary membership can be viewed as the dual problem of regional estimation using adjacency information. Such a method permits the data (and perhaps other observed covariate information) to help determine the neighborhood structure and the degree and nature of spatial smoothing, while simultaneously offering a new stochastic definition of the boundary elements.

Recall that for the standard CAR model, the spatial random effects $\boldsymbol{\phi}$ have conditional distribution

$$\phi_i \mid \boldsymbol{\phi}_{(-i)} \sim N\left(\frac{\sum_j w_{ij}\phi_j}{\sum_j w_{ij}}\ ,\ \frac{1}{\tau \sum_j w_{ij}}\right)\ , \tag{13.31}$$

where N denotes the normal distribution. Here the neighborhood weights w_{ij} are traditionally fixed at 1 or 0 based on whether regions i and j share a common geographic boundary; this is the approach taken above. Now, rather than fix the w_{ij}, we model them as

$$w_{ij}|p_{ij} \overset{indep}{\sim} \text{Bernoulli}(p_{ij})\ ,\ \text{where}\ \ \log\left(\frac{p_{ij}}{1-p_{ij}}\right) = \mathbf{z}'_{ij}\boldsymbol{\gamma}\ . \tag{13.32}$$

Here $\mathbf{z}_{ij}$ is a set of known features of the $(i,j)^{th}$ pair of regions, with corresponding parameter vector $\boldsymbol{\gamma}$. Regions i and j are thus considered to be neighbors with probability p_{ij} provided they share a common geographical boundary; $w_{ij} = 0$ for all non-adjacent regions. Now that the w_{ij} are random, one can also draw crisp wombled maps based on the $1 - w_{ij}$ posteriors, offering a third, clearly variance-based alternative to the mean- and residual-based methods illustrated in Figure 13.4.

A fairly standard formulation of the CAR prior (Best et al., 1999; c.f. Section 4.3 of this book) takes the form

$$\boldsymbol{\phi} \sim N(0, (I - C)^{-1}M) \ , \tag{13.33}$$

where $C_{ij} = \frac{w_{ij}}{\sum_{j\sim i}^{m_i} w_{ij}}$, and M is a diagonal matrix with elements $M_{ii} = \frac{1}{\tau \sum_{j\sim i}^{m_i} w_{ij}}$, the conditional variances of the ϕ_i. The precision matrix $B = M^{-1}(I - C)$ is often rewritten as $B = \tau Q$ with $Q = (D_W - W)$, W being the $N \times N$ matrix of weights w_{ij}, and $D_W = \text{Diag}(\sum_{j\sim i}^{m_i} w_{ij})$. Unfortunately, because $(D_W - W)\mathbf{1} = 0$, i.e., the precision matrix is singular, the distribution is improper. As seen above, the common remedies for this problem are to constrain $\sum_{i=1}^{N} \phi_i = 0$ to identify the overall intercept term in the model, or to include a "propriety parameter" α in the precision matrix B, i.e.,

$$\boldsymbol{\phi} \sim N(\mu, (I - \alpha C)^{-1}M) \ . \tag{13.34}$$

For $|\alpha| < 1$, the covariance matrix $(I - \alpha C)^{-1}M$ is positive definite, ensuring $\boldsymbol{\phi}$ has a proper distribution provided every region has at least one neighbor. However, when some region has no neighbors (i.e., $\sum_{j\sim i}^{m_i} w_{ij} = 0$ for some i), the covariance matrix is again not well defined.

A wide variety of covariates might be considered for $\mathbf{z}_{ij}$ in (13.32), containing information about the difference between regions i and j that might impact neighbor status. For instance, $\mathbf{z}_{ij}$ could be purely map-based, such as the distance between the centroids of regions i and j, or the difference between the percentage of common boundaries shared by the two regions among their own total geographical boundaries. Auxiliary topological information, such as the presence of a mountain range or river across which travel is difficult, might also be important here. Alternatively, $\mathbf{z}_{ij}$ may include sociodemographic information, such as the difference between the regions' percentage of urban area, or the absolute difference of some regional covariate (say, the percent of residents who are smokers, or even the region's expected age-adjusted disease count).

Vague priors are typically chosen for $\boldsymbol{\beta}$ in the mean structure, and a conventional gamma prior can be used for the precision parameter τ. However, it is not clear how to specify the prior for $\boldsymbol{\gamma}$. If the covariates $\mathbf{z}_{ij}$ measure dissimilarity between regions i and j, common sense might suggest an informative prior on $\boldsymbol{\gamma}$ that favors negative values, because the more dissimilar two regions are, the less likely that they should be thought of as neighbors. The induced prior on p_{ij} could be fairly flat even if $\boldsymbol{\gamma}$'s prior has small variance, due to the logit transformation. If the proper CAR prior (13.34) is used, either a noninformative (e.g., Unif$(0, 1)$) or an informative (e.g., Beta$(18, 2)$) prior could be used for α, depending in part on one's tolerance for slow MCMC convergence. The covariate effect $\boldsymbol{\gamma}$ is identified by the data, even under a noninformative prior distribution for $\boldsymbol{\gamma}$; however, as with α, moderately informative priors are often used.

13.10.3 Joint site-edge areal wombling

The method of Lu and Carlin (2005) is very easy to understand and implement, but suffers from the oversmoothing problems mentioned above. It also generally fails to produce the long series of connected boundary segments often desired by practitioners. In this subsection, we suggest a variety of hierarchical models for areal boundary analysis that hierarchically or jointly parameterize *both* the areas and the edge segments. The approach uses a compound Gaussian Markov random field model that adopts an *Ising* distribution as the prior on the edges. This leads to conceptually appealing solutions for our data that remain computationally feasible. While our approaches parallel similar developments in statistical image restoration using Markov random fields, important differences arise due to the irregular

(a) Area Adjacency, Regular Lattice

(b) Edge Adjacency, Regular Lattice

Figure 13.5 *Illustration of area and edge domain neighborhood structures: (a) areal neighborhood structure, regular idealized map; (b) edge neighborhood structure, regular idealized map. In panel (b), note that edges (such as (6,10) and (7,11)) can be neighbors even though their index sets have no areal units in common.*

nature of our lattices, the sparseness and high variability of our data, the existence of important covariate information, and most importantly, our desire for full posterior inference on the boundary. We will illustrate the performance of these methods compared to the basic Lu and Carlin approach in the context of a zip code-level hospice utilization dataset. Our description basically follows that in Ma et al. (2010), but will bring in other references as appropriate.

13.10.3.1 Edge smoothing and random neighborhood structure

We begin by extending our notion of adjacency to the edge domain. Figure 13.5(a) illustrates this neighborhood structure on an idealized regular lattice. The dark square (Region 7) has 4 neighbors (Regions 3, 6, 8, and 11, shaded light gray). In this case we have $w_{i+} = m_i$, the number of neighbors for region i, so the conditional distribution has mean $\bar{\phi}_i$, the average of the neighboring ϕ_j's, and variance inversely proportional to m_i.

Ma et al. (2006) proposed direct modeling in the edge domain, where the basic data elements are assumed to arise on the edge segments themselves. A CAR model for the edge segments is adopted to favor connected boundaries. For example, in Figure 13.5(b), the thick black boundary corresponding to edge (7,11) has six "neighboring" edges, highlighted as thick gray lines. Thus edge segments are adjacent if and only if they connect to one another. Note that edges (6,10) and (8,12) are adjacent to edge (7,11) even though these segments have no areal units in common.

13.10.3.2 Two-level CAR model

As mentioned in the previous subsection,the edge elements in the adjacency matrix can be modeled as random, potentially offering a natural framework for areal wombling. Since we prefer connected boundaries, given that a particular edge segment is part of the boundary, we would like our model to favor the inclusion of neighboring edge segments in the boundary as well. The standard, 0-1 adjacency-based CAR model appears naturally suited to this task: all we require is a *second* CAR model on the edge space (in addition to the original CAR on the areal unit space) with edge adjacency matrix W^* determined by the regional map as illustrated in Figure 13.5.

Let us explore this *two-level hierarchical CAR* (CAR2) model in the case of Poisson data. Similar to the approach in LC, we start with

$$\begin{aligned} Y_i \mid \boldsymbol{\beta}, \phi_i &\overset{ind}{\sim} Poisson(\mu_i) \\ \text{where } \log(\mu_i) &= \log(E_i) + \mathbf{x}_i'\boldsymbol{\beta} + \phi_i \ , \ \ i = 1, \ldots, n \ , \qquad (13.35) \\ \text{and } p(\boldsymbol{\phi} \mid \tau_\phi, W) &= C(\tau_\phi, W) \exp\left\{-\frac{\tau_\phi}{2}\boldsymbol{\phi}'(D_w - W)\boldsymbol{\phi}\right\} \ , \qquad (13.36) \end{aligned}$$

where $C(\tau_\phi, W)$ is an unknown normalizing constant. We then augment model (13.32) to

$$w_{ij}|p_{ij} \sim Bernoulli(p_{ij}) \ \text{ and } \ logit(p_{ij}) = \mathbf{z}_{ij}'\boldsymbol{\gamma} + \theta_{ij} \ , \qquad (13.37)$$

where θ_{ij} is a spatial random effect associated with the edge separating areas i and j. Note that in our random W setting, if two regions i and j are neighbors (i.e., $w_{ij} = 1$) then they must also be adjacent, but the converse need not be true. Because of the symmetry of W, we need only be concerned with its upper triangle.

For consistency with previous notation, we reorder the w_{ij} into the singly-indexed vector $\boldsymbol{\xi} = (\xi_1, \ldots \xi_K)'$, where K is the number of regional adjacencies in the map. We also carry out a corresponding reordering of the θ_{ij} into a vector $\boldsymbol{\psi} = (\psi_1, \ldots \psi_K)'$. We then place the second-level CAR as the prior on the edge random effects, i.e.,

$$\boldsymbol{\psi} \mid \tau_\psi \sim CAR(\tau_\psi, W^*) \ , \qquad (13.38)$$

so that ψ_k has conditional distribution $N(\bar{\psi}_k \ , \ 1/(\tau_\psi w_{k+}^*))$, where $\tau_\psi > 0$ and W^* is the fixed $K \times K$ 0-1 adjacency matrix for $\boldsymbol{\psi}$, determined as in Figure 13.5(b).

Equations (13.36)–(13.38) comprise the CAR2 model. Vague conjugate gamma prior distributions for the precision hyperparameters τ_ϕ and τ_ψ, along with normal or flat priors for $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$, complete the hierarchical specification. The posterior distribution of the parameters can be estimated via MCMC techniques; the "inhomogeneous model" of Aykroyd (1998) is a Gaussian-response variant of the CAR2 for image data over a regular grid.

13.10.3.3 Site-edge (SE) models

A primary issue in implementing the CAR2 method is the determination of good "discrepancy" covariates $\mathbf{z}_{ij}$. Although $\boldsymbol{\gamma}$ is estimable even under a noninformative prior distribution, these second-level regression coefficients are often hard to estimate. Also, p_{ij} (and correspondingly w_{ij}) can be sensitive to the prior specification of $\boldsymbol{\gamma}$. Since the edge parameters enter the model only to specify the variances of the first-level random effects, they may be "too far away from the data" in the hierarchical model. This motivates a model with fewer levels or more direct modeling of edge effects.

As such, in this subsection we now consider "site-edge" (SE) models, where both the areal units (sites) *and* the edges between them contribute random effects to the mean structure. Let $\mathcal{G} = (\mathcal{S}, \mathcal{E})$, where $\mathcal{S} = \{1, \ldots, n\}$ is a set of sites/areas, and $\mathcal{E} = \{(i,j) : i \sim j\}$ is a set of edges, where $\sim$ indicates the symmetric "adjacency" relation. Suppose the data follow an exponential family likelihood, and let $\boldsymbol{\phi} = (\boldsymbol{\phi}^S, \boldsymbol{\phi}^E)$ be a vector of site- and edge-level effects, respectively. The general class of SE models is then given by $g(\mu_i) = f(\mathbf{x}_i, \boldsymbol{\beta}, \boldsymbol{\phi})$ and $\boldsymbol{\phi} = (\boldsymbol{\phi}^S, \boldsymbol{\phi}^E) \sim MRF$, where μ_i is the mean of the likelihood and $g(\cdot)$ stands for the canonical link function (e.g., the log link for the Poisson).

Here we take f to be linear, but this is not required for posterior propriety. To facilitate parameter identification and data information flow while encouraging sensible interaction between the areal and edge random effects, we now propose the hierarchical model,

$$Y_i \mid \boldsymbol{\beta}, \phi_i^S \ \sim \ Poisson(\mu_i), \ \text{ with}$$

$$\log \mu_i = \log E_i + \mathbf{x}_i'\boldsymbol{\beta} + \phi_i^S, \quad i = 1, \ldots, n,$$

$$p(\boldsymbol{\phi}^S \mid \boldsymbol{\phi}^E, \tau_\phi) = C(\tau_\phi, \boldsymbol{\phi}^E)\exp\left\{-\frac{\tau_\phi}{2}\sum_{i\sim j}(1-\phi_{ij}^E)(\phi_i^S - \phi_j^S)^2\right\} \tag{13.39}$$

$$\text{and} \quad p(\boldsymbol{\phi}^E) \propto \exp\left\{-\nu \sum_{ij\sim kl} \phi_{ij}^E \phi_{kl}^E\right\}, \tag{13.40}$$

where $\phi_i^S \in \Re$ as before, but now $\phi_{ij}^E \in \{0, 1\}$ for all edges $(i, j) \in \mathcal{E}$. The conditional distribution in (13.39) is IAR, with $(1 - \phi_{ij}^E)$ playing the roles of the w_{ij} in (13.36). That is, $\phi_{ij}^E = 1$ if edge (i, j) is a boundary element, and 0 otherwise. Thus smoothing of neighboring ϕ_i^S and ϕ_j^S is only encouraged if there is no boundary between them. The prior for $\boldsymbol{\phi}^E$ in (13.40) is an *Ising* model with tuning parameter ν, often used in image restoration (e.g. Geman and Geman, 1984, p. 725). This prior yields a binary MRF that allows binary variables (the ϕ_{ij}^E's) to directly borrow strength across their neighbors, avoiding the need for a link-function to introduce continuous spatial effects as in (13.37). The ν here is interpreted as measuring "binding strength" between the edges; smaller values of ν lead to more connected boundary elements, hence more separated areal units. We recommend comparing results under different fixed ν values. We refer to model (13.39)–(13.40) as an *SE-Ising* model. Strictly speaking, this model is not an MRF. Based upon our discussion in Chapter 4, an MRF would require not only $\phi_{ij}^E \phi_{kl}^E$, but also $(1 - \phi_{ij}^E)(1 - \phi_{kl}^E)$. The latter is absent in $p(\boldsymbol{\phi}^E)$. Nevertheless, the SE-Ising model is a legitimate probability model that supplies legitimate posteriors.

The improper CAR prior for $\boldsymbol{\phi}^S$ makes the joint prior $p(\boldsymbol{\phi}^S, \boldsymbol{\phi}^E) \equiv p(\boldsymbol{\phi}^S|\boldsymbol{\phi}^E)p(\boldsymbol{\phi}^E)$ improper regardless of the choice of $p(\boldsymbol{\phi}^E)$, but the joint *posterior* of these parameters will still be proper. To see this, note that $p(\boldsymbol{\phi}^S|\boldsymbol{\phi}^E, \mathbf{y})$ is proper under the usual improper CAR prior, and the discrete support of $p(\boldsymbol{\phi}^E)$ in (13.40) means it too is proper by construction. Since $p(\boldsymbol{\phi}^S, \boldsymbol{\phi}^E|\mathbf{y}) \propto p(\boldsymbol{\phi}^S|\boldsymbol{\phi}^E, \mathbf{y})p(\boldsymbol{\phi}^E)$, the joint posterior is proper as well.

While the SE-Ising model is quite sensible for boundary analysis, it does not explicitly encourage long strings of connected boundary segments of the sort that would be needed to separate a hospice service area from an unserved area. As such, we further propose a *penalized* SE-Ising distribution

$$p(\boldsymbol{\phi}^E) \propto \exp\left\{-\nu \sum_{ij\sim kl} \phi_{ij}^E \phi_{kl}^E + \kappa M\right\}, \tag{13.41}$$

where M is the number of strings of connected "on" edges ($\phi_{ij}^E = 1$) and $\kappa < 0$ is a second tuning parameter. Adding this additional penalty on edge arrangements that do not favor series of connected boundary segments helps to impose the kind of structure we want on our fitted boundaries.

Example 13.5 Minnesota Medicare hospice utilization data. Here, interest lies in identifying unserved areas in the state of Minnesota. Our data consist of ZIP code area-level Medicare beneficiary death counts from 2000 to 2002, as well as the number of these deaths among patients served by each hospice, both based on Medicare billing records. The use of zip code areas as our zonal system has inherent problems (Grubesic, 2008), including that it evolves over time at the whim of the U.S. Postal Service. We focus on the two hospice systems headquartered in the city of Duluth that serve rural northeast and north-central Minnesota, St. Luke's and St. Mary's/Duluth Clinic (SMDC). Figure 13.6(a) and (c) give raw data maps for St. Luke's, while those for SMDC appear in Figure 13.6(b) and (d). The first row of the figure maps the numbers of hospice deaths during the three-year

Figure 13.6 *St. Luke's and SMDC hospice system usage data, northeastern Minnesota zip codes: (a) St. Luke's hospice death counts; (b) SMDC hospice death counts; (c) St. Luke's internally standardized mortality ratios; (d) SMDC internally standardized mortality ratios; (e) St. Luke's self-reported service area boundaries; (f) SMDC self-reported service area boundaries. In (e) and (f), hospice home bases are marked (D = Duluth, GR = Grand Rapids, V = Virginia). Note that the self-reported service area for St. Luke's is entirely contained by the self-reported service area for SMDC.*

period by zip code for the two hospice systems, while the second row maps the internally standardized mortality ratios, i.e., actual hospice death count divided by expected deaths (taken as proportional to the total Medicare death count) in each zip code area. Using either definition of "service," St. Luke's service area appears much smaller and more tightly clustered than SMDC's.

Determining the "service area" for each hospice system based only on the zip code area-specific hospice and total death counts is not as easy as simply drawing boundaries that separate zip code areas with zero counts from those with nonzero counts, since a patient's actual and billing addresses may not coincide. While calling every hospice in the country is infeasible, one might wonder if this would at least provide a "gold standard" to help validate a statistical model in a few cases. To check this, the two hospices were contacted and lists of zip code areas that each said it served were obtained. These results are shown in Figure 13.6(e) and (f) for St. Luke's and SMDC, respectively. The former self-reported service area appears congruent with the observed data for St. Luke's, a smaller hospice focusing on zip codes in or just northeast of Duluth (indicated with a "D" in panels (e) and (f)). But this is not the case for SMDC, a fast-developing hospice system with home base in Duluth and two satellite bases in Grand Rapids ("GR") and Virginia ("V"), both north of Duluth. Comparing the SMDC self-report to the actual hospice death counts in

Figure 13.6(b), it does not appear that its service region extended quite as far south or west as claimed during the three years covered by our data.

We apply the LC, CAR2, SE-Ising, and penalized SE-Ising models to these data. With the latter three models we use edge correction, while for all four methods we use a *thresholding* approach designed to detect a boundary only when differences in the means of adjacent ZIP codes lie on opposite sides of some predetermined minimum service level.

Our analysis considers a single covariate x_i, the intercentroidal (geodetic) distance from the patient's zip code area to the nearest relevant hospice home base zip code area (again see Figure 13.6(e) and (f) for locations). Since hospice services are provided in the patient's home, increasing this distance should decrease the probability of that zip code area being served.

We use vague $N(0, 10^4)$ priors for both the intercept β_0 and distance effect β_1. All of our models for both hospices also employ gamma priors for τ_ϕ having mean 1 and variance 1; this prior still permits significant prior-to-posterior Bayesian learning for this parameter while delivering acceptable MCMC convergence. For the SE-Ising model, we begin by setting the binding strength parameter ν equal to 0.5, and additionally set $\kappa = -3$ in the penalized SE-Ising model. For the CAR2 model, we were unable to identify satisfactory areal discrepancy covariates $\mathbf{z}_{ij}$ at the zip code area level, though in a previous, county-based analysis Ma and Carlin (2007) used median income, local business pattern, and health insurance coverage. While median income would likely be useful here as well, the logit in (13.37) contains only the random effects θ_{ij}, assigned the second stage zero-centered CAR in (13.38). For τ_ψ, the precision parameter of this second-level CAR, we use the same gamma prior as that for τ_ϕ. We tried different gamma distributions and even fixed τ_ψ at its MLE based on the self-reported boundaries. Although τ_ψ and $\boldsymbol{\psi}$ estimates are sensitive to prior choices, the lower-level parameter estimates are fairly robust. Finally, our summary displays are based on the more robust posterior medians (not means), acknowledging possible skewness in the posterior densities.

Ma et al. (2010) show that the SE-Ising models perform best with respect to DIC. The corresponding effects are significant: for example, using the penalized SE-Ising model, the posterior medians and 95% equal-tail credible sets for β are –2.93 (–4.26,–1.58) for St. Luke's and –3.06 (–5.03, –1.50) for SMDC. The negative signs indicate that the farther away a zip code is from the nearest hospice home base, the less likely it is to be served by that hospice. As such, we include the distance covariate as x in all of our subsequent analyses.

Figure 13.7(a)–(b) show μ-based boundary maps for St. Luke's, while Figure 13.7(c)–(d) give them for SMDC. All four panels in these figures are based on absolute posterior medians of $\Delta_{\mu,ij} = \mu_i - \mu_j$. Panels (a) and (c) give results from the LC model, which appears to do a credible job for both hospices. However, even in the easier St. Luke's case, the LC map does include some "clutter" (identified boundary segments apparently internal to the service area) near the bottom of the map.

Panels (b) and (d) in Figure 13.7 give the hierarchically smoothed boundaries from the penalized SE-Ising model. For St. Luke's, the penalized SE-Ising boundaries in Figure 13.7(b) are quite satisfactory, showing essentially no internal clutter and offering a good match with the self-reported boundaries in Figure 13.6(e). However, for SMDC the boundaries are less well connected, perhaps owing to the more complex nature of the data, which features a much larger service region and comprises three offices shown in Figure 13.6(f).

The wombled boundaries in Figure 13.7(c) and (d) are quite similar, as the μ_i are fairly well-estimated by any reasonable spatial model. However, none of our SMDC wombled maps provide a very good match to the self-reported boundaries in the south, since the data do not support the claim of service coverage there. This disagreement between our results and the self-report could be the result of reporting lags or migration in and out of service areas, but is more likely due to the focus of some hospices (especially larger ones, like SMDC) on urban patients, at the expense of harder-to-reach rural ones.

Figure 13.7 *Maps of St. Luke's and SMDC's service area boundaries: (a) St. Luke's service area boundaries given by the LC (usually CAR) model; (b) St. Luke's service area boundaries given by the penalized SE-Ising model; (c) SMDC's service area boundaries given by the LC (usually CAR) model; (d) SMDC's service area boundaries given by the penalized SE-Ising model.*

Ma et al. (2010) also consider boundaries based not on the $\Delta_{\mu,ij}$ but on the $\Delta_{\phi^S,ij} \equiv \phi_i^S - \phi_j^S$ *or* on the ϕ_{ij}^E themselves, to provide information about boundaries separating areas having significantly different *residuals*. This analysis reveals a diagonal edge separating urban Duluth from the rest of the service area, and a potential need for an indicator of whether a ZIP code area is a largely uninhabited, protected area to be added as an areal covariate $\mathbf{x}_i$ in (13.39)–(13.40).

13.10.4 FDR-based areal wombling

The approaches of the previous subsections do not reckon with the multiplicity issues afflicting inference from marginal posterior estimates. Li et al. (2012) pursue a simpler formulation that attempts to resolve the multiplicities using false discovery rates (FDR). These authors formulate the problem of areal wombling as one of testing different boundary hypotheses. A boundary hypothesis posits whether a pair of neighbors have equal spatial random effects or not. We want to test, for each pair of adjacent geographical regions in a map, a null model that posits equal spatial effects for the two regions against an alternative model that allows unconstrained, but spatially correlated, regional effects. As such, we will have as many hypothesis as there are geographical boundary segments on our map. For example, there are 211 such segments in the county map for the state of Minnesota. Each hypothesis corresponds to a two-component mixture distribution that assigns a point mass to the null hypothesis and distributes the remaining mass to the alternative.

When multiple hypotheses are tested simultaneously, classical inference is usually concerned about controlling the overall Type I error rate. Benjamini and Hochberg (1995) introduced the FDR as an error criterion in multiple testing and described procedures to control it. The FDR is the expected proportion of falsely rejected null hypotheses among all rejected null hypotheses. Bayesian versions of FDR have been proposed and discussed by several authors including Storey (2002; 2003), Genovese and Waserman (2002), and Newton et al. (2004). Mueller et al. (2008) used a decision theoretic perspective and set up decision problems that lead to the use of FDR-based rules and generalizations. Li et al. (2012) adapt this framework to our "areal wombling" problem. They depart from the more traditional

conditionally autoregressive (CAR) and simultaneous autoregressive (SAR) models used for areal data analysis as they create problems in implementing the mixture models in the model framework.

13.11 Wombling with point process data

In disease mapping and public health, a spatial layer for which boundary analysis would be of considerable interest is the pattern of disease incidence. In particular, we would seek to identify transition from areas with low incidence to areas with elevated incidence. For cases aggregated to counts for areal units, e.g., census blocks or zip codes (in order to protect confidentiality), this would require obtaining standardized incidence rates for the units. Wombling for such data is discussed in Jacquez and Greiling (2003).

Taking a step back from the areal wombling analysis, with observations as points, we are interested in wombling for a spatial point pattern. Hence, whether we analyze the random locations or the aggregated counts, we assume that the point pattern is driven by an intensity surface, $\lambda(\mathbf{s})$ (see Chapter 8). Wombling for the observed point pattern would be achieved by wombling the estimated $\lambda(\mathbf{s})$. Hence, we have the machinery of Chapter 8 available to go forward. That is, with an intensity of the form, $\lambda_{\mathbf{s}}) = \exp(\mathbf{X}(\mathbf{s})^T\boldsymbol{\beta} + Z(\mathbf{s}))$ (see Subsection 8.4.2) with $Z(\mathbf{s})$ a Gaussian process, upon model fitting, we can study gradient behavior of $Z(\mathbf{s})$. Since, from Section 13.3 the gradient for $g(Z(\mathbf{s}))$, with g continuous and monotonic increasing, results in $D_{\mathbf{u}}g(Z(\mathbf{s})) = g'(Z(\mathbf{s}))D_{\mathbf{u}}Z(\mathbf{s})$, we can directly explore the gradient surface of $\exp(Z(\mathbf{s}))$ as well. Details, including an extension to *marked point processes* is discussed in Liang, Banerjee and Carlin (2008).

13.12 Concluding remarks

Boundary analysis is related to the problem of spatial cluster analysis. However, the latter is focused on detecting clusters of homogeneous regions, identifying their shape and range, and often on identifying whether a particular region is or is not part of a particular "cluster" (say, of infected or exposed counties). By contrast, boundary analysis is more focused on detecting and locating rapid changes, which are typically thought of as "lines" or "curves" on the spatial surface. Substantive interest focuses on the boundary itself and what distinguishes the regions on either side, rather than on any particular region. As such, methods for spatial clustering (as summarized for instance by Lawson and Denison, 2002) are also often not directly applicable here.

In this chapter we have discussed recently proposed theories concerning the use of spatial gradients to detect points and curves that represent rapid change. Here we have confined ourselves to curves that track zones of rapid change. However, as we have alluded to above, zones of rapid change are areal notions; description by a curve may be an unsatisfying simplification. To describe zones as areal quantities, i.e., as sets of nonzero Lebesgue measure in $\Re^2$ is an alternative. To proceed, the crucial issue is to formalize shape-oriented definitions of a wombling boundary. While much work has been done on statistical shape analysis, its use in the point-referenced spatial data context we have set out is unexplored. There are many possibilities, using formal differential geometry and calculus of variations, providing directions for future research. Finally, we also note that current approaches are built entirely upon the specification of a point-referenced spatial process model. One might examine the boundary analysis problem from an alternative modelling perspective, where curves or zones arise as random processes. Possibilities include line processes, random tessellations, and random areal units.

Chapter 14

Spatial survival models

The use of survival models involving a random effect or "frailty" term is becoming more common. Usually the random effects are assumed to represent different clusters, and clusters are assumed to be independent. In this chapter, we consider random effects corresponding to clusters that are spatially arranged, such as clinical sites or geographical regions. That is, we might suspect that random effects corresponding to strata in closer proximity to each other might also be similar in magnitude.

Survival models have a long history in the biostatistical and medical literature (see, e.g., Cox and Oakes, 1984). Very often, time-to-event data will be grouped into *strata* (or *clusters*), such as clinical sites, geographic regions, and so on. In this setting, a hierarchical modeling approach using stratum-specific parameters called *frailties* is often appropriate. Introduced by Vaupel, Manton, and Stallard (1979), this is a mixed model with random effects (the frailties) that correspond to a stratum's overall health status.

To illustrate, let t_{ij} be the time to death or censoring for subject j in stratum i, $j = 1, \ldots, n_i$, $i = 1, \ldots, I$. Let $\mathbf{x}_{ij}$ be a vector of individual-specific covariates. The usual assumption of proportional hazards $h(t_{ij}; \mathbf{x}_{ij})$ enables models of the form

$$h(t_{ij}; \mathbf{x}_{ij}) = h_0(t_{ij}) \exp(\boldsymbol{\beta}^T \mathbf{x}_{ij}) \ , \tag{14.1}$$

where h_0 is the *baseline hazard*, which is affected only multiplicatively by the exponential term involving the covariates. In the frailty setting, model (14.1) is extended to

$$\begin{aligned} h(t_{ij}; x_{ij}) &= h_0(t_{ij})\, \omega_i \exp(\boldsymbol{\beta}^T \mathbf{x}_{ij}) \\ &= h_0(t_{ij}) \exp(\boldsymbol{\beta}^T \mathbf{x}_{ij} + W_i) \ , \end{aligned} \tag{14.2}$$

where $W_i \equiv \log \omega_i$ is the stratum-specific frailty term, designed to capture differences among the strata. Typically a simple i.i.d. specification for the W_i is assumed, e.g.,

$$W_i \overset{iid}{\sim} N(0, \sigma^2) \ . \tag{14.3}$$

With the advent of MCMC computational methods, the Bayesian approach to fitting hierarchical frailty models such as these has become increasingly popular (see, e.g., Carlin and Louis, 2000, Sec. 7.6). Perhaps the simplest approach is to assume a *parametric* form for the baseline hazard h_0. While a variety of choices (gamma, lognormal, etc.) have been explored in the literature, in Section 14.1 we adopt the Weibull, which seems to represent a good tradeoff between simplicity and flexibility. This then produces

$$h(t_{ij}; x_{ij}) = \rho t_{ij}^{\rho - 1} \exp(\boldsymbol{\beta}^T \mathbf{x}_{ij} + W_i) \ . \tag{14.4}$$

Now, placing prior distributions on $\rho, \boldsymbol{\beta}$, and σ^2 completes the Bayesian model specification. Such models are by now a standard part of the literature, and easily fit (at least in the univariate case) using `WinBUGS`. Carlin and Hodges (1999) consider further extending model

(14.4) to allow stratum-specific baseline hazards, i.e., by replacing ρ by ρ_i. MCMC fitting is again routine given a distribution for these new random effects, say, $\rho_i \overset{iid}{\sim} Gamma(\alpha, 1/\alpha)$, so that the ρ_i have mean 1 (corresponding to a constant hazard over time) but variance $1/\alpha$.

A richer but somewhat more complex alternative is to model the baseline hazard *non-parametrically*. In this case, letting γ_{ij} be a death indicator (0 if alive, 1 if dead) for patient ij, we may write the likelihood for our model $L(\beta, \mathbf{W}; \mathbf{t}, \mathbf{x}, \boldsymbol{\gamma})$ generically as

$$\prod_{i=1}^{I} \prod_{j=1}^{n_i} \{h(t_{ij}; \mathbf{x}_{ij})\}^{\gamma_{ij}} \exp\left\{-H_{0i}(t_{ij}) \exp\left(\boldsymbol{\beta}^T \mathbf{x}_{ij} + W_i\right)\right\} ,$$

where $H_{0i}(t) = \int_0^t h_{0i}(u)\, du$, the integrated baseline hazard. A frailty distribution parametrized by λ, $p(\mathbf{W}|\lambda)$, coupled with prior distributions for λ, $\boldsymbol{\beta}$, and the hazard function h complete the hierarchical Bayesian model specification.

In this chapter we consider both parametric and semiparametric hierarchical survival models for data sets that are spatially arranged. Such models might be appropriate anytime we suspect that frailties W_i corresponding to strata in closer proximity to each other might also be similar in magnitude. This could arise if, say, the strata corresponded to hospitals in a given region, to counties in a given state, and so on. The basic assumption here is that "expected" survival times (or hazard rates) will be more similar in proximate regions, due to underlying factors (access to care, willingness of the population to seek care, etc.) that vary spatially. We hasten to remind the reader that this does not imply that the observed survival times from subjects in proximate regions must be similar, since they include an extra level of randomness arising from their variability around their (spatially correlated) underlying model quantities.

14.1 Parametric models

14.1.1 Univariate spatial frailty modeling

While it is possible to identify centroids of geographic regions and employ spatial process modeling for these locations, the effects in our examples are more naturally associated with areal units. As such we work exclusively with CAR models for these effects, i.e., we assume that

$$\mathbf{W} \mid \lambda \sim CAR(\lambda) . \tag{14.5}$$

Also, we note that the resulting model for, say, (14.2) is an extended example of a generalized linear model for areal spatial data (Section 6.5). That is, (14.2) implies that

$$f(t_{ij}|\boldsymbol{\beta}, x_{ij}, W_i) = h_0(t_{ij}) e^{\boldsymbol{\beta}^T \mathbf{x}_{ij} + W_i} e^{-H_0(t_{ij}) \exp(\boldsymbol{\beta}^T \mathbf{x}_{ij} + W_i)} . \tag{14.6}$$

In other words, $U_{ij} = H_0(t_{ij}) \sim \text{Exponential}\Big(\exp[-(\boldsymbol{\beta}^T \mathbf{x}_{ij} + W_i)]\Big)$ so $-\log EH_0(t_{ij}) = \boldsymbol{\beta}^T \mathbf{x}_{ij} + W_i$. The analogy with (6.29) and $g(\eta_i)$ is clear. The critical difference is that in Section 6.5 the link g is assumed known; here the link to the linear scale requires h_0, which is unknown (and will be modeled parametrically or nonparametrically).

Finally, we remark that it would certainly be possible to include both spatial and non-spatial frailties, which as already seen (Subsection 6.4.3) is now common practice in areal data modeling. Here, this would mean supplementing our spatial frailties W_i with a collection of nonspatial frailties, say, $V_i \overset{iid}{\sim} N(0, 1/\tau)$. The main problem with this approach is again that the frailties now become identified only by the prior, and so the proper choice of priors for τ and λ (or $\boldsymbol{\theta}$) becomes problematic. Another problem is the resultant decrease in algorithm performance wrought by the addition of so many additional, weakly identified parameters.

14.1.1.1 Bayesian implementation

As already mentioned, the models outlined above are straightforwardly implemented in a Bayesian framework using MCMC methods. In the parametric case, say (14.4), the joint posterior distribution of interest is

$$p\left(\boldsymbol{\beta}, \mathbf{W}, \rho, \lambda \mid \mathbf{t}, \mathbf{x}, \boldsymbol{\gamma}\right) \propto L\left(\boldsymbol{\beta}, \mathbf{W}, \rho\,;\, \mathbf{t}, \mathbf{x}, \boldsymbol{\gamma}\right) p\left(\mathbf{W}|\lambda\right) p(\boldsymbol{\beta}) p\left(\rho\right) p(\lambda)\ , \tag{14.7}$$

where the first term on the right-hand side is the Weibull likelihood, the second is the CAR distribution of the random frailties, and the remaining terms are prior distributions. In (14.7), $\mathbf{t} = \{t_{ij}\}$ denotes the collection of times to death, $\mathbf{x} = \{\mathbf{x}_{ij}\}$ the collection of covariate vectors, and $\boldsymbol{\gamma} = \{\gamma_{ij}\}$ the collection of death indicators for all subjects in all strata.

For our investigations, we retain the parametric form of the baseline hazard given in (14.4). Thus $L\left(\boldsymbol{\beta}, \mathbf{W}, \rho\,;\, \mathbf{t}, \mathbf{x}, \boldsymbol{\gamma}\right)$ is proportional to

$$\prod_{i=1}^{I} \prod_{j=1}^{n_i} \left\{\rho t_{ij}^{\rho-1} \exp\left(\boldsymbol{\beta}^T \mathbf{x}_{ij} + W_i\right)\right\}^{\gamma_{ij}} \exp\left\{-t_{ij}^{\rho} \exp\left(\boldsymbol{\beta}^T \mathbf{x}_{ij} + W_i\right)\right\}. \tag{14.8}$$

The model specification in the Bayesian setup is completed by assigning prior distributions for $\boldsymbol{\beta}, \rho$, and λ. Typically, a flat (improper uniform) prior is chosen for $\boldsymbol{\beta}$, while vague but proper priors are chosen for ρ and λ, such as a $G(\alpha, 1/\alpha)$ prior for ρ and a $G(a, b)$ prior for λ. Hence the only extension beyond the disease mapping illustrations of Section 6.4 is the need to update ρ.

Example 14.1 *(Application to Minnesota infant mortality data).* We apply the methodology above to the analysis of infant mortality in Minnesota, originally considered by Banerjee, Wall, and Carlin (2003). The data were obtained from the linked birth-death records data registry kept by the Minnesota Department of Health. The data comprise 267,646 live births occurring during the years 1992–1996 followed through the first year of life, together with relevant covariate information such as birth weight, sex, race, mother's age, and the mother's total number of previous births. Because of the careful linkage connecting infant death certificates with birth certificates (even when the death occurs in a separate state), we assume that each baby in the data set that is not linked with a death must have been alive at the end of one year. Of the live births, only 1,547 babies died before the end of their first year. The number of days they lived is treated as the response t_{ij} in our models, while the remaining survivors were treated as "censored," or in other words, alive at the end of the study period. In addition to this information, the mother's Minnesota county of residence prior to the birth is provided. We implement the areal frailty model (14.5), the nonspatial frailty model (14.3), and a simple nonhierarchical ("no-frailty") model that sets $W_i = 0$ for all i.

For all of our models, we adopt a flat prior for $\boldsymbol{\beta}$, and a $G(\alpha, 1/\alpha)$ prior for ρ, setting $\alpha = 0.01$. Metropolis random walk steps with Gaussian proposals were used for sampling from the full conditionals for $\boldsymbol{\beta}$, while Hastings independence steps with gamma proposals were used for updating ρ. As for λ, in our case we are fortunate to have a data set that is large relative to the number of random effects to be estimated. As such, we simply select a vague (mean 1, variance 1000) gamma specification for λ, and rely on the data to overwhelm the priors.

Table 14.1 compares our three models in terms of two of the criteria discussed in Subsection 5.2.3, DIC and effective model size p_D. For the no-frailty model, we see a p_D of 8.72, very close to the actual number of parameters, 9 (8 components of $\boldsymbol{\beta}$ plus the Weibull parameter ρ). The random effects models have substantially larger p_D values, though much smaller than their actual parameter counts (which would include the 87 random frailties

Model	p_D	DIC
No-frailty	8.72	511
Nonspatial frailty	39.35	392
CAR frailty	34.52	371

Table 14.1 *DIC and effective number of parameters p_D for competing parametric survival models.*

Covariate	2.5%	50%	97.5%
Intercept	−2.135	−2.024	−1.976
Sex (boys = 0)			
girls	−0.271	−0.189	−0.105
Race (white = 0)			
black	−0.209	−0.104	−0.003
Native American	0.457	0.776	1.004
unknown	0.303	0.871	1.381
Mother's age	−0.005	−0.003	−0.001
Birth weight in kg	−1.820	−1.731	−1.640
Total births	0.064	0.121	0.184
ρ	0.411	0.431	0.480
σ	0.083	0.175	0.298

Table 14.2 *Posterior summaries for the nonspatial frailty model.*

W_i); apparently there is substantial shrinkage of the frailties toward their grand mean. The DIC values suggest that each of these models is substantially better than the no-frailty model, despite their increased size. Though the spatial frailty model has the best DIC value, plots of the full estimated posterior deviance distributions (not shown) suggest substantial overlap. On the whole we seem to have modest support for the spatial frailty model over the ordinary frailty model.

Tables 14.2 and 14.3 provide 2.5, 50, and 97.5 posterior percentiles for the main effects in our two frailty models. In both cases, all of the predictors are significant at the .05 level. Since the reference group for the sex variable is boys, we see that girls have a lower hazard of death during the first year of life. The reference group for the race variables is white; the Native American beta coefficient is rather striking. In the CAR model, this covariate increases the posterior median hazard rate by a factor of $e^{0.782} = 2.19$. The effect of "unknown" race is also significant, but more difficult to interpret: in this group, the race of the infant was not recorded on the birth certificate. Separate terms for Hispanics, Asians, and Pacific Islanders were also originally included in the model, but were eliminated after emerging as not significantly different from zero. Note that the estimate of ρ is quite similar across models, and suggests a decreasing baseline hazard over time. This is consistent with the fact that a high proportion (495, or 32%) of the infant deaths in our data set occurred in the first *day* of life: the force of mortality (hazard rate) is very high initially, but drops quickly and continues to decrease throughout the first year.

A benefit of fitting the spatial CAR structure is seen in the reduction of the length of the 95% credible intervals for the covariates in the spatial models compared to the i.i.d. model. As we might expect, there are modest efficiency gains when the model that better specifies the covariance structure of its random effects is used. That is, since the spatial dependence priors for the frailties are in better agreement with the likelihood than is the independence prior, the prior-to-posterior learning afforded by Bayes' Rule leads to smaller posterior variances in the former cases. Most notably, the 95% credible set for the effect of "unknown" race is (0.303, 1.381) under the nonspatial frailty model (Table 14.2), but

Covariate	2.5%	50%	97.5%
Intercept	−2.585	−2.461	−2.405
Sex (boys = 0)			
girls	−0.224	−0.183	−0.096
Race (white = 0)			
black	−0.219	−0.105	−0.007
Native American	0.455	0.782	0.975
unknown	0.351	0.831	1.165
Mother's age	−0.005	−0.004	−0.003
Birth weight in kg	−1.953	−1.932	−1.898
Total births	0.088	0.119	0.151
ρ	0.470	0.484	0.497
λ	12.62	46.07	100.4

Table 14.3 *Posterior summaries for the CAR frailty model.*

Figure 14.1 *Posterior median frailties, i.i.d. model with covariates, Minnesota county-level infant mortality data.*

(0.351, 1.165) under the CAR frailty model (Table 14.3), a reduction in length of roughly 25%.

Figures 14.1 and 14.2 map the posterior medians of the W_i under the nonspatial (i.i.d. frailties) and CAR models, respectively, where the models include all of the covariates listed in Tables 14.2 and 14.3. As expected, no clear spatial pattern is evident in the i.i.d. map, but from the CAR map we are able to identify two clusters of counties having somewhat higher hazards (in the southwest following the Minnesota River, and in the northeast "arrowhead" region), and two clusters with somewhat lower hazards (in the northwest, and the southeastern corner). Thus, despite the significance of the covariates now in these models, Figure 14.2 suggests the presence of some still-missing, spatially varying covariate(s) relevant for infant mortality. Such covariates might include location of birth (home or hospital),

Figure 14.2 *Posterior median frailties, CAR model with covariates, Minnesota county-level infant mortality data.*

Figure 14.3 *Boxplots of posterior median frailties, i.i.d. and CAR models with and without covariates.*

overall quality of available health or hospital care, mother's economic status, and mother's number of prior abortions or miscarriages.

In addition to the improved appearance and epidemiological interpretation of Figure 14.2, another reason to prefer the CAR model is provided in Figure 14.3, which shows boxplots of the posterior median frailties for the two cases corresponding to Figures 14.1 and 14.2, plus two preliminary models in which *no* covariates $\mathbf{x}$ are included. The tightness

of the full CAR boxplot suggests this model is best at reducing the need for the frailty terms. This is as it should be, since these terms are essentially spatial residuals, and represent lingering lack of fit in our spatial model (although they may well also account for some excess *non*spatial variability, since our current models do not include nonspatial frailty terms). Note that all of the full CAR residuals are in the range (–0.15, 0.10), or (0.86, 1.11) on the hazard scale, suggesting that missing spatially varying covariates have only a modest (10 to 15%) impact on the hazard; from a practical standpoint, this model fits quite well.

14.1.2 Spatial frailty versus logistic regression models

In many contexts (say, a clinical trial enrolling and following patients at spatially proximate clinical centers), a spatial survival model like ours may be the only appropriate model. However, since the Minnesota infant mortality data does not have any babies censored because of loss to followup, competing risks, or any reason other than the end of the study, there is no ambiguity in defining a *binary* survival outcome for use in a random effects logistic regression model. That is, we replace the event time data t_{ij} with an indicator of whether the subject did ($Y_{ij} = 0$) or did not ($Y_{ij} = 1$) survive the first year. Letting $p_{ij} = Pr(Y_{ij} = 1)$, our model is then

$$logit(p_{ij}) = \widetilde{\boldsymbol{\beta}}^T \mathbf{x}_{ij} + \widetilde{W}_i \ , \tag{14.9}$$

with the usual flat prior for $\widetilde{\boldsymbol{\beta}}$ and an i.i.d. or CAR prior for the $\widetilde{W}_i$. As a result, (14.9) is exactly an example of a generalized linear model for areal spatial data.

Other authors (Doksum and Gasko, 1990; Ingram and Kleinman, 1989) have shown that in this case of no censoring before followup (and even in cases of equal censoring across groups), it is possible to get results for the $\widetilde{\boldsymbol{\beta}}$ parameters in the logistic regression model very similar to those obtained in the proportional hazards model (14.1), except of course for the differing interpretations (log odds versus log relative risk, respectively). Moreover when the probability of death is very small, as it is in the case of infant mortality, the log odds and log relative risk become even more similar. Since it uses more information (i.e., time to death rather than just a survival indicator), intuitively, the proportional hazards model should make gains over the logistic model in terms of power to detect significant covariate effects. Yet, consistent with the simulation studies performed by Ingram and Kleinman (1989), our experience with the infant mortality data indicate that only a marginal increase in efficiency (decrease in variance) is exhibited by the posterior distributions of the parameters.

On the other hand, we did find some difference in terms of the estimated random effects in the logistic model compared to the proportional hazards model. Figure 14.4 shows a scatterplot of the estimated posterior medians of W_i versus $\widetilde{W}_i$ for each county obtained from the models where there were no covariates, and the random effects were assumed to i.i.d. The sample correlation of these estimated random effects is 0.81, clearly indicating that they are quite similar. Yet there are still some particular counties that result in rather different values under the two models. One way to explain this difference is that the hazard functions are not exactly proportional across the 87 counties of Minnesota. A close examination of the counties that had differing $\widetilde{W}_i$ versus W_i shows that they had different average times at death compared to other counties with similar overall death rates. Consider for example County 70, an outlier circled in Figure 14.4, and its comparison to circled Counties 73, 55, and 2, which have similar death rates (and hence roughly the same horizontal position in Figure 14.4). We find County 70 has the smallest mean age at death, implying that it has more early deaths, explaining its smaller frailty estimate. Conversely, County 14 has a higher average time at death but overall death rates similar to Counties 82, 48, and 5 (again note the horizontal alignment in Figure 14.4), and as a result has higher estimated

Figure 14.4 *Posterior medians of the frailties W_i (horizontal axis) versus posterior medians of the logistic random effects $\widetilde{W}_i$ (vertical axis). Plotting character is county number; significance of circled counties is described in the text.*

frailty. A lack of proportionality in the baseline hazard rates across counties thus appears to manifest as a departure from linearity in Figure 14.4.

We conclude this subsection by noting that previous work by Carlin and Hodges (1999) suggests a generalization of our basic model (14.4) to

$$h(t_{ij}; x_{ij}) = \rho_i t_{ij}^{\rho_i - 1} \exp(\boldsymbol{\beta}^T \mathbf{x}_{ij} + W_i) \ .$$

That is, we allow two sets of random effects: the existing frailty parameters W_i, and a new set of shape parameters ρ_i. This then allows both the overall level and the shape of the hazard function over time to vary from county to county. Either i.i.d. or CAR priors could be assigned to these two sets of random effects, which could themselves be correlated within county. In the latter case, this might be fit using the MCAR model of Section 10.1; see Jin and Carlin (2003), as well as Section 14.4.

14.2 Semiparametric models

While parametric models are easily interpretable and often afford a surprisingly good fit to survival data, many practitioners continue to prefer the additional richness of the nonparametric baseline hazard offered by the celebrated Cox model. In this section we turn to nonparametric models for the baseline hazard. Such models are often referred to as *semiparametric*, since we continue to assume proportional hazards of the form (14.1) in which the covariate effects are still modeled parametrically, While Li and Ryan (2002) address this problem from a classical perspective, in this section we follow the hierarchical Bayesian approach of Banerjee and Carlin (2002).

Within the Bayesian framework, several authors have proposed treating the Cox partial likelihood as a full likelihood, to obtain a posterior distribution for the treatment effect. However, this approach does not allow fully hierarchical modeling of stratum-specific baseline hazards (with stratum-specific frailties) because the baseline hazard is implicit in the

partial likelihood computation. In the remainder of this section, we describe two possible methodological approaches to modeling the baseline hazard in Cox regression, which thus lead to two semiparametric spatial frailty techniques. We subsequently revisit the Minnesota infant mortality data.

14.2.1 Beta mixture approach

Our first approach uses an idea of Gelfand and Mallick (1995) that flexibly models the integrated baseline hazard as a mixture of monotone functions. In particular, these authors use a simple transformation to map the integrated baseline hazard onto the interval $[0,1]$, and subsequently approximate this function by a weighted mixture of incomplete beta functions. Implementation issues are discussed in detail by Gelfand and Mallick (1995) and also by Carlin and Hodges (1999) for stratum-specific baseline hazards. The likelihood and Bayesian hierarchical setup remain exactly as above.

Thus, we let $h_{0i}(t)$ be the baseline hazard in the ith region and $H_{0i}(t)$ be the corresponding integrated baseline hazard, and define

$$J_{0i}(t) = a_0 H_{0i}(t) \,/\, [a_0 H_{0i}(t) + b_0] \ ,$$

which conveniently takes values in $[0,1]$. We discuss below the choice of a_0 and b_0 but note that this is not as much a modeling issue as a computational one, important only to ensure appropriate coverage of the interval $[0,1]$. We next model $J_{0i}(t)$ as a mixture of $Beta(r_l, s_l)$ cdfs, for $l = 1, \ldots, m$. The r_l and s_l are chosen so that the beta cdfs have evenly spaced means and are centered around $\widetilde{J_0}(t)$, a suitable function transforming the time scale to $[0,1]$. We thus have

$$J_{0i}(t) = \sum_{l=1}^{m} v_{il}\, IB\left(\widetilde{J_0}(t); r_l, s_l\right) \ ,$$

where $\sum_{l=1}^{m} v_{il} = 1$ for all i, and $IB(\cdot; a, b)$ denotes the incomplete beta function (i.e., the cdf of a $Beta\,(a,b)$ distribution). Since any distribution function on $[0,1]$ can be approximated arbitrarily well by a finite mixture of beta cdfs, the same is true for J_{0i}, an increasing function that maps $[0,1]$ onto itself. Thus, working backward, we find the following expression for the cumulative hazard in terms of the above parameters:

$$H_{0i}(t) = \frac{b_0 \sum_{l=1}^{m} v_{il}\, IB\left(\widetilde{J_0}(t); r_l, s_l\right)}{a_0 \left\{1 - \sum_{l=1}^{m} v_{il}\, IB\left(\widetilde{J_0}(t); r_l, s_l\right)\right\}} \ .$$

Taking derivatives, we have for the hazard function,

$$h_{0i}(t) = \frac{b_0 \frac{\partial}{\partial t}\widetilde{J_0}(t) \sum_{l=1}^{m} v_{il} Beta\left(\widetilde{J_0}(t); r_l, s_l\right)}{a_0 \left\{1 - \sum_{l=1}^{m} v_{il}\, IB\left(\widetilde{J_0}(t); r_l, s_l\right)\right\}^2} \ .$$

Typically m, the number of mixands of the beta cdfs, is fixed, as are the $\{(r_l, s_l)\}_{l=1}^{m}$, so chosen that the resulting beta densities cover the interval $[0,1]$. For example, we might fix $m = 5$, $\{r_l\} = (1,2,3,4,5)$ and $\{s_l\} = (5,4,3,2,1)$, producing five evenly-spaced beta cdfs.

Regarding the choice of a_0 and b_0, we note that it is intuitive to specify $\widetilde{J_0}(t)$ to represent a plausible central function around which the J_{0i}'s are distributed. Thus, if we consider the cumulative hazard function of an exponential distribution to specify $\widetilde{J_0}(t)$, then we get $\widetilde{J_0}(t) = a_0 t/(a_0 t + b_0)$. In our Minnesota infant mortality data set, since the survival times ranged between 1 day and 365 days, we found $a_0 = 5$ and $b_0 = 100$ lead to values for

$\widetilde{J_0}(t)$ that largely cover the interval $[0,1]$, and so fixed them as such. The likelihood is thus a function of the regression coefficients $\boldsymbol{\beta}$, the stratum-specific weight vectors $\mathbf{v}_i = (v_{i1}, ..., v_{im})^T$, and the spatial effects W_i. It is natural to model the $\mathbf{v}_i$'s as draws from a Dirichlet$(\phi_1, \ldots, \phi_m)$ distribution, where for simplicity we often take $\phi_1 = \cdots = \phi_m = \phi$.

14.2.2 Counting process approach

The second nonparametric baseline hazard modeling approach we investigate is that of Clayton (1991, 1994). While the method is less transparent theoretically, it is gaining popularity among Bayesian practitioners due to its ready availability within `WinBUGS`. Here we give only the essential ideas, referring the reader to Andersen and Gill (1982) or Clayton (1991) for a more complete treatment. The underlying idea is that the number of failures up to time t is assumed to arise from a *counting process* $N(t)$. The corresponding *intensity process* is defined as

$$I(t)\,dt = E\left(dN(t)\,|F_{t-}\right) \ ,$$

where $dN(t)$ is the increment of N over the time interval $[t, t+dt)$, and F_{t-} represents the available data up to time t. For each individual, $dN(t)$ therefore takes the value 1 if the subject fails in that interval, and 0 otherwise. Thus $dN(t)$ may be thought of as the "death indicator process," analogous to γ in the model of the previous subsection. For the jth subject in the ith region, under the proportional hazards assumption, the intensity process (analogous to our hazard function $h(t_{ij}; \mathbf{x}_{ij})$) is modeled as

$$I_{ij}(t) = Y_{ij}(t)\lambda_0(t)\exp\left(\boldsymbol{\beta}^T\mathbf{x}_{ij} + W_i\right) \ ,$$

where $\lambda_0(t)$ is the baseline hazard function and $Y_{ij}(t)$ is an indicator process taking the value 1 or 0 according to whether or not subject i is observed at time t. Under the above formulation and keeping the same notation as above for $\mathbf{W}$ and $\mathbf{x}$, a Bayesian hierarchical model may be formulated as:

$$\begin{aligned} dN_{ij}(t) &\sim Poisson\left(I_{ij}(t)\,dt\right) \ , \\ I_{ij}(t)\,dt &= Y_{ij}(t)\exp\left(\boldsymbol{\beta}^T\mathbf{x}_{ij} + W_i\right)d\Lambda_0(t) \ , \\ d\Lambda_0(t) &\sim Gamma\left(c\,d\Lambda_0^*(t), c\right) \ . \end{aligned}$$

As before, priors $p(\mathbf{W}|\lambda)$, $p(\lambda)$, and $p(\boldsymbol{\beta})$ are required to completely specify the Bayesian hierarchical model. Here, $d\Lambda_0(t) = \lambda_0(t)\,dt$ may be looked upon as the increment or jump in the integrated baseline hazard function occurring during the time interval $[t, t+dt)$. Since the conjugate prior for the Poisson mean is the gamma distribution, $\Lambda_0(t)$ is conveniently modeled as a process whose increments $d\Lambda_0(t)$ are distributed according to gamma distributions. The parameter c in the above setup represents the degree of confidence in our prior guess for $d\Lambda_0(t)$, given by $d\Lambda_0^*(t)$. Typically, the prior guess $d\Lambda_0^*(t)$ is modeled as $r\,dt$, where r is a guess at the failure rate per unit time. The `LeukFr` example in the `WinBUGS` examples manual offers an illustration of how to code the above formulation.

Example 14.2 *(Application to Minnesota infant mortality data, continued).* We now apply the methodology above to the reanalysis of our Minnesota infant mortality data set. For both the CAR and nonspatial models we implemented the Cox model with the two semiparametric approaches outlined above. We found very similar results, and so in our subsequent analysis we present only the results with the beta mixture approach (Subsection 14.2.1). For all of our models, we adopt vague Gaussian priors for $\boldsymbol{\beta}$. Since the full conditionals for each component of $\boldsymbol{\beta}$ are log-concave, adaptive rejection sampling was used for sampling from the $\boldsymbol{\beta}$ full conditionals. As in Section 14.1, we again simply select a

Model	p_D	DIC
No-frailty	6.82	507
Nonspatial frailty	27.46	391
CAR frailty	32.52	367

Table 14.4 *DIC and effective number of parameters p_D for competing nonparametric survival models.*

Covariate	2.5%	50%	97.5%
Intercept	−2.524	−1.673	−0.832
Sex (boys = 0)			
girls	−0.274	−0.189	−0.104
Race (white = 0)			
black	−0.365	−0.186	−0.012
Native American	0.427	0.737	1.034
unknown	0.295	0.841	1.381
Mother's age	−0.054	−0.035	−0.014
Birth weight in kg	−1.324	−1.301	−1.280
Total births	0.064	0.121	0.184

Table 14.5 *Posterior summaries for the nonspatial semiparametric frailty model.*

vague $G(0.001, 1000)$ (mean 1, variance 1000) specification for CAR smoothness parameter λ, though we maintain more informative priors on the other variance components.

Table 14.4 compares our three models in terms of DIC and effective model size p_D. For the no-frailty model, we see a p_D of 6.82, reasonably close to the actual number of parameters, 8 (the components of $\boldsymbol{\beta}$). The other two models have substantially larger p_D values, though much smaller than their actual parameter counts (which would include the 87 random frailties W_i); apparently there is substantial shrinkage of the frailties toward their grand mean. The DIC values suggest that both of these models are substantially better than the no-frailty model, despite their increased size. As in Table 14.1, the spatial frailty model has the best DIC value.

Tables 14.5 and 14.6 provide 2.5, 50, and 97.5 posterior percentiles for the main effects in our two frailty models, respectively. In both tables, all of the predictors are significant at the .05 level. Overall, the results are broadly similar to those from our earlier parametric analysis in Tables 14.2 and 14.3. For instance, the effect of being in the Native American group is again noteworthy. Under the CAR model, this covariate increases the posterior median hazard rate by a factor of $e^{0.599} = 1.82$. The benefit of fitting the spatial CAR structure is also seen again in the reduction of the length of the 95% credible intervals for the spatial model compared to the i.i.d. model. Most notably, the 95% credible set for the effect of "mother's age" is $(-0.054, -0.014)$ under the nonspatial frailty model (Table 14.5), but $(-0.042, -0.013)$ under the CAR frailty model (Table 14.6), a reduction in length of roughly 28%. Thus overall, adding spatial structure to the frailty terms appears to be reasonable and beneficial. Maps analogous to Figures 14.1 and 14.2 (not shown) reveal a very similar story.

14.3 Spatiotemporal models

In this section we follow Banerjee and Carlin (2003) to develop a semiparametric (Cox) hierarchical Bayesian frailty model for capturing spatiotemporal heterogeneity in survival data. We then use these models to describe the pattern of breast cancer in the 99 counties

Covariate	2.5%	50%	97.5%
Intercept	−1.961	−1.532	−0.845
Sex (boys = 0)			
girls	−0.351	−0.290	−0.217
Race (white = 0)			
black	−0.359	−0.217	−0.014
Native American	0.324	0.599	0.919
unknown	0.365	0.863	1.316
Mother's age	−0.042	−0.026	−0.013
Birth weight in kg	−1.325	−1.301	−1.283
Total births	0.088	0.135	0.193

Table 14.6 *Posterior summaries for the CAR semiparametric frailty model.*

of Iowa while accounting for important covariates, spatially correlated differences in the hazards among the counties, and possible space-time interactions.

We begin by extending the framework of the preceding section to incorporate temporal dependence. Here we have t_{ijk} as the response (time to death) for the jth subject residing in the ith county who was diagnosed in the kth year, while the individual-specific vector of covariates is now denoted by $\mathbf{x}_{ijk}$, for $i = 1, 2, ..., I$, $k = 1, \ldots, K$, and $j = 1, 2, ..., n_{ik}$. We note that "time" is now being used in two ways. The measurement or response is a survival time, but these responses are themselves observed at different areal units *and* different times (years). Furthermore, the spatial random effects W_i in the preceding section are now modified to W_{ik}, to represent spatiotemporal frailties corresponding to the ith county for the kth diagnosis year. Our spatial frailty specification in (14.1) now becomes

$$h\left(t_{ijk}; \mathbf{x}_{ijk}\right) = h_{0i}\left(t_{ijk}\right) \exp\left(\boldsymbol{\beta}^T \mathbf{x}_{ijk} + W_{ik}\right) \ . \tag{14.10}$$

Our CAR prior would now have conditional representation $W_{ik}| \quad W_{(i' \neq i)k} \sim N(\overline{W}_{ik}, 1/(\lambda_k m_i))$.

Note that we can account for temporal correlation in the frailties by assuming that the λ_k are themselves identically distributed from a common hyperprior (Subsection 11.7.1). A gamma prior (usually vague but proper) is often selected here, since this is particularly convenient for MCMC implementation. A flat prior for $\boldsymbol{\beta}$ is typically chosen, since this still admits a proper posterior distribution. Adaptive rejection (Gilks and Wild, 1992) or Metropolis-Hastings sampling are usually required to update the $\mathbf{W}_k$ and $\boldsymbol{\beta}$ parameters in a hybrid Gibbs sampler.

We remark that it would certainly be possible to include both spatial and nonspatial frailties, as mentioned in Subsection 14.1.1. This would mean supplementing our spatial frailties W_{ik} with a collection of nonspatial frailties, say, $V_{ik} \overset{iid}{\sim} N(0, 1/\tau_k)$. We summarize our full hierarchical model as follows:

$$\begin{aligned} L\left(\beta, \mathbf{W}; \mathbf{t}, \mathbf{x}, \gamma\right) \quad &\propto \quad \prod_{k=1}^{K} \prod_{i=1}^{I} \prod_{j=1}^{n_{ik}} \left\{h_{0i}\left(t_{ijk}; \mathbf{x}_{ijk}\right)\right\}^{\gamma_{ijk}} \\ &\qquad \times \exp\left\{-H_{0i}\left(t_{ijk}\right) \exp\left(\boldsymbol{\beta}^T \mathbf{x}_{ijk} + W_{ik} + V_{ik}\right)\right\}, \\ \text{where } \ p(\mathbf{W}_k|\lambda_k) \quad &\sim \quad CAR(\lambda_k) \quad p\left(\mathbf{V}_k|\tau_k\right) \ \sim \ N_I\left(\mathbf{0}, \tau_k \mathbf{I}\right) \\ \text{and } \ \lambda_k \quad &\sim \quad G\left(a, b\right) \ , \ \tau_k \ \sim \ G\left(c, d\right) \ \text{ for } \ k = 1, 2, ..., K \ . \end{aligned}$$

In the sequel we adopt the beta mixture approach of Subsection 14.2.1 to model the baseline hazard functions $H_{0i}(t_{ijk})$ nonparametrically.

Covariate	2.5%	50%	97.5%
Age at diagnosis	0.0135	0.0148	0.0163
Number of primaries	–0.43	–0.40	–0.36
Race (white = 0)			
black	–0.14	0.21	0.53
other	–2.25	–0.30	0.97
Stage (local = 0)			
regional	0.30	0.34	0.38
distant	1.45	1.51	1.58

Table 14.7 *Posterior summaries for the spatiotemporal frailty model.*

Example 14.3 *(Analysis of Iowa SEER breast cancer data).* The National Cancer Institute's SEER program (`seer.cancer.gov`) is the most authoritative source of cancer data in the U.S., offering county-level summaries on a yearly basis for several states in various parts of the country. In particular, the database provides a cohort of 15,375 women in Iowa who were diagnosed with breast cancer starting in 1973, and have been undergoing treatment and have been progressively monitored since. Only those who have been identified as having died from metastasis of cancerous nodes in the breast are considered to have failed, while the rest (including those who might have died from metastasis of other types of cancer, or from other causes of death) are considered censored. By the end of 1998, 11,912 of the patients had died of breast cancer while the remaining were censored, either because they survived until the end of the study period, dropped out of the study, or died of causes other than breast cancer. For each individual, the data set records the time in months (1 to 312) that the patient survived, and her county of residence at diagnosis. Several individual-level covariates are also available, including race (black, white, or other), age at diagnosis, number of primaries (i.e., the number of other types of cancer diagnosed for this patient), and the stage of the disease (local, regional, or distant).

14.3.1 Results for the full model

We begin by summarizing our results for the spatiotemporal frailty model described above, i.e., the full model having both spatial frailties W_{ik} and nonspatial frailties V_{ik}. We chose vague $G(0.01, 0.01)$ hyperpriors for the λ_k and τ_k (having mean 1 but variance 100) in order to allow maximum flexibility in the partitioning of the frailties into spatial and nonspatial components. Best et al. (1999) suggest that a higher variance prior for the τ_k (say, a $G(0.001, 0.001)$) may lead to better prior "balance" between the spatial and nonspatial random effects, but there is controversy on this point and so we do not pursue it here. While overly diffuse priors (as measured for example as in Weiss, 1996) may result in weak identifiability of these parameters, their posteriors remain proper, and the impact of these priors on the posterior for the well-identified subset of parameters (including $\boldsymbol{\beta}$ and the log-relative hazards themselves) should be minimal (Daniels and Kass, 1999; Eberly and Carlin, 2000).

Table 14.7 provides 2.5, 50, and 97.5 posterior percentiles for the main effects (components of $\boldsymbol{\beta}$) in our model. All of the predictors *except* those having to do with race are significant at the .05 level. Since the reference group for the stage variable is local, we see that women with regional and distant (metastasized) diagnoses have higher and much higher hazard of death, respectively; the posterior median hazard rate increases by a factor of $e^{1.51} = 4.53$ for the latter group. Higher age at diagnosis also increases the hazard, but a larger number of primaries (the number of other types of cancer a patient is be suffering

Figure 14.5 *Fitted spatiotemporal frailties, Iowa counties, 1986.*

from) actually leads to a *lower* hazard, presumably due to the competing risk of dying from one of these other cancers.

Figure 14.5 maps the posterior medians of the frailties $W_{ik} + V_{ik}$ for the representative year 1986. We see clusters of counties with lower median frailties in the north-central and south-central parts of the state, and also clusters of counties with higher median frailties in the central, northeastern, and southeastern parts of the state.

Maps for other representative years showed very similar patterns, as well as an overall decreasing pattern in the frailties over time (see Banerjee and Carlin, 2003, for details). Figure 14.6 clarifies this pattern by showing boxplots of the posterior medians of the W_{ik} over time (recall our full model does not have year-specific intercepts; the average of the W_{ik} for year k plays this role). We see an essentially horizontal trend during roughly the first half of our observation period, followed by a decreasing trend that seems to be accelerating. Overall the total decrease in median log hazard is about 0.7 units, or about a 50% reduction in hazard over the observation period. A cancer epidemiologist would likely be unsurprised by this decline, since it coincides with the recent rise in the use of mammography by American women.

14.3.2 Bayesian model choice

For model choice, we again turn to the DIC criterion. The first six lines of Table 14.8 provide p_D and DIC values for our full model and several simplications thereof. Note the full model (sixth line) is estimated to have only just over 150 effective parameters, a substantial reduction (recall there are $2 \times 99 \times 26 = 5148$ random frailty parameters alone). Removing the spatial frailties W_{ik} from the log-relative hazard has little impact on p_D, but substantial negative impact on the DIC score. By contrast, removing the nonspatial frailties V_{ik} reduces (i.e., improves) both p_D and DIC, consistent with our findings in the previous subsection. Further simplifying the model to having a single set of spatial frailties W_i that do not vary with time (but now also reinserting year-specific intercepts α_k) has little effect on p_D but does improve DIC a bit more (though this improvement appears only slightly larger than the order of Monte Carlo error in our calculations). Even more drastic simplifications (eliminating the W_i, and perhaps even the α_k) lead to further drops in p_D, but at the cost of

Figure 14.6 *Boxplots of posterior medians for the spatial frailties W_{ik} over the Iowa counties for each year, k=1973, . . . , 1998.*

Baseline hazard	Log-relative hazard	p_D	DIC
Semiparametric mixture	$\boldsymbol{\beta}^T\mathbf{x}_{ijk}$	6.17	780
Semiparametric mixture	$\boldsymbol{\beta}^T\mathbf{x}_{ijk} + \alpha_k$	33.16	743
Semiparametric mixture	$\boldsymbol{\beta}^T\mathbf{x}_{ijk} + \alpha_k + W_i$	80.02	187
Semiparametric mixture	$\boldsymbol{\beta}^T\mathbf{x}_{ijk} + W_{ik}$	81.13	208
Semiparametric mixture	$\boldsymbol{\beta}^T\mathbf{x}_{ijk} + V_{ik}$	149.45	732
Semiparametric mixture	$\boldsymbol{\beta}^T\mathbf{x}_{ijk} + W_{ik} + V_{ik}$	151.62	280
Weibull	$\boldsymbol{\beta}^T\mathbf{x}_{ijk} + \alpha_k + W_i$	79.22	221
Weibull	$\boldsymbol{\beta}^T\mathbf{x}_{ijk} + W_{ik}$	80.75	239
Weibull	$\boldsymbol{\beta}^T\mathbf{x}_{ijk} + W_{ik} + V_{ik}$	141.67	315

Table 14.8 *DIC and effective number of parameters p_D for the competing models.*

unacceptably large increases in DIC. Thus our county-level breast cancer survival data seem to have strong spatial structure that is still unaccounted for by the covariates in Table 14.7, but structure that is fairly similar for all diagnosis years.

The last three lines of Table 14.8 reconsider the best three log-relative hazard models above, but where we now replace the semiparametric mixture baseline hazard with a Weibull hazard having region-specific baseline hazards $h_{0i}(t_{ijk};\rho_i) = \rho_i t_{ijk}^{\rho_i - 1}$ (note the spatial frailties play the role of the second parameter customarily associated with the Weibull model). These fully parametric models offer small advantages in terms of parsimony (smaller p_D), but these gains are apparently more than outweighed by a corresponding degradation in fit (much larger DIC score).

14.4 Multivariate models ⋆

In this section we extend to multivariate spatial frailty modeling, using the MCAR model introduced in Section 10.1. In particular, we use a semiparametric model, and consider MCAR structure on both residual (spatial frailty) and regression (space-varying coefficient) terms. We also extend to the spatiotemporal case by including temporally correlated cohort effects (say, one for each year of initial disease diagnosis) that can be summarized and plotted over time. Example 14.4 illustrates the utility of our approach in an analysis of survival times of patients suffering from one or more types of cancer. We obtain posterior estimates of key fixed effects, smoothed maps of both frailties and spatially varying coefficients, and compare models using the DIC criterion.

14.4.1 Static spatial survival data with multiple causes of death

Consider the following multivariate survival setting. Let t_{ijk} denote the time to death or censoring for the kth patient having the jth type of primary cancer living in the ith county, $i = 1, \ldots, n$, $j = 1, \ldots, p$, $k = 1, \ldots, s_{ij}$, and let γ_{ijk} be the corresponding death indicator. Let us write $\mathbf{x}_{ijk}$ as the vector of covariates for the above individual, and let $\mathbf{z}_{ijk}$ denote the vector of cancer indicators for this individual. That is, $\mathbf{z}_{ijk} = (z_{ijk1}, z_{ijk2}, ..., z_{ijkp})^T$ where $z_{ijkl} = 1$ if patient ijk suffers from cancer type l, and 0 otherwise (note that $z_{ijkj} = 1$ by definition). Then we can write the likelihood of our proportional hazards model $L(\boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{\Phi}; \mathbf{t}, \mathbf{x}, \boldsymbol{\gamma})$ as

$$\begin{aligned} &\textstyle\prod_{i=1}^{n} \prod_{j=1}^{p} \prod_{k=1}^{s_{ij}} \left\{h\left(t_{ijk}; \mathbf{x}_{ijk}, \mathbf{z}_{ijk}\right)\right\}^{\gamma_{ijk}} \\ &\quad \times \exp\left\{-H_{0i}\left(t_{ijk}\right) \exp\left(\mathbf{x}_{ijk}^T \boldsymbol{\beta} + \mathbf{z}_{ijk}^T \boldsymbol{\theta} + \phi_{ij}\right)\right\}, \end{aligned} \tag{14.11}$$

where

$$h\left(t_{ijk}; \mathbf{x}_{ijk}, \mathbf{z}_{ijk}\right) = h_{0i}\left(t_{ijk}\right) \exp\left(\mathbf{x}_{ijk}^T \boldsymbol{\beta} + \mathbf{z}_{ijk}^T \boldsymbol{\theta} + \phi_{ij}\right) \ . \tag{14.12}$$

Here, $H_{0i}\left(t_{ijk}\right) = \int_0^{t_{ijk}} h_{0i}(u)\, du$, $\boldsymbol{\phi}_i = (\phi_{i1}, \phi_{i2}, ..., \phi_{in})^T$, $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$ are given flat priors, and

$$\boldsymbol{\Phi} \equiv \left(\boldsymbol{\phi}_1^T, \ldots, \boldsymbol{\phi}_n^T\right)^T \sim MCAR(\rho, \boldsymbol{\Sigma}) \ ,$$

using the notation of Section 10.1. The region-specific baseline hazard functions $h_{0i}\left(t_{ijk}\right)$ are modeled using the beta mixture approach (Subsection 14.2.1) in such a way that the intercept in $\boldsymbol{\beta}$ remains estimable. We note that we could extend to a county *and* cancer-specific baseline hazard h_{0ij}; however, preliminary exploratory analyses of our data suggest such generality is not needed here.

Several alternatives to model formulation (14.12) immediately present themselves. For example, we could convert to a space-varying coefficients model (Assunção, 2003), replacing the log-relative hazard $\mathbf{x}_{ijk}^T \boldsymbol{\beta} + \mathbf{z}_{ijk}^T \boldsymbol{\theta} + \phi_{ij}$ in (14.12) with

$$\mathbf{x}_{ijk}^T \boldsymbol{\beta} + \mathbf{z}_{ijk}^T \boldsymbol{\theta}_i \ , \tag{14.13}$$

where $\boldsymbol{\beta}$ again has a flat prior, but $\boldsymbol{\Theta} \equiv \left(\boldsymbol{\theta}_1^T, \ldots, \boldsymbol{\theta}_n^T\right)^T \sim MCAR(\rho, \boldsymbol{\Sigma})$. In Example 14.4 we apply this method to our cancer data set; we defer mention of still other log-relative hazard modeling possibilities until after this illustration.

14.4.2 MCAR specification, simplification, and computing

To efficiently implement the $MCAR(\rho, \boldsymbol{\Sigma})$ as a prior distribution for our spatial process, suppose that we are using the usual 0-1 adjacency weights in W. Then recall from equation

(10.4) that we may express the MCAR precision matrix $\mathbf{B} \equiv \mathbf{\Sigma}_{\boldsymbol{\phi}}^{-1}$ in terms of the $n \times n$ adjacency matrix W as

$$\mathbf{B} = (Diag(m_i) - \rho W) \otimes \Lambda \ ,$$

where we have added a propriety parameter ρ. Note that this is a Kronecker product of an $n \times n$ and a $p \times p$ matrix, thereby rendering $\mathbf{B}$ as $np \times np$ as required. In fact, $\mathbf{B}$ may be looked upon as the Kronecker product of two partial precision matrices: one for the spatial components, $(Diag(m_i) - \rho W)$ (depending upon their adjacency structure and number of neighbors), and another for the variation across diseases, given by Λ. We thus alter our notation slightly to $MCAR\,(\rho, \Lambda)$.

Also as a consequence of this form, a sufficient condition for positive definiteness of the dispersion matrix for this MCAR model becomes $|\rho| < 1$ (as in the univariate case). Negative smoothness parameters are not desirable, so we typically take $0 < \rho < 1$. We can now complete the Bayesian hierarchical formulation by placing appropriate priors on ρ (say, a $Unif\,(0,1)$ or $Beta(18,2)$) and Λ (say, a $Wishart\,(\rho, \Lambda_0)$).

The Gibbs sampler is the MCMC method of choice here, particularly because, as in the univariate case, it takes advantage of the MCAR's conditional specification. Adaptive rejection sampling may be used to sample the regression coefficients $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$, while Metropolis steps with (possibly multivariate) Gaussian proposals may be employed for the spatial effects $\mathbf{\Phi}$. The full conditional for ρ is nicely suited for slice sampling (see Subsection 5.3.3), given its bounded support. Finally, the full conditional for Λ^{-1} emerges in closed form as an inverted Wishart distribution.

We conclude this subsection by recalling that our model can be generalized to admit different propriety parameters ρ_j for different diseases (recall Section 10.1 and, in particular, the discussion surrounding (10.8) and (10.9)). We notate this model as $MCAR\,(\boldsymbol{\rho}, \Lambda)$, where $\boldsymbol{\rho} = (\rho_1, \ldots, \rho_p)^T$.

14.4.3 Spatiotemporal survival data

Here we extend our model to allow for cohort effects. Let r index the year in which patient ijk entered the study (i.e., the year in which the patient's primary cancer was diagnosed). Extending model (14.12) we obtain the log-relative hazard,

$$\mathbf{x}_{ijkr}^T \boldsymbol{\beta} + \mathbf{z}_{ijkr}^T \boldsymbol{\theta} + \phi_{ijr} \ , \tag{14.14}$$

with the obvious corresponding modifications to the likelihood (14.11). Here, $\boldsymbol{\phi}_{ir} = (\phi_{i1r}, \phi_{i2r}, ..., \phi_{ipr})^T$ and $\mathbf{\Phi}_r = \left(\boldsymbol{\phi}_{1r}^T, \ldots, \boldsymbol{\phi}_{nr}^T\right)^T \overset{ind}{\sim} MCAR\,(\rho_r, \Lambda_r)$. This permits addition of an exchangeable prior structure,

$$\rho_r \overset{iid}{\sim} Beta(a,b) \text{ and } \Lambda_r \overset{iid}{\sim} Wishart(\rho, \Lambda_0) \ ,$$

where we may choose fixed values for a, b, ρ, and Λ_0, or place hyperpriors on them and estimate them from the data. Note also the obvious extension to disease-specific ρ_{jr}, as mentioned at the end of the previous subsection.

Example 14.4 *(Application to Iowa SEER multiple cancer survival data).* We illustrate the approach with an analysis of SEER data on 17,146 patients from the 99 counties of the state of Iowa who have been diagnosed with cancer between 1992 and 1998, and who have a well-identified primary cancer. Our covariate vector $\mathbf{x}_{ijk}$ consists of a constant (intercept), a gender indicator, the age of the patient, indicators for race with "white" as the baseline, indicators for the stage of the primary cancer with "local" as the baseline, and indicators for year of primary cancer diagnosis (cohort) with the first year (1992) as the baseline. The vector $\mathbf{z}_{ijk}$ comprises the indicators of which cancers the patient has; the corresponding

Variable	2.5%	50%	97.5%
Intercept	0.102	0.265	0.421
Sex (female = 0)	0.097	0.136	0.182
Age	0.028	0.029	0.030
Stage of primary cancer (local = 0)			
regional	0.322	0.373	0.421
distant	1.527	1.580	1.654
Type of primary cancer			
colorectal	0.112	0.252	0.453
gallbladder	1.074	1.201	1.330
pancreas	1.603	1.701	1.807
small intestine	0.128	0.287	0.445
stomach	1.005	1.072	1.141

Table 14.9 *Posterior quantiles for the fixed effects in the MCAR frailty model.*

parameters will thus capture the effect of these cancers on the hazards regardless of whether they emerge as primary or secondary.

With regard to modeling details, we used five separate (cancer-specific) propriety parameters ρ_j having an exchangeable $Beta(18, 2)$ prior, and a vague $Wishart\,(\rho, \Lambda_0)$ for Λ, where $\rho = 5$ and $\Lambda_0 = 0.01 I_{5\times 5}$. (Results for $\boldsymbol{\beta}, \boldsymbol{\theta}$, and $\boldsymbol{\Phi}$ under a $U(0, 1)$ prior for the ρ_j were broadly similar.) Table 14.9 gives posterior summaries for the main effects $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$; note that $\boldsymbol{\theta}$ is estimable despite the presence of the intercept since many individuals have more than one cancer. No race or cohort effects emerged as significantly different from zero, so they have been deleted; all remaining effects are shown here. All of these effects are significant and in the directions one would expect. In particular, the five cancer effects are consistent with results of previous modeling of this and similar data sets, with pancreatic cancer emerging as the most deadly (posterior median log relative hazard 1.701) and colorectal and small intestinal cancer relatively less so (.252 and .287, respectively).

Table 14.10 gives posterior variance and correlation summaries for the frailties ϕ_{ij} among the five cancers for two representative counties, Dallas (urban; Des Moines area) and Clay (rural northwest). Note that the correlations are as high as 0.528 (pancreas and stomach in Dallas County), suggesting the need for the multivariate structure inherent in our MCAR frailty model. Note also that summarizing the posterior distribution of Λ^{-1} would be inappropriate here, since despite the Kronecker structure here, Λ^{-1} cannot be directly interpreted as a primary cancer covariance matrix across counties.

Turning to geographic summaries, Figure 14.7 shows `ArcView` maps of the posterior means of the MCAR spatial frailties ϕ_{ij}. Recall that in this model, the ϕ_{ij} play the role of spatial residuals, capturing any spatial variation not already accounted for by the spatial main effects $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$. The lack of spatial pattern in these maps suggest there is little additional spatial "story" in the data beyond what is already being told by the fixed effects. However, the map scales reveal that one cancer (gallbladder) is markedly different from the others, both in terms of total range of the mean frailties (rather broad) and their center (negative; the other four are centered near 0).

Next, we change from the MCAR spatial frailty model to the MCAR spatially varying coefficients model (14.13). This model required a longer burn-in period (20,000 instead of 10,000), but otherwise our prior and MCMC control parameters remain unchanged. Figure 14.8 shows `ArcView` maps of the resulting posterior means of the spatially varying coefficients θ_{ij}. Unlike the ϕ_{ij} in the previous model, these parameters are not "residuals," but the effects of the presence of the primary cancer indicated on the death rate in each county. Clearly these maps show a strong spatial pattern, with (for example) southwestern

Dallas County	Colo-rectal	Gall-bladder	Pancreas	Small intestine	Stomach
Colorectal	0.852	0.262	0.294	0.413	0.464
Gallbladder		1.151	0.314	0.187	0.175
Pancreas			0.846	0.454	0.528
Small intestine				1.47	0.413
Stomach					0.908

Clay County	Colo-rectal	Gall-bladder	Pancreas	Small intestine	Stomach
Colorectal	0.903	0.215	0.273	0.342	0.352
Gallbladder		1.196	0.274	0.128	0.150
Pancreas			0.852	0.322	0.402
Small intestine				1.515	0.371
Stomach					1.068

Table 14.10 *Posterior variances and correlation summaries, Dallas and Clay Counties, MCAR spatial frailty model. Diagonal elements are estimated variances, while off-diagonal elements are estimated correlations.*

Log-relative hazard model	p_D	DIC
$\mathbf{x}_{ijk}^T\boldsymbol{\beta} + \mathbf{z}_{ijk}^T\boldsymbol{\theta}$	10.97	642
$\mathbf{x}_{ijk}^T\boldsymbol{\beta} + \mathbf{z}_{ijk}^T\boldsymbol{\theta} + \phi_i,\ \boldsymbol{\phi} \sim CAR(\rho, \lambda)$	103.95	358
$\mathbf{x}_{ijk}^T\boldsymbol{\beta} + \mathbf{z}_{ijk}^T\boldsymbol{\theta} + \phi_{ij},\ \boldsymbol{\Phi} \sim MCAR(\rho = 1, \Lambda)$	172.75	247
$\mathbf{x}_{ijk}^T\boldsymbol{\beta} + \mathbf{z}_{ijk}^T\boldsymbol{\theta} + \phi_{ij},\ \boldsymbol{\Phi} \sim MCAR(\rho, \Lambda)$	172.40	246
$\mathbf{x}_{ijk}^T\boldsymbol{\beta} + \mathbf{z}_{ijk}^T\boldsymbol{\theta} + \phi_{ij},\ \boldsymbol{\Phi} \sim MCAR(\rho_1, \ldots, \rho_5, \Lambda)$	175.71	237
$\mathbf{x}_{ijk}^T\boldsymbol{\beta} + \mathbf{z}_{ijk}^T\boldsymbol{\theta} + \phi_{ij} + \epsilon_{ij},\ \boldsymbol{\Phi} \sim MCAR(\rho_1, \ldots, \rho_5, \Lambda),$ $\epsilon_{ij} \overset{iid}{\sim} N(0, \tau^2)$	177.25	255
$\mathbf{x}_{ijk}^T\boldsymbol{\beta} + \mathbf{z}_{ijk}^T\boldsymbol{\theta}_i,\ \boldsymbol{\Theta} \sim MCAR(\rho, \Lambda)$	169.42	235
$\mathbf{x}_{ijk}^T\boldsymbol{\beta} + \mathbf{z}_{ijk}^T\boldsymbol{\theta}_i,\ \boldsymbol{\Theta} \sim MCAR(\rho_1, \ldots, \rho_5, \Lambda)$	171.46	229

Table 14.11 *DIC comparison, spatial survival models for the Iowa cancer data.*

Iowa counties having relatively high fitted values for pancreatic and stomach cancer, while southeastern counties fare relatively poorly with respect to colorectal and small intestinal cancer. The overall levels for each cancer are consistent with those given for the corresponding fixed effects $\boldsymbol{\theta}$ in Table 14.9 for the spatial frailty model.

Table 14.11 gives the effective model sizes p_D and DIC scores for a variety of spatial survival models. The first two listed (fixed effects only and standard CAR frailty) have few effective parameters, but also poor (large) DIC scores. The MCAR spatial frailty models (which place the MCAR on $\boldsymbol{\Phi}$) fare better, especially when we add the disease-specific ρ_j (the model summarized in Table 14.9, Table 14.10, and Figure 14.7). However, adding heterogeneity effects ϵ_{ij} to this model adds essentially no extra effective parameters, and is actually harmful to the overall DIC score (since we are adding complexity for little or no benefit in terms of fit). Finally, the two spatially varying coefficients models enjoy the best (smallest) DIC scores, but only by a small margin over the best spatial frailty model.

Finally, we fit the spatiotemporal extension (14.14) of our MCAR frailty model to the data where the cohort effect (year of study entry r) is taken into account. Year-by-year boxplots of the posterior median frailties (Figure 14.9) reveal the expected steadily decreasing trend for all five cancers, though it is not clear how much of this decrease is simply an artifact of the censoring of survival times for patients in more recent cohorts. The spatiotemporal

Figure 14.7 *Posterior mean spatial frailties, Iowa cancer data, static spatial MCAR model.*

extension of the spatially varying coefficients model (14.13) (i.e., $\mathbf{x}_{ijkr}^T\boldsymbol{\beta} + \mathbf{z}_{ijkr}^T\boldsymbol{\theta}_{ir}$) might well produce results that are temporally more interesting in this case. Incorporating change points, cancer start date measurement errors, and other model enhancements (say, interval censoring) might also be practically important model enhancements here.

14.5 Spatial cure rate models ⋆

In Section 14.1 we investigated spatially correlated frailties in traditional parametric survival models, choosing a random effects distribution to reflect the spatial structure in the problem. Sections 14.2 and 14.3 extended this approach to spatial and spatiotemporal settings within a semiparametric model.

In this section our ultimate goal is the proper analysis of a geographically referenced smoking cessation study, in which we observe subjects periodically through time to check for relapse following an initial quit attempt. Each patient is observed once each year for five consecutive years, whereupon the current average number of cigarettes smoked at each visit is recorded, along with the zip code of residence and several other potential explanatory variables. This data set requires us to extend the work of Carlin and Hodges (1999) in a number of ways. The primary extension involves the incorporation of a *cure fraction* in our models. In investigating the effectiveness of quitting programs, data typically reveal many former smokers having successfully given up smoking, and as such may be thought

Figure 14.8 *Posterior mean spatially varying coefficients, Iowa cancer data, static spatial MCAR model.*

of as "cured" of the deleterious habit. Incorporating such cure fractions in survival models leads to *cure rate models*, which are often applied in survival settings where the endpoint is a particular disease (say, breast cancer) which the subject may never reexperience. These models have a long history in the biostatistical literature, with the most popular perhaps being that of Berkson and Gage (1952). This model has been extensively studied in the statistical literature by a number of authors, including Farewell (1982, 1986), Goldman (1984), and Ewell and Ibrahim (1997). Recently, cure rates have been studied in somewhat more general settings by Chen, Ibrahim, and Sinha (1999) following earlier work by Yakovlev and Tsodikov (1996).

In addition, while this design can be analyzed as an ordinary right-censored survival model (with relapse to smoking as the endpoint), the data are perhaps more accurately viewed as *interval-censored*, since we actually observe only approximately annual intervals within which a failed quitter resumed smoking. We will consider both right- and interval-censored models, where in the former case we simply approximate the time of relapse by the midpoint of the corresponding time interval. Finally, we capture spatial variation through zip code-specific spatial random effects in the cure fraction or the hazard function, which in either case may act as spatial frailties. We find that incorporating the covariates and frailties into the hazard function is most natural (both intuitively and methodologically), especially after adopting a Weibull form for the baseline hazard.

Figure 14.9 *Boxplots of posterior medians for the spatial frailties ϕ_{ijr} over the 99 Iowa counties for each year, $r = 1973, \ldots, 1998$.*

14.5.1 Models for right- and interval-censored data

14.5.1.1 Right-censored data

Our cure rate models are based on those of Chen et al. (1999) and derived assuming that some latent biological process is generating the observed data. Suppose there are I regions and n_i patients in the ith region. We denote by T_{ij} the random variable for time to event (relapse, in our case) of the jth person in the ith region, where $j = 1, 2, ..., n_i$ and $i = 1, 2, ..., I$. (While acknowledging the presence of the regions in our notation, we postpone explicit spatial modeling to the next section.) Suppose that the (i, j)th individual has N_{ij} potential latent (unobserved) risk factors, the presence of any of which (i.e., $N_{ij} \geq 1$) will ultimately lead to the event. For example, in cancer settings these factors may correspond to metastasis-competent tumor cells within the individual. Typically, there will be a number of subjects who do not undergo the event during the observation period, and are therefore considered censored. Thus, letting U_{ijk}, $k = 1, 2, ..., N_{ij}$ be the time to an event arising from the kth latent factor for the (i, j)th individual, the observed time to event for an uncensored individual is generated by $T_{ij} = \min\{U_{ijk},\ k = 1, 2, ..., N_{ij}\}$. If the (i, j)th individual is right-censored at time t_{ij}, none of the latent factors has led to an event by that time, and clearly $T_{ij} > t_{ij}$ (and in fact $T_{ij} = \infty$ if $N_{ij} = 0$).

Given N_{ij}, the U_{ijk}'s are independent with survival function $S(t|\Psi_{ij})$ and corresponding density function $f(t|\Psi_{ij})$. The parameter Ψ_{ij} is a collection of all the parameters

(including possible regression parameters) that may be involved in a parametric specification for the survival function S. In this section we will work with a two-parameter Weibull distribution specification for the density function $f(t|\Psi_{ij})$, where we allow the Weibull scale parameter ρ to vary across the regions, and η, which may serve as a link to covariates in a regression setup, to vary across individuals. Therefore $f(t|\rho_i, \eta_{ij}) = \rho_i t^{\rho_i - 1} \exp(\eta_{ij} - t^{\rho_i} \exp(\eta_{ij}))$.

In terms of the hazard function h, $f(t|\rho_i, \eta_{ij}) = h(t|\rho_i, \eta_{ij}) S(t|\rho_i, \eta_{ij})$, with $h(t; \rho_i, \eta_{ij}) = \rho_i t^{\rho_i - 1} \exp(\eta_{ij})$ and $S(t|\rho_i, \eta_{ij}) = \exp(-t^{\rho_i} \exp(\eta_{ij}))$. Note we implicitly assume proportional hazards, with baseline hazard function $h_0(t|\rho_i) = \rho_i t^{\rho_i - 1}$. Thus an individual ij who is censored at time t_{ij} before undergoing the event contributes $(S(t_{ij}|\rho_i, \eta_{ij}))^{N_{ij}}$ to the likelihood, while an individual who experiences the event at time t_{ij} contributes $N_{ij}(S(t_{ij}|\rho_i, \eta_{ij}))^{N_{ij}-1} f(t_{ij}|\rho_i, \eta_{ij})$. The latter expression follows from the fact that the event is experienced when any one of the latent factors occurs. Letting ν_{ij} be the observed event indicator for individual ij, this person contributes

$$\begin{aligned} &L(t_{ij}|N_{ij}, \rho_i, \eta_{ij}, \nu_{ij}) \\ &\quad = (S(t_{ij}|\rho_i, \eta_{ij}))^{N_{ij}(1-\nu_{ij})} \left(N_{ij} S(t_{ij}|\rho_i, \eta_{ij})^{N_{ij}-1} f(t_{ij}|\rho_i, \eta_{ij}) \right)^{\nu_{ij}} , \end{aligned}$$

and the joint likelihood for all the patients can now be expressed as

$$\begin{aligned} &L(\{t_{ij}\} \,|\, \{N_{ij}\}, \{\rho_i\}, \{\eta_{ij}\}, \{\nu_{ij}\}) \\ &\quad = \textstyle\prod_{i=1}^{I} \prod_{j=1}^{n_i} L(t_{ij}|N_{ij}, \rho_i, \eta_{ij}, \nu_{ij}) \\ &\quad = \textstyle\prod_{i=1}^{I} \prod_{j=1}^{n_i} (S(t_{ij}|\rho_i, \eta_{ij}))^{N_{ij}(1-\nu_{ij})} \\ &\qquad\qquad \times \left(N_{ij} S(t_{ij}\rho_i, \eta_{ij})^{N_{ij}-1} f(t_{ij}|\rho_i, \eta_{ij}) \right)^{\nu_{ij}} \\ &\quad = \textstyle\prod_{i=1}^{I} \prod_{j=1}^{n_i} (S(t_{ij}|\rho_i, \eta_{ij}))^{N_{ij}-\nu_{ij}} (N_{ij} f(t_{ij}|\rho_i, \eta_{ij}))^{\nu_{ij}} . \end{aligned}$$

This expression can be rewritten in terms of the hazard function as

$$\prod_{i=1}^{I} \prod_{j=1}^{n_i} (S(t_{ij}|\rho_i, \eta_{ij}))^{N_{ij}} (N_{ij} h(t_{ij}|\rho_i, \eta_{ij}))^{\nu_{ij}} . \tag{14.15}$$

A Bayesian hierarchical formulation is completed by introducing prior distributions on the parameters. We will specify independent prior distributions $p(N_{ij}|\theta_{ij})$, $p(\rho_i|\psi_\rho)$ and $p(\eta_{ij}|\psi_\eta)$ for $\{N_{ij}\}$, $\{\rho_i\}$, and $\{\eta_{ij}\}$, respectively. Here, ψ_ρ, ψ_η, and $\{\theta_{ij}\}$ are appropriate hyperparameters. Assigning independent hyperpriors $p(\theta_{ij}|\psi_\theta)$ for $\{\theta_{ij}\}$ and assuming the hyperparameters $\psi = (\psi_\rho, \psi_\eta, \psi_\theta)$ to be fixed, the posterior distribution for the parameters, $p(\{\theta_{ij}\}, \{\eta_{ij}\}, \{N_{ij}\}, \{\rho_i\} \,|\, \{t_{ij}\}, \{\nu_{ij}\})$, is easily found (up to a proportionality constant) using (14.15) as

$$\begin{aligned} &\textstyle\prod_{i=1}^{I} \Big\{ p(\rho_i|\psi_\rho) \prod_{j=1}^{n_i} [S(t_{ij}|\rho_i, \eta_{ij})]^{N_{ij}} [N_{ij} h(t_{ij}|\rho_i, \eta_{ij})]^{\nu_{ij}} \\ &\qquad \times p(N_{ij}|\theta_{ij}) \, p(\eta_{ij}|\psi_\eta) \, p(\theta_{ij}|\psi_\theta) \Big\} . \end{aligned}$$

Chen et al. (1999) assume that the N_{ij} are distributed as independent Poisson random variables with mean θ_{ij}, i.e., $p(N_{ij}|\theta_{ij})$ is $Poisson(\theta_{ij})$. In this setting it is easily seen that the survival distribution for the (i,j)th patient, $P(T_{ij} \geq t_{ij}|\rho_i, \eta_{ij})$, is given by $\exp\{-\theta_{ij}(1 - S(t_{ij}|\rho_i, \eta_{ij}))\}$. Since $S(t_{ij}|\rho_i, \eta_{ij})$ is a proper survival function (corresponding to the latent factor times U_{ijk}), as $t_{ij} \to \infty$, $P(T_{ij} \geq t_{ij}|\rho_i, \eta_{ij}) \to \exp(-\theta_{ij}) > 0$. Thus we have a subdistribution for T_{ij} with a *cure fraction* given by $\exp(-\theta_{ij})$. Here a hyperprior on the θ_{ij}'s would have support on the positive real line.

While there could certainly be multiple latent factors that increase the risk of smoking relapse (age started smoking, occupation, amount of time spent driving, tendency toward addictive behavior, etc.), this is rather speculative and certainly not as justifiable as in

the cancer setting for which the multiple factor approach was developed (where $N_{ij} > 1$ is biologically motivated). As such, we instead form our model using a single, omnibus, "propensity for relapse" latent factor. In this case, we think of N_{ij} as a *binary* variable, and specify $p(N_{ij}|\theta_{ij})$ as Bernoulli $(1-\theta_{ij})$. In this setting it is easier to look at the survival distribution after marginalizing out the N_{ij}. In particular, note that

$$P(T_{ij} \geq t_{ij}|\rho_i, \eta_{ij}, N_{ij}) = \begin{cases} S(t_{ij}|\rho_i, \eta_{ij}) \ , & N_{ij} = 1 \\ 1 \ , & N_{ij} = 0 \end{cases} .$$

That is, if the latent factor is absent, the subject is cured (does not experience the event). Marginalizing over the Bernoulli distribution for N_{ij}, we obtain for the (i,j)th patient the survival function $S^*(t_{ij}|\theta_{ij}, \rho_i, \eta_{ij}) \equiv P(T_{ij} \geq t_{ij}|\rho_i, \eta_{ij}) = \theta_{ij} + (1-\theta_{ij}) S(t_{ij}|\rho_i, \eta_{ij})$, which is the classic cure-rate model attributed to Berkson and Gage (1952) with cure fraction θ_{ij}. Now we can write the likelihood function for the data marginalized over $\{N_{ij}\}$, $L(\{t_{ij}\} \mid \{\rho_i\}, \{\theta_{ij}\}, \{\eta_{ij}\}, \{\nu_{ij}\})$, as

$$\begin{aligned} &\textstyle\prod_{i=1}^{I} \prod_{j=1}^{n_i} \left[S^*(t_{ij}|\theta_{ij}, \rho_i, \eta_{ij})\right]^{1-\nu_{ij}} \left(-\frac{d}{dt_{ij}} S^*(t_{ij}|\theta_{ij}, \rho_i, \eta_{ij})\right)^{\nu_{ij}} \\ &= \textstyle\prod_{i=1}^{I} \prod_{j=1}^{n_i} \left[S^*(t_{ij}|\theta_{ij}, \rho_i, \eta_{ij})\right]^{1-\nu_{ij}} \left[(1-\theta_{ij}) f(t_{ij}|\rho_i, \eta_{ij})\right]^{\nu_{ij}} \ , \end{aligned}$$

which in terms of the hazard function becomes

$$\prod_{i=1}^{I} \prod_{j=1}^{n_i} \left[S^*(t_{ij}|\theta_{ij}, \rho_i, \eta_{ij})\right]^{1-\nu_{ij}} \left[(1-\theta_{ij}) S(t_{ij}|\rho_i, \eta_{ij}) h(t_{ij}|\rho_i, \eta_{ij})\right]^{\nu_{ij}} \ , \tag{14.16}$$

where the hyperprior for θ_{ij} has support on $(0,1)$. Now the posterior distribution of the parameters is proportional to

$$L(\{t_{ij}\} \mid \{\rho_i\}, \{\theta_{ij}\}, \{\eta_{ij}\}, \{\nu_{ij}\}) \prod_{i=1}^{I} \left\{ p(\rho_i|\psi_\rho) \prod_{j=1}^{n_i} p(\eta_{ij}|\psi_\eta) p(\theta_{ij}|\psi_\theta) \right\} . \tag{14.17}$$

Turning to the issue of incorporating covariates, in the general setting with N_{ij} assumed to be distributed Poisson, Chen et al. (1999) propose their introduction in the cure fraction through a suitable link function g, so that $\theta_{ij} = g\left(\mathbf{x}_{ij}^T \widetilde{\boldsymbol{\beta}}\right)$, where g maps the entire real line to the positive axis. This is sensible when we believe that the risk factors affect the probability of an individual being cured. Proper posteriors arise for the regression coefficients $\widetilde{\boldsymbol{\beta}}$ even under improper priors. Unfortunately, this is no longer true when N_{ij} is Bernoulli (i.e., in the Berkson and Gage model). Vague but proper priors may still be used, but this makes the parameters difficult to interpret, and can often lead to poor MCMC convergence.

Since a binary N_{ij} seems most natural in our setting, we instead introduce covariates into $S(t_{ij}|\rho_i, \eta_{ij})$ through the Weibull link η_{ij}, i.e., we let $\eta_{ij} = \mathbf{x}_{ij}^T \boldsymbol{\beta}$. This seems intuitively more reasonable anyway, since now the covariates influence the underlying factor that brings about the smoking relapse (and thus the rapidity of this event). Also, proper posteriors arise here for $\boldsymbol{\beta}$ under improper posteriors even though N_{ij} is binary. As such, henceforth we will only consider the situation where the covariates enter the model in this way (through the Weibull link function). This means we are unable to separately estimate the effect of the covariates on both the *rate* of relapse and the *ultimate level* of relapse, but "fair" estimation here (i.e., allocating the proper proportions of the covariates' effects to each component) is not clear anyway since flat priors could be selected for $\boldsymbol{\beta}$, but not for $\widetilde{\boldsymbol{\beta}}$. Finally, all of our subsequent models also assume a constant cure fraction for the entire population (i.e., we set $\theta_{ij} = \theta$ for all i, j).

Note that the posterior distribution in (14.17) is easily modified to incorporate covariates. For example, with $\eta_{ij} = \mathbf{x}_{ij}^T \boldsymbol{\beta}$, we replace $\prod_{ij} p(\eta_{ij}|\psi_\eta)$ in (14.17) with $p(\boldsymbol{\beta}|\psi_\beta)$, with ψ_β as a fixed hyperparameter. Typically a flat or vague Gaussian prior may be taken for $p(\boldsymbol{\beta}|\psi_\beta)$.

14.5.1.2 Interval-censored data

The formulation above assumes that our observed data are right-censored. This means that we are able to observe the actual relapse time t_{ij} when it occurs prior to the final office visit. In reality, our study (like many others of its kind) is only able to determine patient status at the office visits themselves, meaning we observe only a time *interval* (t_{ijL}, t_{ijU}) within which the event (in our case, smoking relapse) is known to have occurred. For patients who did not resume smoking prior to the end of the study we have $t_{ijU} = \infty$, returning us to the case of right-censoring at time point t_{ijL}. Thus we now set $\nu_{ij} = 1$ if subject ij is interval-censored (i.e., experienced the event), and $\nu_{ij} = 0$ if the subject is right-censored.

Following Finkelstein (1986), the general interval-censored cure rate likelihood, $L\left(\{(t_{ijL}, t_{ijU})\} \mid \{N_{ij}\}, \{\rho_i\}, \{\eta_{ij}\}, \{\nu_{ij}\}\right)$, is given by

$$\prod_{i=1}^{I}\prod_{j=1}^{n_i}\left[S\left(t_{ijL}|\rho_i,\eta_{ij}\right)\right]^{N_{ij}-\nu_{ij}}\left\{N_{ij}\left[S\left(t_{ijL}|\rho_i,\eta_{ij}\right)-S\left(t_{ijU}|\rho_i,\eta_{ij}\right)\right]\right\}^{\nu_{ij}}$$
$$=\prod_{i=1}^{I}\prod_{j=1}^{n_i}\left[S\left(t_{ijL}|\rho_i,\eta_{ij}\right)\right]^{N_{ij}}\left\{N_{ij}\left(1-\frac{S\left(t_{ijU}|\rho_i,\eta_{ij}\right)}{S\left(t_{ijL}|\rho_i,\eta_{ij}\right)}\right)\right\}^{\nu_{ij}}.$$

As in the previous section, in the Bernoulli setup after marginalizing out the $\{N_{ij}\}$ the foregoing becomes $L\left(\{(t_{ijL}, t_{ijU})\} \mid \{\rho_i\}, \{\theta_{ij}\}, \{\eta_{ij}\}, \{\nu_{ij}\}\right)$, and can be written as

$$\prod_{i=1}^{I}\prod_{j=1}^{n_i}S^*\left(t_{ijL}|\theta_{ij},\rho_i,\eta_{ij}\right)\left\{1-\frac{S^*\left(t_{ijU}|\theta_{ij},\rho_i,\eta_{ij}\right)}{S^*\left(t_{ijL}|\theta_{ij},\rho_i,\eta_{ij}\right)}\right\}^{\nu_{ij}}. \tag{14.18}$$

We omit details (similar to those in the previous section) arising from the Weibull parametrization and subsequent incorporation of covariates through the link function η_{ij}.

14.5.2 Spatial frailties in cure rate models

The development of the hierarchical framework in the preceding section acknowledged the data as coming from I different geographical regions (clusters). Such clustered data are common in survival analysis and often modeled using cluster-specific frailties ϕ_i. As with the covariates, we will introduce the frailties ϕ_i through the Weibull link as intercept terms in the log-relative risk; that is, we set $\eta_{ij} = \mathbf{x}_{ij}^T\boldsymbol{\beta} + \phi_i$.

Here we allow the ϕ_i to be spatially correlated across the regions; similarly we would like to permit the Weibull baseline hazard parameters, ρ_i, to be spatially correlated. A natural approach in both cases is to use a univariate CAR prior. While one may certainly employ separate, independent CAR priors on $\boldsymbol{\phi}$ and $\boldsymbol{\zeta} \equiv \{\log \rho_i\}$, another option is to allow these two spatial priors to themselves be correlated. In other words, we may want a bivariate spatial model for the $\delta_i = (\phi_i, \zeta_i)^T = (\phi_i, \log \rho_i)^T$. As mentioned in Sections 10.1 and 14.4, we may use the MCAR distribution for this purpose. In our setting, the MCAR distribution on the concatenated vector $\boldsymbol{\delta} = (\boldsymbol{\phi}^T, \boldsymbol{\zeta}^T)^T$ is Gaussian with mean $\mathbf{0}$ and precision matrix $\Lambda^{-1} \otimes (Diag\,(m_i) - \rho W)$, where Λ is a 2×2 symmetric and positive definite matrix, $\rho \in (0, 1)$, and m_i and W remain as above. In the current context, we may also wish to allow different smoothness parameters (say, ρ_1 and ρ_2) for $\boldsymbol{\phi}$ and $\boldsymbol{\zeta}$, respectively, as in Section 14.4. Henceforth, in this section we will denote the proper MCAR with a common smoothness parameter by $MCAR\,(\rho, \Lambda)$, and the multiple smoothness parameter generalized MCAR by $MCAR\,(\rho_1, \rho_2, \Lambda)$. Combined with independent (univariate) CAR models for $\boldsymbol{\phi}$ and $\boldsymbol{\zeta}$, these offer a broad range of potential spatial models.

14.5.3 Model comparison

Suppose we let Ω denote the set of all model parameters, so that the deviance statistic (5.10) becomes

$$D(\Omega) = -2\log f(\mathbf{y}|\Omega) + 2\log h(\mathbf{y}) \ . \tag{14.19}$$

When DIC is used to compare nested models in standard exponential family settings, the unnormalized likelihood $L(\Omega;\mathbf{y})$ is often used in place of the normalized form $f(\mathbf{y}|\Omega)$ in (14.19), since in this case the normalizing function $m(\Omega) = \int L(\Omega;\mathbf{y})d\mathbf{y}$ will be free of Ω and constant across models, hence contribute equally to the DIC scores of each (and thus have no impact on model selection). However, in settings where we require comparisons across different likelihood distributional forms, it appears one must be careful to use the properly scaled joint density $f(\mathbf{y}|\Omega)$ for each model.

We argue that use of the usual proportional hazards likelihood (which of course is not a joint density function) *is* in fact appropriate for DIC computation here, provided we make a fairly standard assumption regarding the relationship between the survival and censoring mechanisms generating the data. Specifically, suppose the distribution of the censoring times is independent of that of the survival times *and* does not depend upon the survival model parameters (i.e., independent, noninformative censoring). Let $g(t_{ij})$ denote the density of the censoring time for the ijth individual, with corresponding survival (1-cdf) function $R(t_{ij})$. Then the right-censored likelihood (14.16) can be extended to the joint likelihood specification,

$$\begin{aligned}&\textstyle\prod_{i=1}^{I}\prod_{j=1}^{n_i}\left[S^*\left(t_{ij}|\theta_{ij},\rho_i,\eta_{ij}\right)\right]^{1-\nu_{ij}}\\ &\quad\times\left[\left(1-\theta_{ij}\right)S\left(t_{ij}|\rho_i,\eta_{ij}\right)h\left(t_{ij}|\rho_i,\eta_{ij}\right)\right]^{\nu_{ij}}\left[R\left(t_{ij}\right)\right]^{\nu_{ij}}\left[g\left(t_{ij}\right)\right]^{1-\nu_{ij}} \ ,\end{aligned}$$

as for example in Le (1997, pp. 69–70). While not a joint probability density, this likelihood is still an everywhere nonnegative and integrable function of the survival model parameters Ω, and thus suitable for use with the Kullback-Leibler divergences that underlie DIC (Spiegelhalter et al., 2002, p. 586). But by assumption, $R(t)$ and $g(t)$ do not depend upon Ω. Thus, like an $m(\Omega)$ that is free of Ω, they may be safely ignored in both the p_D and DIC calculations. Note this same argument implies that we can use the unnormalized likelihood (14.16) when comparing not only nonnested parametric survival models (say, Weibull versus gamma), but even parametric and semiparametric models (say, Weibull versus Cox) provided our definition of "likelihood" is comparable across models.

Note also that here our "focus" (in the nomenclature of Spiegelhalter et al., 2002) is solely on Ω. An alternative would be instead to use a missing data formulation, where we include the likelihood contribution of $\{s_{ij}\}$, the collection of latent survival times for the right-censored individuals. Values for both Ω and the $\{s_{ij}\}$ could then be imputed along the lines given by Cox and Oakes (1984, pp. 165–166) for the EM algorithm or Spiegelhalter et al. (1995b, the "mice" example) for the Gibbs sampler. This would alter our focus from Ω to $(\Omega, \{s_{ij}\})$, and p_D would reflect the correspondingly larger effective parameter count.

Turning to the interval-censored case, here matters are only a bit more complicated. Converting the interval-censored likelihood (14.18) to a joint likelihood specification yields

$$\begin{aligned}&\prod_{i=1}^{I}\prod_{j=1}^{n_i}S^*\left(t_{ijL}|\theta_{ij},\rho_i,\eta_{ij}\right)\left(1-\frac{S^*\left(t_{ijU}|\theta_{ij},\rho_i,\eta_{ij}\right)}{S^*\left(t_{ijL}|\theta_{ij},\rho_i,\eta_{ij}\right)}\right)^{\nu_{ij}}\\ &\qquad\times\left[R\left(t_{ijL}\right)\right]^{\nu_{ij}}\left(1-\frac{R\left(t_{ijU}\right)}{R\left(t_{ijL}\right)}\right)^{\nu_{ij}}\left[g\left(t_{ijL}\right)\right]^{1-\nu_{ij}} \ .\end{aligned}$$

Now $\left[R\left(t_{ijL}\right)\right]^{\nu_{ij}}\left(1-R\left(t_{ijU}\right)/R\left(t_{ijL}\right)\right)^{\nu_{ij}}\left[g\left(t_{ijL}\right)\right]^{1-\nu_{ij}}$ is the function absorbed into $m(\Omega)$, and is again free of Ω. Thus again, use of the usual form of the interval-censored likelihood

Figure 14.10 *Map showing missingness pattern for the smoking cessation data: lightly shaded regions are those having no responses.*

presents no problems when comparing models within the interval-censored framework (including nonnested parametric models, or even parametric and semiparametric models).

Note that it does *not* make sense to compare a particular right-censored model with a particular interval-censored model. The form of the available data is different; model comparison is only appropriate to a given data set.

Example 14.5 *(Smoking cessation data).* We illustrate our methods using the aforementioned study of smoking cessation, a subject of particular interest in studies of lung health and primary cancer control. Described more fully by Murray et al. (1998), the data consist of 223 subjects who reside in 53 zip codes in the southeastern corner of Minnesota. The subjects, all of whom were smokers at study entry, were randomized into either a smoking intervention (SI) group, or a usual care (UC) group that received no special antismoking intervention. Each subject's smoking habits were monitored at roughly annual visits for five consecutive years. The subjects we analyze are actually the subset who are known to have quit smoking at least once during these five years, and our event of interest is whether they relapse (resume smoking) or not. Covariate information available for each subject includes sex, years as a smoker, and the average number of cigarettes smoked per day just prior to the quit attempt.

To simplify matters somewhat, we actually fit our spatial cure rate models over the 81 contiguous zip codes shown in Figure 14.10, of which only the 54 dark-shaded regions are those contributing patients to our data set. This enables our models to produce spatial predictions even for the 27 unshaded regions in which no study patients actually resided. All of our MCMC algorithms ran 5 initially overdispersed sampling chains, each for 20,000 iterations. Convergence was assessed using correlation plots, sample trace plots, and Gelman-Rubin (1992) statistics. In every case a burn-in period of 15,000 iterations appeared satisfactory. Retaining the remaining 5,000 samples from each chain yielded a final sample of 25,000 for posterior summarization.

Table 14.12 provides the DIC scores for a variety of random effects cure rate models in the interval-censored case. Models 1 and 2 have only random frailty terms ϕ_i with i.i.d. and CAR priors, respectively. Models 3 and 4 add random Weibull shape parameters $\zeta_i = \log \rho_i$, again with i.i.d. and CAR priors, respectively, independent of the priors for the ϕ_i. Finally, Models 5 and 6 consider the full MCAR structure for the (ϕ_i, ζ_i) pairs, assuming common and distinct spatial smoothing parameters, respectively. The DIC scores do not suggest that the more complex models are significantly better; apparently the data encourage a high degree of shrinkage in the random effects (note the low p_D scores). In what follows we present results for the "full" model (Model 6) in order to preserve complete generality, but emphasize that any of the models in Table 14.12 could be used with equal confidence.

Model	Log-relative risk	pD	DIC
1	$\mathbf{x}_{ij}^T\boldsymbol{\beta} + \phi_i;\ \phi_i \overset{iid}{\sim} N(0, \tau_\phi),\ \rho_i = \rho\ \forall\ i$	10.3	438
2	$\mathbf{x}_{ij}^T\boldsymbol{\beta} + \phi_i;\ \{\phi_i\} \sim CAR(\lambda_\phi),\ \rho_i = \rho\ \forall\ i$	9.4	435
3	$\mathbf{x}_{ij}^T\boldsymbol{\beta} + \phi_i;\ \phi_i \overset{iid}{\sim} N(0, \tau_\phi),\ \zeta_i \overset{iid}{\sim} N(0, \tau_\zeta)$	13.1	440
4	$\mathbf{x}_{ij}^T\boldsymbol{\beta} + \phi_i;\ \{\phi_i\} \sim CAR(\lambda_\phi),\ \{\zeta_i\} \sim CAR(\lambda_\zeta)$	10.4	439
5	$\mathbf{x}_{ij}^T\boldsymbol{\beta} + \phi_i;\ (\{\phi_i\}, \{\zeta_i\}) \sim MCAR(\rho, \Lambda)$	7.9	434
6	$\mathbf{x}_{ij}^T\boldsymbol{\beta} + \phi_i;\ (\{\phi_i\}, \{\zeta_i\}) \sim MCAR(\rho_\phi, \rho_\zeta, \Lambda)$	8.2	434

Table 14.12 *DIC and p_D values for various competing interval-censored models.*

Parameter	Median	(2.5%, 97.5%)
Intercept	–2.720	(–4.803, –0.648)
Sex (male = 0)	0.291	(–0.173, 0.754)
Duration as smoker	–0.025	(–0.059, 0.009)
SI/UC (usual care = 0)	–0.355	(–0.856, 0.146)
Cigarettes smoked per day	0.010	(–0.010, 0.030)
θ (cure fraction)	0.694	(0.602, 0.782)
ρ_ϕ	0.912	(0.869, 0.988)
ρ_ζ	0.927	(0.906, 0.982)
Λ_{11} (spatial variance component, ϕ_i)	0.005	(0.001, 0.029)
Λ_{22} (spatial variance component, ζ_i)	0.007	(0.002, 0.043)
$\Lambda_{12}/\sqrt{\Lambda_{11}\Lambda_{22}}$	0.323	(–0.746, 0.905)

Table 14.13 *Posterior quantiles, full model, interval-censored case.*

Table 14.13 presents estimated posterior quantiles (medians, and upper and lower .025 points) for the fixed effects $\boldsymbol{\beta}$, cure fraction θ, and hyperparameters in the interval-censored case. The smoking intervention does appear to produce a decrease in the log relative risk of relapse, as expected. Patient sex is also marginally significant, with women more likely to relapse than men, a result often attributed to the (real or perceived) risk of weight gain following smoking cessation. The number of cigarettes smoked per day does not seem important, but duration as a smoker is significant, and in a possibly counterintuitive direction: shorter-term smokers relapse sooner. This may be due to the fact that people are better able to quit smoking as they age (and are thus confronted more clearly with their own mortality).

The estimated cure fraction in Table 14.13 is roughly .70, indicating that roughly 70% of smokers in this study who attempted to quit have in fact been "cured." The spatial smoothness parameters ρ_ϕ and ρ_ζ are both close to 1, again suggesting we would lose little by simply setting them both equal to 1 (as in the standard CAR model). Finally, the last lines of both tables indicate only a moderate correlation between the two random effects, again consistent with the rather weak case for including them in the model at all.

We compared our results to those obtained from the `R` function `survreg` using a Weibull link, and also to Weibull regression models fit in a Bayesian fashion using the `WinBUGS` package. While neither of these alternatives featured a cure rate (and only the `WinBUGS` analysis included spatial random effects), both produced fixed effect estimates quite consistent with those in Table 14.13.

Turning to graphical summaries, Figure 14.11 maps the posterior medians of the frailty (ϕ_i) and shape (ρ_i) parameters in the full spatial MCAR (Model 6) case. The maps reveal some interesting spatial patterns, though the magnitudes of the differences appear relatively small across zip codes. The south-central region seems to be of some concern, with its

Figure 14.11 *Maps of posterior means for the ϕ_i (a) and the ρ_i (b) in the full spatial MCAR model, assuming the data to be interval-censored.*

high values for both ϕ_i (high overall relapse rate) and ρ_i (increasing baseline hazard over time). By contrast, the four zip codes comprising the city of Rochester, MN (home of the Mayo Clinic, and marked with an "R" in each map) suggest slightly better than average cessation behavior. Note that a nonspatial model cannot impute anything other than the "null values" ($\phi_i = 0$ and $\rho_i = 1$) for any zip code contributing no data (all of the unshaded regions in Figure 14.10). Our spatial model however is able to impute nonnull values here, in accordance with the observed values in neighboring regions.

14.6 Exercises

1. The data located at `www.biostat.umn.edu/~brad/data/MAC.dat`, and also shown in Table 14.14, summarize a clinical trial comparing two treatments for *Mycobacterium avium* complex (MAC), a disease common in late-stage HIV-infected persons. Eleven clinical centers ("units") have enrolled a total of 69 patients in the trial, 18 of which have died; see Cohn et al. (1999) and Carlin and Hodges (1999) for full details regarding this trial.

   As in Section 14.1, let t_{ij} be the time to death or censoring and x_{ij} be the treatment indicator for subject j in stratum i ($j = 1, \ldots, n_i$, $i = 1, \ldots, k$). With proportional hazards and a Weibull baseline hazard, stratum i's hazard is then

$$\begin{aligned} h(t_{ij}; x_{ij}) &= h_0(t_{ij})\omega_i \exp(\beta_0 + \beta_1 x_{ij}) \\ &= \rho_i t_{ij}^{\rho_i - 1} \exp(\beta_0 + \beta_1 x_{ij} + W_i) \ , \end{aligned}$$

   where $\rho_i > 0$, $\boldsymbol{\beta} = (\beta_0, \beta_1)' \in \Re^2$, and $W_i = \log \omega_i$ is a clinic-specific frailty term.

   (a) Assume i.i.d. specifications for these random effects, i.e.,

$$W_i \overset{iid}{\sim} N(0, 1/\tau) \quad \text{and} \quad \rho_i \overset{iid}{\sim} G(\alpha, \alpha) \ .$$

   Then as in the `mice` example (`WinBUGS` Examples Vol. 1),

$$\mu_{ij} = \exp(\beta_0 + \beta_1 x_{ij} + W_i) \ ,$$

   so that $t_{ij} \sim Weibull(\rho_i, \mu_{ij})$. Use `WinBUGS` to obtain posterior summaries for the main and random effects in this model. Use vague priors on β_0 and β_1, a moderately informative $G(1, 1)$ prior on τ, and set $\alpha = 10$. (You might also recode the drug covariate from (1,2) to (–1,1), in order to ease collinearity between the slope β_1 and the intercept β_0.)

Unit	Drug	Time	Unit	Drug	Time	Unit	Drug	Time
A	1	74+	E	1	214	H	1	74+
A	2	248	E	2	228+	H	1	88+
A	1	272+	E	2	262	H	1	148+
A	2	344				H	2	162
			F	1	6			
B	2	4+	F	2	16+	I	2	8
B	1	156+	F	1	76	I	2	16+
			F	2	80	I	2	40
C	2	100+	F	2	202	I	1	120+
			F	1	258+	I	1	168+
D	2	20+	F	1	268+	I	2	174+
D	2	64	F	2	368+	I	1	268+
D	2	88	F	1	380+	I	2	276
D	2	148+	F	1	424+	I	1	286+
D	1	162+	F	2	428+	I	1	366
D	1	184+	F	2	436+	I	2	396+
D	1	188+				I	2	466+
D	1	198+	G	2	32+	I	1	468+
D	1	382+	G	1	64+			
D	1	436+	G	1	102	J	1	18+
			G	2	162+	J	1	36+
E	1	50+	G	2	182+	J	2	160+
E	2	64+	G	1	364+	J	2	254
E	2	82						
E	1	186+	H	2	22+	K	1	28+
E	1	214+	H	1	22+	K	1	70+
						K	2	106+

Table 14.14 *Survival times (in half-days) from the MAC treatment trial, from Carlin and Hodges (1999). Here, "+" indicates a censored observation.*

(b) From Table 14.15, we can obtain the lattitude and longitude of each of the 11 sites, hence the distance d_{ij} between each pair. These distances are included in `www.biostat.umn.edu/~brad/data/MAC.dat`; note they have been scaled so that the largest (New York-San Francisco) equals 1. (Note that since sites F and H are virtually coincident (both in Detroit, MI), we have recoded them as a single clinic (#6) and now think of this as a 10-site model.) Refit the model in `WinBUGS` assuming the frailties to have spatial correlation following the isotropic exponential kriging model,

$$\mathbf{W} \sim N_k(\mathbf{0}, H), \text{ where } H_{ij} = \sigma^2 \exp(-\phi d_{ij}) \ ,$$

where as usual $\sigma^2 = 1/\tau$, and where we place a $G(3, 0.1)$ (mean 30) prior on ϕ.

2. The file `www.biostat.umn.edu/~brad/data/smoking.dat` contains the southeastern Minnesota smoking cessation data discussed in Section 14.5. At each of up to five office visits, the smoking status of persons who had recently quit smoking was assessed. We define relapse to smoking as the endpoint, and denote the failure or censoring time of person j in county i by t_{ij}. The data set (already in `WinBUGS` format) also contains the adjacency matrix for the counties in question.

(a) Assuming that smoking relapses occurred on the day of the office visit when they were detected, build a hierarchical spatial frailty model to analyze these data. Code your

Unit	Number	City
A	1	Harlem (New York City), NY
B	2	New Orleans, LA
C	3	Washington, DC
D	4	San Francisco, CA
E	5	Portland, OR
F	6a	Detroit, MI (Henry Ford Hospital)
G	7	Atlanta, GA
H	6b	Detroit, MI (Wayne State University)
I	8	Richmond, VA
J	9	Camden, NJ
K	10	Albuquerque, NM

Table 14.15 *Locations of the clinical sites in the MAC treatment data set.*

model in `WinBUGS`, run it, and summarize your results. Use the DIC tool to compare a few competing prior or likelihood specifications.

(b) When we observe a subject who has resumed smoking, all we really know is that his failure (relapse) point occurred somewhere between his last office visit and this one. As such, improve your model from part (a) by building an interval-censored version.

3. Consider the extension of the Section 14.4 model in the single endpoint, multiple cause case to the *multiple* endpoint, multiple cause case — say, for analyzing times until diagnosis of each cancer (if any), rather than merely a single time until death. Write down a model, likelihood, and prior specification (including an appropriately specified MCAR distribution) to handle this case.

Chapter 15

Special topics in spatial process modeling

Earlier chapters have developed the basic theory and the general hierarchical Bayesian modeling approach for handling spatial and spatiotemporal point-referenced data. In this chapter, we consider some special topics that are of interest in the context of such models.

15.1 Data assimilation

Data assimilation for spatial and spatiotemporal data can arise in various forms. For instance, there is a substantial literature on data assimilation for climate modeling which we briefly review below. There is a growing literature on data fusion for environmental exposure assessment (which is the main focus of this section). All of this work is embedded, in a sense, in the bigger world of data assimilation with computer model output. The overall goal of data fusion is to synthesize multiple data sources which are informing about spatial and spatiotemporal responses over the same region and the same time period. The anticipated benefit will be improved kriging and, in the temporal case, improved short term forecasting. It will almost always be the case that the data layers are misaligned in space and in time, e.g., areal units for one, point-referenced locations for another; hourly availability for one, daily collection for another. Evidently, this recalls the misalignment discussion of Chapter 7.

Consistent with our philosophy in this volume, we will consider the foregoing objective within a Bayesian framework, eschewing algorithmic and pseudo-statistical approaches, which we briefly mention in the next subsection. We confine ourselves to the case of two data layers, one of which will be monitoring station data, i.e., the point-referenced source and the other will be an environmental computer model which provides *output* (rather than real data) for grid cells at some spatial scale.

15.1.1 Algorithmic and pseudo-statistical approaches in weather prediction

A convenient framework within to review algorithmic and pseudo-statistical approaches to spatial data assimilation is in the context of numerical weather prediction. Kalnay (2003) provides a development of this material. Earliest work created local polynomial interpolations using quadratic trend surfaces in locations in order to interpolate observed values to grid values. Eventually, what emerged in the meteorology community was the recognition that a first guess (or background field or prior information) was needed (Bergthorsson and Döös, 1955), supplying the *initial conditions*. The climatological intuition here is worth articulating. Over "data-rich" areas the observational data dominates while in "data-poor" regions the forecast facilitates transport of information from the data-rich areas.

We review several numerical approaches using, illustratively, temperature as the variable of interest. At time t, we let $T_{obs}(t)$ be an observed measurement, $T_b(t)$ a background level, $T_a(t)$ an assimilated value, and $T_{true}(t)$ the true value. An early scheme is known as the successive corrections method (SCM) which obtains $T_{i,a}(t)$ iteratively through $T_{i,a}^{(r+1)}(t) = T_{i,a}^{(r)}(t) + \frac{\sum_k w_{ik}(T_{k,obs}(t) - T_{k,a}^{(r)}(t))}{\sum_k w_{ik} + \epsilon^2}$. Here, i indexes the grid cells for the interpolation while

k indexes the observed data locations. $T_{k,a}^{(r)}(t)$ is the value of the assimilator at the r-th iteration at the observation point k (obtained from interpolating the surrounding grid points). The weights can be defined in various ways but usually as a decreasing function of the distance between the grid point and the observation point. In fact, they can vary with iteration, perhaps becoming increasingly local. (See, e.g., Cressman, 1959 and Bratseth, 1986.)

A second empirical approach is called *nudging* or Newtonian relaxation. Suppose, suppressing location, we think about a differential equation driving temperature, e.g., $\frac{dT(t)}{dt} = a(T(t), t, \theta(t))$. If we write $a(\cdot)$ as an additive form, say, $a(T(t), t) + \theta(t)$ and let $\theta(t) = (T_{obs}(t) - T(t))/\tau$ then τ controls the relaxation. Small τ implies that the $\theta(t)$ term dominates while large τ implies that the nudging effect will be negligible.

Next might be a least squares approach. Again, suppressing location, suppose we assume that $T_{obs}^{(1)}(t) = T_{true}(t) + \epsilon_1(t)$ and $T_{obs}^{(2)}(t) = T_{true}(t) + \epsilon_2(t)$ where we envision two sources of observational data on the true temperature at t. The ϵ_l have mean 0 and variance $\sigma_l^2, l = 1, 2,$. Then, with the variances known, it is a familiar exercise to obtain the best unbiased estimator of $T_{true}(t)$ based upon these two pieces of information. That is, $T_a(t) = a_1 T_{obs}^{(1)}(t) + a_2 T_{obs}^{(2)}(t)$ where $a_1 = \sigma_2^2/(\sigma_1^2 + \sigma_2^2)$ and $a_2 = \sigma_1^2/(\sigma_1^2 + \sigma_2^2)$. Of course, we obtain the same solution as the MLE if we use independent normal likelihoods for the $T_{obs}^{(l)}(t)$'s.

A last idea here is simple sequential assimilation and its connection to the Kalman filter. In the univariate case suppose we write $T_a(t) = T_b(t) + \gamma(T_{obs}(t) - T_b(t))$. Here, $T_{obs}(t) - T_b(t)$ is referred to as the observational innovation or observational increment relative to the background. The optimal weight $\gamma = \sigma_{obs}^2/(\sigma_{obs}^2 + \sigma_b^2)$, analogous to the previous paragraph. Hence, we only need an estimate of the ratio of the observational variance to the background variance in order to obtain $T_a(t)$. To make this scheme dynamic, suppose the background is updated through the assimilation, i.e., $T_b(t+1) = h(T_a(t))$ where $h(\cdot)$ denotes some choice of forecast model. Then we will also need a revised background variance; this is usually taken to be a scalar (> 1) multiple of the variance of $T_a(t)$.

15.1.2 Fusion modeling using stochastic integration

The fusion approach proposed by Fuentes and Raftery (2005) builds upon earlier Bayesian melding work in Poole and Raftery (2000). It conceptualizes a true exposure surface and views the monitoring station data as well as the model output data as varying in a suitable way around the true surface. In particular, the average exposure in a grid cell A, denoted by $Z(A)$, differs from the exposure at any particular location $\mathbf{s}$, $Z(\mathbf{s})$. The so-called change of support problem in this context addresses converting the point level $Z(\mathbf{s})$ to the grid level $Z(A)$ through the stochastic integral,

$$Z(A) = \frac{1}{|A|} \int_A Z(\mathbf{s})\, d\mathbf{s}, \tag{15.1}$$

where $|A|$ denotes the area of grid cell A and $Z(A)$ is the block average (see Chapter 7). Fusion modeling, working with *block averaging* as in (15.1) has been considered by, e.g., Fuentes and Raftery (2005).

Let $Z(\mathbf{s})$ and $Y(\mathbf{s})$ denote the observed and the true exposure respectively at a station $\mathbf{s}$. The first model assumption is

$$Z(\mathbf{s}) = Y(\mathbf{s}) + \epsilon(\mathbf{s}) \tag{15.2}$$

where $\epsilon(\mathbf{s}) \sim N(0, \sigma_\epsilon)$ represents the measurement error at location $\mathbf{s}$. The true exposure process is assumed to be

$$Y(\mathbf{s}) = \mu(\mathbf{s}) + \eta(\mathbf{s}) \tag{15.3}$$

where $\mu(\mathbf{s})$ provides the spatial mean surface, typically characterized by a trend, by elevation, etc. The error term $\eta(\mathbf{s})$ is a spatially colored process assumed to be the zero mean Gaussian process (GP) with a specified covariance function. The output of the computer model at areal unit scale is usually assumed to be biased/uncalibrated. Hence, the conceptual point level output from this model, denoted by $Q(\mathbf{s})$, usually is related to the true surface as

$$Q(\mathbf{s}) = a(\mathbf{s}) + b(\mathbf{s})Y(\mathbf{s}) + \delta(\mathbf{s}) \tag{15.4}$$

where $a(\mathbf{s})$ denotes the additive bias and $b(\mathbf{s})$ denotes the multiplicative bias.[1] The error term, $\delta(\mathbf{s})$, is assumed to be a white noise process given by $N(0, \sigma_\delta^2)$. Then, since the computer model output is provided on a grid, $A_1, \ldots, A_J$, the point level process is converted to grid level by stochastic integration of (15.4), i.e.,

$$Q(A_j) = \int_{A_j} a(\mathbf{s})\, d\mathbf{s} + \int_{A_j} b(\mathbf{s})Y(\mathbf{s})\, d\mathbf{s} + \int_{A_j} \delta(\mathbf{s})\, d\mathbf{s}.$$

It has been observed that unstable model fitting (identifiability problems) accrues to the case where we have spatially varying $b(\mathbf{s})$ so $b(\mathbf{s}) = b$ is adopted. Spatial prediction at a new location $\mathbf{s}_0$ is done through the posterior predictive distribution $f(Y(\mathbf{s}_0) \mid \mathbf{Z}, \mathbf{Q})$ where $\mathbf{Z}$ denote all the station data and $\mathbf{Q}$ denote all the grid-level computer output $Q(A_1), \ldots, Q(A_J)$.

This fusion strategy becomes computationally infeasible in the setting below, where we have more than 40,000 grid cells for the computer model output over the Eastern U.S. We have a very large number of grid cells with a relatively sparse number of monitoring sites. An enormous amount of stochastic integration is required. Moreover, a dynamic implementation over many time periods becomes even more infeasible. The emergent suggestion is that we should be scaling down rather than integrating up.

While the Fuentes and Raftery (2005) approach models at the point level, the strategy in McMillan et al. (2008) models at the grid cell level. In this fashion, computation is simplified and fusion with space-time data is manageable. In particular, as before, suppose that we have, say, n monitoring stations with, say, J grid cells. Let $Q(A_j)$ denote the CMAQ output value for cell A_j while $Z_{A_j}(\mathbf{s}_k)$ denotes the station data for site $\mathbf{s}_k$ within cell A_j. Of course, for most of the j's, k will be 0 since $n << J$. Let $Y(A_j)$ denote the true value for cell A_j.

Then, paralleling (15.2) and (15.4),

$$Z_{A_j}(\mathbf{s}_k) = Y(A_j) + \epsilon_{A_j}(\mathbf{s}_k) \tag{15.5}$$

and

$$Q(A_j) = Y(A_j) + b(A_j) + \gamma(A_j). \tag{15.6}$$

In (15.6), the CMAQ output is modeled as varying around the true value with a bias term, denoted by $b(A_j)$, specified using a B-spline model. Also, the ϵ's are assumed to be independently and identically distributed and so are the γ's, each with a respective variance component. So, the station data and the CMAQ data are conditionally independent given the true surface. Finally, the true surface is modeled analogously to (15.3) but at grid scale. Operating at areal scale, the η's are given a CAR specification (see Chapter 4). For space-time data, McMillan et al. (2008) offer a dynamic version of this approach, assuming a dynamic CAR specification for the η's.

Wikle and Berliner (1995) are concerned with two sources of wind data — daily wind satellite data and computer model output data supplied by a weather center. The satellite data are at 0.5^o resolution, not on a regular grid while the the computer model output is on a 2.5^o regular grid. The objective is to predict a windstream surface at 1.0^o resolution. They work within the familiar hierarchical framework,

$$[\text{data} \mid \text{process}][\text{process} \mid \text{parameters}][\text{parameters}]\ .$$

[1] Recall the spatially varying coefficient model from Section 9.6.

Here, the process is the true underlying wind surface and the two data sources are assumed observed with measurement error.

Their data structure implies three scales: subgrid (0.5^o resolution), supergrid (2.5^o resolution) and prediction grid (1.0^o resolution). They use block averaging on each of the three scales, with the aforementioned measurement error, to infer about the blocks on the prediction grid from the "observed" blocks on the supergrids and subgrids. Formally, the model introduces a latent Gaussian process to do the block averaging and local scaling factors to move from one grid to another. With temporal wind data, Wikle et al. (2001) implement a space-time data fusion. Here, they account for temporal dependence by using dynamic coefficients in the process model in order to avoid computation of stochastic integrals.

15.1.3 The downscaler

The downscaler model was proposed in Berrocal et al. (2010a). First, we show a static spatial version which can be used at any temporal scale, e.g., for daily data or for annual averages. Then, we present several ways in which the static model can be extended to handle spatiotemporal data.

Though it is more broadly applicable, for ease of interpretation, we present it with regard to an ozone application. So, we consider two sources of ground-level ozone data (daily 8-hour maxima in ppb). The first one comes from the National Air Monitoring Stations/State and Local Air Monitoring Stations (NAMS/SLAMS; *http://www.epa.gov/monitor/programs/namsslam.html*) network for the Eastern U.S. Figure 15.1 shows a map of the NAMS/SLAMS monitoring sites used in our analysis. Panel (a) displays all the monitoring sites (N=803) used to fit and validate the spatiotemporal version of our downscaling model. Panel (b) shows the monitoring sites (N=69) enclosed in the smaller region highlighted by the square in panel (a) that have been used to enable comparison of the static version of our downscaling model with the Bayesian melding model of Fuentes and Raftery (2005) and with ordinary kriging.

The second source of data in our study is the Models-3/Community Multiscale Air Quality (CMAQ; *http://www.epa.gov/asmdnerl/CMAQ*) model. CMAQ is a numerical model that estimates daily 8-hour maximum concentration of ozone by integrating information coming from three different components: a meteorology component which accounts for the

(a) (b)

Figure 15.1 *(a): Ozone monitoring sites in the Eastern U.S.; (b): Subset region used to compare ordinary kriging, Bayesian melding and the downscaler. The black points represent monitoring sites, the black dots represent the centroids of the CMAQ grid cells.*

state and the evolution in time of the atmosphere, an emissions component which deals with emissions injected in the atmosphere by both chemical plants and natural processes, and a component that accounts for the chemical and physical interactions occurring in the atmosphere. We use daily CMAQ predictions gridded to 12km spatial resolution. There are 40,044 grid cells covering the portion of the eastern U.S. displayed in Figure 15.1(a) resulting in $40,044 \times 168 = 6,727,392$ daily modeled output measurements. 651 grid cells cover the smaller region shown in Figure 15.1(b). There, gray dots display the centroids of each 12km CMAQ grid cell. Again, we have 69 monitoring sites in the smaller region. We immediately see that we have many more grid cells than monitoring sites; the enormous amount of block averaging that would be needed to implement the fusion model of Fuentes and Raftery (2005) makes it computationally overwhelming.

With regard to normality assumptions, it is better to model ozone on the square root scale. Therefore, we denote with $Y(\mathbf{s})$ the square root of the observed ozone concentration at $\mathbf{s}$. We use $x(B)$ to denote the square root of the numerical model output over grid cell B. Each point $\mathbf{s}$ is associated with the CMAQ grid cell B in which it lies. So, all the points $\mathbf{s}$ falling in the same 12-km square region are assigned the same CMAQ output value. This is the usual interpretation of the CMAQ model output, i.e., it is viewed as a tiled surface at grid cell resolution, providing a tile height for each grid cell.

We relate the observed data to the CMAQ output in the following way. For each $\mathbf{s}$ in B, we assume that

$$Y(\mathbf{s}) = \tilde{\beta}_0(\mathbf{s}) + \tilde{\beta}_1(\mathbf{s})\, x(B) + \epsilon(\mathbf{s}), \qquad \epsilon(\mathbf{s}) \overset{ind}{\sim} \mathrm{N}(0, \tau^2) \tag{15.7}$$

where

$$\begin{aligned} \tilde{\beta}_0(\mathbf{s}) &= \beta_0 + \beta_0(\mathbf{s}) \\ \tilde{\beta}_1(\mathbf{s}) &= \beta_1 + \beta_1(\mathbf{s}), \end{aligned} \tag{15.8}$$

and $\epsilon(\mathbf{s})$ is a white noise process with nugget variance τ^2. β_0 and β_1 represent the overall intercept and slope bias of the CMAQ model, while $\beta_0(\mathbf{s})$ and $\beta_1(\mathbf{s})$ are local adjustments to the additive and multiplicative bias, respectively. The spatially-varying coefficients $\beta_0(\mathbf{s})$ and $\beta_1(\mathbf{s})$ are in turn modeled as bivariate mean-zero Gaussian spatial processes using the method of coregionalization (see Chapter 9).

Therefore, we suppose that there exist two mean-zero unit-variance independent Gaussian processes $w_0(\mathbf{s})$ and $w_1(\mathbf{s})$ such that, for convenience, $\mathrm{Cov}(w_j(\mathbf{s}), w_j(\mathbf{s}')) = \exp(-\phi_j|\mathbf{s}-\mathbf{s}'|)$, i.e., ϕ_j is the spatial decay parameter for Gaussian process $w_j(\mathbf{s})$, $j = 0, 1$. Therefore,

$$\begin{pmatrix} \beta_0(\mathbf{s}) \\ \beta_1(\mathbf{s}) \end{pmatrix} = A \begin{pmatrix} w_0(\mathbf{s}) \\ w_1(\mathbf{s}) \end{pmatrix} \tag{15.9}$$

where the unknown A matrix in (15.9) is assumed to be a lower-triangular. We note that there is no identifiability problem introduced when we have, say, multiple $\mathbf{s}$'s within a given B since the associated $Y(\mathbf{s})$'s will vary over the $\mathbf{s}$'s in B. Hence, the data informs locally about the $\beta_0(\mathbf{s})$ and $\beta_1(\mathbf{s})$ surfaces. Then the spatial dependence introduced by (15.9) enables interpolation of these surfaces over the region of interest.

We complete the specification of the Bayesian hierarchical model with the following prior specifications. We use a bivariate normal distribution for the overall bias terms β_0 and β_1, lognormal distributions for the two diagonal entries, a_{11} and a_{22}, of the coregionalization matrix A, a normal distribution for the off-diagonal entry, a_{21}, of A, and an inverse gamma distribution for τ^2. Since it is not possible to estimate consistently all of the covariance parameters (Zhang, 2004), under weak prior specifications we find weak identifiability of these parameters in the MCMC chains. Hence, we use discrete uniform priors on m values for the decay parameters ϕ_j, $j = 0, 1$.

Our model specification is rather simple yet it provides calibration at the local level and endows the spatial process $Y(\mathbf{s})$ with a flexible non-stationary covariance structure. The model is much easier to fit than Bayesian melding since we eliminate the need to handle stochastic integrals. Moreover, to fit the model we only need to work with the responses associated with the $Y(\mathbf{s}_i)$, i.e., with the set of monitoring sites, a much smaller number compared to the number of grid cells. Arguably, our model is also preferable to that of McMillan et al. (2008) since we downscale to the point level rather than up to the grid cell level. Spatial interpolation to a new location is based upon the predictive distribution, i.e., at location $\mathbf{s}_0$, we sample $f(Y(\mathbf{s}_0)|\{Y(\mathbf{s}_i)\},\{x(B_j)\})$.

15.1.4 Spatiotemporal versions

Our downscaling model can be extended to accommodate data collected over time in several ways. With time denoted by t, $t = 1, \ldots, T$, let $Y(\mathbf{s}, t)$ be the square root of observed daily 8-hour maximum ozone concentration at $\mathbf{s}$ at time t and let $x(B, t)$ be the square root of the CMAQ predicted average ozone concentration for grid cell B at time t. Again, we associate to each point $\mathbf{s}$ the CMAQ grid cell B in which it lies.

We start by assuming that the overall bias terms, β_0 and β_1, change in time, while the local adjustments to β_0 and β_1, the spatially varying coefficients $\beta_0(\mathbf{s})$ and $\beta_1(\mathbf{s})$, remain constant in time. This means that

$$Y(\mathbf{s}, t) = \beta_{0t} + \beta_{1t}x(B, t) + \beta_0(\mathbf{s}) + \beta_1(\mathbf{s})x(B, t) + \epsilon(\mathbf{s}, t) \tag{15.10}$$

where $\epsilon(\mathbf{s}, t) \overset{ind}{\sim} \mathrm{N}(0, \tau^2)$. There are two customary ways in which the β_{0t} and β_{1t} terms could be specified. The first is to assume that β_{0t} and β_{1t} are nested within time, or, in other words, they are independent across time:

$$\begin{pmatrix} \beta_{0t} \\ \beta_{1t} \end{pmatrix} \overset{ind}{\sim} \mathrm{MVN}_2(\mu, V). \tag{15.11}$$

The second is to assume that the two overall bias terms β_{0t} and β_{1t} evolve dynamically in time. That is,

$$\begin{pmatrix} \beta_{0t} \\ \beta_{1t} \end{pmatrix} = \begin{pmatrix} \rho_0 & 0 \\ 0 & \rho_1 \end{pmatrix} \begin{pmatrix} \beta_{0\,t-1} \\ \beta_{1\,t-1} \end{pmatrix} + \begin{pmatrix} \eta_{0t} \\ \eta_{1t} \end{pmatrix} \tag{15.12}$$

where $\begin{pmatrix} \eta_{0t} \\ \eta_{1t} \end{pmatrix} \overset{ind}{\sim} \mathrm{MVN}_2(\mathbf{0}, V_\eta)$ and $\begin{pmatrix} \beta_{00} \\ \beta_{10} \end{pmatrix} \sim \mathrm{MVN}_2(\mu_0, V_0)$.

A more general way to introduce time in our downscaling model is by assuming that both the overall bias terms β_0 and β_1, and the local adjustments, $\beta_0(\mathbf{s})$ and $\beta_1(\mathbf{s})$, vary with time, that is,

$$Y(\mathbf{s}, t) = \beta_{0t} + \beta_{1t}x(B, t) + \beta_0(\mathbf{s}, t) + \beta_1(\mathbf{s}, t)x(B, t) + \epsilon(\mathbf{s}, t). \tag{15.13}$$

As with β_{0t} and β_{1t} in model (15.10), there are two ways in which we can specify the $\beta_0(\mathbf{s}, t)$ and $\beta_1(\mathbf{s}, t)$ terms. In the first case, we can have

$$\begin{pmatrix} \beta_0(\mathbf{s}, t) \\ \beta_1(\mathbf{s}, t) \end{pmatrix} = A \begin{pmatrix} w_{0t}(\mathbf{s}) \\ w_{1t}(\mathbf{s}) \end{pmatrix} \tag{15.14}$$

with A lower-triangular. In (15.14) $w_{0t}(\mathbf{s})$ and $w_{1t}(\mathbf{s})$ are serially independent replicates of two independent unit-variance Gaussian processes with mean zero and exponential covariance function having spatial decay parameters ϕ_0 and ϕ_1, respectively.

In the second case, that is, if $\beta_0(\mathbf{s},t)$ and $\beta_1(\mathbf{s},t)$ evolve dynamically in time, we follow Section 11.4. Therefore,

$$\begin{aligned}\begin{pmatrix}\beta_0(\mathbf{s},t)\\ \beta_1(\mathbf{s},t)\end{pmatrix} &= \begin{pmatrix}\rho_0 & 0\\ 0 & \rho_1\end{pmatrix}\begin{pmatrix}\beta_0(\mathbf{s},t-1)\\ \beta_1(\mathbf{s},t-1)\end{pmatrix}+\begin{pmatrix}v_{0t}(\mathbf{s})\\ v_{1t}(\mathbf{s})\end{pmatrix}\\ \begin{pmatrix}v_{0t}(\mathbf{s})\\ v_{1t}(\mathbf{s})\end{pmatrix} &= A\begin{pmatrix}w_{0t}(\mathbf{s})\\ w_{1t}(\mathbf{s})\end{pmatrix}\end{aligned} \tag{15.15}$$

where $\begin{pmatrix}\beta_0(\mathbf{s},0)\\ \beta_1(\mathbf{s},0)\end{pmatrix} \equiv \begin{pmatrix}0\\ 0\end{pmatrix}$. Again, A is lower-triangular and does not depend on t, while $w_{0t}(\mathbf{s})$ while $w_{1t}(\mathbf{s})$ are as above.

15.1.5 An illustration

Ozone poses a threat primarily during summer months, i.e., the ozone season, May 1 - October 15, 2001, when it is present at higher concentrations. So, to demonstrate the predictive performance of the static downscaler on ozone data, we consider three illustrative dates, randomly chosen: one in June, one in July, and one in August. For each day, we fit our downscaler using ozone observations coming from 69 monitoring sites and estimates of average ozone concentration over the associated grid cell provided by CMAQ. For comparison, we fit Bayesian melding using the 69 observations and CMAQ model output for the 651 grid cells covering the subset region displayed in Figure 15.1(b). For ordinary kriging, we estimate parameters of the exponential covariance function using only the 69 observations. The mean and standard deviation of the observed daily 8-hour maximum ozone concentration for the three days were, respectively, 69.94 and 14.28 ppb, for June 21 2001, 72.23 and 9.61 ppb, for July 10 2001, and 41.78 and 13.96 ppb for August 11 2001.

After estimating the model parameters, we proceeded to predict ozone concentration at all of the 69 monitoring sites using the three different methods. We evaluate the quality of the predictions by computing the MSE and MAE of the predictions, and by looking at the empirical coverage and average length of the 90% predictive intervals. Table 15.1 reports these summary statistics for all three selected days. We see that the downscaler produces predictions that are far better calibrated than the other methods. Moreover, the empirical coverage of the downscaler is always above the nominal value, thus indicating that it is

Day	Method	MSE	MAE	Empirical coverage of 90% PI	Average length of 90% PI	Average predictive variance
06/21/2001	Ordinary kriging	37.38	4.87	89.9%	20.52	39.81
	Bayesian melding	35.35	4.81	78.26%	14.22	19.36
	Downscaler	9.79	2.45	98.55%	19.34	35.47
07/10/2001	Ordinary kriging	39.15	4.66	95.65%	22.75	48.49
	Bayesian melding	32.26	4.26	88.41%	16.70	26.36
	Downscaler	16.33	2.90	97.10%	21.29	42.75
08/11/2001	Ordinary kriging	30.70	4.38	90.90%	20.53	41.44
	Bayesian melding	14.94	2.99	78.79%	11.74	13.39
	Downscaler	6.06	1.89	96.97 %	17.49	29.48

Table 15.1 *Mean Square Error (MSE), Mean Absolute Error (MAE), empirical coverage, average length of 90% predictive intervals (PI) and predictive variance for ordinary kriging, Bayesian melding and the downscaler method for three days in the 2001 high-ozone season.*

more conservative than Bayesian melding, whose empirical coverage is always below 90%. Alternatively, the predictive intervals constructed using the downscaler are wider than those obtained using Bayesian melding, which are too narrow.

Ordinary kriging performs as well as the downscaler model in terms of coverage and average length of the 90% predictive intervals, but its predictions are much more biased than the ones produced using our downscaling model. This might be expected since ordinary kriging does not exploit the additional information contained in the high resolution CMAQ output, relying only on the 69 ozone observations.

We conclude by noting that the downscaler has been extended to accommodate bivariate data at locations in Berrocal et al. (2010b). Also, there can be measurement error concerns when downscaling. Berrocal et al. (2012) propose new models that are intended to address two such concerns with the computer model output. One is potential spatial displacement in the computer model values assigned to a grid cell. Possibly, this output is appropriate for a displacement of the grid cell. The second recognizes that, with regard to improving predictive performance of the fusion at a location, there may be useful information in the outputs for grid cells that are neighbors of the one in which the location lies.

15.2 Space-time modeling for extremes

Extreme value analysis is frequently applied to environmental science data (e.g., Thompson et al., 2001). Extremes in exposure to environmental contaminants are of interest with regard to public health outcomes. Extremes in weather are of interest with regard to performance of plants and animals. In particular, for plants, it is suggested that extreme weather events, such as drought, heavy rainfall and very high or low temperatures, might be more significant factors in explaining plant performance with regard to survival, growth, reproductivity, etc., than trends in the mean climate.

Here, we illustrate the analysis of spatio-temporal weather extremes with data derived from precipitation surfaces in the Cape Floristic Region (CFR), the smallest but, arguably, the richest of the world's six floral kingdoms, encompassing a region of roughly $90,000km^2$ in southwestern South Africa. The daily precipitation surfaces we employ arise via interpolation to grid cells at 10km resolution based on records reported by up to 3000 stations over South Africa over the period from 1950-1999. We have 50 derived surfaces of annual maxima surfaces from daily rainfalls for 1332 grid cells. Eventually, we hold out 1999 for model validation. Figure 15.2 shows the CFR and the data for the year 1999.

By now there is an enormous literature on the modeling of extremes. At present, the standard approach utilizes Generalized Extreme Value (GEV) distribution families. Alternatively, daily precipitation exceedances for a given threshold are modelled with the Generalized Pareto Distribution (GPD). The book by Coles (2001) provides an accessible introduction. Relevant to the subject matter presented here, recently there has been some work focusing on spatial (or spatiotemporal) characterization of extreme values (see, e.g., Kharin and Zwiers, 2005, Cooley et al., 2007, and Sang and Gelfand, 2009), including several papers discussing spatial interpolation for extreme values (see, e.g., Ribatet et al., 2012, and Buishand et al., 2008). Here, we review the work of Sang and Gelfand (2009, 2011) which develops hierarchical models for rainfall that reflect dependence in space and in time. In particular, they use the GEV which is characterized by a location, a scale and a shape parameter. Conceptually, each of these could vary in space and time and they could be mutually dependent. As a result, one can envision a range of such models, fitting them, and comparing them. Initially, we discuss the grid cell data, using Markov random field models. Then, we move to a dataset at station level and work with Gaussian processes.

Figure 15.2 *The Cape Floristic Region (CFR) and the data for the year 1999.*

15.2.1 Possibilities for modeling maxima

Let us consider possibilities for modeling maxima. To make things concrete, suppose at site $\mathbf{s}$, daily data j in year t, e.g., precipitation or temperature, ozone or $PM_{2.5}$, That is, we envision the collection of surfaces $W_{t,j}(\mathbf{s})$ which are observed at $\{s_i, i = 1, 2, ..., n\}$. We seek inference regarding $Y_t(\mathbf{s}) = \max_j W_{t,j}(s)$. To be more explicit, we are considering maxima over time at a particular location, not maxima over space at a given time. And, with interest in a surface of extremes, we are evidently in the setting of multivariate extreme value theory. It seems clear that spatial dependence in the maxima will be weaker than spatial dependence in the daily data. Inference focuses on parametric issues, e.g., trends in time, dependence in space, as well as on prediction at new locations and at future (one-step ahead) times. Also of interest is risk assessment with regard to return time. That is, if $P(Y_t(\mathbf{s}) > y) = p$, the expected return time until an exceedance of y is $1/p$ time units.

With regard to stochastic specification, in principle, we could model the W's, which will induce inference for the Y's. In this case, we have an enormous literature, using hierarchical space-time modeling that has been much discussed in this book. Summarily, let θ denote all of the parameters in a particular model. Suppose we fit the model and obtain the posterior, $[\theta|\{W_{t,j}(\mathbf{s}_i)\}]$. At a new $\mathbf{s}_0$, we can obtain the predictive distribution for $W_{t,j}(\mathbf{s}_0)$ in the form of posterior samples. These posterior samples can be converted into posterior samples of $Y_t(\mathbf{s}_0)$, i.e., "derived" quantities. We can do similarly for projection to time $T+1$. However, there are several reasons why this path is not of interest here. First, it is not modeling or explaining the maxima and second, it is extremely computationally demanding, generating hundreds and hundreds of daily data samples and discarding them all, retaining just the maxima as a summary. Most importantly, there will be greater uncertainty in predictions made using derived quantities than in prediction doing direct modeling of the maxima.

There are still two paths to directly model the maxima: we can model the maxima or model the "process" that drives the maxima. More precisely, under the first path, the usual assumptions are that the $Y_t(\mathbf{s})$ are spatially dependent but temporally independent, i.e.,

they are viewed as replicates. The argument here is that there will be strong temporal dependence at the scale of days, that is, in the $W_{t,j}(\mathbf{s})$'s but temporal dependence will be negligible between yearly values, that is, for the $Y_t(\mathbf{s})$'s. Usually, this is not an explanatory model, no covariates are introduced. The alternative or second path is to provide space-time modeling for the parameters in the GEV's for the $\{Y_t(\mathbf{s})\}$. We refer to this as the process path and it is what we develop in the sequel. First, we review a little bit of basic extreme value theory.

15.2.2 Review of extreme value theory

Formal extreme value theory begins with a sequence $Y_1, Y_2, ...$ of independent and identically distributed random variables and, for a given n, asks about parametric models for $M_n = \max(Y_1, ..., Y_n)$. If the distribution of the Y_i is specified, the exact distribution of M_n is known. In the absence of such specification, extreme value theory considers the existence of $\lim_{n\to\infty} Pr((M_n - b_n)/a_n \leq y) \equiv F(y)$ for two sequences of real numbers $a_n > 0, b_n$. If $F(y)$ is a non-degenerate distribution function, it belongs to either the Gumbel, the Fréchet or the Weibull class of distributions, which can all be usefully expressed under the umbrella of the GEV. That is,

$$G(y; \mu, \sigma, \xi) = \exp\left\{-\left[1 + \xi\left(\frac{y-\mu}{\sigma}\right)\right]^{-1/\xi}\right\} \tag{15.16}$$

for $\{y : 1 + \xi(y - \mu)/\sigma > 0\}$. Here, $\mu \in \Re$ is the location parameter, $\sigma > 0$ is the scale parameter and $\xi \in \Re$ is the shape parameter. The "residual" $V = (Y - \mu)/\sigma$ follows a GEV$(0, 1, \xi)$. Let $Z = (1 + \xi V)^{\frac{1}{\xi}} \Leftrightarrow V = \frac{Z^{\frac{1}{\xi}}-1}{\xi}$. Then Z follows a standard Fréchet distribution, with distribution function $\exp(-z^{-1})$. (For us, this distribution is more familiar as an inverse gamma, $IG(1, 1)$.) The GEV distribution is heavy-tailed and its probability density function decreases at a slow rate when the shape parameter, ξ, is positive. On the other hand, the GEV distribution has a bounded upper tail for a negative shape parameter. Note that n is not specified; the GEV is viewed as an approximate distribution to model the maximum of a sufficiently long sequence of random variables. In our setting, we will assign daily precipitation into annual blocks and, initially, assume the maxima are conditionally independent (but not identically distributed) across years given their respective, parametrically modeled, μ, σ, and ξ.

We introduce the GEV distribution as a first stage model for annual precipitation maxima, specifying μ, σ, and ξ at the second stage to reflect underlying spatiotemporal structure. In particular, working with grid cells, let $Y_{i,t}$ denote the annual maximum of daily rainfall for cell i in year t. We assume the $Y_{i,t}$ are conditionally independent, each following a GEV distribution with parameters $\mu_{i,t}$, $\sigma_{i,t}$, and $\xi_{i,t}$, respectively. Attention focuses on specification of the models for $\mu_{i,t}$, $\sigma_{i,t}$ and $\xi_{i,t}$.

The conditional independence assumption is interpreted as interest in smoothing the surfaces around which the interpolated data is centered rather than smoothing the data surface itself. As a formal assumption, it is defendable in time since, at a site, the annual maxima likely occur with sufficient time between them to be assumed independent. In space, we would expect small scale dependence in the data at a given time. However, with observations assigned to grid cells at 10 km resolution, we can not hope to learn about fine scale dependence. Below, when we work at the point level, we will restore this spatial dependence.

Exploratory analysis for the foregoing CFR precipitation data, using an available MLE package (Gilleland, E. and Katz, R.W., Extremes Toolkit (extRemes): Weather and Climate Applications of Extreme Value Statistics, http://www.isse.ucar.edu/extremevalues/tutorial.

pdf or Coles, S., *S-plus functions for extreme value modeling: An accompaniment to the book: An introduction to statistical modeling of extreme values*, 2001, `http://www.stats.bris.ac.uk/masgc/ismev/uses.ps`), suggests $\xi_{i,t,} = \xi$ for all i and t and $\sigma_{i,t} = \sigma_i$ for all t so modeling focuses on the $\mu_{i,t}$. But, in addition, we want the $\mu_{i,t}$ and σ_i to be dependent at the same site. This requires the introduction of an association model for a collection of spatially co-varying parameters over a collection of grid cells. We adopt the coregionalization method, as developed in Chapter 9, making a random linear transformation of conditionally autoregressive (CAR) models in order to greatly reduce the computational burden in model fitting.

With regard to modeling the $\mu_{i,t}$, we have many options. With spatial covariates $\mathbf{X}_i$, it is natural to specify a regression model with random effects, $[\mu_{i,t} \mid \boldsymbol{\beta}, W_{i,t}, \tau^2] = N(\mathbf{X}_i^T\boldsymbol{\beta} + W_{i,t}, \tau^2)$. Here, for example, the $\mathbf{X}_i$ could include altitude or specify a trend surface, with coefficient vector $\boldsymbol{\beta}$ while $W_{i,t}$ is a spatiotemporal random effect.

Illustrative possibilities for modeling $W_{i,t}$ include: (i) an additive form, **Model A**: $W_{i,t} = \psi_i + \delta_t$, $\delta_t = \phi\delta_{t-1} + \omega_t$, where $\omega_t \sim N(0, W_0^2) \quad i.i.d$; (ii) a linear form in time with spatial random effects, **Model B**: $W_{i,t} = \psi_i + \rho(t - t_0)$; (iii) a linear form in time with local slope, **Model C**: $W_{i,t} = \psi_i + (\rho + \rho_i)(t - t_0)$; (iv) a multiplicative form in space and time, **Model D**: $W_{i,t} = \psi_i\delta_t$, $\delta_t = \phi\delta_{t-1} + \omega_t$, where $\omega_t \sim N(0, W_0^2) \quad i.i.d.$

The additive form in Model A might appear to over-simplify spatiotemporal structure. However, the data may not be rich enough to find space-time interaction in the $\mu_{i,t}$. Model B and Model C provide evaluations of temporal trends in terms of global and local assessments respectively. The coefficient $\rho + \rho_i$ in Model C represents the spatial trend in location parameters, where ρ could be interpreted as the global change level in the CFR per year. Finally, Model D provides a multiplicative representation of $W_{i,t}$, similar in spirit to work of Huerta and Sansó (2007). Models A and D yield special cases of a dynamic linear model (West and Harrison, 1997). Again, we want to model the dependence between location and scale parameters in the GEV model. In models A, B, and D, we do this by specifying $\log\sigma_i$ and ψ_i to be dependent. We work with $\sigma_i = \sigma_0\exp(\lambda_i)$ and a coregionalized bivariate CAR model. In model C, we specify $\log\sigma_i, \psi_i$ and ρ_i to be dependent and use a coregionalized trivariate CAR specification.

In the foregoing models, temporal evolution in the extreme rainfalls is taken into account in the model for $W_{i,t}$. More precisely, each of the models enables prediction for any grid cell for any year. In fact, with the CFR precipitation data, Sang and Gelfand (2009) held out the annual maximum rainfalls in 1999 (Figure 15.2) for validation purposes, in order to compare models in terms of the predictive performance; see the paper for model fitting details. Posterior medians are adopted as the point estimates of the predicted annual maxima because of the skewness of the predictive distributions. Predictive performance was assessed first by computing the averaged absolute predictive errors (AAPE) for each model. A second comparison among models is to study the proportion of the true annual maximum rainfalls in the year 1999 which lie in the estimated 95% credible intervals for each model. A third model selection criterion which is easily calculated from the posterior samples is the deviance information criterion (DIC). Using these criteria, Model C emerged as best (we omit details here).

15.2.3 A continuous spatial process model

Again, we focus on spatial extremes for precipitation events. However, now the motivating data are annual maxima of daily precipitations derived from daily station records at 281 sites over South Africa in 2006. Often, extreme climate events are driven by multi-scale spatial forcings, say, large regional forcing and small scale local forcing. Therefore, attractive modeling in such cases would have the potential to characterize the multi-scale dependence between locations for extreme values of the spatial process. Additionally, with

point-referenced station data, spatial interpolation is of interest to learn about the predictive distribution of the unobserved extreme value at unmonitored locations.

So, we extend the hierarchical modeling approach developed in the previous subsection to accommodate a collection of point-referenced extreme values in order to achieve multi-scale dependence along with spatial smoothing for realizations of the surface of extremes. In particular, we assume annual maxima follow GEV distributions, with parameters μ, σ , and ϕ specified in the latent stage to reflect underlying spatial structure. We relax the conditional independence assumption previously imposed on the first stage hierarchical specifications. Again, in space, despite the fact that large scale spatial dependence may be accounted for in the latent parameter specifications, there may still remain unexplained small scale spatial dependence in the extreme data. So, we propose a (mean square) continuous spatial process model for the actual extreme values to account for spatial dependence which is unexplained by the latent spatial specifications for the GEV parameters. In other words, we imagine a scale of space-time dependence which is captured at a second stage of a hierarchical model specification with additional proposed first stage spatial smoothing at a much shorter range. This first stage process model is created through a copula approach where the copula idea is applied to a Gaussian spatial process using suitable transformation.

More formally, the first stage of the hierarchical model can be written as:

$$Y(\mathbf{s}) \quad = \quad \mu(\mathbf{s}) + \frac{\sigma(\mathbf{s})}{\xi(\mathbf{s})}(Z(\mathbf{s})^{\xi(\mathbf{s})} - 1) \tag{15.17}$$

where $Z(\mathbf{s})$ follows a standard Fréchet distribution. We may view $Z(\mathbf{s})$ as the "standardized residual" in the first stage GEV model. The conditional independence assumption is equivalent to the assumption that the $Z(\mathbf{s})$ are *i.i.d.* So, again, even if the surface for each model parameter is smooth, realizations of the predictive surface will be everywhere discontinuous under the conditional independence assumption.

In the extreme value literature there is a general class of multivariate extreme value processes which introduce desired dependence between pairs of maxima, the so-called max-stable processes. From Coles (2001), let $W_1(s), W_2(s), ...,$ be i.i.d. replicates of a process on R^2. Let the process $Y^{(n)}(s) = \max_{j=1,...,n} W_j(s) \sim nW_1(s)\ \forall n$. There is a limiting characterization of this class of processes due to De Haan and specific versions have been presented by Smith and by Schlather (see discussion in Coles, 2001). Unfortunately, model fitting using these max-stable processes is challenging since joint distributions are intractable for more than two locations — we would have hundreds of locations requiring explicit forms for high dimensional joint distributions in order to work with likelihoods. Currently, two approximation approaches have emerged: the use of pairwise likelihoods and an approximate Bayesian computation approach These limitations with the use of max-stable processes for spatial setting leads us to look to a different path using Gaussian processes and copulas.

15.2.4 Using copulas

So, we seek to modify the assumption that the $z(\mathbf{s})$ are i.i.d. Fréchet. We wish to introduce spatial dependence to the $z(\mathbf{s})$ while retaining Fréchet marginal distributions. A frequent strategy for introducing dependence subject to specified marginal distributions is through copula models. We need to apply the copula approach to a stochastic process. The Gaussian process, which is determined through its finite dimensional distributions, offers the most convenient mechanism for doing this. With a suitable choice of correlation function, mean square continuous surface realizations result for the Gaussian process, hence for the transformed surfaces under monotone transformation. In other words, through transformation of a Gaussian process, we can obtain a continuous spatial process of extreme values with standard Fréchet marginal distributions.

Copulas have received much attention and application in the past two decades (see Nelsen, 2006, for a review). Consider a random two-dimensional vector distributed according to a standard bivariate Gaussian distribution with correlation ρ. The Gaussian copula function is defined as follows: $C_\rho(u,v) = \Phi_\rho(\Phi^{-1}(u), \Phi^{-1}(v))$ where the $u, v \in [0,1]$, Φ denotes the standard normal cumulative distribution function and Φ_ρ denotes the cumulative distribution function of the standard bivariate Gaussian distribution with correlation ρ. The bivariate random vector (X,Y) having GEV distributions as marginals, denoted as G_x and G_y, respectively, can be given a bivariate extreme value distribution using the Gaussian copula as follows. Let $(X,Y) = (G_x^{-1}(\Phi(X')), G_y^{-1}(\Phi(Y')))$, where G_x^{-1} and G_y^{-1} are the inverse marginal distribution functions for X and Y and $(X', Y') \sim \Phi_\rho$. Then the distribution function of (X,Y) is given by $H(X,Y;\rho) = C_\rho(\Phi(X'), \Phi(Y'))$ and the marginal distributions of X and Y remain to be G_x and G_y.

Now, we can directly propose a spatial extreme value process which is transformed from a standard spatial Gaussian process with mean 0, variance 1, and correlation function $\rho(\mathbf{s}, \mathbf{s}'; \theta)$. The standard Fréchet spatial process is the transformed Gaussian process defined as $z(\mathbf{s}) = G^{-1}(\Phi(z'(\mathbf{s}))$ where now G is the distribution function of a standard Fréchet distribution. It is clear that z(**s**) is a valid stochastic process since it is induced by a strictly monotone transformation of a Gaussian process. Indeed, this standard Fréchet process is completely determined by $\rho(\mathbf{s}, \mathbf{s}; \theta)$. More precisely, suppose we observe extreme values at a set of sites $\{s_i, i = 1, 2, ..., n\}$. The realizations $\mathbf{z}^T = (z^T(\mathbf{s}_1), ..., z^T(\mathbf{s}_n))$ follow a multivariate normal distribution which is determined by ρ and induces the joint distribution for $\mathbf{z} = (z(\mathbf{s}_1), ..., z(\mathbf{s}_n))$.

Properties of the transformed Gaussian process include: (i) joint, marginal and conditional distributions for $\mathbf{z}$ are all immediate; dependence in the Fréchet process is inherited through $\rho(\cdot)$; (ii) a Matérn with smoothness parameter greater than 0 assures mean square continuous realizations; (iii) efficient model fitting approaches for Gaussian processes can be utilized after inverse transformation; (iv) since the $z^T(\mathbf{s})$ process is strongly stationary, then so is the $z(\mathbf{s})$ process; (v) despite the inherited dependence through ρ, G has no moments so non-moment based dependence metrics for $z(\mathbf{s})$ are needed, e.g., $\int([z(\mathbf{s}), z(\mathbf{s}')] - [z(\mathbf{s}][z(\mathbf{s}')])^2 d\mathbf{s}d\mathbf{s}'$; and (vi) evidently, the transformed Gaussian approach is not limited to extreme value analysis; it can create other classes of means square continuous processes.

15.2.5 Hierarchical modeling for spatial extreme values

Returning to (15.17), we now assume that the $z(\mathbf{s})$ follow a standard Fréchet process. That is, we have a hierarchical model in which the first stage conditional independence assumption is removed. Now, we turn to specification of the second stage, creating *latent* spatial models, following Coles and Tawn (1996). Specifications for $\mu(\mathbf{s})$, $\sigma(\mathbf{s})$, and $\xi(\mathbf{s})$ have to be made with care. A simplification to facilitate model fitting is to assume there is spatial dependence for the $\mu(\mathbf{s})$ but that $\sigma(\mathbf{s})$ and $\xi(\mathbf{s})$ are constant across the study region. In fact, the data is not likely to be able to inform about processes for $\sigma(\mathbf{s})$ and for $\xi(\mathbf{s})$.

More precisely, suppose we propose the specification $\mu(\mathbf{s}) = X(\mathbf{s})^T\beta + W(\mathbf{s})$. $X(\mathbf{s})$ is the site-specific vector of potential explanatory variables. The $W(\mathbf{s})$ are spatial random effects, capturing the effect of unmeasured or unobserved covariates with large operational scale spatial pattern. A natural specification for $W(\mathbf{s})$ is a zero-centered Gaussian process determined by a valid covariance function $C(\mathbf{s}_i, \mathbf{s}_j)$. Here C is apart from ρ introduced at the first stage. In fact, from (15.17), plugging in the model for $\mu(\mathbf{s})$ we obtain

$$Y(\mathbf{s}) = \mathbf{x}^T(\mathbf{s})\boldsymbol{\beta} + W(\mathbf{s}) + \frac{\sigma}{\xi}(z(\mathbf{s})^\xi - 1) \tag{15.18}$$

and $z(\mathbf{s}) = G^{-1}\Phi(z^{'}(\mathbf{s}))$. We clearly see the two sources of spatial dependence, the $z(\mathbf{s})$ and the $W(\mathbf{s})$ with two associated scales of dependence. The foregoing interpretation as well as the need for identifiability imply that we assign the shorter range spatial dependence to the $z(\mathbf{s})$ process. Of course, we can compare this specification with the case where we have $z(\mathbf{s})$ i.i.d. Fréchet.

Again, we often have space-time data over long periods of time, e.g., many years, and we seek to study, say, annual spatial maxima. Now, we are given a set of extremes $\{Y(\mathbf{s}_i, t), i = 1, ..., n; t = 1, ..., T\}$. Now, the first stage of the hierarchical model specifies a space-time standard Fréchet process for the $z(\mathbf{s}, t)$, built from a space-time Gaussian process. If time is viewed as continuous, we need only specify a valid space-time covariance function. If time is discretized then we need only provide a dynamic Gaussian process model (see Chapter 10). In fact, we might assume $z_t(\mathbf{s})$ and $z_{t'}(\mathbf{s})$ are two independent Gaussian processes when $t \neq t'$. That is, it may be plausible to assume temporal independence since annual block size may be long enough to yield approximately independent annual maximum observation surfaces. Model specifications for $\mu(\mathbf{s}, t)$, $\sigma(\mathbf{s}, t)$, and $\xi(\mathbf{s}, t)$ should account for dependence structures both within and across location and time. In Section 15.2.2 above, we offered suggested forms with regard to these latent parameter specifications.

Finally, Sang and Gelfand (2011) work with the foregoing annual maxima of daily rainfalls from the station data in South Africa. They consider 200 monitoring sites and fit each of the years 1956, 1976, 1996, and 2006 separately; see the paper for model fitting details. They offer comparison between the conditionally independent first stage specification and the smoothed first stage using the Gaussian copula for the annual maximum rainfall station data for the year 2006. The smoothed first stage specification was superior, achieving 20% reduction in AAPE and more accurate empirical coverage.

15.3 Spatial CDF's

In this section, we review the essentials of spatial cumulative distribution functions (SCDF's), including a hierarchical modeling approach for inference. We then extend the basic definition to allow covariate weighting of the SCDF estimate, as well as versions arising under a bivariate random process.

15.3.1 Basic definitions and motivating data sets

Suppose that $X(\mathbf{s})$ is the log-ozone concentration at location $\mathbf{s}$ over a particular time period. Thinking of $X(\mathbf{s})$, $\mathbf{s} \in D$ as a spatial process, we might wish to find the proportion of area in D that has ozone concentration below some level w (say, a level above which exposure is considered to be unhealthful). This proportion is the random variable,

$$F(w) = Pr\left[\mathbf{s} \in D : \; X(\mathbf{s}) \leq w\right] = \frac{1}{|D|} \int_D Z_w(\mathbf{s}) d\mathbf{s} \; , \tag{15.19}$$

where $|D|$ is the area of D, and $Z_w(\mathbf{s}) = 1$ if $X(\mathbf{s}) \leq w$, and 0 otherwise. Since $X(\mathbf{s})$, $\mathbf{s} \in D$ is random, (15.19) is a random function of $w \in \Re$ that increases from 0 to 1 and is right-continuous. Thus while $F(w)$ is not the usual cumulative distribution function (CDF) of X at $\mathbf{s}$ (which would be given by $Pr[X(\mathbf{s}) \leq x]$, and is not random), it does have all the properties of a CDF, and so is referred to as the *spatial cumulative distribution function*, or SCDF. For a constant mean stationary process, all $X(\mathbf{s})$ have the same marginal distribution, whence $E[F(w)] = Pr\left(X(\mathbf{s}) \leq w\right)$. It is also easy to show that $Var[F(w)] = \frac{1}{|D|^2} \int_D \int_D Pr\left(X(\mathbf{s}) \leq w, X(\mathbf{s}') \leq w\right) d\mathbf{s} d\mathbf{s}' \; - [Pr\left(X(\mathbf{s}) \leq w\right)]^2$. Overton (1989) introduced the idea of an SCDF, and used it to analyze data from the National Surface Water Surveys. Lahiri et al. (1999) developed a subsampling method that provides (among other things)

large-sample prediction bands for the SCDF, which they show to be useful in assessing the foliage condition of red maple trees in the state of Maine.

The *empirical* SCDF based upon data $\mathbf{X}_s = (X(\mathbf{s}_1), \ldots, X(\mathbf{s}_n))'$ at w is the proportion of the $X(\mathbf{s}_i)$ that take values less than or equal to w. Large-sample investigation of the behavior of the empirical SCDF requires care to define the appropriate asymptotics; see Lahiri et al. (1999) and Zhu, Lahiri, and Cressie (2002) for details. When n is not large, as is the case in our applications, the empirical SCDF may become less attractive. Stronger inference can be achieved if one is willing to make stronger distributional assumptions regarding the process $X(\mathbf{s})$. For instance, if $X(\mathbf{s})$ is assumed to be a Gaussian process, the joint distribution of $\mathbf{X}_s$ is multivariate normal. Given a suitable prior specification, a Bayesian framework provides the predictive distribution of $F(w)$ given $X(\mathbf{s})$.

Though (15.19) can be studied analytically, it is difficult to work with in practice. However, approximation of (15.19) via Monte Carlo integration is natural (and may be more convenient than creating a grid of points over D), i.e., replacing $F(w)$ by

$$\widehat{F}(w) = \frac{1}{L}\sum_{\ell=1}^{L} Z_w(\tilde{\mathbf{s}}_\ell) \ , \tag{15.20}$$

where the $\tilde{\mathbf{s}}_\ell$ are chosen randomly in D, and $Z_w(\tilde{\mathbf{s}}_\ell) = 1$ if $X(\tilde{\mathbf{s}}_\ell) \leq w$, and 0 otherwise. Suppose we seek a realization of $\widehat{F}(w)$ from the predictive distribution of $\widehat{F}(w)$ given $\mathbf{X}_s$. In Section 7.1 we showed how to sample from the predictive distribution $p(\mathbf{X}_{\tilde{s}} \mid \mathbf{X}_s)$ for $\mathbf{X}_{\tilde{s}}$ arising from new locations $\tilde{\mathbf{s}} = (\tilde{\mathbf{s}}_1, \ldots, \tilde{\mathbf{s}}_L)'$. In fact, samples $\{\mathbf{X}_{\tilde{s}}^{(g)},\ g = 1, \ldots, G\}$ from the posterior predictive distribution,

$$p(\mathbf{X}_{\tilde{s}} \mid \mathbf{X}_s) = \int p(\mathbf{X}_{\tilde{s}} \mid \mathbf{X}_s, \boldsymbol{\beta}, \boldsymbol{\theta}) p(\boldsymbol{\beta}, \boldsymbol{\theta} \mid \mathbf{X}_s) d\boldsymbol{\beta} d\boldsymbol{\theta} \ ,$$

may be obtained one for one from posterior samples by composition.

The predictive distribution of (15.19) can be sampled at a given w by obtaining $X(\tilde{\mathbf{s}}_\ell)$ using the above algorithm, hence $Z_w(\tilde{\mathbf{s}}_\ell)$, and then calculating $\widehat{F}(w)$ using (15.20). Since interest is in the entire function $F(w)$, we would seek realizations of the approximate function $\widehat{F}(w)$. These are most easily obtained, up to an interpolation, using a grid of w values $\{w_1 < \cdots < w_k < \cdots < w_K\}$, whence each $(\boldsymbol{\beta}^{(g)}, \boldsymbol{\theta}^{(g)})$ gives a realization at grid point w_k,

$$\widehat{F}^{(g)}(w_k) = \frac{1}{L}\sum_{\ell=1}^{L} Z_{w_k}^{(g)}(\tilde{\mathbf{s}}_\ell) \ , \tag{15.21}$$

where now $Z_{w_k}^{(g)}(\tilde{\mathbf{s}}_\ell) = 1$ if $X^{(g)}(\tilde{\mathbf{s}}_\ell) \leq w_k$, and 0 otherwise. Handcock (1999) describes a similar Monte Carlo Bayesian approach to estimating SCDF's in his discussion of Lahiri et al. (1999).

Expression (15.21) suggests placing all of our $\widehat{F}^{(g)}(w_k)$ values in a $K \times G$ matrix for easy summarization. For example, a histogram of all the $\widehat{F}^{(g)}(w_k)$ in a particular row (i.e., for a given grid point w_k) provides an estimate of the predictive distribution of $\widehat{F}(w_k)$. On the other hand, each column (i.e., for a given Gibbs draw g) provides (again up to, say, linear interpolation) an approximate draw from the predictive distribution of the SCDF. Hence, averaging these columns provides, with interpolation, essentially the posterior predictive mean for F and can be taken as an *estimated* SCDF. But also, each draw from the predictive distribution of the SCDF can be inverted to obtain any quantile of interest (e.g., the median exposure $d_{.50}^{(g)}$). A histogram of these inverted values in turn provides an estimate of the posterior distribution of this quantile (in this case, $\widehat{p}(d_{.50}|\mathbf{X}_s)$). While this algorithm provides general inference for SCDF's, for most data sets it will be computationally very demanding, since a large L will be required to make (15.21) sufficiently accurate.

Figure 15.3 *Locations of 67 NO and NO_2 monitoring sites, California air quality data; 500 randomly selected target locations are also shown as dots.*

Our interest in this methodology is motivated by two environmental data sets; we describe both here but only present the inference for the second. The first is the Atlanta 8-hour maximum ozone data, which exemplifies the case of an air pollution variable measured at points, with a demographic covariate measured at a block level. Recall that its first component is a collection of ambient ozone levels in the Atlanta, GA, metropolitan area, as reported by Tolbert et al. (2000). Ozone measurements X_{itr} are available at between 8 and 10 fixed monitoring sites i for day t of year r, where $t = 1, \ldots, 92$ (the summer days from June 1 through August 31) and $r = 1, 2, 3$, corresponding to years 1993, 1994, and 1995. The reader may wish to flip back to Figure 1.3, which shows the 8-hour daily maximum ozone measurements in parts per million at the 10 monitoring sites for one of the days (July 15, 1995). This figure also shows the boundaries of the 162 zip codes in the Atlanta metropolitan area, with the 36 zips falling within the city of Atlanta encircled by the darker boundary on the map. An environmental justice assessment of exposure to potentially harmful levels of ozone would be clarified by examination of the predictive distribution of a *weighted* SCDF that uses the racial makeups of these city zips as the weights. This requires generalizing our SCDF simulation in (15.21) to accommodate covariate weighting in the presence of misalignment between the response variable (at point-referenced level) and the covariate (at areal-unit level).

SCDF's adjusted with point-level covariates present similar challenges. Consider the spatial data setting of Figure 15.3, recently presented and analyzed by Gelfand, Schmidt, and Sirmans (2002). These are the locations of several air pollutant monitoring sites in central and southern California, all of which measure ozone, carbon monoxide, nitric oxide (NO), and nitrogen dioxide (NO_2). For a given day, suppose we wish to compute an SCDF for the log of the daily median NO exposure adjusted for the log of the daily median NO_2 level (since the health effects of exposure to high levels of one pollutant may be exacerbated by further exposure to high levels of the other). Here the data are all point level, so

that Bayesian kriging methods of the sort described above may be used. However, we must still tackle the problem of *bivariate* kriging (for both NO and NO_2) in a computationally demanding setting (say, to the $L = 500$ randomly selected points shown as dots in Figure 15.3). In some settings, we must also resolve the misalignment in the data itself, which arises when NO or NO_2 values are missing at some of the source sites.

15.3.2 Derived-process spatial CDF's

15.3.2.1 Point- versus block-level spatial CDF's

The spatial CDF in (15.19) is customarily referred to as *the* SCDF associated with the spatial process $X(\mathbf{s})$. In fact, we can formulate many other useful SCDF's under this process. We proceed to elaborate choices of possible interest.

Suppose for instance that our data arrive at areal unit level, i.e., we observe $X(B_j)$, $j = 1, \ldots, J$ such that the B_j are disjoint with union D, the entire study region. Let $Z_w(B_j) = 1$ if $X(B_j) \leq w$, and 0 otherwise. Then

$$\widetilde{F}(w) = \frac{1}{|D|} \sum_{j=1}^{J} |B_j| \, Z_w(B_j) \tag{15.22}$$

again has the properties of a CDF and thus can also be interpreted as a spatial CDF. In fact, this CDF is a step function recording the proportion of the area of D that (at block-level resolution) lies below w. Suppose in fact that the $X(B_j)$ can be viewed as block averages of the process $X(\mathbf{s})$, i.e., $X(B_j) = \frac{1}{|B_j|} \int_{B_j} X(\mathbf{s}) d\mathbf{s}$. Then (15.19) and (15.22) can be compared: write (15.19) as $\frac{1}{|D|} \sum_j |B_j| \left[\frac{1}{|B_j|} \int_{B_j} I(X(\mathbf{s}) \leq w) d\mathbf{s} \right]$ and (15.22) as $\frac{1}{|D|} \sum_j |B_j| \; I \left[\left(\frac{1}{|B_j|} \int_{B_j} X(\mathbf{s}) d\mathbf{s} \right) \leq w \right]$. Interpreting $\mathbf{s}$ to have a uniform distribution on B_j, the former is $\frac{1}{|D|} \sum_j |B_j| \; E_{B_j} \left[I(X(\mathbf{s}) \leq w) \right]$ while the latter is $\frac{1}{|D|} \sum_j |B_j| I \left[E_{B_j}(X(\mathbf{s})) \leq w \right]$. In fact, if $X(\mathbf{s})$ is stationary, while $E[F(w)] = P(X(\mathbf{s}) \leq w)$, $E[\widetilde{F}(w)] = \frac{1}{|D|} \sum_j |B_j| \; P(X(B_j) \leq w)$. For a Gaussian process, under weak conditions $X(B_j)$ is normally distributed with mean $E[X(\mathbf{s})]$ and variance $\frac{1}{|B_j|^2} \int_{B_j} \int_{B_j} c(\mathbf{s} - \mathbf{s}'; \boldsymbol{\theta}) d\mathbf{s} d\mathbf{s}'$, so $E[\widetilde{F}(w)]$ can be obtained explicitly. Note also that since $\frac{1}{|B_j|} \int_{B_j} I(X(\mathbf{s}) \leq w) d\mathbf{s}$ is the customary spatial CDF for region B_j, then by the alternate expression for (15.19) above, $F(w)$ is an areally weighted average of *local* SCDF's.

Thus (15.19) and (15.22) differ, but (15.22) should neither be viewed as "incorrect" nor as an approximation to (15.19). Rather, it is an alternative SCDF derived under the $X(\mathbf{s})$ process. Moreover, if only the $X(B_j)$ have been observed, it is arguably the most sensible empirical choice. Indeed, the Multiscale Advanced Raster Map (MARMAP) analysis system project (`www.stat.psu.edu/~gpp/marmap_system_partnership.htm`) is designed to work with "empirical cell intensity surfaces" (i.e., the tiled surface of the $X(B_j)$'s over D) and calculates the "upper level surfaces" (variants of (15.22)) for description and inference regarding multicategorical maps and cellular surfaces.

Next we seek to introduce covariate weights to the spatial CDF, as motivated in Subsection 15.3.1. For a nonnegative function $r(\mathbf{s})$ that is integrable over D, define the SCDF associated with $X(\mathbf{s})$ weighted by r as

$$F_r(w) = \frac{\int_D r(\mathbf{s}) Z_w(\mathbf{s}) d\mathbf{s}}{\int_D r(\mathbf{s}) d\mathbf{s}} \, . \tag{15.23}$$

Evidently (15.23) satisfies the properties of a CDF and generalizes (15.19) (i.e., (15.19) is restored by taking $r(\mathbf{s}) \equiv 1$). But as (15.19) suggests expectation with respect to a uniform

density for $\mathbf{s}$ over D, (15.23) suggests expectation with respect to the density $r(\mathbf{s})/\int_D r(\mathbf{s})d\mathbf{s}$. Under a stationary process, $E[F(w)] = P(X(\mathbf{s}) \leq w)$ and $Var[F(w)]$ is

$$\begin{aligned}&\tfrac{1}{(\int_D r(\mathbf{s})d\mathbf{s})^2}\int_D\int_D r(\mathbf{s})r(\mathbf{s}')P(X(\mathbf{s}) \leq w, X(\mathbf{s}') \leq w)d\mathbf{s}d\mathbf{s}' \\ &\quad -[P(X(\mathbf{s}) \leq w)]^2 \ .\end{aligned}$$

There is an empirical SCDF associated with (15.23) that extends the empirical SCDF in Subsection 15.3.1 using weights $r(\mathbf{s}_i)/\sum_i r(\mathbf{s}_i)$ rather than $1/n$. This random variable is mentioned in Lahiri et al. (1999, p. 87). Following Subsection 15.3.1, we adopt a Bayesian approach and seek a predictive distribution for $F_r(w)$ given $\mathbf{X}_s$. This is facilitated by Monte Carlo integration of (15.23), i.e.,

$$\widehat{F}_r(w) = \frac{\sum_{\ell=1}^L r(\mathbf{s}_\ell)Z_w(\mathbf{s}_\ell)}{\sum_{\ell=1}^L r(\mathbf{s}_\ell)} \ . \tag{15.24}$$

15.3.2.2 Covariate weighted SCDF's for misaligned data

In the environmental justice application described in Subsection 15.3.1, the covariate is only available (indeed, only meaningful) at an areal level, i.e., we observe only the population density associated with B_j. How can we construct a covariate weighted SCDF in this case? Suppose we make the assignment $r(\mathbf{s}) = r_j$ for all $\mathbf{s} \in B_j$, i.e., that the density surface is constant over the areal unit (so that $r_j|B_j|$ is the observed population density for B_j). Inserting this into (15.23) we obtain

$$F_r^*(w) = \frac{\sum_{j=1}^J r_j|B_j|\left[\frac{1}{|B_j|}\int_{B_j} Z_w(\mathbf{s})d\mathbf{s}\right]}{\sum_{j=1}^J r_j|B_j|} \ . \tag{15.25}$$

As a special case of (15.23), (15.25) again satisfies the properties of a CDF and again has mean $P(X(\mathbf{s}) \leq w)$. Moreover, as below (15.22), the bracketed expression in (15.25) is the spatial CDF associated with $X(\mathbf{s})$ restricted to B_j. Monte Carlo integration applied to (15.25) can use the same set of $\mathbf{s}_\ell$'s chosen randomly over D as in Subsection 15.3.1 or as in (15.24). In fact (15.24) becomes

$$\widehat{F}_r(w) = \frac{\sum_{j=1}^J r_j L_j\left[\frac{1}{L_j}\sum_{\mathbf{s}_\ell \in B_j} Z_w(\mathbf{s}_\ell)\right]}{\sum_{j=1}^J r_j L_j} \ , \tag{15.26}$$

where L_j is the number of $\mathbf{s}_\ell$ falling in B_j. Equation (15.25) suggests the alternative expression,

$$\widehat{F}_r^*(w) = \frac{\sum_{j=1}^J r_j|B_j|\left[\frac{1}{L_j}\sum_{\mathbf{s}_\ell \in B_j} Z_w(\mathbf{s}_\ell)\right]}{\sum_{j=1}^J r_j|B_j|} \ . \tag{15.27}$$

Expression (15.27) may be preferable to (15.26), since it uses the exact $|B_j|$ rather than the random L_j.

15.3.3 Randomly weighted SCDF's

If we work solely with the r_j's, we can view (15.24)–(15.27) as *conditional* on the r_j's. However if we work with $r(\mathbf{s})$'s, then we will need a probability model for $r(\mathbf{s})$ in order to interpolate to $r(\mathbf{s}_\ell)$ in (15.24). Since $r(\mathbf{s})$ and $X(\mathbf{s})$ are expected to be associated, we may

conceptualize them as arising from a spatial process, and develop, say, a bivariate Gaussian spatial process model for both $X(\mathbf{s})$ and $h(r(\mathbf{s}))$, where h maps the weights onto $\Re^1$.

Let $\mathbf{Y}(\mathbf{s}) = (X(\mathbf{s}), h(r(\mathbf{s}))^T$ and $\mathbf{Y} = (\mathbf{Y}(\mathbf{s}_1), \ldots, \mathbf{Y}(\mathbf{s}_n))^T$. Analogous to the univariate situation in Subsection 15.3.1, we need to draw samples $\mathbf{Y}_{\tilde{s}}^{(g)}$ from $p(\mathbf{Y}_{\tilde{s}} \mid \mathbf{Y}, \boldsymbol{\beta}^{(g)}, \boldsymbol{\theta}^{(g)}, T^{(g)})$. Again this is routinely done via composition from posterior samples. Since the $\mathbf{Y}_{\tilde{s}}^{(g)}$ samples have marginal distribution $p(\mathbf{Y}_{\tilde{s}} \mid \mathbf{Y})$, we may use them to obtain predictive realizations of the SCDF, using either the unweighted form (15.21) or the weighted form (15.24).

The bivariate structure also allows for the definition of a *bivariate SCDF*,

$$F_{U,V}(w_u, w_v) = \frac{1}{|D|} \int_D I(U(\mathbf{s}) \leq w_u, V(\mathbf{s}) \leq w_v) d\mathbf{s} \ , \tag{15.28}$$

which gives $Pr\left[\mathbf{s} \in D : \ U(\mathbf{s}) \leq w_u, V(\mathbf{s}) \leq w_v\right]$, the proportion of the region having values below the given thresholds for, say, two pollutants. Finally, a sensible *conditional* SCDF might be

$$\begin{aligned} F_{U|V}(w_u|w_v) &= \frac{\int_D I(U(\mathbf{s}) \leq w_u, V(\mathbf{s}) \leq w_v) d\mathbf{s}}{\int_D I(V(\mathbf{s}) \leq w_v) d\mathbf{s}} \\ &= \frac{\int_D I(U(\mathbf{s}) \leq w_u) I(V(\mathbf{s}) \leq w_v) d\mathbf{s}}{\int_D I(V(\mathbf{s}) \leq w_v) d\mathbf{s}} \ . \end{aligned} \tag{15.29}$$

This expression gives $Pr\left[\mathbf{s} \in D : \ U(\mathbf{s}) \leq w_u \mid V(\mathbf{s}) \leq w_v\right]$, the proportion of the region having second pollutant values below the threshold w_v that *also has* first pollutant values below the threshold w_u. Note that (15.29) is again a weighted SCDF, with $r(\mathbf{s}) = I(V(\mathbf{s}) \leq w_v)$. Note further that we could easily alter (15.29) by changing the directions of either or both of its inequalities, if conditional statements involving high (instead of low) levels of either pollutant were of interest.

Example 15.1 *(California air quality data).* We illustrate in the case of a bivariate Gaussian process using data collected by the California Air Resources Board, available at `www.arb.ca.gov/aqd/aqdcd/aqdcddld.htm`. The particular subset we consider is the mean NO and NO_2 values for July 6, 1999, as observed at the 67 monitoring sites shown as solid dots in Figure 15.3. Recall that in our notation, U corresponds to log(mean NO) while V corresponds to log(mean NO_2). A WSCDF based on these two variables is of interest since persons already at high NO risk may be especially vulnerable to elevated NO_2 levels. Figure 15.4 shows interpolated perspective, image, and contour plots of the raw data. Association of the pollutant levels is apparent; in fact, the sample correlation coefficient over the 67 pairs is 0.74.

We fit a separable, Gaussian bivariate model using the simple exponential spatial covariance structure $\rho(d_{ii'}, \boldsymbol{\theta}) = \exp(-\lambda d_{ii'})$, so that $\boldsymbol{\theta} \equiv \lambda$; no σ^2 parameter is required (nor identifiable) here due to the multiplicative presence of the T matrix. For prior distributions, we first assumed $T^{-1} \sim W((\nu R)^{-1}, \nu)$ where $\nu = 2$ and $R = 4I$. This is a reasonably vague specification, both in its small degrees of freedom ν and in the relative size of R (roughly the prior mean of T), since the entire range of the data (for both log-NO and log-NO_2) is only about 3 units. Next, we assume $\lambda \sim G(a, b)$, with the parameters chosen so that the effective spatial range is half the maximum diagonal distance M in Figure 15.3 (i.e., $3/E(\lambda) = .5M$), and the standard deviation is one half of this mean. Finally, we assume constant means $\mu_U(\mathbf{s}; \boldsymbol{\beta}) = \beta_U$ and $\mu_V(\mathbf{s}; \boldsymbol{\beta}) = \beta_V$, and let β_U and β_V have vague normal priors (mean 0, variance 1000).

Our initial Gibbs algorithm sampled over $\boldsymbol{\beta}$, T^{-1}, and λ. The first two of these may be sampled from closed-form full conditionals (normal and inverse Wishart, respectively) while λ is sampled using Hastings independence chains with $G(\frac{2}{3}, \frac{3}{2})$ proposals. We used 3 parallel chains to check convergence, followed by a "production run" of 2000 samples from a

Figure 15.4 *Interpolated perspective, image, and contour plots of the raw log-NO (first row) and log-NO_2 (second row), California air quality data, July 6, 1999.*

Figure 15.5 *Perspective plots of kriged log-NO and log-NO_2 surfaces, California air quality data. First column, raw data; second column, based on a single Gibbs sample; third column, average over 2000 post-convergence Gibbs samples.*

single chain for posterior summarization. Histograms (not shown) of the posterior samples for the bivariate kriging model are generally well behaved and consistent with the results in Figure 15.4.

Figure 15.5 shows perspective plots of raw and kriged log-NO and log-NO_2 surfaces, where the plots in the first column are the (interpolated) raw data (as in the first column of Figure 15.4), those in the second column are based on a single Gibbs sample, and those

Figure 15.6 *Weighted SCDF's for the California air quality data: solid line, ordinary SCDF for log-NO; dotted line, weighted SCDF for log-NO using NO_2 as the weight; dashed lines, weighted SCDF for log-NO using various indicator functions of log-NO_2 as the weights.*

in the third column represent the average over 2000 post-convergence Gibbs samples. The plots in this final column are generally consistent with those in the first, except that they exhibit the spatial smoothness we expect of our posterior means.

Figure 15.6 shows several SCDFs arising from samples from our bivariate kriging algorithm. First, the solid line shows the ordinary SCDF (15.21) for log-NO. Next, we computed the weighted SCDF for two choices of weight function in (15.24). In particular, we weight log-NO exposure U by $h^{-1}(V)$ using $h^{-1}(V) = \exp(V)$ and $h^{-1}(V) = \exp(V)/(\exp(V)+1)$ (the exponential and inverse logit functions). Since V is log-NO_2 exposure, this amounts to weighting by NO_2 itself, and by $NO_2/(NO_2+1)$. The results from these two h^{-1} functions turn out to be visually indistinguishable, and are shown as the dotted line in Figure 15.6. This line is shifted to the right from the unweighted version, indicating higher harmful exposure when the second (positively correlated) pollutant is accounted for.

Also shown as dashed lines in Figure 15.6 are several WSCDF's that result from using a particular indicator of whether log-NO_2 remains below a certain threshold. These WSCDF's are thus also conditional SCDF's, as in equation (15.29). Existing EPA guidelines and expertise could be used to inform the choice of clinically meaningful thresholds; here we simply demonstrate the procedure's behavior for a few illustrative thresholds. For example, when the threshold is set to –3.0 (a rather high value for this pollutant on the log scale), nearly all of the weights equal 1, and the WSCDF differs little from the unweighted SCDF. However, as this threshold moves lower (to –4.0 and –5.0), the WSCDF moves further to the left of its unweighted counterpart. This movement is understandable, since our indicator functions are *decreasing* functions of log-NO_2; movement to the right could be obtained simply by reversing the indicator inequality.

Appendices

Appendix A

Spatial computing methods

A.1 Fast Fourier transforms

The Fast Fourier Transform (FFT) constitutes one of the major breakthroughs in computational mathematics. The FFT can be looked upon as a fast algorithm to compute the Discrete Fourier Transform (DFT), which enjoys wide applications in physics, engineering, and the mathematical sciences. The DFT implements discrete Fourier analysis and is used in time series and periodogram analysis. Here we briefly discuss the computational framework for DFTs; further details may be found in Monahan (2001, pp. 386–400) or Press et al. (1992, pp. 496–532).

For computational purposes, we develop the DFT in terms of a matrix transformation of a vector. For this discussion we let all indices range from 0 to $N-1$ (as in the `C` programming language). Thus if $\mathbf{x} = (x_0, \ldots, x_{N-1})^T$ is a vector representing a sequence of order N, then the DFT of $\mathbf{x}$ is given by

$$y_j = \sum_{k=0}^{N-1} \exp\left(-2\pi i j k/N\right) x_k \ , \tag{A.1}$$

where $i = \sqrt{-1}$, and j and k are indices. Let $w = \exp\left(-2\pi i/N\right)$ (the Nth root of unity) and let W be the $N \times N$ matrix with (j,k)th element given by w^{jk}. Then the relationship in (A.1) can be represented as the linear tranformation $\mathbf{y} = W\mathbf{x}$, with $\mathbf{y} = (y_0, \ldots, y_{N-1})^T$. The matrix of the inverse transformation is given by W^{-1}, whose (j,k)th element is easily verified to be w^{-jk}.

Direct computation of this linear transformation involves $O(N^2)$ arithmetic operations (additions, multiplications and complex exponentiations). The FFT (Cooley and Tukey, 1965) is a modified algorithm that computes the above in only $O(N \log N)$ operations. Note that the difference in these complexities can be immense in terms of CPU time. Press et al. (1992) report that with $N = 10^6$, this difference can be between 2 weeks and 30 seconds of CPU time on a microsecond-cycle computer.

To illustrate the above modification, let us consider a composite $N = N_1 N_2$, where N_1 and N_2 are integers. Using the remainder theorem, we write the indices $j = q_1 N_1 + r_1$ and $k = q_2 N_2 + r_2$ with $q_1, r_2 \in [0, N_2 - 1]$ and $q_2, r_1 \in [0, N_1 - 1]$. It then follows that

$$\begin{aligned}
y_j &= \sum_{k=0}^{N-1} w^{jk} x_j = \sum_{k=0}^{N-1} w^{(q_1 N_1 + r_1)k} x_k \\
&= \sum_{q_2=0}^{N_1-1} \sum_{r_2=0}^{N_2-1} w^{(q_1 N_1 + r_1)(q_2 N_2 + r_2)} x_{q_2 N_2 + r_2} \\
&= \sum_{r_2}^{N_2-1} w^{(q_1 N_1 + r_1) r_2} \sum_{q_2=0}^{N_1-1} (w^{N_2})^{q_2 r_1} x_{q_2 N_2 + r_2}
\end{aligned}$$

$$= \sum_{r_2}^{N_2-1} (w^{N_1})^{q_1 r_2} \left(w^{r_1 r_2} \sum_{q_2=0}^{N_1-1} (w^{N_2})^{q_2 r_1} x_{q_2 N_2 + r_2} \right) ,$$

where the equality in the third line arises from the fact that $w^{N_1 N_2 q_1 q_2} = 1$. This shows that each inner sum is a DFT of length N_1, while each of the N_2 outer sums is a DFT of length N_2. Therefore, to compute the above, we perform a DFT of length $N = N_1 N_2$ by first performing N_2 DFTs of length N_1 to obtain the inner sum, and then N_1 DFTs of length N_2 to obtain the outer sum. Effectively, the new algorithm involves $N_2 O(N_1) + N_1 O(N_2)$ arithmetic operations, which, when N is a power of 2, boils down to an $O(N \log_2 N)$ algorithm. The details of this setting may be found in Monahan (2001).

In spatial statistics, the FFT is often used for computing covariance functions and their spectral densities (Stein, 1999a). Recall that valid correlation functions are related to probability densities via Bochner's theorem. Restricting our attention to isotropic functions on $\Re^1$, we have

$$f(u) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \exp(-itu) C(t) dt , \tag{A.2}$$

where $f(u)$ is the *spectral density* obtained by a Fourier transform of the correlation function, $C(t)$. For example, the Matérn correlation function arises as a transform of the spectral density $f(u) = (\phi_2 + |u|^2)^{-(\nu + r/2)}$, up to a proportionality constant; see also equation (3.5).

To take advantage of the FFT in computing the Matérn correlation function, we first replace the continuous version of Bochner's integral in (A.2) by a finite sum over an evenly spaced grid of points between $-T$ and T, with T large enough so that

$$f(u) \approx \frac{\Delta_t}{2\pi} \sum_{j=0}^{N-1} \exp(-it_j u) C(t_j) , \tag{A.3}$$

where $t_j = j\Delta_t$. Note that we have used the fact that C is an even function, so $\Delta_t = 2T/N$. Next, we select evenly spaced evaluation points for the spectral density $f(u)$, and rewrite equation (A.3) as

$$f(u_k) \approx \frac{\Delta_t}{2\pi} \sum_{j=0}^{N-1} \exp(-it_j u_k) K(t_j) .$$

This, in fact, is a DFT and is cast in the matrix equation, $\mathbf{y} = W\mathbf{x}$, with $\mathbf{y} = (f(u_0), ..., f(u_{M-1}))^T$, $\mathbf{x} = (C(t_0), ..., C(t_{N-1}))^T$, and W is the matrix of the transformation with (j,k)th element given by $\exp(-it_j u_k)$. Now this DFT is made "fast" (into an FFT) by appropriate choice of the evaluation points for f and C. Let Δ_u denote the spacings of u. The FFT implementation is obtained by ensuring the product of the two spacings to equal $2\pi/N$. That is, we ensure that $\Delta_t \Delta_u = 2\pi/N$. This results in $W_{jk} = \exp(-ijk2\pi/N)$, which is the exactly the FFT matrix.

The FFT enables fast conversion between the spectral domain and the frequency domain. Since the Matérn function has an easily computed spectral density, the inverse FFT is used to approximate $C(t)$ from $f(t)$, using W^{-1}. Note that $W_{jk}^{-1} = \exp(ijk2\pi/N)$, which enables a direct efficient computation of the inverse, instead of the usual $O(n^3)$ inversion algorithms.

A.2 Slice Gibbs sampling for spatial process model fitting

Auxiliary variable methods are receiving increased attention among those who use MCMC algorithms to simulate from complex nonnormalized multivariate densities. Work in the statistical literature includes Tanner and Wong (1987), Besag and Green (1993), Besag et al. (1995), and Higdon (1998). The particular version we focus on here introduces a single

auxiliary variable to "knock out" or "slice" the likelihood. Employed in the context of spatial modeling for georeferenced data using a Bayesian formulation with commonly used proper priors, in this section we show that convenient Gibbs sampling algorithms result. Our approach thus finds itself as a special case of work by Damien, Wakefield, and Walker (1999), who view methods based on multiple auxiliary variables as a general approach to constructing Markov chain samplers for Bayesian inference problems. We are also close in spirit to recent work of Neal (2003), who also employs a single auxiliary variable, but prefers to slice the entire nonnormalized joint density and then do a single multivariate updating of all the variables. Such updating requires sampling from a possibly high-dimensional uniform distribution with support over a very irregular region. Usually, a bounding rectangle is created and then rejection sampling is used. As a result, a single updating step will often be inefficient in practice.

Currently, with the wide availability of cheap computing power, Bayesian spatial model fitting typically turns to MCMC methods. However, most of these algorithms are hard to automate since they involve tuning tailored to each application. In this section we demonstrate that a *slice Gibbs sampler*, done by knocking out the likelihood and implemented with a Gibbs updating, enables essentially an automatic MCMC algorithm for fitting Gaussian spatial process models. Additional advantages over other simulation-based model fitting schemes accrue, as we explain below. In this regard, we could instead slice the product of the likelihood and the prior, yielding uniform draws to implement the Gibbs updates. However, the support for these conditional uniform updates changes with iteration. The conditional interval arises through matrix inverse and determinant functions of model parameters with matrices of dimension equal to the sample size. Slicing only the likelihood and doing Gibbs updates using draws from the prior along with rejection sampling is truly "off the shelf," requiring no tuning at all. Approaches that require first and second derivatives of the log likelihood or likelihood times prior, e.g., the MLE approach of Mardia and Marshall (1984) or Metropolis-Hastings proposal approaches within Gibbs samplers will be very difficult to compute, particularly with correlation functions such as those in the Matérn class.

Formally, if $L(\boldsymbol{\theta};\boldsymbol{Y})$ denotes the likelihood and $\pi(\boldsymbol{\theta})$ is a proper prior, we introduce the single auxiliary variable U, which, given $\boldsymbol{\theta}$ and $\boldsymbol{Y}$, is distributed uniformly on $(0, L(\boldsymbol{\theta};\boldsymbol{Y}))$. Hence the joint posterior distribution of $\boldsymbol{\theta}$ and U is given by

$$p(\boldsymbol{\theta}, U|\boldsymbol{Y}) \propto \pi(\boldsymbol{\theta})\, I(U < L(\boldsymbol{\theta};\boldsymbol{Y}))\ , \tag{A.4}$$

where I denotes the indicator function. The Gibbs sampler updates U according to its full conditional distribution, which is the above uniform. A component θ_i of $\boldsymbol{\theta}$ is updated by drawing from its prior subject to the indicator restriction given the other θ's and U. A standard distribution is sampled and only L needs to be evaluated. Notice that, if hyperparameters are introduced into the model, i.e., $\pi(\boldsymbol{\theta})$ is replaced with $\pi(\boldsymbol{\theta}|\boldsymbol{\eta})\pi(\boldsymbol{\eta})$, the foregoing still applies and $\boldsymbol{\eta}$ is updated without restriction. Though our emphasis here is spatial model fitting, it is evident that slice Gibbs sampling algorithms are more broadly applicable. With regard to computation, for large data sets often evaluation of $L(\boldsymbol{\theta};\boldsymbol{Y})$ will produce an underflow, preventing sampling from the uniform distribution for U given $\boldsymbol{\theta}$ and $\boldsymbol{Y}$. However, $\log L(\boldsymbol{\theta};\boldsymbol{Y})$ will typically not be a problem to compute. So, if $V = -\log U$, given $\boldsymbol{\theta}$ and $\boldsymbol{Y}$, $V + \log L(\boldsymbol{\theta};\boldsymbol{Y}) \sim Exp(mean = 1.0)$, and we can transform (A.4) to $p(\boldsymbol{\theta}, V|\boldsymbol{Y}) \propto exp(-V)\, I(-\log L(\boldsymbol{\theta};\boldsymbol{Y}) < V < \infty)$.

In fact, in some cases we can implement a more efficient slice sampling algorithm than the slice Gibbs sampler. We need only impose constrained sampling on a subset of the components of $\boldsymbol{\theta}$. In particular, suppose we write $\boldsymbol{\theta} = (\boldsymbol{\theta_1}, \boldsymbol{\theta_2})$ and suppose that the full conditional distribution for $\boldsymbol{\theta_1}$, $p(\boldsymbol{\theta_1}|\boldsymbol{\theta_2},\boldsymbol{Y}) \propto L(\boldsymbol{\theta_1},\boldsymbol{\theta_2};\boldsymbol{Y})\pi(\boldsymbol{\theta_1}|\boldsymbol{\theta_2})$, is a standard distribution. Then consider the following iterative updating scheme: sample U given $\boldsymbol{\theta_1}$ and $\boldsymbol{\theta_2}$ as above; then, update $\boldsymbol{\theta_2}$ given $\boldsymbol{\theta_1}$ and U with a draw from $\pi(\boldsymbol{\theta_2}|\boldsymbol{\theta_1})$ subject to the constraint $U < L(\boldsymbol{\theta_1},\boldsymbol{\theta_2};\boldsymbol{Y})$; finally, update $\boldsymbol{\theta_1}$ with an unconditional draw from $p(\boldsymbol{\theta_1}|\boldsymbol{\theta_2},\boldsymbol{Y})$.

Formally, this scheme is not a Gibbs sampler. Suppressing $\boldsymbol{Y}$, we are updating $p(U|\boldsymbol{\theta_1}, \boldsymbol{\theta_2})$, then $p(\boldsymbol{\theta_2}|\boldsymbol{\theta_1}, U)$, and finally $p(\boldsymbol{\theta_1}|\boldsymbol{\theta_2})$. However, the first and third distribution uniquely determine $p(U, \boldsymbol{\theta_1}|\boldsymbol{\theta_2})$ and, this, combined with the second, uniquely determine the joint distribution. The Markov chain iterated in this fashion still has $p(\boldsymbol{\theta}, U|\boldsymbol{Y})$ as its stationary distribution. In fact, if $p(\boldsymbol{\theta_1}|\boldsymbol{\theta_2}, \boldsymbol{Y})$ is a standard distribution, this implies that we can marginalize over $\boldsymbol{\theta_1}$ and run the slice Gibbs sampler on $\boldsymbol{\theta_2}$ with U. Given posterior draws of $\boldsymbol{\theta_2}$, we can sample $\boldsymbol{\theta_1}$ one for one from its posterior using $p(\boldsymbol{\theta_1}|\boldsymbol{\theta_2}, \boldsymbol{Y})$ and the fact that $p(\boldsymbol{\theta_1}|\boldsymbol{Y}) = \int p(\boldsymbol{\theta_1}|\boldsymbol{\theta_2}, \boldsymbol{Y})p(\boldsymbol{\theta_2}|\boldsymbol{Y})$. Moreover, if $p(\boldsymbol{\theta_1}|\boldsymbol{\theta_2}, \boldsymbol{Y})$ is not a standard distribution, we can add Metropolis updating of $\boldsymbol{\theta_1}$ either in its entirety or through its components (we can also use Gibbs updating here). We employ these modified schemes for different choices of $(\boldsymbol{\theta_1}, \boldsymbol{\theta_2})$ in the remainder of this section.

We note that the performance of the algorithm depends critically on the distribution of the number of draws needed from $\pi(\boldsymbol{\theta_2}|\boldsymbol{\theta_1})$ to update $\boldsymbol{\theta_2}$ given $\boldsymbol{\theta_1}$ and U subject to the constraint $U < L(\boldsymbol{\theta_1}, \boldsymbol{\theta_2}; \boldsymbol{Y})$. Henceforth, this will be referred to as "getting a point in the slice." A naive rejection sampling scheme (repeatedly sample from $\pi(\boldsymbol{\theta_2}|\boldsymbol{\theta_1})$ until we get to a point in the slice) may not always give good results. An algorithm that shrinks the support of $\pi(\boldsymbol{\theta_2}|\boldsymbol{\theta_1})$ so that it gives a better approximation to the slice whenever there is a rejection is more appropriate.

We consider one such scheme called "shrinkage sampling" described in Neal (2003). In this context, it results in the following algorithm. For simplicity, let us assume $\boldsymbol{\theta_2}$ is one-dimensional. If a point $\hat{\boldsymbol{\theta}}_2$ drawn from $\pi(\boldsymbol{\theta_2}|\boldsymbol{\theta_1})$ is not in the slice and is larger (smaller) than the current value $\boldsymbol{\theta_2}$ (which is of course in the slice), the next draw is made from $\pi(\boldsymbol{\theta_2}|\boldsymbol{\theta_1})$ truncated with the upper (lower) bound being $\hat{\boldsymbol{\theta}}_2$. The truncated interval keeps shrinking with each rejection until a point in the slice is found. The multidimensional case works by shrinking hyperrectangles. As mentioned in Neal (2003), this ensures that the expected number of points drawn will not be too large, making it a more appropriate method for general use. However, intuitively it might result in higher autocorrelations compared to the simple rejection sampling scheme. In our experience, the shrinkage sampling scheme has performed better than the naive version in most cases.

Following Agarwal and Gelfand (2005), and suppressing $\boldsymbol{Y}$ in our notation, we summarize the main steps in our slice Gibbs sampling algorithm as follows:

(a) Partition $\boldsymbol{\theta} = (\boldsymbol{\theta_1}, \boldsymbol{\theta_2})$ so that samples from $p(\boldsymbol{\theta_1}|\boldsymbol{\theta_2})$ are easy to obtain;

(b) Draw $V = -\log L(\boldsymbol{\theta}) + Z$, where $Z \sim Exp(\text{mean} = 1)$;

(c) Draw $\boldsymbol{\theta_2}$ from $p(\boldsymbol{\theta_2}|\boldsymbol{\theta_1}, V)\, I(-\log L(\boldsymbol{\theta}) < V < \infty)$ using shrinkage sampling;

(d) Draw $\boldsymbol{\theta_1}$ from $p(\boldsymbol{\theta_1}|\boldsymbol{\theta_2})$;

(e) Iterate (b) through (d) until we get the appropriate number of MCMC samples.

The spatial models on which we focus arise through the specification of a Gaussian process for the data. With, for example, an isotropic covariance function, proposals for simulating the range parameter for, say, an exponential choice, or the range and smoothness parameters for a Matérn choice can be difficult to develop. That is, these parameters appear in the covariance matrix for $\boldsymbol{Y}$ in a nonstructured way (unless the spatial locations are on a regular grid). They enter the likelihood through the determinant and inverse of this matrix. And, for large n, the fewer matrix inversion and determinant computations, the better. As a result, for a noniterative sampling algorithm, it is very difficult to develop an effective importance sampling distribution for all of the model parameters. Moreover, as overall model dimension increases, resampling typically yields a very "spiked" discrete distribution.

Alternative Metropolis algorithms require effective proposal densities with careful tuning. Again, these densities are difficult to obtain for parameters in the correlation function. Morover, in general, such algorithms will suffer slower convergence than the Gibbs samplers we suggest, since full conditional distributions are not sampled. Furthermore, in our

experience, with customary proposal distributions we often encounter serious autocorrelation problems. When thinning to obtain a sample of roughly uncorrelated values, high autocorrelation necessitates an increased number of iterations. Additional iterations require additional matrix inversion and determinant calculation and can substantially increase run times. Discretizing the parameter spaces has been proposed to expedite computation in this regard, but it too has problems. The support set is arbitrary, which may be unsatisfying, and the support will almost certainly be adaptive across iterations, diminishing any computational advantage.

A.2.1 Constant mean process with nugget

Suppose $Y(\mathbf{s}_1), \ldots, Y(\mathbf{s}_n)$ are observations from a constant mean spatial process over $s \in D$ with a nugget. That is,

$$Y(\mathbf{s}_i) = \mu + w(\mathbf{s}_i) + \epsilon(\mathbf{s}_i)\,, \tag{A.5}$$

where the $\epsilon(\mathbf{s}_i)$ are realizations of a white noise process with mean 0 and variance τ^2. In (A.5), the $w(\mathbf{s}_i)$ are realizations from a second-order stationary Gaussian process with covariance function $\sigma^2 C(\boldsymbol{h};\boldsymbol{\rho})$ where C is a valid two-dimensional correlation function with parameters $\boldsymbol{\rho}$ and separation vector $\boldsymbol{h}$. Below we work with the Matérn class (2.8), so that $\boldsymbol{\rho} = (\phi, \nu)$. Thus (A.5) becomes a five-parameter model: $\boldsymbol{\theta} = (\mu, \sigma^2, \tau^2, \phi, \nu)^T$.

Note that though the $Y(\mathbf{s}_i)$ are conditionally independent given the $w(\mathbf{s}_i)$, a Gibbs sampler that also updates the latent $w(\mathbf{s}_i)$'s will be sampling an $(n+5)$-dimensional posterior density. However, it is possible to marginalize explicitly over the $w(\mathbf{s}_i)$'s (see Section 6.1), and it is almost always preferable to implement iterative simulation with a lower-dimensional distribution. The marginal likelihood associated with $\boldsymbol{Y} = (Y(\mathbf{s}_1), \ldots, Y(\mathbf{s}_n))$ is

$$\begin{aligned} L(\mu, \sigma^2, \tau^2, \phi, \nu; \boldsymbol{Y}) =& \mid \sigma^2 H(\boldsymbol{\rho}) + \tau^2 I \mid^{-\frac{1}{2}} \\ &\times \exp\{-(\boldsymbol{Y} - \mu\mathbf{1})^T (\sigma^2 H(\boldsymbol{\rho}) + \tau^2 I)^{-1} (\boldsymbol{Y} - \mu\mathbf{1})/2\}\,, \end{aligned} \tag{A.6}$$

where $(H(\boldsymbol{\rho}))_{ij} = \sigma^2 C(d_{ij};\boldsymbol{\rho})$ (d_{ij} being the distance between $\mathbf{s}_i$ and $\mathbf{s}_j$). Suppose we adopt a prior of the form $\pi_1(\mu)\pi_2(\tau^2)\pi_3(\sigma^2)\pi_4(\phi)\pi_5(\nu)$. Then (A.4) becomes $\pi_1(\mu)\pi_2(\tau^2)\pi_3(\sigma^2)\pi_4(\phi)\pi_5(\nu)\, I(U < L(\mu, \sigma^2, \tau^2, \phi, \nu; \boldsymbol{Y}))$. The Gibbs sampler is most easily implemented if, given ϕ and ν, we diagonalize $H(\boldsymbol{\rho})$, i.e., $H(\boldsymbol{\rho}) = P(\boldsymbol{\rho})D(\boldsymbol{\rho})(P(\boldsymbol{\rho}))^T$ where $P(\boldsymbol{\rho})$ is orthogonal with the columns of $P(\boldsymbol{\rho})$ giving the eigenvectors of $H(\boldsymbol{\rho})$ and $D(\boldsymbol{\rho})$ is a diagonal matrix with diagonal elements λ_i, the eigenvalues of $H(\boldsymbol{\rho})$. Then (A.6) simplifies to

$$\prod_{i=1}^{n} (\sigma^2 \lambda_i + \tau^2))^{-\frac{1}{2}} \exp\{-\frac{1}{2}(\boldsymbol{Y} - \mu\mathbf{1})^T P(\boldsymbol{\rho})(\sigma^2 D(\boldsymbol{\rho}) + \tau^2 I)^{-1} P^T(\boldsymbol{\rho})(\boldsymbol{Y} - \mu\mathbf{1})\}.$$

As a result, the constrained updating of σ^2 and τ^2 at a given iteration does not require repeated calculation of a matrix inverse and determinant. To minimize the number of diagonalizations of $H(\boldsymbol{\rho})$ we update ϕ and ν together. If there is interest in the $w(\mathbf{s}_i)$, their posteriors can be sampled straightforwardly after the marginalized model is fitted. For instance, $p(w(\mathbf{s}_i)|\boldsymbol{Y}) = \int p(w(\mathbf{s}_i)|\boldsymbol{\theta}, \boldsymbol{Y}) p(\boldsymbol{\theta}|\boldsymbol{Y}) d\boldsymbol{\theta}$ so each posterior sample $\boldsymbol{\theta}^\star$, using a draw from $p(w(\mathbf{s}_i)|\boldsymbol{\theta}^\star, \boldsymbol{Y})$ (which is a normal distribution), yields a sample from the posterior for $w(\mathbf{s}_i)$.

We remark that (A.5) can also include a parametric transformation of $Y(s)$. For instance, we could employ a power transformation to find a scale on which the Gaussian process assumption is comfortable. This only requires replacing $Y(\mathbf{s})$ with $Y^p(\mathbf{s})$ and adds one more parameter to the likelihood in (A.6). Lastly, we note that other dependence structures for $\boldsymbol{Y}$ can be handled in this fashion, e.g., equicorrelated forms, Toeplitz forms, and circulants.

A.2.2 Mean structure process with no pure error component

Now suppose $Y(\mathbf{s}_1), \ldots, Y(\mathbf{s}_n)$ are observations from a spatial process over $\mathbf{s} \in \mathrm{D}$ such that

$$Y(\mathbf{s}_i) = \boldsymbol{X}^T(\mathbf{s}_i)\boldsymbol{\beta} + w(\mathbf{s}_i)\,. \tag{A.7}$$

Again, the $w(\mathbf{s}_i)$ are realizations from a second-order stationary Gaussian process with covariance parameters σ^2 and $\boldsymbol{\rho}$. In (A.7), $\boldsymbol{X}(\mathbf{s}_i)$ could arise as a vector of site level covariates or $\boldsymbol{X}^T(\mathbf{s}_i)\boldsymbol{\beta}$ could be a trend surface specification as in the illustration below. To complete the Bayesian specification we adopt a prior of the form $\pi_1(\boldsymbol{\beta})\pi_2(\sigma^2)\pi_3(\boldsymbol{\rho})$ where $\pi_1(\boldsymbol{\beta})$ is $N(\boldsymbol{\mu}_\beta, \boldsymbol{\Sigma}_\beta)$ with $\boldsymbol{\mu}_\beta$ and $\boldsymbol{\Sigma}_\beta$ known.

This model is not hierarchical in the sense of our earlier forms, but we can marginalize explicitly over $\boldsymbol{\beta}$, obtaining

$$\begin{aligned} L(\sigma^2, \boldsymbol{\rho}; \boldsymbol{Y}) =& \mid \sigma^2 H(\boldsymbol{\rho}) + X\Sigma_{\boldsymbol{\beta}} X^T \mid^{-\frac{1}{2}} \\ &\times \exp\{-(\boldsymbol{Y} - X\boldsymbol{\mu}_\beta)^T(\sigma^2 H(\boldsymbol{\rho}) + X\Sigma_\beta X^T)^{-1}(\boldsymbol{Y} - X\boldsymbol{\mu}_\beta)/2\}\,, \end{aligned} \tag{A.8}$$

where the rows of X are the $\boldsymbol{X}^T(\mathbf{s}_i)$. Here, $H(\boldsymbol{\rho})$ is positive definite while $X\Sigma_{\boldsymbol{\beta}}X^T$ is symmetric positive semidefinite. Hence, there exists a nonsingular matrix $Q(\rho)$ such that $(Q^{-1}(\boldsymbol{\rho}))^T Q^{-1}(\boldsymbol{\rho}) = H(\boldsymbol{\rho})$ and also satisfying $(Q^{-1}(\boldsymbol{\rho}))^T \Omega Q^{-1}(\boldsymbol{\rho}) = X\Sigma_{\boldsymbol{\beta}}X^T$, where Ω is diagonal with diagonal elements that are eigenvalues of $X\Sigma_{\boldsymbol{\beta}}X^T H^{-1}(\boldsymbol{\rho})$. Therefore, (A.8) simplifies to

$$\begin{aligned} &|Q(\boldsymbol{\rho})| \prod_{i=1}^n (\sigma^2 + \lambda_i))^{-\frac{1}{2}} \\ &\quad \times \exp\left\{-\tfrac{1}{2}(\boldsymbol{Y} - X\mu_{\boldsymbol{\beta}})^T Q(\boldsymbol{\rho})^T(\sigma^2 I + \Omega)^{-1} Q(\boldsymbol{\rho})(\boldsymbol{Y} - X\mu_{\boldsymbol{\beta}})\right\}\,. \end{aligned}$$

As in the previous section, we run a Gibbs sampler to update U given $\sigma^2, \boldsymbol{\rho}$, and $\boldsymbol{Y}$, then σ^2 given $\boldsymbol{\rho}, U$, and $\boldsymbol{Y}$, and finally $\boldsymbol{\rho}$ given σ^2, U, and $\boldsymbol{Y}$. Then, given posterior samples $\{\sigma_l^{2*}, \boldsymbol{\rho}_l^*,\ l = 1, \ldots, L\}$ we can obtain posterior samples for $\boldsymbol{\beta}$ one for one given σ_l^{2*} and $\boldsymbol{\rho}_l^*$ by drawing $\boldsymbol{\beta}_l^*$ from a $N(A\mathbf{a}, A)$ distribution, where

$$\begin{aligned} A^{-1} &= \tfrac{1}{\sigma_l^{2*}} X^T H^{-1}(\boldsymbol{\rho}_l^*) X + \Sigma_{\boldsymbol{\beta}} \\ \text{and } \mathbf{a} &= \tfrac{1}{\sigma_l^{2*}} X^T H^{-1}(\boldsymbol{\rho}_l^*)\boldsymbol{Y} + \Sigma_{\boldsymbol{\beta}}^{-1}\boldsymbol{\mu}_{\boldsymbol{\beta}}\,. \end{aligned} \tag{A.9}$$

In fact, using standard identities (see, e.g., Rao, 1973, p. 29),

$$\left(\frac{1}{\sigma^2} X^T H^{-1}(\boldsymbol{\rho}) X + \Sigma_{\boldsymbol{\beta}}^{-1}\right)^{-1} = \Sigma_{\boldsymbol{\beta}} - \Sigma_{\boldsymbol{\beta}} X^T Q(\boldsymbol{\rho})(\sigma^2 I + \Omega)^{-1} Q^T(\boldsymbol{\rho}) X \Sigma_{\boldsymbol{\beta}}\,,$$

facilitating sampling from (A.9). Finally, if $\mu_{\boldsymbol{\beta}}$ and $\Sigma_{\boldsymbol{\beta}}$ were viewed as unknown we could introduce hyperparameters. In this case $\Sigma_{\boldsymbol{\beta}}$ would typically be diagonal and $\mu_{\boldsymbol{\beta}}$ might be $\mu_0\mathbf{1}$, but the simultaneous diagonalization would still simplify the implementation of the slice Gibbs sampler.

We note an alternate strategy that does not marginalize over $\boldsymbol{\beta}$ and does not require simultaneous diagonalization. The likelihood of $(\boldsymbol{\beta}, \sigma^2, \boldsymbol{\rho})$ is given by

$$L(\boldsymbol{\beta}, \sigma^2, \boldsymbol{\rho}; \boldsymbol{Y}) \propto |\sigma^2 H(\boldsymbol{\rho})|^{-\frac{1}{2}} \exp\left\{-(\boldsymbol{Y} - X\boldsymbol{\beta})^T H(\boldsymbol{\rho})^{-1}(\boldsymbol{Y} - X\boldsymbol{\beta})/2\sigma^2\right\}. \tag{A.10}$$

Letting $\boldsymbol{\theta_1} = (\boldsymbol{\beta}, \sigma^2)$ and $\boldsymbol{\theta_2} = \boldsymbol{\rho}$ with normal and inverse gamma priors on $\boldsymbol{\beta}$ and σ^2, respectively, we can update $\boldsymbol{\beta}$ and σ^2 componentwise conditional on $\boldsymbol{\theta_2}, \boldsymbol{Y}$, since $\boldsymbol{\beta}|\sigma^2, \boldsymbol{\theta_2}, \boldsymbol{Y}$ is normal while $\sigma^2|\boldsymbol{\beta}, \boldsymbol{\theta_2}, \boldsymbol{Y}$ is inverse gamma. $\boldsymbol{\theta_2}|\boldsymbol{\theta_1}, U, \boldsymbol{Y}$ is updated using the slice Gibbs sampler with shrinkage as described earlier.

A.2.3 Mean structure process with nugget

Extending (A.5) and (A.7) we now assume that $Y(\mathbf{s}_1), \ldots, Y(\mathbf{s}_n)$ are observations from a spatial process over $\mathbf{s} \in D$ such that

$$Y(\mathbf{s}_i) = \boldsymbol{X}^T(\mathbf{s}_i)\boldsymbol{\beta} + w(\mathbf{s}_i) + \epsilon(\mathbf{s}_i) \ . \tag{A.11}$$

As above, we adopt a prior of the form $\pi_1(\boldsymbol{\beta})\pi_2(\tau^2)\pi_3(\sigma^2)\pi_4(\boldsymbol{\rho})$, where $\pi_1(\boldsymbol{\beta})$ is $N(\boldsymbol{\mu}_\beta, \Sigma_\beta)$. Note that we could again marginalize over $\boldsymbol{\beta}$ and the $\mathrm{w}(\mathrm{s}_i)$ as in the previous section, but the resulting marginal covariance matrix is of the form $\sigma^2 H(\boldsymbol{\rho}) + X\Sigma_\beta X^T + \tau^2 I$. The simultaneous diagonalization trick does not help here since $Q(\boldsymbol{\rho})$ is not orthogonal. Instead we just marginalize over the $\mathrm{w}(\mathrm{s}_i)$, obtaining the joint posterior $p(\boldsymbol{\beta}, \tau^2, \sigma^2, \boldsymbol{\rho}, U|\boldsymbol{Y})$ proportional to

$$\begin{aligned}\pi_1(\boldsymbol{\beta})\pi_2(\tau^2)\pi_3(\sigma^2)\pi_4(\boldsymbol{\rho})\, I\Big(&U < |\sigma^2 H(\boldsymbol{\rho}) + \tau^2 I|^{-\frac{1}{2}} \\ &\times \exp\{-(\boldsymbol{Y} - X\boldsymbol{\beta})^T(\sigma^2 H(\boldsymbol{\rho}) + \tau^2 I)^{-1}(\boldsymbol{Y} - X\boldsymbol{\beta})/2\}\Big) \ .\end{aligned}$$

We employ the modified scheme suggested below (A.4) taking $\boldsymbol{\theta_1} = \boldsymbol{\beta}$ and $\boldsymbol{\theta_2} = (\tau^2, \sigma^2, \boldsymbol{\rho})$. The required full conditional distribution $p(\boldsymbol{\beta}|\tau^2, \sigma^2, \boldsymbol{\rho}, \boldsymbol{Y})$ is $N(A\mathbf{a}, A)$, where

$$\begin{aligned} A^{-1} &= X^T(\sigma^2 H(\boldsymbol{\rho}) + \tau^2 I)^{-1} X + \Sigma_{\boldsymbol{\beta}}^{-1} \\ \text{and } \mathbf{a} &= X^T(\sigma^2 H(\boldsymbol{\rho}) + \tau^2 I)^{-1}\boldsymbol{Y} + \Sigma_{\boldsymbol{\beta}}^{-1}\mu_{\boldsymbol{\beta}} \ . \end{aligned}$$

A.3 Structured MCMC sampling for areal model fitting

Structured Markov chain Monte Carlo (SMCMC) was introduced by Sargent, Hodges, and Carlin (2000) as a general method for Bayesian computing in richly parameterized models. Here, "richly parameterized" refers to hierarchical and other multilevel models. SMCMC (pronounced "smick-mick") provides a simple, general, and flexible framework for accelerating convergence in an MCMC sampler by providing a systematic way to update groups of similar parameters in blocks while taking full advantage of the posterior correlation structure induced by the model and data. Sargent (2000) applies SMCMC to several different models, including a hierarchical linear model with normal errors and a hierarchical Cox proportional hazards model.

Blocking, i.e., simultaneously updating multivariate blocks of (typically highly correlated) parameters, is a general approach to accelerating MCMC convergence. Liu (1994) and Liu et al. (1994) confirm its good performance for a broad class of models, though Liu et al. (1994, Sec. 5) and Roberts and Sahu (1997, Sec. 2.4) give examples where blocking slows a sampler's convergence. In this section, we show that spatial models of the kind proposed by Besag, York, and Mollíe (1991) using nonstationary "intrinsic autoregressions" are richly parameterized and lend themselves to the SMCMC algorithm. Bayesian inference via MCMC for these models has generally used single-parameter updating algorithms with often poor convergence and mixing properties. There have been some recent attempts to use blocking schemes for similar models. Cowles (2002, 2003) uses SMCMC blocking strategies for geostatistical and areal data models with normal likelihoods, while Knorr-Held and Rue (2002) implement blocking schemes using algorithms that exploit the sparse matrices that arise out of the areal data model.

We study several strategies for block-sampling parameters in the posterior distribution when the likelihood is Poisson. Among the SMCMC strategies we consider here are blocking using different-sized blocks (grouping by geographical region), updating jointly with and without model hyperparameters, "oversampling" some of the model parameters, reparameterization via hierarchical centering and "pilot adaptation" of the transition kernel.

Our results suggest that our techniques will generally be far more accurate (produce less correlated samples) and often more efficient (produce more effective samples per second) than univariate sampling procedures.

A.3.1 SMCMC algorithm basics

Following Hodges (1998), we consider a hierarchical model expressed in the general form,

$$\left[\begin{array}{c} y \\ \hline 0 \\ \hline M \end{array}\right] = \left[\begin{array}{c|c} X_1 & 0 \\ \hline H_1 & H_2 \\ \hline G_1 & G_2 \end{array}\right] \left[\begin{array}{c} \theta_1 \\ \hline \theta_2 \end{array}\right] + \left[\begin{array}{c} \epsilon \\ \hline \delta \\ \hline \xi \end{array}\right] . \tag{A.12}$$

The first row of this layout is actually a collection of rows corresponding to the "data cases," or the terms in the joint posterior into which the response, the data y, enters directly. The terms in the second row (corresponding to the H_i) are called "constraint cases" since they place stochastic constraints on possible values of θ_1 and θ_2. The terms in the third row, the "prior cases" for the model parameters, have known (specified) error variances for these parameters. Equation (A.12) can be expressed as $Y = \mathrm{X}\mathbf{\Theta} + \mathrm{E}$, where X and Y are known, $\mathbf{\Theta}$ is unknown, and E is an error term with block diagonal covariance matrix $\Gamma = \mathrm{Diag}(\mathrm{Cov}(\epsilon), \mathrm{Cov}(\delta), \mathrm{Cov}(\xi))$. If the error structure for the data is normal, i.e., if the ϵ vector in the constraint case formulation (A.12) is normally distributed, then the conditional posterior density of $\mathbf{\Theta}$ is

$$\mathbf{\Theta}|Y, \Gamma \sim N((X^T\Gamma^{-1}X)^{-1}(X^T\Gamma^{-1}Y)\ ,\ (X^T\Gamma^{-1}X)^{-1})\ . \tag{A.13}$$

The basic SMCMC algorithm is then nothing but the following two-block Gibbs sampler :

(a) Sample $\mathbf{\Theta}$ as a single block from the above normal distribution, using the current value of Γ.

(b) Update Γ using the conditional distribution of the variance components, using the current value of $\mathbf{\Theta}$.

In our spatial model setting, the errors are not normally distributed, so the normal density described above is not the correct conditional posterior distribution for $\mathbf{\Theta}$. Still, a SMCMC algorithm with a Metropolis-Hastings implementation can be used, with the normal density in (A.13) taken as the candidate density.

A.3.2 Applying structured MCMC to areal data

Consider again the Poisson-CAR model of Subsection 6.4.3, with no covariates so that $\mu_i = \theta_i + \phi_i, i = 1, \ldots, N$, where N is the total number of regions, and $\{\theta_1, .., \theta_N\}, \{\phi_1, .., \phi_N\}$ are vectors of random effects. The θ_i's are independent and identically distributed Gaussian normal variables with precision parameter τ_h, while the ϕ_i's are assumed to follow a $CAR(\tau_c)$ distribution. We place conjugate gamma hyperpriors on the precision parameters, namely $\tau_h \sim G(\alpha_h, \beta_h)$ and $\tau_c \sim G(\alpha_c, \beta_c)$ with $\alpha_h = 1.0$, $\beta_h = 100.0$, $\alpha_c = 1.0$ and $\beta_c = 50.0$ (these hyperpriors have means of 100 and 50, and standard deviations of 10,000 and 2,500, respectively, a specification recommended by Bernardinelli et al., 1995).

There is a total of $2N + 2$ model parameters: $\{\theta_i :\ i = 1, \ldots N\}$, $\{\phi_i :\ i = 1, \ldots N\}$, τ_h and τ_c. The SMCMC algorithm requires that we transform the Y_i data points to $\hat{\mu}_i = \log(Y_i/E_i)$, which can be conveniently thought of as the response since they should be roughly linear in the model parameters (the θ_i's and ϕ_i's). For the constraint case formulation, the different levels of the model are written down case by case. The data cases are $\hat{\mu}_i$, $i = 1, \ldots, N$. The constraint cases for the θ_i's are $\theta_i \sim N(0, 1/\tau_h)$, $i = 1, \ldots, N$. For the constraint cases involving the ϕ_i's, the differences between the neighboring ϕ_i's can be

used to get an unconditional distribution for the ϕ_i's using pairwise differences (Besag et al., 1995). Thus the constraint cases can be written as

$$(\phi_i - \phi_j)|\tau_c \sim N(0, 1/\tau_c) \tag{A.14}$$

for each pair of adjacent regions (i, j).

To obtain an estimate of Γ, we need estimates of the variance-covariance matrix corresponding to the $\hat{\mu}_i$'s (the data cases) and initial estimates of the variance-covariance matrix for the constraint cases (the rows corresponding to the θ_i's and ϕ_i's). Using the delta method, we can obtain an approximation as follows: assume $Y_i \sim N(E_i e^{\mu_i}, E_i e^{\mu_i})$ (roughly), so invoking the delta method we can see that $\mathrm{Var}(log(Y_i/E_i))$ is approximately $1/Y_i$. A reasonably good starting value is particularly important here since we never update these variance estimates (the data variance section of Γ stays the same throughout the algorithm). For initial estimates of the variance components corresponding to the θ_i's and the ϕ_i's, we can use the mean of the hyperprior densities on τ_h and τ_c, and substitute these values into Γ.

As a result, the SMCMC candidate generating distribution is thus of the form (A.13), with the Y_i's replaced by $\hat{\boldsymbol{\mu}}$. To compute the Hastings ratio, the distribution of the ϕ_i's is rewritten in the joint pairwise difference form with the appropriate exponent for τ_c (Hodges, Carlin, and Fan, 2003):

$$p(\phi_1, \phi_2, ..., \phi_N|\tau_c) \propto \tau_c^{(N-1)/2} \exp\left\{-\frac{\tau_c}{2}\sum_{i\sim j}(\phi_i - \phi_j)^2\right\}, \tag{A.15}$$

where $i \sim j$ if i and j are neighboring regions. Finally, the joint distribution of the θ_i's is given by

$$p(\theta_1, \theta_2, ..., \theta_N|\tau_h) \propto \tau_h^{N/2} \exp\left\{-\frac{\tau_h}{2}\sum_{i=1}^{N}\theta_i^2\right\}. \tag{A.16}$$

As above, the response vector is $\hat{\boldsymbol{\mu}}^T = \{\log(Y_1/E_1), \ldots, \log(Y_N/E_N)\}$. The $(2N + C) \times 2N$ design matrix for the spatial model is defined by

$$X = \left[\begin{array}{c|c} I_{N\times N} & I_{N\times N} \\ \hline -I_{N\times N} & 0_{N\times N} \\ \hline 0_{C\times N} & A_{C\times N} \end{array}\right]. \tag{A.17}$$

The design matrix is divided into two halves, the left half corresponding to the N θ_i's and the right half referring to the N ϕ_i's. The top section of this design matrix is an $N \times 2N$ matrix relating $\hat{\mu}_i$ to the model parameters θ_i and ϕ_i. In the ith row, a 1 appears in the ith and $(N + i)$th columns while 0's appear elsewhere. Thus the ith row corresponds to $\mu_i = \theta_i + \phi_i$. The middle section of the design matrix is an $N \times 2N$ matrix that imposes a stochastic constraint on each θ_i separately (θ_i's are i.i.d normal). The bottom section of the design matrix is a $C \times 2N$ matrix with each row having a -1 and 1 in the $(N + k)$th and $(N + l)$th columns, respectively, corresponding to a stochastic constraint being imposed on $\phi_l - \phi_k$ (using the pairwise difference form of the prior on the ϕ_i's as described in (A.14) with regions l and k being neighbors). The variance-covariance matrix Γ is a diagonal matrix with the top left section corresponding to the variances of the data cases, i.e., the $\hat{\mu}_i$'s. Using the variance approximations described above, the $(2N + C) \times (2N + C)$ block diagonal variance-covariance matrix is

$$\Gamma = \left[\begin{array}{c|c|c} Diag(1/Y_1, 1/Y_2, \ldots, 1/Y_N) & 0_{N\times N} & 0_{N\times C} \\ \hline 0_{N\times N} & \frac{1}{\tau_h} I_{N\times N} & 0_{N\times C} \\ \hline 0_{C\times N} & 0_{C\times N} & \frac{1}{\tau_c} I_{C\times C} \end{array}\right]. \tag{A.18}$$

Note that the exponent on τ_c in (A.15) would actually be $C/2$ (instead of $(N-1)/2$) if obtained by taking the product of the terms in (A.14). Thus, (A.14) is merely a form we use to describe the distribution of the ϕ_is for our constraint case specification. The formal way to incorporate the distribution of the ϕ_is in the constraint case formulation is by using an alternate specification of the joint distribution of the ϕ_i's, as described in Besag and Kooperberg (1995). This form is a $N \times N$ Gaussian density with precision matrix, Q,

$$p(\phi_1, \phi_2, ..., \phi_N | \tau_c) \propto \exp\left(-\frac{\tau_c}{2}\boldsymbol{\phi}^T Q \boldsymbol{\phi}\right), \text{ where } \boldsymbol{\phi}^T = (\phi_1, \phi_2, ..., \phi_N), \tag{A.19}$$

and

$$Q_{ij} = \begin{cases} c & \text{if } i = j \text{ where } c = \text{ number of neighbors of region } i \\ 0 & \text{if } i \text{ is not adjacent to } j \\ -1 & \text{if } i \text{ is not adjacent to } j \end{cases}.$$

However, it is possible to show that this alternate formulation (using the corresponding design and Γ matrices) results in the same SMCMC candidate mean and covariance matrix for $\boldsymbol{\Theta}$ given τ_h and τ_c as the one described in (A.13); see Haran, Hodges, and Carlin (2003) for details.

A.3.3 Algorithmic schemes

Univariate MCMC (UMCMC): For the purpose of comparing the different blocking schemes, one might begin with a univariate (updating one variable at a time) sampler. This can be done by sampling τ_h and τ_c from their gamma full conditional distributions, and then, for each i, sampling each θ_i and ϕ_i from its full conditional distribution, the latter using a Metropolis step with univariate Gaussian random walk proposals.

Reparameterized Univariate MCMC (RUMCMC): One can also reparameterize from $(\theta_1, \ldots, \theta_N, \phi_1, \ldots, \phi_N)$ to $(\mu_1, \ldots, \mu_N, \phi_1, \ldots, \phi_N)$, where $\mu_i = \theta_i + \phi_i$. The (new) model parameters and the precision parameters can be sampled in a similar manner as for UMCMC. This "hierarchical centering" was suggested by Besag et al. (1995) and Waller et al. (1997) for the spatial model, and discussed in general by Gelfand et al. (1995, 1996).

Structured MCMC (SMCMC): A first step here is *pilot adaptation*, which involves sampling $(\tau_h,\ \tau_c)$ from their gamma full conditionals, updating the Γ matrix using the averaged $(\tau_h,\ \tau_c)$ sampled so far, updating the SMCMC candidate covariance matrix and mean vector using the Γ matrix, and then sampling $(\boldsymbol{\theta}, \boldsymbol{\phi})$ using the SMCMC candidate in a Metropolis-Hastings step. We may run the above steps for a "tuning" period, after which we fix the SMCMC candidate mean and covariance, sampled $(\tau_h,\ \tau_c)$ as before, and use the Metropolis-Hastings to sample $(\boldsymbol{\theta}, \boldsymbol{\phi})$ using SMCMC proposals. Some related strategies include adaptation of the Γ matrix more or less frequently, adaptation over shorter and longer periods of time, and pilot adaptation while blocking on groups of regions.

Our experience with pilot adaptation schemes indicates that a single proposal, regardless of adaptation period length, will probably be unable to provide a reasonable acceptance rate for the many different values of (τ_h, τ_c) that will be drawn in realistic problems. As such, we typically turn to oversampling $\boldsymbol{\Theta}$ relative to (τ_h, τ_c); that is, the SMCMC proposal is always based on the current (τ_h, τ_c) value. In this algorithm, we sample τ_h and τ_c from their gamma full conditionals, then compute the SMCMC proposal based on the Γ matrix using the generated τ_h and τ_c. For each (τ_h, τ_c) pair, we run a Hastings independence subchain by sampling a sequence of length 100 (say) of $\boldsymbol{\Theta}$'s using the SMCMC proposal. Further implementational details for this algorithm are given in Haran (2003).

Reparameterized Structured MCMC (RSMCMC): This final algorithm is the SMCMC analogue of the reparametrized univariate algorithm (RUMCMC). It follows exactly the same steps as the SMCMC algorithm, with the only difference being that $\boldsymbol{\Theta}$ is

now $(\boldsymbol{\mu}, \boldsymbol{\phi})$ instead of $(\boldsymbol{\theta}, \boldsymbol{\phi})$, and the proposal distribution is adjusted according to the new parameterization.

Haran, Hodges, and Carlin (2003) compare these schemes in the context of two areal data examples, using the notion of *effective sample size*, or ESS (Kass et al., 1998). ESS is defined for each parameter as the number of MCMC samples drawn divided by the parameter's so-called autocorrelation time, $\kappa = 1 + 2\sum_{k=1}^{\infty} \rho(k)$, where $\rho(k)$ is the autocorrelation at lag k. One can estimate κ from the MCMC chain, using the initial monotone positive sequence estimator as given by Geyer (1992). Haran et al. (2003) find UMCMC to perform poorly, though the reparameterized univariate algorithm (RUMCMC) does provide a significant improvement in this case. However, SMCMC and RSMCMC still perform better than both univariate algorithms. Even when accounting for the amount of time taken by the SMCMC algorithm (in terms of effective samples per second), the SMCMC scheme results in a far more efficient sampler than the univariate algorithm; for some parameters, SMCMC produced as much as 64 times more effective samples per second.

Overall, experience with applying several SMCMC blocking schemes to real data sets suggests that SMCMC provides a standard, systematic technique for producing samplers with far superior mixing properties than simple univariate Metropolis-Hastings samplers. The SMCMC and RSMCMC schemes appear to be reliable ways of producing good ESSs, irrespective of the data sets and parameterizations. In many cases, the SMCMC algorithms are also competitive in terms of ES/s. In addition since the blocked SMCMC algorithms mix better, their convergence should be easier to diagnose and thus lead to final parameter estimates that are less biased. These estimates should also have smaller associated Monte Carlo variance estimates.

A.4 `spBayes`: Under the hood

Finally, we conclude with some remarks on computing behind the algorithms in `spBayes`. `spBayes` version 0.3-7 (CRAN 6/01/13) comprises a substantial reformulation and rewrite of core functions for model fitting, with a focus on improving computational efficiency, flexibility, and usability. Among other improvements, this and subsequent versions offer: (i) improved sampler convergence rate and efficiency by reducing parameter space; (ii) decreased sampler run-time by avoiding expensive matrix computations; and (iii) increased scalability to large datasets by implementing a class of *predictive process* models that attempt to overcome computational hurdles by representing spatial processes in terms of lower-dimensional realizations. Beyond these general computational improvements for existing models, new functions were added to model data indexed in both space and time. These functions implement a class of dynamic spatio-temporal models for settings where space is viewed as continuous and time is taken as discrete.

With multiple processors, substantial gains in computing efficiency can be realized through parallel processing of matrix operations. Most of the functions in `spBayes` are written in `C/C++` and leverage `R`'s *Foreign Language Interface* to call `FORTRAN BLAS` (Basic Linear Algebra Subprograms; see Blackford et al., 2002) and `LAPACK` (Linear Algebra Package; see Anderson et al., 1999) libraries for efficient matrix computations. Table A.4 offers a list of BLAS and LAPACK functions used to implement the MCMC samplers. Referring to Table A.4, Cholesky decompositions are computed using `dpotrf` and triangular systems are solved using either `dtrsv` or `dtrsm`, depending on the form of the equation's right-hand side. We try to avoid dense matrix-matrix multiplication, i.e., calls to `dgemm`, due to its computational overhead. Often careful formulation of the problem can result in fewer calls to `dgemm` and other *expensive* BLAS level-2 and LAPACK functions.

A heavy reliance on BLAS and LAPACK functions for matrix operations allows us to leverage multi-processor/core machines via threaded implementations of BLAS and LAPACK, e.g., Intel's Math Kernel Library (MKL; `http://software.intel.com/en-us/`

Function	Description
`dpotrf`	LAPACK for Cholesky factorization.
`dtrsv`	Level 2 BLAS for solving $A\mathbf{x} = \mathbf{b}$, where A is triangular.
`dtrsm`	Level 3 BLAS for solving $AX = B$, where A is triangular.
`dgemv`	Level 2 BLAS matrix-vector multiplication.
`dgemm`	Level 3 BLAS matrix-matrix multiplication.

Table A.1 *Common BLAS and LAPACK functions used in* `spBayes` *function calls.*

`intel-mkl`). With the exception of `dtrsv`, all functions in Table A.4 are threaded in Intel's MKL. Use of MKL, or similar threaded libraries, can dramatically reduce sampler run-times. For example, the illustrative analyses offered in the earlier sections were conducted using `R`, and hence `spBayes`, compiled with MKL on an Intel Ivy Bridge i7 quad-core processor with hyperthreading. The use of these parallel matrix operations results in a near linear speadup in the MCMC sampler's run-time with the number of CPUs — at least 4 CPUs were in use in each function call.

Appendix B

Answers to selected exercises

Chapter 1

3 As hinted in the problem statement, level of urbanicity might well explain the poverty pattern evident in Figure 1.2. Other regional spatially oriented covariates to consider might include percent of minority residents, percent with high school diploma, unemployment rate, and average age of the housing stock. The point here is that spatial patterns can often be explained by patterns in existing covariate data. Accounting for such covariates in a statistical model may result in residuals that show little or no spatial pattern, thus obviating the need for formal spatial modeling.

7(a) The appropriate R code is as follows:

```
#  R program to compute geodesic distance
#  see also www.auslig.gov.au/geodesy/datums/distance.htm

#  input:   point1=(long,lat) and point2=(long,lat)
              in degrees
#  output:  distance in km between the two points
#  example:
point1 <- c(87.65,41.90)  # Chicago (downtown)
point2 <- c(87.90,41.98)  # Chicago (O'Hare airport)
point3 <- c(93.22,44.88)  # Minneapolis (airport)
#  geodesic(point1,point3) returns 558.6867

geodesic <- function(point1, point2){
  R <- 6371
    point1 <- point1 * pi/180
    point2 <- point2 * pi/180
      d <- sin(point1[2]) * sin(point2[2]) +
              cos(point1[2]) * cos(point2[2]) *
              cos(abs(point1[1] - point2[1]))
    R*acos(d)
}
```

(b) Chicago to Minneapolis, 562 km; New York to New Orleans, 1897.2 km.

8 Chicago to Minneapolis, 706 km; New York to New Orleans, 2172.4 km. This overestimation is expected since the approach stretches the meridians and parallels, or equivalently, presses the curved domain onto a plane, thereby stretching the domain (and hence the distances). As the geodesic distance increases, the quality of the naive estimates deteriorates.

9 Chicago to Minneapolis, 561.8 km; New York to New Orleans, 1890.2 km. Here, the slight underestimation is expected, since it finds the straight line by penetrating (burrowing

through) the spatial domain. Still, this approximation seems quite good even for distances close to 2000 km (e.g., New York to New Orleans).

10(a) Chicago to Minneapolis, 562.2 km; New York to New Orleans, 1901.5 km.

(b) Whenever all of the points are located along a parallel or a meridian, this projection will not be defined.

Chapter 2

8(a) Conditional on $\mathbf{u}$, finite realizations of Y are clearly Gaussian since $(Y(\mathbf{s}_i))_{i=1}^n = (W(x_i))_{i=1}^n$, where $x_i = \mathbf{s}_i^T\mathbf{u}$, and W is Gaussian. The covariance function is given by $Cov(Y(\mathbf{s}), Y(\mathbf{s}+\mathbf{h})) = c\left(\mathbf{h}^T\mathbf{u}\right)$, where c is the (stationary) covariance function of W.

(b) For the marginal process, we need to take expectation over the distribution of $\mathbf{u}$, which is uniform over the n-dimensional sphere. Note that

$$\begin{aligned} Cov(Y(\mathbf{s}), Y(\mathbf{s}+\mathbf{h})) &= E_{\mathbf{u}}\left[Cov\left(W\left(\mathbf{s}^T\mathbf{u}\right), W\left((\mathbf{s}+\mathbf{h})^T\mathbf{u}\right)\right)\right] \\ &= E_{\mathbf{u}}\left[c\left(\mathbf{h}^T\mathbf{u}\right)\right] . \end{aligned}$$

Then, we need to show that $E_{\mathbf{u}}\left[c\left(\mathbf{h}^T\mathbf{u}\right)\right]$ is a function of $||\mathbf{h}||$. Now, $\mathbf{h}^T\mathbf{u} = ||\mathbf{h}||\cos\theta$, so $E_{\mathbf{u}}\left[c\left(\mathbf{h}^T\mathbf{u}\right)\right] = E_\theta\left[c\left(||\mathbf{h}||\cos\theta\right)\right]$. But θ, being the angle made by a uniformly distributed random vector $\mathbf{u}$, has a distribution that is invariant over the choice of $\mathbf{h}$. Thus, the marginal process $Y(\mathbf{s})$ has isotropic covariance function $K(r) = E_\theta\left[c\left(r\cos\theta\right)\right]$.

Note: The above covariance function (in $\Re^n$) can be computed using spherical integrals as

$$K(r) = \frac{2\Gamma(n/2)}{\sqrt{\pi}\Gamma((n-1)/2)}\int_0^1 c(r\nu)\left(1-\nu^2\right)^{(n-3)/2} d\nu.$$

9 If $\tau^2 = 0$, then $\Sigma = \sigma^2 H(\phi)$. If $\mathbf{s}_0 = \mathbf{s}_k$, where $\mathbf{s}_k$ is a monitored site, we have $\boldsymbol{\gamma}^T = \sigma^2\left[H(\phi)\right]_{k*}$, the kth row of $\sigma^2 H(\phi)$. Thus, $\mathbf{e}_k^T H(\phi) = \left(1/\sigma^2\right)\boldsymbol{\gamma}^T$, where $\mathbf{e}_k = (0,\ldots,1,\ldots,0)^T$ is the kth coordinate vector. So $\mathbf{e}_k^T = \left(1/\sigma^2\right)\boldsymbol{\gamma}^T H^{-1}(\phi)$. Substituting this into equation (2.15), we get

$$E\left[Y(\mathbf{s}_k)|\mathbf{y}\right] = \mathbf{x}_k^T\boldsymbol{\beta} + \mathbf{e}_k^T(\mathbf{y} - X\boldsymbol{\beta}) = \mathbf{x}_k^T\boldsymbol{\beta} + y(\mathbf{s}_k) - \mathbf{x}_k^T\boldsymbol{\beta} = y(\mathbf{s}_k) .$$

When $\tau^2 > 0$, $\Sigma = \sigma^2 H(\phi) + \tau^2 I$, so the Σ^{-1} in equation (2.15) does not simplify, and we do not have the above result.

Chapter 4

1 Brook's Lemma, equation (4.7), is easily verified as follows: Starting with the extreme right-hand side, observe that

$$\frac{p(y_{10},\ldots,y_{n0})}{p(y_{n0}|y_{10},\ldots,y_{n-1,0})} = p(y_{10},\ldots,y_{n-1,0}) .$$

Now observe that

$$p(y_n|y_{10},\ldots,y_{n-1,0})p(y_{10},\ldots,y_{n-1,0}) = p(y_{10},\ldots,y_{n-1,0},y_n) .$$

The result follows by simply repeating these two steps, steadily moving leftward through (4.7).

3 We provide two different approaches to solving the problem. The first approach is a direct manipulative approach, relying upon elementary algebraic simplifications, and might seem a bit tedious. The second approach relies upon some relatively advanced concepts in matrix analysis, yet does away with most of the manipulations of the first approach.

Method 1: In the first method we derive the following identity:

$$\mathbf{u}^T D^{-1}(I-B)\mathbf{u} = \sum_{i=1}^{n} \frac{u_i^2}{\tau_i^2}\left(1-\sum_{j=1}^{n} b_{ij}\right) + \sum_{i<j} \frac{b_{ij}}{\tau_i^2}\left(u_i-u_j\right)^2 \ , \tag{B.1}$$

where $\mathbf{u} = (u_1, ..., u_n)^T$. Note that if this identity is indeed true, the right-hand side must be strictly positive; all the terms in the r.h.s. are strictly positive by virtue of the conditions on the elements of the B matrix unless $\mathbf{u} = \mathbf{0}$. This would imply the required positive definiteness.

We may derive the above identity either by starting with the l.h.s. and eventually obtaining the r.h.s., or vice versa. We adopt the former. So,

$$\begin{aligned}
\mathbf{u}^T D^{-1}(I-B)\mathbf{u} &= \sum_i \frac{u_i^2}{\tau_i^2} - \sum_i \sum_j \frac{b_{ij}}{\tau_i^2} u_i u_j \\
&= \sum_i \frac{u_i^2}{\tau_i^2} - \sum_i \frac{b_{ii}}{\tau_i^2} u_i^2 - \sum_i \sum_{j\neq i} \frac{b_{ij}}{\tau_i^2} u_i u_j \\
&= \sum_i \frac{u_i^2}{\tau_i^2}\left(1-b_{ii}\right) - \sum_i \sum_{j\neq i} \frac{b_{ij}}{\tau_i^2} u_i u_j \ .
\end{aligned}$$

Adding and subtracting $\sum_i \sum_{j\neq i} \left(u_i^2/\tau_i^2\right) b_{ij}$ to the last line of the r.h.s., we write

$$\begin{aligned}
\mathbf{u}^T D^{-1}(I-B)\mathbf{u} &= \sum_i \frac{u_i^2}{\tau_i^2}\left(1-\sum_j b_{ij}\right) + \sum_i \sum_{j\neq i} \frac{u_i^2}{\tau_i^2} b_{ij} \\
&\qquad - \sum_i \sum_{j\neq i} \frac{b_{ij}}{\tau_i^2} u_i u_j \\
&= \sum_i \frac{u_i^2}{\tau_i^2}\left(1-\sum_j b_{ij}\right) + \sum_i \sum_{j\neq i} \frac{b_{ij}}{\tau_i^2}\left(u_i^2 - u_i u_j\right) \\
&= \sum_i \frac{u_i^2}{\tau_i^2}\left(1-\sum_j b_{ij}\right) + \sum_{i<j} \frac{b_{ij}}{\tau_i^2}\left(u_i-u_j\right)^2 \ .
\end{aligned}$$

To explain the last manipulation,

$$\sum_i \sum_{j\neq i} \frac{b_{ij}}{\tau_i^2}\left(u_i^2 - u_i u_j\right) = \sum_{i<j} \frac{b_{ij}}{\tau_i^2}\left(u_i-u_j\right)^2 , \tag{B.2}$$

note that the sum on the l.h.s. of (B.2) extends over the $2 \times \binom{n}{2}$ (unordered) pairs of (i, j). Consider any particular pair, say, (k, l) with $k < l$, and its "reflection" (l, k). Using the symmetry condition, $b_{kl}/\tau_k^2 = b_{lk}/\tau_l^2$, we may combine the two terms from this pair as

$$\frac{b_{kl}}{\tau_k^2}\left(u_k^2 - u_k u_l\right) + \frac{b_{lk}}{\tau_l^2}\left(u_l^2 - u_l u_k\right) = \frac{b_{kl}}{\tau_k^2}\left(u_k-u_l\right)^2 .$$

Performing the above trick for each of the $\binom{n}{2}$ pairs, immediately results in (B.2).

Method 2: The algebra above may be skipped using the following argument, based on eigenanalysis. First, note that, with the given conditions on B, the matrix $D^{-1}(I-B)$ is (weakly) diagonally dominant. This means that, if $A = D^{-1}(I-B)$, and $R_i(A) = \sum_{j \neq i} |a_{ij}|$ (the sum of the absolute values of the ith row less that of the diagonal element), then $|a_{ii}| \geq R_i(A)$, for all i, with strict inequality for at least one i. Now, using the Gershgorin Circle Theorem (see, e.g., Theorem 7.2.1 in Golub and Van Loan, p. 320), we immediately see that 0 cannot be an interior point of Gershgorin circle. Therefore, all the eigenvalues of A must be nonnegative. But note that all the elements of B are strictly positive. This means that all the elements of A are nonzero, which means that 0 cannot be a boundary point of the Gershgorin circle. Therefore, 0 must be an exterior point of the circle, proving that all the eigenvalues of A must be strictly positive. So A, being symmetric, must be positive definite.

Note: It is important that the matrix D be chosen so as to ensure $D^{-1}(I-B)$ is symmetric. To see that this condition cannot be relaxed, consider the following example. Let us take $B = \begin{pmatrix} 0.3 & 0.5 \\ 0.1 & 0.9 \end{pmatrix}$. Clearly the matrix satisfies the conditions laid down in the problem statement. If we are allowed to choose an arbitrary D, we may take $D = I_2$, the 2×2 identity matrix, and so $D^{-1}(I-B) = \begin{pmatrix} 0.7 & -0.5 \\ -0.1 & 0.1 \end{pmatrix}$. But this is not positive definite, as is easily seen by noting that with $\mathbf{u}^T = (1, 2)$, we obtain $\mathbf{u}^T D^{-1}(I-B)\,\mathbf{u} = -0.1 < 0$.

4 Using the identity in (B.1), it is immediately seen that, taking B to be the scaled proximity matrix (as in the text just above equation (4.15)), we have $\sum_{j=1}^{n} b_{ij} = 1$, for each i. This shows that the first term on the r.h.s. of (B.1) vanishes, leading to the second term, which is a pairwise difference prior.

Chapter 5

1 The complete `WinBUGS` code to fit this model is given below. Recall "#" is a comment in `WinBUGS`, so this version actually corresponds model for part (c).

```
model
{
    for (i in 1:N) {
        y[i] ~ dbern(p[i])
#        logit(p[i]) <- b0 + b1*kieger[i] + b2*team[i]
#        logit(p[i]) <- b0 + b2*(team[i]-mean(team[]))
        logit(p[i]) <- b0 + b1*(pct[i]-mean(pct[]))
        pct[i] <- kieger[i]/(kieger[i]+team[i])
}
   b0 ~ dnorm(0, 1.E-3)
   b1 ~ dnorm(0, 1.E-3)
   b2 ~ dnorm(0, 1.E-3)
}

HERE ARE INITS:
list(b0=0, b1=0, b2=0)


HERE ARE THE DATA:
list(N = 9,                          # number of observations
```

	95% Credible intervals			
Model	β_1	β_2	DIC	p_D
(a)	(–3.68, 1.21)	(.152, 2.61)	8.82	1.69
(b)	—	(.108, 1.93)	9.08	1.61
(c)	(–70.8, –3.65)	—	8.07	1.59

Table B.1 *Posterior summaries, Carolyn Kieger prep basketball logit model.*

```
y = c(1,1,1,1,0,1,1,1,0),                # team win/loss
kieger = c(31,31,36,30,32,33,31,33,32),  # Kieger points
team = c(31,16,35,42,19,37,29,23,15))    # team points
```

Running a single Gibbs sampling chain for 20,000 iterations after a 1,000-iteration burn-in period, Table B.1 gives the resulting 95% equal tail posterior credible intervals for β_1 and β_2 for each model, as well as the corresponding DIC and p_D scores.

(a) Running this model produces MCMC chains with slowly moving sample traces and very high autocorrelations and cross-correlations (especially between β_0 and β_1, since Kieger's uncentered scores are nearly identical). The 95% equal-tail confidence interval for β_1 includes 0, suggesting Kieger's score is not a significant predictor of game outcome; the p_D score of just 1.69 also suggests there are not 3 "effective" parameters in the model (although none of these posterior summaries are very trustworthy due to the high autocorrelations, hence low effective sample MCMC sample size). Thus, the model is not acceptable either numerically (poor convergence; unstable estimates due to low effective sample size) or statistically (model is overparametrized).

(b) Since "kieger" was not a significant predictor in part (a), we delete it, and center the remaining covariate ("team") around its own mean. This helps matters immensely: numerically, convergence is much better and parameter and other estimates are much more stable. Statistically, the DIC score is not improved (slightly higher), but the p_D is virtually unchanged at 1.6 (so both of the remaining parameters in the model are needed), and β_2 is more precisely estimated.

(c) Again convergence is improved, and now the DIC score is also better. β_1 is significant and negative, since the higher the proportion of points scored by Kieger (i.e., the lower the output by the rest of the team), the less likely a victory becomes.

(d) The p_i themselves have posteriors implied by the β_j posteriors and reveal that the team was virtually certain to win Games 1, 3, 4, 6, and 7 (where Kieger scored a lower percentage of the points), but could well have lost the others, especially Games 2 and 9 (the former of which the team was fortunate to win anyway). This implies that the only thing that might still be missing from our model is some measure of how *few* points the opponent scores, which is of course governed by how well Kieger and the other team members play on *defense*. But fitting such a model would obviously require defensive statistics (blocked shots, etc.) that we currently lack.

5(a) In our implementation of `WinBUGS`, we obtained a slightly better DIC score with Model XI (7548.3, versus 7625.2 for Model XII), suggesting that the full time-varying complexity is not required in the survival model. The fact that the 95% posterior credible interval for γ_3, (–0.43, .26), includes 0 supports this conclusion.

(b) We obtained point and 95% interval estimates of –0.20 and (–0.25, –0.14) for γ_1, and –1.61 and (–2.13, –1.08) for γ_2.

(c) Figure B.1 plots the estimated posteriors (smoothed histograms of `WinBUGS` output). In both the separate (panel a) and joint (panel b) analyses, this patient's survival is clearly better if he receives ddC instead of ddI. However, the joint analysis increases the estimated median survival times by roughly 50% in both groups.

Figure B.1 *Median survival time for a hypothetical patient (male, negative AIDS diagnosis at study entry, intolerant of AZT): (a) estimated posterior density of median survival time of the patient from separate analysis; (b) estimated posterior density of median survival time of the patient from joint analysis.*

(d) Estimation of the random effects in `NLMIXED` is via empirical Bayes, with associated standard errors obtained by the delta method. Approximate 95% prediction intervals can then be obtained by assuming asymptotic normality. We obtained point and interval estimates in rough agreement with the above `WinBUGS` results, and for broadly comparable computer runtimes (if anything, our `NLMIXED` code ran slower). However, the asymmetry of some of the posteriors in Figure B.1 (recall they are truncated at 0) suggests traditional confidence intervals based on asymptotic normality and approximate standard errors will not be very accurate. Only the fully Bayesian-MCMC (`WinBUGS`) approach can produce exact results and corresponding full posterior inference.

Chapter 6

1 The calculations for the full conditionals for $\boldsymbol{\beta}$ and $\mathbf{W}$ follow from the results of the general linear model given in Example 4.2. Thus, with a $N(A\boldsymbol{\alpha}, V)$ prior on $\boldsymbol{\beta}$, (i.e., $p(\boldsymbol{\beta}) = N(A\boldsymbol{\alpha}, V)$) the full conditional for $\boldsymbol{\beta}$ is $N(D\mathbf{d}, D)$, where

$$\begin{aligned} D^{-1} &= \left(\frac{1}{\tau^2}X^T X + V^{-1}\right)^{-1} \\ \text{and } \mathbf{d} &= \frac{1}{\tau^2}X^T(\mathbf{Y} - \mathbf{W}) + V^{-1}A\boldsymbol{\alpha}. \end{aligned}$$

Note that with a flat prior on $\boldsymbol{\beta}$, we set $V^{-1} = 0$ to get

$$\boldsymbol{\beta}|\mathbf{Y}, \mathbf{W}, X, \tau^2 \sim N\left(\left(X^T X\right)^{-1} X^T(\mathbf{Y} - \mathbf{W}), \tau^2\left(X^T X\right)^{-1}\right).$$

Similarly for $\mathbf{W}$, since $p(\mathbf{W}) = N\left(\mathbf{0}, \sigma^2 H(\phi)\right)$, the full conditional distribution is again of the form $N(D\mathbf{d}, D)$, but where this time

$$\begin{aligned} D^{-1} &= \left(\frac{1}{\tau^2}I + \frac{1}{\sigma^2}H^{-1}(\phi)\right)^{-1} \\ \text{and } \mathbf{d} &= \frac{1}{\tau^2}(\mathbf{Y} - X\boldsymbol{\beta}). \end{aligned}$$

Next, with $p\left(\tau^2\right) = IG(a_\tau, b_\tau)$, we compute the full conditional distribution for τ^2, $p\left(\tau^2|\mathbf{Y}, X, \boldsymbol{\beta}, \mathbf{W}\right)$, as proportional to

$$\begin{aligned} &\tfrac{1}{(\tau^2)^{a_\tau+1}}\exp\left(-b_\tau/\tau^2\right) \\ &\qquad \times \tfrac{1}{(\tau^2)^{n/2}}\exp\left(-\tfrac{1}{2\tau^2}(\mathbf{Y} - X\boldsymbol{\beta} - \mathbf{W})^T(\mathbf{Y} - X\boldsymbol{\beta} - \mathbf{W})\right) \\ &\propto \tfrac{1}{(\tau^2)^{a_\tau+n/2}}\exp\left(-\tfrac{1}{\tau^2}\left(b_\tau + \tfrac{1}{2}(\mathbf{Y} - X\boldsymbol{\beta} - \mathbf{W})^T(\mathbf{Y} - X\boldsymbol{\beta} - \mathbf{W})\right)\right), \end{aligned}$$

where n is the number of sites. Thus we have the conjugate distribution

$$IG\left(a_\tau + \frac{n}{2}\,,\, b_\tau + \frac{1}{2}(\mathbf{Y} - X\boldsymbol{\beta} - \mathbf{W})^T(\mathbf{Y} - X\boldsymbol{\beta} - \mathbf{W})\right).$$

Similar calculations for the spatial variance parameter, σ^2, yield a conjugate full conditional when $p\left(\sigma^2\right) = IG(a_\sigma, b_\sigma)$, namely

$$\sigma^2 \mid \mathbf{W}, \phi \sim IG\left(a_\sigma + \frac{n}{2}\,,\, b_\sigma + \frac{1}{2}\mathbf{W}^T H^{-1}(\phi)\,\mathbf{W}\right).$$

Finally, for the spatial correlation function parameter ϕ, no closed form solution is available, and one must resort to Metropolis-Hastings or slice sampling for updating. Here we would need to compute

$$p\left(\phi|\mathbf{W}, \sigma^2\right) \propto p(\phi) \times \exp\left(-\frac{1}{2\sigma^2}\mathbf{W}^T H^{-1}(\phi)\,\mathbf{W}\right).$$

Typically the prior $p(\phi)$ is taken to be uniform or gamma.

5(a) These relationships follow directly from the definition of $w(\mathbf{s})$ in equation (3.9):

$$\begin{aligned} &Cov\left(w(\mathbf{s}), w(\mathbf{s}')\right) \\ &\quad = Cov\left(\int_{\Re^2} k(\mathbf{s} - \mathbf{t})\, z(\mathbf{t})\, d\mathbf{t}\,,\, \int_{\Re^2} k(\mathbf{s}' - \mathbf{t})\, z(\mathbf{t})\, d\mathbf{t}\right) \\ &\quad = \sigma^2 \int_{\Re^2} k(\mathbf{s} - \mathbf{t})\, k(\mathbf{s}' - \mathbf{t})\, d\mathbf{t}\,. \end{aligned}$$

and

$$var\left(w\left(\mathbf{s}\right)\right) = \sigma^2 \int_{R^2} k^2\left(\mathbf{s}-\mathbf{t}\right) d\mathbf{t},$$

obtained by setting $\mathbf{s} = \mathbf{s}'$ above.

(b) This follows exactly as above, except that we adjust for the covariance in the stationary $z(\mathbf{t})$ process:

$$\begin{aligned} Cov\left(w\left(\mathbf{s}\right), w\left(\mathbf{s}'\right)\right) &= \int_{\Re^2}\int_{\Re^2} k\left(\mathbf{s}-\mathbf{t}\right) k\left(\mathbf{s}'-\mathbf{t}\right) \\ &\qquad \times Cov\left(z\left(\mathbf{t}\right), z\left(\mathbf{t}'\right)\right) d\mathbf{t}d\mathbf{t}' \\ &= \sigma^2 \int_{\Re^2}\int_{\Re^2} k\left(\mathbf{s}-\mathbf{t}\right) k\left(\mathbf{s}'-\mathbf{t}\right) \rho\left(\mathbf{t}-\mathbf{t}'\right) d\mathbf{t}d\mathbf{t}' \\ \text{and } var\left(w\left(\mathbf{s}\right)\right) &= \sigma^2 \int_{\Re^2}\int_{\Re^2} k\left(\mathbf{s}-\mathbf{t}\right) k\left(\mathbf{s}-\mathbf{t}'\right) \rho\left(\mathbf{t}-\mathbf{t}'\right) d\mathbf{t}d\mathbf{t}', \end{aligned}$$

obtained by setting $\mathbf{s} = \mathbf{s}'$ above.

10 From (6.25), the full conditional $p(\phi_i|\boldsymbol{\phi}_{j\neq i}, \boldsymbol{\theta}, \boldsymbol{\beta}, \mathbf{y})$ is proportional to the product of a Poisson and a normal density. On the log scale we have

$$\log p(\phi_i|\boldsymbol{\phi}_{j\neq i}, \boldsymbol{\theta}, \boldsymbol{\beta}, \mathbf{y}) \propto -E_i e^{\mathbf{x}_i'\boldsymbol{\beta}+\theta_i+\phi_i} + \phi_i y_i - \frac{\tau_c m_i}{2}(\phi_i - \bar{\phi}_i)^2 \ .$$

Taking two derivatives of this expression, it is easy to show that in fact $(\partial^2/\partial\phi_i^2)\log p(\phi_i|\boldsymbol{\phi}_{j\neq i}, \boldsymbol{\theta}, \boldsymbol{\beta}, \mathbf{y}) < 0$, meaning that the log of the full conditional is a concave function, as required for ARS sampling.

Chapter 7

6(a) Denoting the likelihood by L, the prior by p, and writing $\mathbf{y} = (y_1, y_2)$, the joint posterior distribution of m_1 and m_2 is given as

$$\begin{aligned} p(m_1, m_2|\mathbf{y}) &\propto L(m_1, m_2; \mathbf{y})p(m_1, m_2) \\ &\propto (7m_1 + 5m_2)^{y_1} e^{-(7m_1+5m_2)} \\ &\quad \times (6m_1 + 2m_2)^{y_2} e^{-(6m_1+2m_2)} \\ &\quad \times m_1^{a-1} e^{-m_1/b} m_2^{a-1} e^{-m_2/b} \ , \end{aligned}$$

so that the resulting full conditional distributions for m_1 and m_2 are

$$\begin{aligned} p(m_1|m_2, \mathbf{y}) &\propto (7m_1 + 5m_2)^{y_1}(6m_1 + 2m_2)^{y_2} m_1^{a-1} e^{-m_1(13+b^{-1})}; \\ p(m_2|m_1, \mathbf{y}) &\propto (7m_1 + 5m_2)^{y_1}(6m_1 + 2m_2)^{y_2} m_2^{a-1} e^{-m_2(7+b^{-1})} \ . \end{aligned}$$

We see immediately that conjugacy is absent; these two expressions are not proportional to any standard distributional form. As such, one might think of univariate Metropolis updating to obtain samples from the joint posterior distribution $p(m_1, m_2|\mathbf{y})$, though since this is a very low-dimensional problem, the use of MCMC methods here probably constitutes overkill!

Drawing our Metropolis candidates from Gaussian distributions with means equal to the current chain value and variances $(0.3)^2$ and $(0.1)^2$ for δ_1 and δ_2, respectively, for each parameter we ran five independent sampling chains with starting points overdispersed with respect to the suspected target distribution for 2000 iterations. The observed Metropolis acceptance rates were 45.4% and 46.4%, respectively, near the 50%

Figure B.2 *Posterior histograms of sampled* $\mathbf{m}$ *values, motivating example.*

rate suggested by Gelman et al. (1996) as well as years of Metropolis "folklore." The vagueness of the prior distributions coupled with the paucity of the data in this simple example (in which we are estimating two parameters from just two data points, y_1 and y_2) leads to substantial autocorrelation in the observed chains. However, plots of the observed chains as well as the convergence diagnostic of Gelman and Rubin (1992) suggested that a suitable degree of algorithm convergence obtains after 500 iterations. The histograms of the remaining $5 \times 1500 = 7500$ iterations shown in Figure B.2(a) and (b) provide estimates of the marginal posterior distributions $p(m_1|\mathbf{y})$ and $p(m_2|\mathbf{y})$. We see that point estimates for m_1 and m_2 are 18.5 and 100.4, respectively, implying best guesses for $7m_1+5m_2$ and $6m_1+2m_2$ of 631.5 and 311.8, respectively, quite consistent with the observed data values $y_1 = 632$ and $y_2 = 311$. Also shown are 95% Bayesian credible intervals (denoted "95% BCI" in the figure legends), available simply as the 2.5 and 97.5 empirical percentiles in the ordered samples.

(b) By the Law of Iterated Expectation, $E(Y_{3a}|\mathbf{y}) = E[E(Y_{3a}|\mathbf{m}, \mathbf{y})]$. Now we need the following well-known result from distribution theory:

Lemma: If $X_1 \sim Po(\lambda_1)$, $X_2 \sim Po(\lambda_2)$, and X_1 and X_2 are independent, then

$$X_1 \mid (X_1 + X_2 = n) \;\sim\; Bin\left(n\,,\; \frac{\lambda_1}{\lambda_1 + \lambda_2}\right) . \quad \blacksquare$$

We apply this lemma in our setting with Y_{3a} playing the role of X_1, y_1 playing the role of n, and the calculation conditional on $\mathbf{m}$. The result is

$$\begin{aligned} E(Y_{3a}|\mathbf{y}) &= E[E(Y_{3a}|\mathbf{m}, \mathbf{y})] \;=\; E[E(Y_{3a}|m_1, y_1)] \\ &= E\left[y_1 \left(\frac{2m_1 + 2m_2}{7m_1 + 5m_2}\right) \Big| y_1\right] \\ &\approx \frac{y_1}{G} \sum_{g=1}^{G} \frac{2m_1^{(g)} + 2m_2^{(g)}}{7m_1^{(g)} + 5m_2^{(g)}} \;\equiv\; \hat{E}(Y_{3a}|\mathbf{y}) \,, \end{aligned} \tag{B.3}$$

where $\{(m_1^{(g)}, m_2^{(g)}),\, g = 1, \ldots, G\}$ are the Metropolis samples drawn above. A similar calculation produces a Monte Carlo estimate of $E(Y_{3b}|\mathbf{y})$, so that our final estimate of $E(Y_3|\mathbf{y})$ is the sum of these two quantities. In our problem this turns out to be $\hat{E}(Y_3|\mathbf{y}) = 357.0$.

Figure B.3 *Posterior histogram and kernel density estimate, sampled Y_3 values, motivating example.*

(c) Again using Monte Carlo integration, we write

$$p(y_3|\mathbf{y}) \quad = \quad \int p(y_3|\mathbf{m},\mathbf{y})p(\mathbf{m}|\mathbf{y})d\mathbf{m} \;\approx\; \frac{1}{G}\sum_{g=1}^{G} p(y_3|\mathbf{m}^{(g)},\mathbf{y}) \; .$$

Using the lemma again, $p(y_3|\mathbf{m},\mathbf{y})$ is the convolution of two independent binomials,

$$Y_{3a}|\mathbf{m},\mathbf{y} \sim Bin\left(y_1\,,\; \frac{2m_1+2m_2}{7m_1+5m_2}\right) \; , \tag{B.4}$$

$$\text{and} \quad Y_{3b}|\mathbf{m},\mathbf{y} \sim Bin\left(y_2\,,\; \frac{m_1+m_2}{6m_1+2m_2}\right) \; . \tag{B.5}$$

Since these two binomials do not have equal success probabilities, this convolution is a complicated (though straightforward) calculation that unfortunately will not emerge as another binomial distribution. However, we may perform the sampling analog of this calculation simply by drawing $Y_{3a}^{(g)}$ from $p(y_{3a}|\mathbf{m}^{(g)},y_1)$ in (B.4), $Y_{3b}^{(g)}$ from $p(y_{3b}|\mathbf{m}^{(g)},y_2)$ in (B.5), and defining $Y_3^{(g)} = Y_{3a}^{(g)} + Y_{3b}^{(g)}$. The resulting pairs $\{(Y_3^{(g)},\mathbf{m}^{(g)}),\; g=1,\ldots,G\}$ are distributed according to the joint posterior distribution $p(y_3,\mathbf{m}|\mathbf{y})$, so that marginally, the $\{Y_3^{(g)},\; g=1,\ldots,G\}$ values have the desired distribution, $p(y_3|\mathbf{y})$.

In our setting, we actually drew 25 $Y_{3a}^{(g)}$ and $Y_{3b}^{(g)}$ samples for each $\mathbf{m}^{(g)}$ value, resulting in 25(7500) = 187,500 $Y_3^{(g)}$ draws from the convolution distribution. A histogram of these values (and a corresponding kernel density estimate) is shown in Figure B.3. The mean of these samples is 357.2, which agrees quite well with our earlier mean estimate of 357.0 calculated just below equation (B.3).

Chapter 9

4(a) This setup closely follows that below equation (2.14), so we imitate this argument in the case of a bivariate process, where now $\mathbf{Y}_1 = Y_1(\mathbf{s}_0)$ and $\mathbf{Y}_2 = \mathbf{y}$. Then, as in equation (2.15),

$$E\left[Y_1(\mathbf{s}_0)|\mathbf{y}\right] = \mathbf{x}^T(\mathbf{s}_0)\boldsymbol{\beta} + \boldsymbol{\gamma}^T\Sigma^{-1}(\mathbf{y}-X\boldsymbol{\beta}) \; ,$$

parameter	2.5%	50%	97.5%
θ_1	-0.437	-0.326	-0.216
β_1	3.851	5.394	6.406
β_2	-2.169	2.641	7.518
σ_1	0.449	0.593	2.553
σ_2	0.101	1.530	6.545
ϕ_1	0.167	0.651	0.980
ϕ_2	0.008	0.087	0.276
τ	4.135	5.640	7.176

Table B.2 *Posterior quantiles for the conditional LMC model.*

where $\boldsymbol{\gamma}^T = \left(\boldsymbol{\gamma}_1^T, \boldsymbol{\gamma}_2^T\right)$, where $\boldsymbol{\gamma}_1^T = \left(c_{11}\left(\mathbf{s}_0 - \mathbf{s}_1\right), \ldots, c_{11}\left(\mathbf{s}_0 - \mathbf{s}_n\right)\right)$ and $\boldsymbol{\gamma}_2^T = \left(c_{12}\left(\mathbf{s}_0 - \mathbf{s}_1\right), \ldots, c_{12}\left(\mathbf{s}_0 - \mathbf{s}_n\right)\right)$. Also,

$$\Sigma_{2n\times 2n} = \begin{pmatrix} C_{11} & C_{12} \\ C_{12}^T & C_{22} \end{pmatrix} + \begin{pmatrix} \tau_1^2 I_n & 0 \\ 0 & \tau_2^2 I_n \end{pmatrix},$$

with $C_{lm} = \left(c_{lm}\left(\mathbf{s}_i - \mathbf{s}_j\right)\right)_{i,j=1,\ldots,n}$ with $l, m = 1, 2$.

(b) The approach in this part is analogous to that of Chapter 2, Exercise 9 Observe that with $\mathbf{s}_0 = \mathbf{s}_k$, $\left(\mathbf{e}_k^T : \mathbf{0}\right)\Sigma = \boldsymbol{\gamma}^T$ if and only if $\tau_1^2 = 0$, where $\mathbf{e}_k^T = (0, \ldots, 1, \ldots, 0)$ is the n-dimensional kth coordinate vector. This immediately leads to $E\left[Y_1\left(\mathbf{s}_k\right)|\mathbf{y}\right] = y_1\left(\mathbf{s}_k\right)$; $E\left[Y_2\left(\mathbf{s}_k\right)|\mathbf{y}\right] = y_2\left(\mathbf{s}_k\right)$ is shown analogously.

8(a) Let $Y_1(\mathbf{s})$ be the temperature at location $\mathbf{s}$, $Y_2(\mathbf{s})$ be the precipitation at location $\mathbf{s}$, and $X(\mathbf{s})$ be the elevation at location $\mathbf{s}$. We then fit the following conditional LMC, as in equation (9.64):

$$\begin{aligned} Y_1\left(\mathbf{s}\right) &= \theta_1 X\left(\mathbf{s}\right) + \sigma_1 w_1\left(\mathbf{s}\right) \\ Y_2\left(\mathbf{s}\right)|Y_1\left(\mathbf{s}\right) &= \beta_1 X\left(\mathbf{s}\right) + \beta_2 Y_1\left(\mathbf{s}\right) + \sigma_2 w_2\left(\mathbf{s}\right) + \epsilon\left(\mathbf{s}\right), \end{aligned}$$

where $\epsilon(\mathbf{s}) \sim N(0, \tau^2)$, $w_i(\mathbf{s}) \sim GP\left(0, \rho\left(\cdot, \phi_i\right)\right)$, for $i = 1, 2$.

The file `www.biostat.umn.edu/~brad/data/ColoradoLMCa.bug` on the web contains the `WinBUGS` code for this problem. Table B.2 gives a brief summary of the results. The results are more or less as expected: temperature is negatively associated with elevation, while precipitation is positively associated. Temperature and precipitation do not seem to be significantly associated with each other. The spatial smoothing parameters ϕ_1 and ϕ_2 were both assigned $U(0, 1)$ priors for this analysis, but it would likely be worth investigating alternate choices in order to check prior robustness.

(b) These results can be obtained simply by switching Y_1 and Y_2 in the data labels for the model and computer code of part (a).

Chapter 10

1(a) This follows directly by noting $var\left(\mathbf{v}_1\right) = \lambda_{11} I_n$, $var\left(\mathbf{v}_2\right) = \lambda_{22} I_n$, and $cov\left(\mathbf{v}_1, \mathbf{v}_2\right) = \lambda_{12} I$.

(b) Note that $var\left(\boldsymbol{\phi}_1\right) = \lambda_{11} A_1 A_1^T$, $var\left(\boldsymbol{\phi}_2\right) = \lambda_{22} A_2 A_2^T$, and also that $cov\left(\boldsymbol{\phi}_1, \boldsymbol{\phi}_2\right) = \lambda_{12} A_1 A_2^T$. So with $A_1 = A_2$, the dispersion of $\boldsymbol{\phi}$ is given by

$$\Sigma_{(\boldsymbol{\phi})} = \begin{pmatrix} \lambda_{11} A A^T & \lambda_{12} A A^T \\ \lambda_{12} A^T A & \lambda_{12} A A^T \end{pmatrix} = \Lambda \otimes A A^T.$$

Taking A as the square root of $\left(D_W - \rho W\right)^{-1}$ yields $\Sigma_{(\boldsymbol{\phi})} = \Lambda \otimes \left(D_W - \rho W\right)^{-1}$.

Figure B.4 *Fitted median lung cancer death rates per 1000 population, nonwhite females.*

Note that the order of the Kronecker product is different from equation (10.4), since we have blocked the $\boldsymbol{\phi}$ vector by components rather than by areal units.

(c) In general, with $A_1 \neq A_2$, we have

$$\Sigma(\boldsymbol{\phi}) = \begin{pmatrix} \lambda_{11} A_1 A_1^T & \lambda_{12} A_1 A_2^T \\ \lambda_{12} A_2 A_1^T & \lambda_{12} A_2 A_2^T \end{pmatrix} = \mathcal{A}\,(\Lambda \otimes I)\,\mathcal{A}^T,$$

where $\mathcal{A} = BlockDiag\,(A_1, A_2)$. For the generalized MCAR, with different spatial smoothness parameters ρ_1 and ρ_2 for the different components, take A_i as the Cholesky square root of $(D_W - \rho_i W)^{-1}$ for $i = 1, 2$.

Chapter 11

2 The code in `www.biostat.umn.edu/~brad/data/ColoradoS-T1.bug` fits model (11.6), the additive space-time model. This is a "direct" solution, where we explicitly construct the temporal process. By contrast, the file `www.biostat.umn.edu/~brad/data/ColoradoS-T2.bug` uses the `spatial.exp` function, tricking it to handle temporal correlations by setting the y-coordinates to 0.

4(a) Running five chains of an MCMC algorithm, we obtained point and 95% interval estimates of –0.01 and [–0.20, 0.18] for β; using the same reparametrization under the chosen model (10) in Waller et al. (1997), the point and interval estimates instead are –0.20 and [–0.26, –0.15]. Thus, using this reparametrization shows that age adjusting has eliminated the statistical significance of the difference between the two female groups.

(b) Figure B.4 shows the fitted age-adjusted lung cancer death rates per 1000 population for nonwhite females for the years 1968, 1978, and 1988. The scales of the three figures show that lung cancer death rates are increasing over time. For 1968, we see a strong spatial pattern of increasing rates as we move from northwest to southeast, perhaps the result of an unmeasured occupational covariate (farming versus mining). Except for persistent low rates in the northwest corner, however, this trend largely disappears over time, perhaps due to increased mixing of the population or improved access to quality health care and health education.

node (unit)	Mean	sd	MC error	2.5%	Median	97.5%
W_1 (A)	–0.0491	0.835	0.0210	–1.775	–0.0460	1.639
W_3 (C)	–0.183	0.9173	0.0178	–2.2	–0.136	1.52
W_5 (E)	–0.0320	0.8107	0.0319	–1.682	–0.0265	1.572
W_6 (F)	0.417	0.8277	0.0407	–1.066	0.359	2.227
W_9 (I)	0.255	0.7969	0.0369	–1.241	0.216	1.968
W_{11} (K)	–0.195	0.9093	0.0209	–2.139	–0.164	1.502
ρ_1 (A)	1.086	0.1922	0.0072	0.7044	1.083	1.474
ρ_3 (C)	0.901	0.2487	0.0063	0.4663	0.882	1.431
ρ_5 (E)	1.14	0.1887	0.0096	0.7904	1.139	1.521
ρ_6 (F)	0.935	0.1597	0.0084	0.6321	0.931	1.265
ρ_9 (I)	0.979	0.1683	0.0087	0.6652	0.971	1.339
ρ_{11} (K)	0.881	0.2392	0.0103	0.4558	0.861	1.394
τ	1.73	1.181	0.0372	0.3042	1.468	4.819
β_0	–7.11	0.689	0.0447	–8.552	–7.073	–5.874
β_1	0.596	0.2964	0.0105	0.0610	0.578	1.245
RR	3.98	2.951	0.1122	1.13	3.179	12.05

Table B.3 *Posterior summaries, MAC survival model (10,000 samples, after a burn-in of 1,000).*

Chapter 14

1(a) Table B.3 summarizes the results from the nonspatial model, which are based on 10,000 posterior samples obtained from a single MCMC chain after a burn-in of 1,000 iterations. Looking at this table and the raw data in Table 14.14, basic conclusions are as follows:

- Units A and E have moderate overall risk ($W_i \approx 0$) but increasing hazards ($\rho > 1$): few deaths, but they occur late.
- Units F and I have high overall risk ($W_i > 0$) but decreasing hazards ($\rho < 1$): several early deaths, many long-term survivors.
- Units C and K have low overall risk ($W_i < 0$) and decreasing hazards ($\rho < 1$): no deaths at all; a few survivors.
- The two drugs differ significantly: CI for β_1 (RR) excludes 0 (1).

2(b) The appropriate interval censored `WinBUGS` code is as follows:

```
model
{
for (i in 1:N) {
  TimeSmoking[i] <- Age[i] - AgeStart[i]
  RelapseT[i] ~ dweib(rho[i],mu[i])I(censored.time1[i],
    censored.time2[i])
  log(mu[i]) <- beta0 + beta[1]*TimeSmoking[i]
    + beta[2]*SexF[i] + beta[3]*SIUC[i]
    + beta[4]*F10Cigs[i] + W[County[i]]
  rho[i] <- exp(lrho[County[i]])
  }

# for (i in 1:regions) {W[i] ~ dnorm(0.0, tau_W)}
# for (i in 1:regions) {lrho[i] ~ dnorm(0.0, tau_rho)}
```

```
for (i in 1:sumnum) {weights[i] <- 1}

W[1:regions] ~ car.normal(adj[], weights[], num[], tau_W)
lrho[1:regions] ~ car.normal(adj[], weights[], num[],
  tau_rho)

for (i in 1:4) { beta[i] ~ dnorm(0.0, 0.0001)}
beta0 ~ dnorm(0.0,0.0001)
tau_W ~ dgamma(0.1,0.1)
tau_rho ~ dgamma(0.1,0.1)
}
```

Bibliography

[1] Abrahamsen, N. (1993). Bayesian kriging for seismic depth conversion of a multilayer reservoir. In *Geostatistics Troia, '92*, ed. A. Soares, Boston: Kluwer Academic Publishers, pp. 385–398.

[2] Abramowitz, M. and Stegun, I.A. (1965). *Handbook of Mathematical Functions.* New York: Dover.

[3] Adler, R. J. (1981), *The Geometry of Random Fields*, Chichester, UK: Wiley.

[4] Agarwal, D.K. and Gelfand, A.E. (2005). Slice Gibbs sampling for simulation based fitting of spatial data models. *Statistics and Computing*, **15**, 61–69.

[5] Agarwal, D.K., Gelfand, A.E., and Silander, J.A. (2002). Investigating tropical deforestation using two-stage spatially misaligned regression models. *J. Agric. Biol. Environ. Statist.*, **7**, 420–439.

[6] Agarwal, D.K., Gelfand, A.E., Sirmans, C.F. and Thibadeau, T.G. (2005). Nonstationary spatial house price models. To appear *J. Statist. Plann. Inf.*

[7] Agresti, A. (2002). *Categorical Data Analysis*, 2nd ed. New York: Wiley.

[8] Aitken, M., Anderson, D., Francis, B., and Hinde, J. (1989). *Statistical Modelling in GLIM.* Oxford: Oxford Statistical Science.

[9] Akima, H. (1978). A method of bivariate interpolation and smooth surface fitting for irregularly distributed data points. *ACM Transactions on Mathematical Software*, **4**, 148-164.

[10] Akima, H. (1996), Algorithm 761: Scattered-data surface fitting that has the accuracy of a cubic polynomial. *ACM Transactions on Mathematical Software*, **22**, 362–371.

[11] Albert, J. (2009). *Bayesian Computation with R.* New York: Springer.

[12] Andersen, P.K. and Gill, R.D. (1982). Cox's regression model for counting processes: A large sample study. *Ann. Statist.*, **10**, 1100–1120.

[13] Andersen, P.K., Borgan, O., Gill, R.D. and Keiding, N. (1995). *Statistical Models Based on Counting Processes.* New York: Springer.

[14] Anderson, E., Bai, Z., Bischof, C., Blackford, S., Demmel, J., Dongarra, J., Du Croz, J., Greenbaum, A., Hammarling, S., McKenney, A., Sorensen, D. (1999). *LAPACK Users' Guide*, 3rd edition. Philadelphia, PA: Society for Industrial and Applied Mathematics.

[15] Anselin, L. (1988). *Spatial Econometrics: Models and Methods.* Dordrecht: Kluwer Academic Publishers.

[16] Anton, H. (1984). *Calculus with Analytic Geometry*, 2nd ed. New York: Wiley.

[17] Apanasovich, T.V. and Genton, M.G. (2010). Cross-covariance functions for multivariate random fields based on latent dimensions. *Biometrika*, **97**, 15–30.

[18] Apanasovich, T.V., Genton, M.G. and Sun, Y. (2012). A valid Matern class of cross-covariance functions for multivariate random fields with any number of components. *Journal of the American Statistical Association*, **107**, 180–193.

[19] Armstrong, M. and Diamond, P. (1984). Testing variograms for positive definiteness. *Mathematical Geology*, **24**, 135–147.

[20] Armstrong, M. and Jabin, R. (1981). Variogram models must be positive definite. *Mathematical Geology*, **13**, 455–459.

[21] Arnold, B.C. and Strauss, D.J. (1991). Bivariate distributions with conditionals in prescribed exponential families. *J. Roy. Statist. Soc., Ser. B*, **53**, 365–375.

[22] Arnold, B.C., Castillo, E. and Sarabia, J.M. (1999). *Conditional Specification of Statistical Models*. New York: Springer-Verlag.

[23] Assunção, R.M. (2003). Space-varying coefficient models for small area data. *Environmetrics*, **14**, 453–473.

[24] Assunção, R.M., Potter, J.E., and Cavenaghi, S.M. (2002). A Bayesian space varying parameter model applied to estimating fertility schedules. *Statistics in Medicine*, **21**, 2057–2075.

[25] Assunção, R.M., Reis, I.A., and Oliveira, C.D.L. (2001). Diffusion and prediction of Leishmaniasis in a large metropolitan area in Brazil with a Bayesian space-time model. *Statistics in Medicine*, **20**, 2319–2335.

[26] Baddeley, A. and Møller J. (1989). Nearest-neighbour Markov point processes and random sets. *International Statistical Review* **57**, 89-121.

[27] Baddeley, A.J., Møller, J., and Waagepetersen, R. (2000). Non- and semi-parametric estimation of interaction in inhomogeneous point patterns. *Statistica Neerlandica*, **54**, 329–350.

[28] Baddeley, A., Gregori P., Mateu, J., Stoica, R. and Stoyan, D. (Eds) (2005). Case Studies in Spatial Point Process Modeling. *Lecture Notes in Statistics*. Springer-Verlag.

[29] Baddeley, A. and Turner, R. (2005). `spatstat`: An `R` package for analyzing spatial point patterns. *Journal of Statistical Software*, **12**, 6.

[30] Baddeley, A., Turner, R., Moller, J. and Hazelton, M. (2005). Residual analysis for spatial point processes (with discussion). *Journal of the Royal Statistical Society, Series B*, **67**, 617–666.

[31] Bailey, T.C. and Gatrell, A.C. (1995). *Interactive Spatial Data Analysis*. Essex: Addison Wesley Longman.

[32] Bailey, M.J., Muth, R.F., and Nourse, H.O. (1963). A regression method for real estate price index construction. *J. Amer. Statist. Assoc.*, **58**, 933–942.

[33] Baillo, A. and Grane, A. (2009). Local linear regression for functional predictor and scalar response. *Journal of Multivariate Analysis*, **100**, 102–111.

[34] Baladandayuthapani, V., Mallick, B., Hong, M., Lupton, J., Turner, N. and Carroll, R. (2008). Bayesian hierarchical spatially correlated functional data analysis with application to colon carcinoginesis. *Biometrics*, **64**, 64–73.

[35] Banerjee, S. (2000). On multivariate spatial modelling in a Bayesian setting. Unpublished Ph.D. dissertation, Department of Statistics, University of Connecticut.

[36] Banerjee, S. (2005). On geodetic distance computations in spatial modelling. To appear *Biometrics*.

[37] Banerjee, S. and Carlin, B.P. (2002). Spatial semiparametric proportional hazards models for analyzing infant mortality rates in Minnesota counties. In *Case Studies in Bayesian Statistics, Volume VI*, eds. C. Gatsonis et al. New York: Springer-Verlag, pp. 137–151.

[38] Banerjee, S. and Carlin, B.P. (2003). Semiparametric spatio-temporal frailty modeling. *Environmetrics*, **14**, 523–535.

[39] Banerjee, S., Finley, A.O., Waldmann, P. and Ericcson, T. (2010). Hierarchical spatial process models for multiple traits in large genetic trials. *Journal of the American Statistical Association*, **105**, 506–521.

[40] Banerjee, S., Gamerman, D., and Gelfand, A.E. (2003). Spatial process modelling for univariate and multivariate dynamic spatial data. Technical report, Division of Biostatistics, University of Minnesota.

[41] Banerjee, S. and Gelfand, A.E. (2002). Prediction, interpolation and regression for spatially misaligned data. *Sankhya, Ser. A*, **64**, 227–245.

[42] Banerjee, S. and Gelfand, A.E. (2003). On smoothness properties of spatial processes. *J. Mult. Anal.*, **84**, 85–100.

[43] Banerjee, S., and Gelfand, A.E. (2006). Bayesian wombling: curvilinear gradient assessment under spatial process models. *Journal of American Statistical Association*, **101**, 1487–1501.

[44] Banerjee, S., Gelfand, A.E., Finley, A.O., and Sang, H. (2008). Gaussian predictive process models for large spatial datasets. *J. Roy. Statist. Soc., Ser. B*, **70**, 825–848.

[45] Banerjee, S., Gelfand, A.E., Knight, J., and Sirmans, C.F. (2004). Spatial modelling of house prices using normalized distance-weighted sums of stationary processes. *J. Bus. Econ. Statist.*, **22**, 206–213.

[46] Banerjee, S., Gelfand, A.E., and Polasek, W. (2000). Geostatistical modelling for spatial interaction data with application to postal service performance. *J. Statist. Plann. Inf.*, **90**, 87–105.

[47] Banerjee, S., Gelfand, A.E. and Sirmans, C.F. (2004). Directional rates of change under spatial process models. *J. Amer. Statist. Assoc.*, **98**, 946–954.

[48] Banerjee and Johnson (2006). Coregionalized single- and multi-resolution spatially-varying growth curve modelling with applications to weed growth. *Biometrics* **61**, 617–625.

[49] Banerjee, S., Wall, M.M., and Carlin, B.P. (2003). Frailty modeling for spatially correlated survival data, with application to infant mortality in Minnesota. *Biostatistics*, **4**, 123–142.

[50] Barber, J., Gelfand, A. and Silander A. (2006). Modeling map positional error to infer true feature location. *Canadian Journal of Statistics*, **34**, **4**, 659-676.

[51] Barbujani, G., Jacquez, G.M. and Ligi, L. (1990). Diversity of some gene frequencies in European and Asian populations V. steep multilocus clines. *American Journal of Human Genetics*, **47**, 867-875.

[52] Barbujani, G., Oden, N.L., and Sokal, R.R. (1989). Detecting areas of abrupt change in maps of biological variables. *Systematic Zoology*, **38**, 376–389.

[53] Barry, R.P. and Ver Hoef, J.M. (1996). Blackbox kriging: Spatial prediction without specifying variogram models. *J. Agric. Biol. Environ. Statist.*, **1**, 297–322.

[54] Bayes, T. (1763). An essay towards solving a problem in the doctrine of chances. *Philos. Trans. Roy. Soc. London*, **53**, 370–418. Reprinted, with an introduction by George Barnard, in 1958 in *Biometrika*, **45**, 293–315.

[55] Becker, R.A. and Wilks, A.R. (1993). Maps in S. Technical report, AT&T Bell Laboratories; website `www.research.att.com/areas/stat/doc/93.2.ps`.

[56] Beneš, V., Bodlák, K., Møller, J., and Waagepetersen, R. (2003). Application of log Gaussian Cox processes in disease mapping. Proceedings of the ISI Conference on Environmental Statistics and Health, Santiago de Compostela, 2003. Eds. J. Mateu, D. Holland and W. Gonzalez-Manetiga, 95-105.

[57] Benjamin, G. and Sirmans, G.S. (1991). Determinants of apartment rent. *Journal of Real Estate Research*, **6**, 357–379.

[58] Berger, J.O. (1985). *Statistical Decision Theory and Bayesian Analysis*, 2nd ed. New York: Springer-Verlag.

[59] Berger, J.O. and Pericchi, L.R. (1996). The intrinsic Bayes factor for linear models. In *Bayesian Statistics 5*, eds. J.M. Bernardo, J.O. Berger, A.P. Dawid, and A.F.M. Smith. Oxford: Oxford University Press, pp. 25–44.

[60] Berger, J.O., De Oliveira, V., and Sansó, B. (2001). Objective Bayesian analysis of spatially correlated data. *Journal of the American Statistical Association*, **96**, 1361–1374.

[61] Bergthorsson, P. and Döös, B. (1955). Numerical weather map analysis. *Tellus*, **7**, 329-340.

[62] Berkson, J. and Gage, R.P. (1952). Survival curve for cancer patients following treatment. *J. Amer. Statist. Assoc.*, **47**, 501–515.

[63] Berliner, L.M. (2000). Hierarchical Bayesian modeling in the environmental sciences. *Allgemeines Statistisches Archiv (Journal of the German Statistical Society)*, **84**, 141–153.

[64] Bernardinelli, L., Clayton, D., and Montomoli, C. (1995). Bayesian estimates of disease maps: how important are priors? *Statistics in Medicine*, **14**, 2411–2431.

[65] Bernardinelli, L. and Montomoli, C. (1992). Empirical Bayes versus fully Bayesian analysis of geographical variation in disease risk. *Statistics in Medicine*, **11**, 983–1007.

[66] Bernardinelli, L., Pascutto, C., Best, N.G. and Gilks, W.R. (1997). Disease mapping with errors in covariates. *Statistics in Medicine*, **16**, 741–752.

[67] Bernardo, J.M. and Smith, A.F.M. (1994). *Bayesian Theory.* New York: Wiley.

[68] Berrocal, V. J., Gelfand, A. E., and Holland, D. M. (2010a). A spatio-temporal downscaler for outputs from numerical models. *Journal of Agricultural, Biological and Environmental Statistics*, **15**, 176–197.

[69] Berrocal, V.J., Gelfand, A.E., and D.M. Holland, D.M. (2010b) A bivariate downscaler under space and time misalignment. *Annals of Applied Statistics*, **4**, 1942–1975.

[70] Berrocal, V.J., Gelfand, A.E., and Holland, D.M. (2012). Space-time data fusion under error in computer model output: an application to modeling air quality. *Biometrics*, **68**, 837–848.

[71] Berthelsen, K.K. and Mller, J. (2003). Likelihood and non-parametric Bayesian MCMC inference for spatial point processes based on perfect simulation and path sampling. *Scandinavian Journal of Statistics* **30**, 549–564.

[72] Berthelsen, K.K and Møller, J. (2004) A Bayesian MCMC method for point process models with intractable normalising constants. In *Spatial point process modelling and its applications* (Eds. A. Baddeley, P. Gregori, J. Mateu, R. Stoica and D. Stoyan), Publicacions de la Universitat Jaume I, pp. 7–15.

[73] Berthelsen, K.K. and Møller, J. (2006). Bayesian analysis of Markov point processes. In *Case Studies in Spatial Point Process Modeling* (Eds. A. Baddeley, P. Gregori, J. Mateu, R. Stoica and D. Stoyan), Springer Lecture Notes in Statistics 185, Springer-Verlag, New York, pp. 85–97.

[74] Berthelsen, K.K. and Møller, J. (2008). Non-parametric Bayesian inference for inhomogeneous Markov point processes. *Australian and New Zealand Journal of Statistics*, **50**, 627–649.

[75] Besag, J. (1974). Spatial interaction and the statistical analysis of lattice systems (with discussion). *J. Roy. Statist. Soc., Ser. B*, **36**, 192–236.

[76] Besag, J. (1981). On a system of two-dimensional recurrence equations. *J. R. Statist. Soc. B*, **43**, 302–309.

[77] Besag, J. and Green, P.J. (1993). Spatial statistics and Bayesian computation (with discussion). *J. Roy. Statist. Soc., Ser. B*, **55**, 25–37.

[78] Besag, J., Green, P., Higdon, D., and Mengersen, K. (1995). Bayesian computation and stochastic systems (with discussion). *Statistical Science*, **10**, 3–66.

[79] Besag, J. and Kooperberg, C. (1995). On conditional and intrinsic autoregressions. *Biometrika*, **82**, 733–746.

[80] Besag, J., York, J.C., and Mollié, A. (1991). Bayesian image restoration, with two applications in spatial statistics (with discussion). *Annals of the Institute of Statistical Mathematics*, **43**, 1–59.

[81] Best, N.G., Ickstadt, K., and Wolpert, R.L. (2000). Spatial Poisson regression for health and exposure data measured at disparate resolutions. *J. Amer. Statist. Assoc.*, **95**, 1076–1088.

[82] Best, N.G., Waller, L.A., Thomas, A., Conlon, E.M. and Arnold, R.A. (1999). Bayesian models for spatially correlated diseases and exposure data. In *Bayesian Statistics 6*, eds. J.M. Bernardo et al. Oxford: Oxford University Press, pp. 131–156.

[83] Billheimer, D. and Guttorp, P. (1996). Spatial models for discrete compositional data. Technical Report, Department of Statistics, University of Washington.

[84] Billingsley, P. (1995). *Probability and Measure* (3rd edition). New York: Wiley-Interscience.

[85] Blackford, S.L., Demmel, J., Dongarra, J., Duff, I., Hammarling, S., Henry, G., Heroux, M., Kaufman, L., Lumsdaine, A., Petitet, A., Pozo, R., Remington, K., Whaley, C.R. (2002). An updated set of Basic Linear Algebra Subprograms (BLAS). *Transactions on Mathematical Software*, **28**, 135–151.

[86] Bocquet-Appel, J.P. and Bacro, J.N. (1994), Generalized wombling. *Systematic Biology*, **43**, 442–448.

[87] Box, G.E.P. and Tiao, G.C. (1992). *Bayesian Inference in Statistical Analysis*. Wiley Classics Library edition: Wiley-Interscience.

[88] Bratseth, A.M. (1986). Statistical interpolation by means of successive corrections. *Tellus*, **38A**, 439–447.

[89] Breslow, N.E. and Clayton, D.G. (1993). Approximate inference in generalized linear mixed models. *J. Amer. Statist. Assoc.*, **88**, 9–25.

[90] Breslow, N.E. and Day, N.E. (1987). *Statistical Methods in Cancer Research, Volume II – The Design and Analysis of Cohort Studies*. Lyon: International Agency for Research on Cancer.

[91] Brook, D. (1964). On the distinction between the conditional probability and the joint probability approaches in the specification of nearest-neighbour systems. *Biometrika*, **51**, 481–483.

[92] Brooks, S.P. and Gelman, A. (1998). General methods for monitoring convergence of iterative simulations. *J. Comp. Graph. Statist.*, **7**, 434–455.

[93] Brooks, S.P., Gelman, A., Jones, G.L. and Meng, X.L. (2011). *Handbook of Markov Chain Monte Carlo*. Boca Raton: Chapman and Hall/CRC.

[94] Brown, P.E., Kåresen, K.F., Roberts, G.O., and Tonellato, S. (2000). Blur-generated nonseparable space-time models. *J. Roy. Statist. Soc., Ser. B*, **62**, 847–860.

[95] Brown, P., Le, N. and Zidek, J. (1994). Multivariate spatial interpolation and exposure to air pollutants. *The Canadian Journal of Statistics*, **22**, 489–509.

[96] Brush, S.G. (1967). History of the Lenz-Ising Model. *Reviews of Modern Physics (American Physical Society)*, **39**, 883–893.

[97] Buishand, T.A., de Haan, L. and Zhou, C. (2008). On spatial extremes: With application to a rainfall problem. *Annals of Applied Statistics*, **2**, 624–642.

[98] California Office of Statewide Health Planning and Development (1997). *Hospital Patient Discharge Data (Public Use Version)*. Sacramento, CA: State of California.

[99] Cancer Surveillance and Control Program (1997). Case completeness and data quality audit: Minnesota Cancer Surveillance System 1994–1995. Technical report, Minnesota Department of Health.

[100] Carlin, B.P. and Banerjee, S. (2003). Hierarchical multivariate CAR models for spatio-temporally correlated survival data (with discussion). In *Bayesian Statistics 7*, eds. J.M. Bernardo, M.J. Bayarri, J.O. Berger, A.P. Dawid, D. Heckerman, A.F.M. Smith, and M. West. Oxford: Oxford University Press, pp. 45–63.

[101] Carlin, B.P., Chaloner, K., Church, T., Louis, T.A., and Matts, J.P. (1993). Bayesian approaches for monitoring clinical trials with an application to toxoplasmic encephalitis prophylaxis. *The Statistician*, **42**, 355–367.

[102] Carlin, B.P. and Hodges, J.S. (1999). Hierarchical proportional hazards regression models for highly stratified data. *Biometrics*, **55**, 1162–1170.

[103] Carlin, B.P. and Louis, T.A. (2008). *Bayesian Methods for Data Analysis*, 3rd ed. Boca Raton, FL: Chapman and Hall/CRC Press.

[104] Carlin, B.P. and Pérez, M.-E. (2000). Robust Bayesian analysis in medical and epidemiological settings. In *Robust Bayesian Analysis (Lecture Notes in Statistics, Vol. 152)*, eds. D.R. Insua and F. Ruggeri. New York: Springer-Verlag, pp. 351–372.

[105] Carlin, B.P., Xia, H., Devine, O., Tolbert, P., and Mulholland, J. (1999). Spatio-temporal hierarchical models for analyzing Atlanta pediatric asthma ER visit rates. In *Case Studies in Bayesian Statistics, Volume IV*, eds. C. Gatsonis et al. New York: Springer-Verlag, pp. 303–320.

[106] Carlin, B.P., Zhu, L., and Gelfand, A.E. (2001). Accommodating scale misalignment in spatio-temporal data. In *Bayesian Methods with Applications to Science, Policy and Official Statistics*, eds. E.I. George et al. Luxembourg: Office for Official Publications of the European Communities (Eurostat), pp. 41–50.

[107] Case, K.E. and Shiller R.J. (1989). The efficiency of the market for single family homes. *American Economic Review*, **79**, 125–137.

[108] Casella, G. and George, E. (1992). Explaining the Gibbs sampler. *The American Statistician*, **46**, 167–174.

[109] Casella, G., Lavine, M., and Robert, C.P. (2001). Explaining the perfect sampler. *The American Statistician*, **55**, 299–305.

[110] Celeux, G., Hurn, M. and Robert, C.P. (2000). Computational and inferential difficulties with mixture posterior distributions. *Journal of the American Statistical Association*, **95**, 957–970.

[111] Centers for Disease Control and Prevention (1987). Underreporting of alcohol-related mortality on death certificates of young U.S. Army veterans. In *Morbidity and Mortality Weekly Report*, U.S. Department of Health and Human Services, Vol. 36, No. 27 (July 1987), pp. 437–440.

[112] Centers for Disease Control and Prevention (1996). Mortality trends for Alzheimer's disease, 1979–1991. In *Vital and Health Statistics*, U.S. Department of Health and Human Services, Series 20, No. 28 (January 1996), p. 3.

[113] Chakraborty, A. and Gelfand, A.E. (2010). Measurement error in spatial point patterns. *Bayesian Analysis*, **5**, 97–122.

[114] Chakraborty, A., Gelfand, A.E., Silander, J. A., Jr., Wilson, A., and Latimer, A. (2011). Point pattern modeling for degraded presence-only data over large regions. *Journal of the Royal Statistical Society, Series C (Applied Statistics)*, **60**, 757–776.

[115] Chambers, J.A., Cleveland, W.S., Kleiner, B., and Tukey, P.A. (1983). *Graphical Methods for Data Analysis.* Belmont, CA: Wadsworth.

[116] Cherry, S., Banfield, J., and Quimby, W.F. (1996). An evaluation of a non-parametric method of estimating semi-variograms of isotropic spatial processes. *Journal of Applied Statistics*, **23**, 435–449.

[117] Chellappa, R. and Jain, A.K., (1993), *Markov Random Fields.* Academic Press.

[118] Chen, M.-H., Ibrahim, J.G., and Sinha, D. (1999). A new Bayesian model for survival data with a surviving fraction. *J. Amer. Statist. Assoc.*, **94**, 909–919.

[119] Chen, M.-H., Shao, Q.-M., and Ibrahim, J.G. (2000). *Monte Carlo Methods in Bayesian Computation.* New York: Springer-Verlag.

[120] Chib, S. and Greenberg, E. (1998). Analysis of multivariate probit models. *Biometrika*, **85**, 347–361.

[121] Chiles, J.P. and Delfiner, P. (1999). *Geostatistics: Modeling Spatial Uncertainty.* New York: Wiley.

[122] Christakos, G. (1984). On the problem of permissible covariance and variogram models. *Water Resources Research*, **20**, 251–265.

[123] Christakos, G. (1992). *Random Field Models in Earth Sciences.* New York: Academic Press.

[124] Clark, I., Basinger, K. L., and Harper, W. V. (1989). MUCK: A novel approach to co-kriging. In *Proceedings of the Conference on Geostatistical, Sensitivity, and Uncertainty Methods for Ground-Water Flow and Radionuclide Transport Modeling*, 473–493, editor: Buxton, B.E. Battelle Press, Columbus, OH.

[125] Clayton, D. (1991). A Monte Carlo method for Bayesian inference in frailty models. *Biometrics*, **47**, 467–485.

[126] Clayton, D. (1994). Some approaches to the analysis of recurrent event data. *Statistics in Medical Research*, **3**, 244–262.

[127] Clayton, D.G. and Kaldor, J.M. (1987). Empirical Bayes estimates of age-standardized relative risks for use in disease mapping. *Biometrics*, **43**, 671-681.

[128] Cliff, A.D. and Ord, J.K. (1973). *Spatial Autocorrelation.* London: Pion.

[129] Clifford, P. (1990). Markov random fields in statistics. In *Disorder in Physical Systems.* Oxford: Oxford University Press, pp. 20–32.

[130] Csillag, F. and Kabos, S. (2002). Wavelets, boundaries and the analysis of landscape pattern. *Ecoscience*, **9**, 177–190.

[131] Coles, S. (2001). *S-plus functions for extreme value modeling: An accompaniment to the book: An introduction to statistical modeling of extreme values.*

`http://www.stats.bris.ac.uk/ masgc/ismev/uses.ps.`

[132] Congdon, P. (2001). *Bayesian Statistical Modelling.* Chichester: Wiley.

[133] Congdon, P. (2003). *Applied Bayesian Modelling.* Chichester: Wiley.

[134] Congdon, P. and Best, N.G. (2000). Small area variation in hospital admission rates: Adjusting for referral and provider variation. *J. Roy. Statist. Soc. Ser. C (Applied Statistics)*, **49**, 207–226.

[135] Cohn, D.L., Fisher, E., Peng, G., Hodges, J., Chesnutt, J., Child, C., Franchino, B., Gibert, C., El-Sadr, W., Hafner, R., Korvick, J., Ropka, M., Heifets, L., Clotfelter, J., Munroe, D., and Horsburgh, R. (1999). A prospective randomized trial of four three-drug regimens in the treatment of disseminated *Mycobacterium avium* complex disease in AIDS patients: Excess mortality associated with high-dose clarithromycin. *Clinical Infectious Diseases*, **29**, 125–133.

[136] Cole, T.J. and Green, P.J. (1992). Smoothing reference centiles; The LMS method and penalized likelihood. *Statistics in Medicine*, **11**, 1305–1319.

[137] Conlon, E.M. and Waller, L.A. (1999). Flexible spatial hierarchical models for mapping disease rates. *Proceedings of the Statistics and the Environment Section of the American Statistical Association*, pp. 82–87.

[138] Coles S.G. and Tawn J.A. (1996). Modeling extremes of the areal rainfall processes. *Journal of the Royal Statistical Society, Series B*, **58**, 329–347.

[139] Coles S.G. (2001). *An introduction to statistical modeling of extreme values.* New York: Springer-Verlag.

[140] Cooley D., Nychka D. and Naveau P. (2007). Bayesian spatial modeling of extreme precipitation return levels. *Journal of the Americal Statistical Association*, **102**, 824–840.

[141] Cooley, J.W. and Tukey, J.W. (1965). An algorithm for the machine computation of complex Fourier series. *Mathematics of Computation*, **19**, 297–301.

[142] Corsten, L.C.A. (1989). Interpolation and optimal linear prediction. *Statistica Neerlandica*, **43**, 69–84.

[143] Cowles, M.K. (2002). MCMC sampler convergence rates for hierarchical normal linear models: A simulation approach. *Statistics and Computing*, **12**, 377–389.

[144] Cowles, M.K. (2003). Efficient model-fitting and model-comparison for high-dimensional Bayesian geostatistical models. *Journal of Statistical Planning and Inference*, **112**, 221–239.

[145] Cowles, M.K. and Carlin, B.P. (1996). Markov chain Monte Carlo convergence diagnostics: A comparative review. *J. Amer. Statist. Assoc.*, **91**, 883–904.

[146] Cox, D.R. and Oakes, D. (1984). *Analysis of Survival Data.* London: Chapman and Hall.

[147] Cramér, H. (1940). On the theory of stationary random processes. *Annals of Mathematics*, **41**, 215–230.

[148] Cramér, H. and Leadbetter M.R. (1967). *Stationary and Related Stochastic Processes.* New York: Wiley.

[149] Cressman, G.P. (1959). An operational objective analysis system. *Monthly Weather Review*, **87**, 367–374.

[150] Cressie, N.A.C. (1993). *Statistics for Spatial Data*, 2nd ed. New York: Wiley.

[151] Cressie, N.A.C. (1996). Change of support and the modifiable areal unit problem. *Geographical Systems*, **3**, 159–180.

[152] Cressie, N. and Huang, H.-C. (1999). Classes of nonseparable spatio-temporal stationary covariance functions. *J. Amer. Statist. Assoc.*, **94**, 1330–1340.

[153] Cressie, N., and Johannesson, G. (2008). Fixed rank kriging for very large spatial data sets. *Journal of the Royal Statistical Society: Series B*, **70**, 209–226.

[154] Cressie, N.A.C. and Wikle, C.K. (1998). The variance-based cross-variogram: You can add apples and oranges. *Mathematical Geology*, **30**, 789–799.

[155] Cressie, N. and Wikle, C. (2011). *Statistics for Spatio-Temporal Data*. New York: Wiley.

[156] Daley, D.J., and Vere-Jones, D. (2003). *An Introduction to the Theory of Point Processes*, 2nd edition. New York: Springer

[157] Daley, D.J and Vere-Jones, D. (2008). *An Introduction to the Theory of Point Processes. Volume II: General Theory and Structure*. New York: Springer.

[158] Damian, D., Sampson, P.D., and Guttorp, P. (2001). Bayesian estimation of semi-parametric non-stationary spatial covariance structures. *Environmetrics*, **12**, 161–178.

[159] Damien, P., Wakefield, J., and Walker, S. (1999). Gibbs sampling for Bayesian non-conjugate and hierarchical models by using auxiliary variables. *J. Roy. Statist. Soc., Ser. B*, **61**, 331–344.

[160] Daniels, M.J. and Kass, R.E. (1999). Nonconjugate Bayesian estimation of covariance matrices and its use in hierarchical models. *J. Amer. Statist. Assoc.*, **94**, 1254–1263.

[161] De Iaco, S., Myers, D.E. and Posa, D. (2002). Nonseparable space-time covariance models: some parametric families *Mathematical Geology* **34**, 23–42.

[162] DeGroot, M.H. (1970). *Optimal Statistical Decisions*. New York: McGraw-Hill.

[163] Delicado, P., Giraldo, R., Comas, C. and Mateu, J. (2010), Statistics for spatial functional data: some recent contributions. *Environmetrics*, **21**, 224–239.

[164] Dempster, A.M. (1972). Covariance selection. *Biometrics*, **28**, 157–175.

[165] Dempster, A.P., Laird, N.M., and Rubin, D.B. (1977). Maximum likelihood estimation from incomplete data via the EM algorithm (with discussion). *J. Roy. Statist. Soc., Ser. B*, **39**, 1–38.

[166] DeOliveira, V. (2000). Bayesian prediction of clipped Gaussian random fields. *Computational Statistics and Data Analysis*, **34**, 299–314.

[167] DeOliveira, V., Kedem, B., and Short, D.A. (1997). Bayesian prediction of transformed Gaussian random fields. *J. Amer. Statist. Assoc.*, **92**, 1422–1433.

[168] Devesa, S.S., Grauman, D.J., Blot, W.J., Pennello, G.A., Hoover, R.N., and Fraumeni, J.F., Jr. (1999). *Atlas of Cancer Mortality in the United States, 1950–94*. NIH Publ. No. 99-4564, Bethesda, MD: National Institutes of Health; website `www-dceg.ims.nci.nih.gov/atlas/index.html`.

[169] Devine, O.J., Qualters, J.R., Morrissey, J.L., and Wall, P.A. (1998). Estimation of the impact of the former Feed Materials Production Center (FMPC) on lung cancer mortality in the surrounding community. Technical report, Radiation Studies Branch, Division of Environmental Hazards and Health Effects, National Center for Environmental Health, Centers for Disease Control and Prevention.

[170] Diggle, P.J., (1993). Point process modeling in environmental epidemiology. In: *Statistics for the Environment*, Eds. Barnett, V. and Feridun Turkman, K., John Wiley and Sons Ltd.: Chichester, pp. 89-110.

[171] Diggle, P.J. (2003). *Statistical Analysis of Spatial Point Patterns*, 2nd ed. London: Arnold.

[172] Diggle, P.J., Gmez-Rubio, V., Brown, P.E., Chetwynd, A.G. and Gooding, S. (2007), Second-order analysis of inhomogeneous spatial point processes using casecontrol data. *Biometrics*, **63**, 550–557.

[173] Diggle, P.J., Menezes, R. and Su, T.L. (2010). Geostatistical inference under preferential sampling. *Journal of the Royal Statistical Society*, Series C, **59**, 191–232.

[174] Diggle, P.J. and Ribeiro, P.J. (2002). Bayesian inference in Gaussian model-based geostatistics. *Geographical and Environmental Modelling*, **6**, 129–146.

[175] Diggle, P.J. and Rowlingson, B.S. (1994). A conditional approach to point process modelling of elevated risk. *Journal of the Royal Statistical Society, Ser. A*, **157**, 433–440.

[176] Diggle, P.J., Tawn, J.A., and Moyeed, R.A. (1998). Model-based geostatistics (with discussion). *J. Roy. Statist. Soc., Ser. C (Applied Statistics)*, **47**, 299–350.

[177] Doksum, K.A. and Gasko, M. (1990). On a correspondence between models in binary regression analysis and in survival analysis. *International Statistical Review*, **58**, 243–252.

[178] Duan, J. and Gelfand, A.E. (2003). Finite mixture model of nonstationary spatial data. Technical report, Institute for Statistics and Decision Sciences, Duke University.

[179] Duan, J., Gelfand, A.E., and Sirmans, C.F. (2009). Space-time point process models using differential equations with application to urban development. *Bayesian Analysis*, **4**, 733–758.

[180] Eberly, L.E. and Carlin, B.P. (2000). Identifiability and convergence issues for Markov chain Monte Carlo fitting of spatial models. *Statistics in Medicine*, **19**, 2279–2294.

[181] Ecker, M.D. and Gelfand, A.E. (1997). Bayesian variogram modeling for an isotropic spatial process. *J. Agric. Biol. Environ. Statist.*, **2**, 347–369.

[182] Ecker, M.D. and Gelfand, A.E. (1999). Bayesian modeling and inference for geometrically anisotropic spatial data. *Mathematical Geology*, **31**, 67–83.

[183] Ecker, M.D. and Gelfand, A.E. (2003). Spatial modeling and prediction under stationary non-geometric range anisotropy. *Environmental and Ecological Statistics*, **10**, 165–178.

[184] Ecker, M.D. and Heltshe, J.F. (1994). Geostatistical estimates of scallop abundance. In *Case Studies in Biometry*, eds. N. Lange, L. Ryan, L. Billard, D. Brillinger, L. Conquest, and J. Greenhouse. New York: Wiley, pp. 107–124.

[185] Efron, B. (2010). *Large-Scale Inference: Empirical Bayes Methods for Estimation, Testing and Prediction*. Cambridge: Cambridge University Press.

[186] Eidsvik, J., Finley, A.O., Banerjee, S. and Rue, H. (2012) Approximate Bayesian inference for large spatial datasets using predictive process models. *Computational Statistics and Data Analysis*, **56**, 1362–1380.

[187] Elliott, P., Wakefield, J.C., Best, N.G., and Briggs, D.J., eds. (2000). *Spatial Epidemiology: Methods and Applications*. Oxford: Oxford University Press.

[188] Ewell, M. and Ibrahim, J.G. (1997). The large sample distribution of the weighted log-rank statistic under general local alternatives. *Lifetime Data Analysis*, **3**, 5–12.

[189] Fagan, W.F., Fortin, M.J. and Soykan, C. (2003). Integrating edge detection and dynamic modeling in quantitative analyses of ecological boundaries. *BioScience*, **53**, 730–738.

[190] Fahrmeir, L., Kneib, T. and Lang, S. (2007). *Regression: Models, Methods and Applications*, Springer Verlag.

[191] Farewell, V.T. (1982). The use of mixture models for the analysis of survival data with long term survivors. *Biometrics*, **38**, 1041–1046.

[192] Farewell, V.T. (1986). Mixture models in survival analysis: Are they worth the risk? *Canadian Journal of Statistics*, **14**, 257–262.

[193] Fick, A. (1855). On liquid diffusion. *Philosophical Magazine and Journal of Science*, **10**, 30–39.

[194] Finkelstein, D.M. (1986). A proportional hazards model for interval-censored failure time data. *Biometrics*, **42**, 845–854.

[195] Finley, A.O., Banerjee, S. and Carlin, B.P. (2007). **spBayes**: an R package for univariate and multivariate hierarchical point-referenced spatial models. *Journal of Statistical Software*, **19**, 4.

[196] Finley, A.O., Banerjee, S. and Gelfand, A.E. (2012). Bayesian dynamic modeling for large space-time datasets using Gaussian predictive processes. *Journal of Geographical Systems*, **14**, 29–47.

[197] Finley, A.O., Banerjee, S., Ek, A.R. and McRoberts, R. (2008). Bayesian multivariate process modeling for predicting forest attributes. *Journal of Agricultural, Biological and Environmental Statistics*, **13**, 1–24.

[198] Finley, A.O., Banerjee, S. and McRoberts, R.E. (2009a). Hierarchical spatial models for predicting tree species assemblages across large domains. *Annals of Applied Statistics*, **3**, 1052–1079.

[199] Finley, A.O., Sang, H., Banerjee, S., and Gelfand, A.E. (2009b). Improving the performance of predictive process modeling for large datasets. *Computational Statistics and Data Analysis*, **53**, 2873–2884.

[200] Fitzmaurice, G. M., Laird, N. M., and Ware, J. W. (2004). *Applied longitudinal analysis.* Hoboken, NJ: Wiley.

[201] Fleming, T.R. and Harrington, D.P. (2005). *Counting Processes and Survival Analysis.* New York: John Wiley and Sons.

[202] Flowerdew, R. and Green, M. (1989). Statistical methods for inference between incompatible zonal systems. In *Accuracy of Spatial Databases*, eds. M. Goodchild and S. Gopal. London: Taylor and Francis, pp. 239–247.

[203] Flowerdew, R. and Green, M. (1992). Developments in areal interpolating methods and GIS. *Annals of Regional Science*, **26**, 67–78.

[204] Flowerdew, R. and Green, M. (1994). Areal interpolation and yypes of data. In *Spatial Analysis and GIS*, S. Fotheringham and P. Rogerson, eds., London: Taylor and Francis, pp. 121–145.

[205] Fortin, M.J. (1994). Edge detection algorithms for two-dimensional ecological data. *Ecology*, **75**, 956–965.

[206] Fortin, M.J. (1997). Effects of data types on vegetation boundary delineation. *Canadian Journal of Forest Research*, **27**, 1851–1858.

[207] Fortin, M.J. and Drapeau, P. (1995). Delineation of ecological boundaries: comparisons of approaches and significance tests. *Oikos*, **72**, 323–332.

[208] Fotheringham, A.S. and Rogerson, P., eds. (1994). *Spatial Analysis and GIS.* London: Taylor and Francis.

[209] Frankel, T. (2003). *The Geometry of Physics: An Introduction.* Cambridge, UK: Cambridge University Press.

[210] Fuentes, M. (2001). A high frequency kriging approach for non-stationary environmental processes. *Environmetrics*, **12**, 469–483.

[211] Fuentes, M. (2002a). Spectral methods for nonstationary spatial processes. *Biometrika*, **89**, 197–210.

[212] Fuentes, M. (2002b). Modeling and prediction of non-stationary spatial processes. *Statistical Modeling*, **2**, 281–298.

[213] Fuentes, M. and Raftery, A.E. (2005), Model evaluation and spatial interpolation by Bayesian combination of observations with outputs from numerical models. *Biometrics*, **61**, 36–45.

[214] Fuentes, M. and Smith, R.L. (2001) Modeling nonstationary processes as a convolution of local stationary processes. Technical report, Department of Statistics, North Carolina State University.

[215] Fuentes, M. and Smith, R.L. (2003). A new class of models for nonstationary processes. Technical report, Department of Statistics, North Carolina State University.

[216] Gabriel, E., Allard, D. and Bacro, J.N. (2011). Estimating and testing zones of abrupt changes for spatial data. *Statistics and Computing*, **21**, 107–120.

[217] Gamerman, D. (1997). *Markov Chain Monte Carlo: Stochastic Simulation for Bayesian Inference*. Boca Raton, FL: Chapman and Hall/CRC Press.

[218] Gamerman, D., Moreira, A.R.B., and Rue, H. (2005). Space-varying regression models: Specifications and simulation. To appear *Computational Statistics and Data Analysis*.

[219] Gaspari, G. and Cohn, S. E. (1999). Construction of correlation functions in two and three dimensions. *Quarterly Journal of the Royal Metereological Society*, **125** 723–757.

[220] Gatsonis, C., Hodges, J.S., Kass, R.E., and Singpurwalla, N.D., eds. (1993). *Case Studies in Bayesian Statistics*. New York: Springer-Verlag.

[221] Gatsonis, C., Hodges, J.S., Kass, R.E., and Singpurwalla, N.D., eds. (1995). *Case Studies in Bayesian Statistics, Volume II*. New York: Springer-Verlag.

[222] Gatsonis, C., Hodges, J.S., Kass, R.E., McCulloch, R.E., Rossi, P., and Singpurwalla, N.D., eds. (1997). *Case Studies in Bayesian Statistics, Volume III*. New York: Springer-Verlag.

[223] Gatsonis, C., Kass, R.E., Carlin, B.P., Carriquiry, A.L., Gelman, A., Verdinelli, I., and West, M., eds. (2002). *Case Studies in Bayesian Statistics, Volume V*. New York: Springer-Verlag.

[224] Gatsonis, C., Kass, R.E., Carriquiry, A.L., Gelman, A., Higdon, D., Pauler, D., and Verdinelli, I., eds. (2003). *Case Studies in Bayesian Statistics, Volume VI*. New York: Springer-Verlag.

[225] Gelfand, A.E. (2010). Misaligned spatial data: The change of support problem. In *Handbook of Spatial Statistics*, eds. A.E. Gelfand, P. Diggle, P. Guttorp and M. Fuentes. Boca Raton, FL: CRC Press.

[226] Gelfand, A.E., Diggle, P., Guttorp, P. and Fuentes, M. (2010). *Handbook of Spatial Statistics*. Boca Raton, FL: CRC Press.

[227] Gelfand, A.E., Ecker, M.D., Knight, J.R., and Sirmans, C.F. (2004). The dynamics of location in home price. *Journal of Real Estate Finance and Economics*, **29**, 149–166..

[228] Gelfand, A.E. and Ghosh, S.K. (1998). Model choice: a minimum posterior predictive loss approach. *Biometrika*, **85**, 1–11.

[229] Gelfand, A.E., Kim, H.-J., Sirmans, C.F., and Banerjee, S. (2003). Spatial modeling with spatially varying coefficient processes. *J. Amer. Statist. Assoc.*, **98**, 387–396.

[230] Gelfand, A.E., Kottas, A. and Mac Eachern, S.N. (2005). Bayesian Nonparametric Spatial Modeling with Dirichlet Process Mixing. *Journal of the American Statistical Association*, **100**, 1021–1035

[231] Gelfand, A.E., Kottas, A., and MacEachern, S.N. (2003). Nonparametric Bayesian spatial modeling using dependent Dirichlet processes. Technical report, Institute for Statistics and Decision Sciences, Duke University.

[232] Gelfand A.E. and Mallick, B.K. (1995). Bayesian analysis of proportional hazards models built from monotone functions. *Biometrics*, **51**, 843–852.

[233] Gelfand, A.E., Sahu, S.K., and Carlin, B.P. (1995). Efficient parametrizations for normal linear mixed models. *Biometrika*, **82**, 479–488.

[234] Gelfand, A.E., Sahu, S.K., and Carlin, B.P. (1996). Efficient parametrizations for generalized linear mixed models (with discussion). In *Bayesian Statistics 5*, eds. J.M. Bernardo, J.O. Berger, A.P. Dawid, and A.F.M. Smith. Oxford: Oxford University Press, pp. 165–180.

[235] Gelfand, A.E., Sahu, S.K., and Holland, D.M. (2012). On the effect of preferential sampling in spatial prediction. *Environmetrics*, **23**, 565–578.

[236] Gelfand, A.E., Schmidt, A.M., Banerjee, S., and Sirmans, C.F. (2004). Nonstationary multivariate process modeling through spatially varying coregionalization (with discussion). *Test*, **13**, 263–312.

[237] Gelfand, A.E., Schmidt, A.M., Wu, S., Silander, J.A., Latimer, A., and Rebelo, A.G. (2005). Explaining Species Diversity Through Species Level Hierarchical Modeling. *Applied Statistics* **54**, 1-20.

[238] Gelfand, A.E. and Smith, A.F.M. (1990). Sampling-based approaches to calculating marginal densities. *J. Amer. Statist. Assoc.*, **85**, 398–409.

[239] Gelfand, A.E. and Vounatsou, P. (2003). Proper multivariate conditional autoregressive models for spatial data analysis. *Biostatistics*, **4**, 11–25.

[240] Gelfand, A.E., Zhu, L., and Carlin, B.P. (2001). On the change of support problem for spatio-temporal data. *Biostatistics*, **2**, 31–45.

[241] Gelfand, A.E., Banerjee, S., and Finley, A. (2012). Spatial design for knot selection in knot-based dimension reduction models. In *Spatio-temporal design: Advances in efficient data acquisition*, eds: J. M. Mateu and W. Mueller. J.Wiley and Sons, pp. 142–169.

[242] Gelman, A., Carlin, J.B., Stern, H.S., Dunson, D.B., Vehtari, A. and Rubin, D.B. (2013). *Bayesian Data Analysis*, 3rd ed. Boca Raton, FL: Chapman and Hall/CRC Press.

[243] Gelman, A., Roberts, G.O., and Gilks, W.R. (1996). Efficient Metropolis jumping rules. In *Bayesian Statistics 5*, eds. J.M. Bernardo, J.O. Berger, A.P. Dawid, and A.F.M. Smith. Oxford: Oxford University Press, pp. 599–607.

[244] Gelman, A. and Rubin, D.B. (1992). Inference from iterative simulation using multiple sequences (with discussion). *Statistical Science*, **7**, 457–511.

[245] Geltner, D. and Miller, N.G. (2001). *Commercial Real Estate Analysis and Investments*. South-Western, Cincinnati.

[246] Geman, S. and Geman, D. (1984). Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, **6**, 721–741.

[247] Geman, S. and McClure, D.E. (1985). Bayesian image analysis: An application to single photon emission tomography. *Proceedings of the Statistical Computing Section, American Statistical Association*. American Statistical Association, DC, pp. 12–18.

[248] Geman, S. and McClure, D.E. (1987). Statistical methods for tomographic image reconstruction. *Bulletin of the International Statistical Institute*, **52**, 5–21.

[249] Geman, D. and Reynolds, G. (1992). Constrained restoration and the recovery of discontinuities. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **14**, 367–383.

[250] Genovese, C. and Wasserman, L. (2002). Operating characteristics and extensions of the false discovery rate procedure. *Journal of the Royal Statistical Society B*, **64**, 499–518.

[251] Geyer, C.J. (1992). Practical Markov Chain Monte Carlo (with discussion). *Statistical Science*, **7**, 473–511.

[252] Ghosh, M. (1992). Constrained Bayes estimates with applications. *J. Amer. Statist. Assoc.*, **87**, 533–540.

[253] Ghosh, S., Gelfand, A.E., and Clark, J.S. (2012). Inference for size demography from point pattern data using Integral Projection Models (with discussion). *Journal of Agricultural, Biological and Environmental Statistics*, **17**, 641–699

[254] Gikhman, I.I. and Skorokhod, A.V. (1974). *Stochastic Differential Equations.* Berlin: Springer-Verlag.

[255] Gilks, W.R., Richardson, S., and Spiegelhalter, D.J., eds. (1996). *Markov Chain Monte Carlo in Practice.* London: Chapman and Hall.

[256] Gilks, W.R. and Wild, P. (1992). Adaptive rejection sampling for Gibbs sampling. *J. Roy. Statist. Soc., Ser. C (Applied Statistics)*, **41**, 337–348.

[257] Gneiting, T. (2002). Nonseparable, stationary covariance functions for space-time data. *J. Amer. Statist. Assoc.*, **97**, 590–600.

[258] Gneiting, T. and Raftery, A.E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, **102**, 359–378.

[259] Goldman, A.I. (1984). Survivorship analysis when cure is a possibility: A Monte Carlo study. *Statistics in Medicine* **3**, 153–163.

[260] Goldman, A.I., Carlin, B.P., Crane, L.R., Launer, C., Korvick, J.A., Deyton, L., and Abrams, D.I. (1996). Response of $CD4^+$ and clinical consequences to treatment using ddI or ddC in patients with advanced HIV infection. *J. Acquired Immune Deficiency Syndromes and Human Retrovirology*, **11**, 161–169.

[261] Goldstein, H. (1995). *Kendall's Library of Statistics 3: Multilevel Statistical Models*, 2nd ed. London: Arnold.

[262] Golub, G.H. and van Loan, C.F. (2012). *Matrix Computations*, 4th edition. Baltimore, MD: Johns Hopkins University Press.

[263] Gotway, C.A. and Young, L.J. (2002). Combining incompatible spatial data. *J. Amer. Statist. Assoc.*, **97**, 632–648.

[264] Goulard and Voltz (1992). Linear coregionalization model: Tools for estimation and choice of cross-variogram matrix. *Mathematical Geology*, **24**, 269–286.

[265] Green, P.J. and Richardson, S. (2002). Hidden Markov models and disease mapping. *J. Amer. Statist. Assoc.*, **97**, 1055–1070.

[266] Greenwood, J.A. (1984). A unified theory of surface roughness. *Proceedings of the Royal Society of London. Series A, Mathematical and Physical Sciences*, **393**, 133–157.

[267] Griffith, D.A. (1988). *Advanced Spatial Statistics.* Dordrecht, the Netherlands: Kluwer.

[268] Grzebyk, M. and Wackernagel, H. (1994). Multivariate analysis and spatial/temporal scales: real and complex models. In *Proceedings of the XVIIth International Biometrics Conference*, Hamilton, Ontario, Canada: International Biometric Society, pp. 19–33.

[269] Guan, Y. (2006). A composite likelihood approach in fitting spatial point process models. *Journal of the American Statistical Association*, **101**, 1502–1512.

[270] Guan, Y. and Loh, J.M. (2007). A thinned block bootstrap variance estimation procedure for inhomogeneous spatial point patterns. *Journal of the American Statistical Association*, **102**, 1377–1386.

[271] Guan, Y., Waagepetersen, R. and Beale, C. (2008). Second-order analysis of inhomogeneous spatial point processes with proportional intensity functions. *Journal of the American Statistical Association*, **103**, 769–777.

[272] Guggenheimer, H.W. (1977). *Differential Geometry.* New York: Dover Publications.

[273] Guhaniyogi, R., Finley, A.O., Banerjee, S. and Gelfand, A.E. (2011). Adaptive Gaussian predictive process models for large spatial datasets. *Environmetrics*, **22**, 997–1007.

[274] Guo, X. and Carlin, B.P. (2004). Separate and joint modeling of longitudinal and event time data using standard computer packages. *The American Statistician*, **58**, 16–24.

[275] Guttman, I. (1982). *Linear Models: An Introduction.* New York: Wiley.

[276] Guyon, X. (1995). *Random Fields on a Network: Modeling, Statistics, and Applications.* New York: Springer-Verlag.

[277] Haining, R. (1990). *Spatial Data Analysis in the Social and Environmental Sciences.* Cambridge: Cambridge University Press.

[278] Hall, P., Fisher, N.I., and Hoffmann, B. (1994). On the nonparametric estimation of covariance functions. *Ann. Statist.*, **22**, 2115–2134.

[279] Handcock, M.S. (1999). Comment on "Prediction of spatial cumulative distribution functions using subsampling." *J. Amer. Statist. Assoc.*, **94**, 100–102.

[280] Handcock, M.S. and Stein, M.L. (1993). A Bayesian analysis of kriging. *Technometrics*, **35**, 403–410.

[281] Handcock, M.S. and Wallis, J. (1994). An approach to statistical spatial-temporal modeling of meteorological fields (with discussion). *J. Amer. Statist. Assoc.*, **89**, 368–390.

[282] Haran, M. (2003). Efficient perfect and MCMC sampling methods for Bayesian spatial and components of variance models. Unpublished Ph.D. dissertation, School of Statistics and Division of Biostatistics, University of Minnesota.

[283] Haran, M., Hodges, J.S., and Carlin, B.P. (2003). Accelerating computation in Markov random field models for spatial data via structured MCMC. *Journal of Computational and Graphical Statistics*, **12**, 249–264.

[284] Harville, D.A. (1997). *Matrix Algebra from a Statistician's Perspective.* New York: Springer-Verlag.

[285] Hastings, W.K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, **57**, 97–109.

[286] Heaton, M.J. and Gelfand, A.E. (2011). Spatial prediction using kernel averaged predictors. *Journal of Agricultural, Biological, and Environmental Statistics*, **10**, 233-252.

[287] Heikkinen, J. and Högmander, H. (1994). Fully Bayesian approach to image restoration with an application in biogeography. *Applied Statistics*, **43**, 569–582.

[288] Helterbrand and Cressie (1994). Universal cokriging under intrinsic coregionalization. *Mathematical Geology*, **26**, 205–226.

[289] Helterbrand, J.D., Cressie, N., and Davidson, J.L. (1994). A statistical approach to identifying closed object boundaries in images. *Advances in Applied Probability*, **26**, 831–854.

[290] Henderson, R., Diggle, P.J., and Dobson, A. (2000). Joint modelling of longitudinal measurements and event time data. *Biostatistics*, **1**, 465–480.

[291] Heywood, P.F., Singleton, N., and Ross, J. (1988). Nutritional status of young children – The 1982/3 National Nutrition Survey. *Papua New Guinea Medical Journal*, **31**, 91–101.

[292] Higdon, D.M. (1998a). Auxiliary variable methods for Markov chain Monte Carlo with applications. *J. Amer. Statist. Assoc.*, **93**, 585–595.

[293] Higdon, D.M. (1998b). A process-convolution approach to modeling temperatures in

the north Atlantic Ocean. *Journal of Environmental and Ecological Statistics*, **5**, 173–190.

[294] Higdon, D.M. (2002). Space and space-time modeling using process convolutions. In *Quantitative Methods for Current Environmental Issues*, eds. C. Anderson, V. Barnett, P.C. Chatwin, and A.H. El-Shaarawi. London: Springer-Verlag, pp. 37–56.

[295] Higdon, D., Lee, H., and Holloman, C. (2003). Markov chain Monte Carlo-based approaches for inference in computationally intensive inverse problems (with discussion). In *Bayesian Statistics 7*, eds. J.M. Bernardo, M.J. Bayarri, J.O. Berger, A.P. Dawid, D. Heckerman, A.F.M. Smith, and M. West. Oxford: Oxford University Press, pp. 181–197.

[296] Higdon, D., Swall, J., and Kern, J. (1999). Non-stationary spatial modeling. In *Bayesian Statistics 6*, eds. J.M. Bernardo, J.O. Berger, A.P. Dawid, and A.F.M. Smith. Oxford: Oxford University Press, pp. 761–768.

[297] Hill, R.C., Knight, J.R., and Sirmans, C.F. (1997). Estimating capital asset price indexes. *The Review of Economics and Statistics*, **79**, 226–233.

[298] Hill, R.C., Sirmans, C.F., and Knight, J.R. (1999). A random walk down main street. *Regional Science and Urban Economics*, **29**, 89–103.

[299] Hjort, N. and Omre, H. (1994). Topics in spatial statistics (with discussion). *Scand. J. Statist.,* **21**, 289–357.

[300] Hoaglin, D.C., Mosteller, F., and Tukey, J.W. (1983). *Understanding Robust and Exploratory Data Analysis*. New York: Wiley.

[301] Hoaglin, D.C., Mosteller, F., and Tukey, J.W. (1985). *Exploring Data Tables, Trends, and Shapes*. New York: Wiley.

[302] Hobert, J.P., Jones, G.L., Presnell, B. and Rosenthal, J.S. (2002). On the applicability of regenerative simulation in Markov chain Monte Carlo. *Biometrika*, **89**, 731–743.

[303] Hodges, J.S. (1998). Some algebra and geometry for hierarchical models, applied to diagnostics (with discussion). *J. Roy. Statist. Soc., Series B*, **60**, 497–536.

[304] Hodges, J.S., Carlin, B.P., and Fan, Q. (2003). On the precision of the conditionally autoregressive prior in spatial models. *Biometrics*, **59**, 317–322.

[305] Hodges, J. S., Cui, Y., Sargent, D. J. and Carlin, B. P. (2007). Smoothing balanced single-error-term analysis of variance. *Technometrics*, **49**, 12–25.

[306] Hoel, P.G., Port, S.C., and Stone, C.J. (1972). *Introduction to Stochastic Processes*. Boston: Houghton Mifflin.

[307] Hogmander, H. and Møller, J. (1995). Estimating distribution maps from atlas data using methods of statistical image analysis. *Biometrics*, **51**, 393–404.

[308] Hoeting, J.A., Leecaster, M., and Bowden, D. (2000). An improved model for spatially correlated binary responses. *J. Agr. Biol. Env. Statist.*, **5**, 102–114.

[309] Hoff, P.D. (2009). *A First Course in Bayesian Statistical Methods*. New York: Springer.

[310] Hohn, M. (1988). *Geostatistics and Petroleum Geology*, New York: Van Nostrand Reinhold.

[311] Hooten, M. B. and C.K. Wikle, (2007) A hierarchical Bayesian non-linear spatio-temporal model for the spread of invasive species with application to the Eurasian Collared-Dove. *Environmental and Ecological Statistics*, **15**, 59–70.

[312] Horn, R.A and Johnson, C.R. (2012). *Matrix Analysis*, Second Edition. Cambridge, UK: Cambridge University Press.

[313] Hrafnkelsson, B. and Cressie, N. (2003). Hierarchical modeling of count data with application to nuclear fall-out. *Environmental and Ecological Statistics*, **10**, 179–200.

[314] Huang, H.-C. and Cressie, N.A.C. (1996). Spatio-temporal prediction of snow water equivalent using the Kalman filter. *Computational Statistics and Data Analysis*, **22**, 159–175.

[315] Huerta, G., and Sansó, B. (2007). Time-varying models for extreme values. *Environmental and Ecological Statistics*, **14**, 285–299.

[316] Huerta, G., Sanso, B., and Stroud, J.R. (2003). A spatiotemporal model for Mexico city ozone levels. *J. Roy. Statist. Soc., Ser. C (Applied Statistics)*, **53**, 231–248.

[317] Ingram, D.D. and Kleinman, J.C. (1989). Empirical comparisons of proportional hazards and logistic regression models. *Statistics in Medicine*, **8**, 525–538.

[318] Isaaks, E.H. and Srivastava, R.M. (1989). *An Introduction to Applied Geostatistics.* Oxford: Oxford University Press.

[319] Jacquez, G.M. and Greiling, D.A. (2003). Geographic boundaries in breast, lung and colorectal cancers in relation to exposure to air toxins in Long Island, New York. *International Journal of Health Geographics*, **2**, 4.

[320] Jin, X. and Carlin, B.P. (2003). Multivariate parametric spatio-temporal models for county level breast cancer survival data. *Lifetime Data Analysis*, **11**, 5–27.

[321] Jin, X., Carlin, B.P., and Banerjee, S. (2005). Generalized hierarchical multivariate CAR models for areal data. *Biometrics*, **61**, 950–961.

[322] Jin, X., Banerjee, S. and Carlin, B.P. (2007). Order-free coregionalized areal data models with application to multiple disease mapping. *Journal of the Royal Statistical Society, Series B*, **69**, 817–838.

[323] Jones, C.B. (1997). *Geographical Information Systems and Computer Cartography.* Harlow, Essex, UK: Addison Wesley Longman.

[324] Journel, A.G. and Froidevaux, R. (1982). Anisotropic hole-effect modelling. *Math. Geology*, **14**, 217–239.

[325] Journel, A.G. and Huijbregts, C.J. (1978). *Mining Geostatistics.* New York: Academic Press.

[326] Kaiser, M.S. and Cressie, N. (2000). The construction of multivariate distributions from Markov random fields. *J. Mult. Anal.*, **73**, 199–220.

[327] Kalnay, E. (2003). *Atmospheric Modeling, Data Assimilation and Predictability.* Cambridge, UK: Cambridge University Press.

[328] Kaluzny, S.P., Vega, S.C., Cardoso, T.P., and Shelly, A.A. (1998). *S+SpatialStats: User's Manual for Windows and UNIX.* New York: Springer-Verlag.

[329] Karson, M.J., Gaudard, M., Linder, E. and Sinha, D. (1999). Bayesian analysis and computations for spatial prediction (with discussion). *Environmental and Ecological Statistics*, **6**, 147–182.

[330] Kashyap, R. and Chellappa, R. (1983). Estimation and choice of neighbors in spatial interaction models of images. *IEEE Transactions on Information Theory*, **IT-29**, 60–72.

[331] Kass, R.E., Carlin, B.P., Gelman, A., and Neal, R. (1998). Markov chain Monte Carlo in practice: A roundtable discussion. *The American Statistician*, **52**, 93–100.

[332] Kass, R.E. and Raftery, A.E. (1995). Bayes factors. *J. Amer. Statist. Assoc.*, **90**, 773–795.

[333] Kass, R.E., Tierney, L. and Kadane, J.B. (1989). The validity of posterior expansions based on Laplace's method. In *Bayesian and Likelihood Methods in Statistics and Economics: Essays in Honor of George A. Barnard*, S. Geisser, J.S. Hodges, S.J. Press and A. Zellner, eds, North-Holland, Amsterdam, pp. 473–488.

[334] Kass, R.E., Tierney, L. and Kadane, J.B. (1991). Laplace's method in Bayesian analysis. In *Statistical Multiple Integration*, N. Fluornoy and R.K. Tsutukawa, eds, American Statistical Association, Providence, pp. 89-99.

[335] Katzfuss, M. (2013). Bayesian nonstationary spatial modeling for very large datasets. *Environmetrics* (forthcoming).

[336] Kaufman, C.G., Schervish, M.J. and Nychka, D.W.. (2008). Covariance tapering for likelihood-based estimation in large spatial data sets. *Journal of the American Statistical Association*, **103**, 1545–1555.

[337] Kennedy, M.C. and O'Hagan, A. (2001). Bayesian calibration of computer models. *Journal of the Royal Statistical Society Series B*, **63**, 425–464.

[338] Kent, J.T. (1989). Continuity properties for random fields. *Annals of Probability*, **17**, 1432–1440.

[339] Kharin, V.V. and Zwiers, F.W. (2005). Estimating extremes in transient climate change simulations. *Journal of Climate*, 1:1.

[340] Killough, G.G., Case, M.J., Meyer, K.R., Moore, R.E., Rope, S.K., Schmidt, D.W., Schleien, B., Sinclair, W.K., Voillequé, P.G., and Till, J.E. (1996). Task 6: Radiation doses and risk to residents from FMPC operations from 1951–1988. Draft report, Radiological Assessments Corporation, Neeses, SC.

[341] Kim, H., Sun, D., and Tsutakawa, R.K. (2001). A bivariate Bayes method for improving the estimates of mortality rates with a twofold conditional autoregressive model. *J. Amer. Statist. Assoc.*, **96**, 1506–1521.

[342] Kinnard, W.N. (1971). *Income Property Valuation.* Lexington, MA: Heath-Lexington Books.

[343] Kot, M. (2001). *Elements of Mathematical Ecology.* Cambridge University Press, Cambridge, UK, (2nd corrected printing in 2003).

[344] Knight, J.R., Dombrow, J., and Sirmans, C.F. (1995). A varying parameters approach to constructing house price indexes. *Real Estate Economics*, **23**, 87–105.

[345] Knorr-Held, L. (2002). Some remarks on Gaussian Markov random field models for disease mapping. In *Highly Structured Stochastic Systems*, eds. N. Hjort, P. Green and S. Richardson. Oxford: Oxford University Press.

[346] Knorr-Held, L. and Best, N.G. (2001). A shared component model for detecting joint and selective clustering of two diseases. *J. Roy. Statist. Soc. Ser. A*, **164**, 73–85.

[347] Knorr-Held, L. and Rue, H. (2002). On block updating in Markov random field models for disease mapping. *Scand. J. Statist.*, **29**, 597–614.

[348] Krige, D.G. (1951). A statistical approach to some basic mine valuation problems on the Witwatersrand. *J. Chemical, Metallurgical and Mining Society of South Africa*, **52**, 119–139.

[349] Kulldorff, M. (1997). A spatial scan statistic. Communications in Statistics: Theory and Methods, **26**, 1481–1496.

[350] Kulldorff, M. (2006). *SaTScan User Guide*, Version 7.0. Available online at `http://www.satscan.org/`.

[351] Lahiri, S.N., Kaiser, M.S., Cressie, N., and Hsu, N.-J. (1999). Prediction of spatial cumulative distribution functions using subsampling (with discussion). *J. Amer. Statist. Assoc.*, **94**, 86–110.

[352] Laird, N.M. and Ware, J.H. (1982). Random-effects models for longitudinal data. *Biometrics*, **38**, 963–974.

[353] Langford, I.H., Leyland, A.H., Rasbash, J., and Goldstein, H. (1999). Multilevel modelling of the geographical distributions of diseases. *Applied Statistics*, **48**, 253–268.

[354] Lawson, A.B. (2001). *Statistical Methods in Spatial Epidemiology.* New York: Wiley.

[355] Lawson, A.B. and Denison, D.G.T., eds. (2002). *Spatial Cluster Modelling.* Boca Raton, FL: Chapman and Hall/CRC Press.

[356] Le, C.T. (1997). *Applied Survival Analysis.* New York: Wiley.

[357] Le, N. and Zidek, J. (1992). Interpolation with uncertain spatial covariances: A Bayesian alternative to kriging. *J. Mult. Anal.*, **43**, 351–374.

[358] Le, N.D., Sun, W., and Zidek, J.V. (1997). Bayesian multivariate spatial interpolation with data missing by design. *J. Roy. Statist. Soc., Ser. B*, **59**, 501–510.

[359] Lee, S., Wolberg, G. and Shin, S.Y. (1997). Scattered data interpolation with multilevel B-splines. *IEEE Trans. Visualization and Computer Graphics*, **3**, 228–244.

[360] Leecaster, M.K. (2002). Geostatistic modeling of subsurface characteristics in the Radioactive Waste Management Complex Region, Operable Unit 7-13/14. Technical report, Idaho National Engineering and Environmental Laboratory, Idaho Falls, ID.

[361] Lele, S. (1995). Inner product matrices, kriging, and nonparametric estimation of the variogram. *Mathematical Geology*, **27**, 673–692.

[362] Leroux BG, Lei, X. and Breslow N. (1999). Estimation of disease rates in small areas: a new mixed model for spatial dependence. In: Halloran ME, Berry D, editors. Statistical models in epidemiology, the environment, and clinical trials. New York: Springer, pp. 135–178.

[363] Li, Y. and Ryan, L. (2002). Modeling spatial survival data using semiparametric frailty models. *Biometrics*, **58**, 287–297.

[364] Li, P., Banerjee, S., McBean, A.M., and Carlin, B.P. (2012). Bayesian areal wombling using false discovery rates, *Statistics and Its Interface*, **5**, 149–158.

[365] Liang, S., Banerjee, S. and Carlin, B.P. (2009). Bayesian wombling for spatial point processes. *Biometrics*, **65**, 1243–1253.
Liang, S., Carlin, B.P. and Gelfand, A.E. (2009). Analysis of marked point patterns with spatial and nonspatial covariate information, *Annals of Applied Statistics*, **3**, 943–962.

[366] Lieshout, Van M.N.M. (2000). Markov point processes and their applications. Imperial College Press: London.

[367] Lindgren, F., Rue, H. and Lindstrom, J. (2011). An explicit link between Gaussian fields and Gaussian Markov random fields: the stochastic partial differential equation approach. *Journal of the Royal Statistical Society, Series B*, **73**, 423–498.

[368] Little, R.J.A. and Rubin, D.B. (2002). *Statistical Analysis with Missing Data.* New York: Wiley-Interscience.

[369] Liu, J.S. (1994). The collapsed Gibbs sampler in Bayesian computations with applications to a gene regulation problem. *J. Amer. Statist. Assoc.*, **89**, 958–966.

[370] Liu, J.S. (2001). *Monte Carlo Strategies in Scientific Computing.* New York: Springer-Verlag.

[371] Liu, J.S., Wong, W.H., and Kong, A. (1994). Covariance structure of the Gibbs sampler with applications to the comparisons of estimators and augmentation schemes. *Biometrika*, **81**, 27–40.

[372] Loh, W-H. (2005). Fixed-domain asymptotics for a subclass of Matérn-type Gaussian random fields. *Annals of Statistics*, **33**, 2344–2394.

[373] Louis, T.A. (1982). Finding the observed information matrix when using the EM algorithm. *J. Roy. Statist. Soc., Ser. B*, **44**, 226–233.

[374] Louis, T.A. (1984). Estimating a population of parameter values using Bayes and empirical Bayes methods. *J. Amer. Statist. Assoc.*, **79**, 393–398.

[375] Lu, H. and Carlin, B.P. (2005). Bayesian areal wombling for geographical boundary analysis. *Geographical Analysis*, **37**, 265–285.

[376] Lu, H., Reilly, C.S., Banerjee, S., and Carlin, B.P. (2007). Bayesian areal wombling via adjacency modeling. *Environmental and Ecological Statistics*, **14**, 433–452.

[377] Lund, J., Penttinen, A. and Rudemo M. (1999). Bayesian analysis of spatial point patterns from noisy observations. *Report, Department of Mathematics and Physics*, The Royal Veterinary and Agricultural University, Copenhagen.

[378] Lund, J., Rudemo, M. (2000). Models for point processes observed with noise. *Biometrika*, **87**, no. 2, 235-249.

[379] Luo, Z. and Wahba, G. (1998). Spatio-temporal analogues of temperature using smoothing spline ANOVA. *Journal of Climatology*, **11**, 18–28.

[380] Lusht, K.M. (1997). *Real Estate Valuation*. Chicago: Irwin.

[381] Ma, H. and Carlin, B.P. (2007). Bayesian multivariate areal wombling for multiple disease boundary analysis. *Bayesian Analysis*, **2**, 281–302.

[382] Ma, H., Carlin, B.P., and Banerjee, S. (2010). Hierarchical and joint site-edge methods for Medicare hospice service region boundary analysis. *Biometrics*, **66**, 355–364.

[383] Ma, H., Virnig, B., and Carlin, B.P. (2006). Spatial methods in areal administrative data analysis. *Italian Journal of Public Health*, **3**, 94–103.

[384] MacEachern, S.N. and Berliner, L.M. (1994). Subsampling the Gibbs sampler. *The American Statistician*, **48**, 188–190.

[385] Majumdar, A. and Gelfand, A.E. (2003). Convolution methods for developing cross-covariance functions. Technical report, Institute for Statistics and Decision Sciences, Duke University.

[386] Majumdar, A. and Gelfand, A.E. (2007). Multivariate Spatial Process Modeling Using Convolved Covariance Functions, *Mathematical Geology*, **79**, 225–245.

[387] Mardia, K.V. (1988). Multi-dimensional multivariate Gaussian Markov random fields with application to image processing. *Journal of Multivariate Analysis*, **24**, 265–284.

[388] Mardia, K.V. and Goodall, C. (1993). Spatio-temporal analyses of multivariate environmental monitoring data. In *Multivariate Environmental Statistics*, eds. G.P. Patil and C.R. Rao. Amsterdam: Elsevier, pp. 347–386.

[389] Mardia, K.V., Goodall, C., Redfern, E.J., and Alonso, F.J. (1998). The kriged Kalman filter (with discussion). *Test*, **7**, 217–285.

[390] Mardia, K.V., Kent, J.T., and Bibby, J.M. (1979). *Multivariate Analysis*. New York: Academic Press.

[391] Mardia, K.V. and Marshall, R.J. (1984). Maximum likelihood estimation of models for residual covariance in spatial regression. *Biometrika*, **71**, 135–146.

[392] Matérn, B. (1960; reprinted 1986). *Spatial Variation, 2nd ed.* Berlin: Springer-Verlag.

[393] Matheron, G. (1963). Principles of geostatistics. *Economic Geology*, **58**, 1246–1266.

[394] Matheron, G. (1973). The intrinsic random functions and their applications. *Advances in Applied Probability*, **5**, 437–468.

[395] Matheron, G. (1979). Recherche de simplification dans un probleme de cokrigeage: Centre de Gostatistique, Fountainebleau, N-698.

[396] Matheron, G. (1982). Pour une analyse krigeante des données regionaliées. Technical report, Ecole Nationale Supérieure des Mines de Paris.

[397] McBratney, A. and Webster, R. (1986). Choosing functions for semi-variograms of soil properties and fitting them to sampling estimates. *Journal of Soil Science*, **37**, 617–639.

[398] MacNab, Y.C. and Dean, C.B. (2000). Parametric bootstrap and penalized quasi-likelihood inference in conditional autoregressive models. *Statistics in Medicine*, **19**, 2421–2435.

[399] McMillan, N., Holland, D.M., Morara, M. and Feng, J. (2008), Combining numerical model output and particulate data using Bayesian space-time modeling. *Environmetrics*, **21**, 48–65.

[400] Mengersen, K.L., Robert, C.P., and Guihenneuc-Jouyaux, C. (1999). MCMC convergence diagnostics: A reviewww (with discussion). In *Bayesian Statistics 6*, eds. J.M. Bernardo, J.O. Berger, A.P. Dawid, and A.F.M. Smith. Oxford: Oxford University Press, pp. 415–440.

[401] Metropolis, N., Rosenbluth, A.W., Rosenbluth, M.N., Teller, A.H., and Teller, E. (1953). Equations of state calculations by fast computing machines. *J. Chemical Physics*, **21**, 1087–1091.

[402] Meyer, T.H., Ericksson, M., and Maggio, R.C. (2001). Gradient estimation from irregularly spaced datasets. *Mathematical Geology*, **33**, 693–717.

[403] Mira, A., Møller, J., and Roberts, G.O. (2001). Perfect slice samplers. *J. Roy. Statist. Soc., Ser. B*, **63**, 593–606.

[404] Mira, A. and Sargent, D.J. (2005). Strategies for speeding Markov Chain Monte Carlo algorithms. To appear *Statistical Methods and Applications*.

[405] Mira, A. and Tierney, L. (2001). Efficiency and convergence properties of slice samplers. *Scandinavian Journal of Statistics*, **29**, 1–12.

[406] Mitas, L. and Mitasova, H. (1999). *Spatial Interpolation*. In: P.Longley, M.F. Goodchild, D.J. Maguire, D.W.Rhind (Eds.) Geographical Information Systems: Principles, Techniques, Management and Applications, GeoInformation International. Wiley, pp. 481–492.

[407] Møller, J., Pettitt, A.N., Berthelsen, K.K. and Reeves, R.W. (2006). An efficient Markov chain Monte Carlo method for distributions with intractable normalising constants. *Biometrika*, **93**, 451–458.

[408] Møller, J. and Waagepetersen, R.P. (2003). An introduction to simulation-based inference for spatial point processes. In *Spatial Statistics and Computational Methods*, *Lecture Notes in Statistics 173*, ed. J. Møller, New York: Springer-Verlag, pp. 143–198.

[409] Møller, J. and Waagepetersen, R. (2004). *Statistical Inference and Simulation for Spatial Point Processes*. Boca Raton, FL: Chapman and Hall/CRC Press.

[410] Møller, J. and Waagepetersen, R.P. (2007). Modern statistics for spatial point processes (with discussion). *Scandinavian Journal of Statistics*, **34**, 643–711.

[411] Monahan, J.F. (2001). *Numerical Methods of Statistics*. Cambridge: Cambridge University Press.

[412] Mueller, I., Vounatsou, P., Allen, B.J., and Smith, T. (2001). Spatial patterns of child growth in Papua New Guinea and their relation to environment, diet, socio-economic status and subsistence activities. *Annals of Human Biology*, **28**, 263–280.

[413] Mueller, P., Parmigiani, G., and Rice, K. (2006). FDR and Bayesian multiple comparisons rules. In *Bayesian Statistics 8*. Ed(s) J.M. Bernardo, S. Bayarri, J.O. Berger, A.P. Dawid, D. Heckerman, A.F.M. Smith and M. West. Oxford University Press.

[414] Mugglin, A.S. and Carlin, B.P. (1998). Hierarchical modeling in Geographic Information Systems: population interpolation over incompatible zones. *J. Agric. Biol. Environ. Statist.*, **3**, 111–130.

[415] Mugglin, A.S., Carlin, B.P., and Gelfand, A.E. (2000). Fully model based approaches for spatially misaligned data. *J. Amer. Statist. Assoc.*, **95**, 877–887.

[416] Mugglin, A.S., Carlin, B.P., Zhu, L., and Conlon, E. (1999). Bayesian areal interpolation, estimation, and smoothing: An inferential approach for geographic information systems. *Environment and Planning A*, **31**, 1337–1352.

[417] Müller, W. G. (2001). *Collecting data: optimum design of experiments for random fields*. New York: Springer.

[418] Murray, I., Adams, R.P. and Mackay, D.J.C. (2010). Elliptical slice sampling. *Journal of Machine Learning Research W&CP*, **9**, 541–548.

[419] Murray, I. and Adams, R.P. (2010). Slice sampling covariance hyperparameters of latent Gaussian models. *Advances in Neural Information Processing Systems*, **23**.

[420] Murray, R., Anthonisen, N.R., Connett, J.E., Wise, R.A., Lindgren, P.G., Greene, P.G., and Nides, M.A. for the Lung Health Study Research Group (1998). Effects of multiple attempts to quit smoking and relapses to smoking on pulmonary function. *J. Clin. Epidemiol.*, **51**, 1317–1326.

[421] Myers, D.E. (1982). Matrix formulation of co-kriging. *Journal of the International Association for Mathematical Geology*, **15**, 633–637.

[422] Myers, D.E. (1991). Pseudo-cross variograms, positive definiteness and cokriging. *Mathematical Geology*, **23**, 805–816

[423] Mykland, P., Tierney, L., and Yu, B. (1995). Regeneration in Markov chain samplers. *J. Amer. Statist. Assoc.*, **90**, 233–241.

[424] Neal, R.M. (2003). Slice sampling (with discussion). *Annals of Statistics*, **31**, 705–767.

[425] Neill, D.B., Moore, A.W. and Cooper, G.F. (2006). A Bayesian spatial scan statistic. *Advances in Neural Information Processing Systems*, **18**, 1003–1010.

[426] Nelsen, R. (2006). *An Introduction to Copulas*. New York: Springer.

[427] Newton, M., Noueriry, A., Sarkar, D. and Ahlquist, P. (2004). Detecting differential gene expression with a semiparametric heirarchical mixture model. *Biostatistics*, **5**, 155–176.

[428] Neyman, J. and Scott, E.L. (1958). Statistical approaches to problems of cosmology (with discussion). *Journal of Royal Statistical Society, Series B*, **20**, 1–43.

[429] Nychka, D. and Saltzman, N. (1998). Design of air-quality monitoring networks. In *Case studies in Environmental Statistics,* ed. Nychka D, Cox L, Piegorsch W. Editors. Lecture Notes in Statistics, Springer Verlag: New York.

[430] Oakley, J.E. and O'Hagan, A. (2004). Probabilistic sensitivity analysis of complex models: a Bayesian approach. *J. Roy. Statist. Soc. Ser. B*, **66**, 751–769.

[431] O'Hagan, A. (1994). *Kendall's Advanced Theory of Statistics Volume 2b: Bayesian Inference*. London: Edward Arnold.

[432] O'Hagan, A. (1995). Fractional Bayes factors for model comparison (with discussion). *J. Roy. Statist. Soc., Ser. B*, **57**, 99–138.

[433] Omre, H. (1987). Bayesian kriging – merging observations and qualified guesses in kriging. *Math. Geology*, **19**, 25–39.

[434] Omre, H. (1988). A Bayesian approach to surface estimation. In *Quantitative Analysis of Mineral and Energy Resources*, eds. C.F. Chung et al., Boston: D. Reidel Publishing Co., pp. 289–306.

[435] Omre, H. and Halvorsen, K.B. (1989). The Bayesian bridge between simple and universal kriging. *Math. Geology*, **21**, 767–786.

[436] Omre, H., Halvorsen, K.B. and Berteig, V. (1989). A Bayesian approach to kriging. In *Geostatistics*, ed. M. Armstrong, Boston: Kluwer Academic Publishers, pp. 109–126.

[437] Overton, W.S. (1989). Effects of measurements and other extraneous errors on estimated distribution functions in the National Surface Water Surveys. Technical Report 129, Department of Statistics, Oregon State University.

[438] Pace, R.K. and Barry, R. (1997a). Sparse spatial autoregressions. *Statistics and Probability Letters*, **33**, 291–297.

[439] Pace, R.K. and Barry, R. (1997b). Fast spatial estimation. *Applied Economics Letters*, **4**, 337–341.

[440] Pace, R.K., Barry, R., Gilley, O.W., and Sirmans, C.F. (2000). A method for spatial-temporal forecasting with an application to real estate prices. *International J. Forecasting*, **16**, 229–246.

[441] Paci, L., Gelfand, A.E., and Holland, D.M. (2013). Spatio-temporal modeling for real-time ozone forecasting. *Spatial Statistics* (in press).

[442] Paciorek and Schervish (2004). Nonstationary covariance functions for Gaussian process regression. *Advances in Neural Information Processing Systems*, **16**, 273–280.

[443] Paciorek, C.J. and Schervish, M. (2006). Spatial modelling using a new class of nonstationary covariance functions. *Environmetrics*, **17**, 483–506

[444] Padoan, S.A., Ribatet, M. and Sisson, S.A. (2010). Likelihood-based inference for max-stable processes. *Journal of the American Statistical Association*, **105**, 263–277

[445] Pardo-Igúzquiza, E. and Dowd, P.A. (1997). AMLE3D: A computer program for the inference of spatial covariance parameters by approximate maximum likelihood estimation. *Computers and Geosciences*, **23**, 793–805.

[446] Pati D., Reich B.J. and Dunson, D.B. (2011). Bayesian geostatistical modelling with informative sampling locations. *Biometrika*, **98**, 35–48.

[447] Pearson, F. (1990). *Map Projections: Theory and Applications.* Boca Raton, FL: CRC Press.

[448] Papangelou, F. (1974). The conditional intensity of general point processes and an application to line processes. *Zeitschrift for Wahrscheinlichkeitstheorie und Verwandte Gebiete*, **28**, 207–226

[449] Penberthy, L. and Stivers, C. (2000). Analysis of cancer risk in district health departments in Virginia, 1992–1995. Technical report, Cancer Prevention and Control Project, Virginia Department of Health.

[450] Peters G. W., S. A. Sisson and Y. Fan (2012). Likelihood-free Bayesian inference for alpha-stable models. *Computational Statistics and Data Analysis*, **56**, 3743–3756

[451] Pickle, L.W., Mungiole, M., Jones G.K., and White, A.A. (1996). *Atlas of United States Mortality.* Hyattsville, MD: National Center for Health Statistics.

[452] Pilz, J. and Spöck, G. (2008). Bayesian spatial sampling design. In *Proc. 8th International Geostatistics Congress* (J. M. Ortiz and X. Emery, Eds.), Gecamin Ltd., Santiago de Chile, 21-30.

[453] Poole, D. and Raftery, A.E. (2000). Inference for deterministic simulation models: the Bayesian melding approach. *Journal of the American Statistical Association*, **95**, 1244–1255.

[454] Potts, R.B. (1952). Some generalized order-disorder transformations. *Mathematical Proceedings*, **48**, 106–109.

[455] Press, W.H., Teukolsky, S.A, Vetterling, W.T., and Flannery, B.P. (1992). *Numerical Recipes in C*. Cambridge: Cambridge University Press.

[456] Ramsay, J.O. and Silverman, B.W. (1997). *Functional Data Analysis*, 2nd ed. New York: Springer.

[457] Rasmussen, C.E. and Williams, C.K.I. (2006). *Gaussian Processes for Machine Learning*. Cambridge, MA: MIT Press.

[458] Rao, C.R. (1973). *Linear Statistical Inference and its Applications*, 2nd ed. New York: Wiley.

[459] Raudenbush, S.W. and Bryk, A.S. (2002). *Hierarchical Linear Models: Applications and Data Analysis Methods*, 2nd ed. Newbury Park, CA: Sage Press.

[460] Rehman, S.U. and Shapiro, A. (1996). An integral transform approach to cross-variograms modeling. *Computational Statistics and Data Analysis*, **22**, 213–233.

[461] Ren, Q. and Banerjee, S. (2013). Hierarchical factor models for large spatially misaligned datasets: A low-rank predictive process approach. *Biometrics*, **69**, 19–30.

[462] Ren, Q., Banerjee, S., Finley, A.O. and Hodges, J.S. (2011). Variational Bayesian methods for spatial data analysis. *Computational Statistics and Data Analysis*, **55**, 3197–3217.

[463] Ribatet M., Cooley D. and Davison A.C. (2012). Bayesian inference from composite likelihoods, with an application to spatial extremes. *Statistica Sinica*, **22**, 813–846

[464] Ribeiro, P.J. and Diggle, P.J. (2001). geoR: a package for geostatistical analysis. *R News*, **1**, 14–18.

[465] Ripley J. (1977). Modeling Spatial Patterns. (with discussion) *Journal of Royal Statistical Society, Series B*, **39**, 172-212.

[466] Ripley, B.D. (1981). *Spatial Statistics*. New York: Wiley.

[467] Robert, C.P. (2007). *The Bayesian Choice: From Decision-Theoretic Foundations to Computational Implementation*. New York: Springer-Verlag.

[468] Robert, C.P. and Casella, G. (2004). *Monte Carlo Statistical Methods*, Second Edition. New York: Springer-Verlag.

[469] Robert, C.P. and Casella, G. (2009). *Introducing Monte Carlo Methods with R*. New York: Springer-Verlag.

[470] Roberts, G.O. and Rosenthal, J.S. (1999). Convergence of slice sampler Markov chains. *J. Roy. Statist. Soc., Ser. B*, **61**, 643–660.

[471] Roberts, G.O. and Sahu, S.K. (1997). Updating schemes, correlation structure, blocking and parameterization for the Gibbs sampler. *J. Roy. Statist. Soc.*, Ser. B, **59**, 291–317.

[472] Roberts, G.O. and Smith, A.F.M. (1993). Simple conditions for the convergence of the Gibbs sampler and Metropolis-Hastings algorithms. *Stochastic Processes and their Applications*, **49**, 207–216.

[473] Robinson, W.S. (1950). Ecological correlations and the behavior of individuals. *American Sociological Review*, **15**, 351–357.

[474] Rogers, J.F. and Killough, G.G. (1997). Historical dose reconstruction project: estimating the population at risk. *Health Physics*, **72**, 186–194.

[475] Royle, J.A. and Berliner, L.M. (1999). A hierarchical approach to multivariate spatial modeling and prediction. *Journal of Agricultural, Biological and Environmental Statistics*, **4**, 1–28.

[476] Rudin, W. (1976). *Principles of Mathematical Analysis*. New York: McGraw-Hill Book Co.

[477] Rue, H. and Held, L. (2005). *Gaussian Markov Random Fields: Theory and Applications*, Boca Raton: Chapman & Hall.

[478] Rue, H. and Tjelmeland, H. (2002). Fitting Gaussian Markov random fields to Gaussian fields. *Scand. J. Statist.*, **29**, 31–49.

[479] Rue, H., Martino, S. and Chopin, N. (2009). Approximate Bayesian inference for latent Gaussian models using integrated nested Laplace approximations (with discussion). *Journal of the Royal Statistical Society, Series B*, **71**, 319–392.

[480] Sahu S.K. and Bakar K.S. (2012). A Comparison of Bayesian Models for Daily Ozone Concentration Levels. *Statistical Methodology*, **9**, 144–157.

[481] Sain, S.R. and Cressie, N. (2002). Multivariate lattice models for spatial environmental data. In *Proc. A.S.A. Section on Statistics and the Environment.* Alexandria, VA: American Statistical Association, pp. 2820–2825.

[482] Sain, S.R. and Cressie, N. (2007), A spatial model for multivariate lattice data. *Journal of Econometrics*, **140**, 226–259

[483] Sang, H. and Gelfand, A.E. (2009). Hierarchical Modeling for extreme values observed over space and time. *Environmental and Ecological Statistics*, **16**, 407–426

[484] Sang, H. and Gelfand, A.E. (2011). Continuous spatial process models for spatial extreme values. *Journal of Agricultural, Biological, and Environmental Statistics*, **15**, 49–65.

[485] Sang, H., Jun, M. and Huang, J.Z. (2011), "Covariance Approximation for Large Multivariate Spatial Data Sets With an Application to Multiple Climate Model Errors," *Annals of Applied Statistics*, **4**, 2519–2548.

[486] Sang, H. and Huang, J.Z. (2012). A full scale approximation of covariance functions for large spatial data sets. *Journal of the Royal Statistical Society: Series B*, **74**, 111–132.

[487] Sampson, P.D. and Guttorp, P. (1992). Nonparametric estimation of nonstationary spatial covariance structure. *J. Amer. Statist. Assoc.*, **87**, 108–119.

[488] Sanso, B. and Guenni, L. (1999). Venezuelan rainfall data analysed using a Bayesian space-time model. *Applied Statistics*, **48**, 345–362.

[489] Santner, T.J., Williams, B.J. and Notz, W.I. (2003). *The Design and Analysis of Computer Experiments*, New York: Springer.

[490] Sargent, D.J., Hodges, J.S., and Carlin, B.P. (2000). Structured Markov chain Monte Carlo. *Journal of Computational and Graphical Statistics*, **9**, 217–234.

[491] Scalf, R. and English, P. (1996). Border Health GIS Project: Documentation for intercensal zip code population estimates. Technical report, Impact Assessment Inc., Environmental Health Investigations Branch, California Department of Health Services.

[492] Schafer, J.L. (2000). *Analysis of Incomplete Multivariate Data.* London: Chapman and Hall/CRC Press.

[493] Schervish, M.J. and Carlin, B.P. (1992). On the convergence of successive substitution sampling. *J. Computational and Graphical Statistics*, **1**, 111–127.

[494] Schmidt, A.M. and Gelfand, A.E. (2003). A Bayesian Coregionalization Approach for Multivariate Pollutant Data, *Journal of Geophysical Research - Atmosphere*, **108, D24**, 8783.

[495] Schmidt, A.M. and O'Hagan, A. (2005). Bayesian inference for nonstationary spatial covariance structure via spatial deformations. To appear *J. Roy. Statist. Soc., Ser. B.*

[496] Schoenberg, I.J. (1942). Positive definite functions on spheres. *Duke Mathematics Journal*, **9**, 96–108.

[497] Shapiro, A. and Botha, J. (1991). Variogram fitting with a general class of conditionally nonnegative definite functions. *Computational Statistics and Data Analysis*, **11**, 87–96.

[498] Sheehan, T.J., Gershman, S.T., MacDougall, L.A., Danley, R.A., Mroszczyk, M., Sorensen, A.M., and Kulldorff, M. (2000). Geographic assessment of breast cancer screening by towns, zip codes and census tracts. *Journal of Public Health Management Practice*, **6(6)**, 48–57.

[499] Short, M., Carlin, B.P., and Bushhouse, S. (2002). Using hierarchical spatial models for cancer control planning in Minnesota. *Cancer Causes and Control*, **13**, 903–916.

[500] Silverman, B. (1986). *Density Estimation for Statistics and Data Analysis*. Boca Raton, FL: Chapman and Hall/CRC Press.

[501] Skeel, R.D. (1980). Iterative refinement implies numerical stability for Gaussian elimination. *Mathematics of Computation*, **35**, 817–832.

[502] Smith, R.L. (1996). Estimating nonstationary spatial correlations. Technical report, Department of Statistics, Cambridge University, UK.

[503] Smith, R.L. (2001). *Environmental Statistics*. Lecture notes for CBMS course at the University of Washington, under revision for publication; website `www.unc.edu/depts/statistics/postscript/rs/envnotes.pdf`.

[504] Snyder, J.P. (1987). *Map Projections: A Working Manual*. Professional Paper 1395, United States Geological Survey.

[505] Solow, A.R. (1986). Mapping by simple indicator kriging. *Mathematical Geology*, **18**, 335–354.

[506] Spiegelhalter, D.J., Best, N., Carlin, B.P., and van der Linde, A. (2002). Bayesian measures of model complexity and fit (with discussion). *J. Roy. Statist. Soc., Ser. B*, **64**, 583–639.

[507] Spiegelhalter, D.J., Thomas, A., Best, N., and Gilks, W.R. (1995a). BUGS: Bayesian inference using Gibbs sampling, Version 0.50. Technical report, Medical Research Council Biostatistics Unit, Institute of Public Health, Cambridge University.

[508] Spiegelhalter, D.J., Thomas, A., Best, N., and Gilks, W.R. (1995b). BUGS examples, Version 0.50. Technical report, Medical Research Council Biostatistics Unit, Institute of Public Health, Cambridge University.

[509] Stein, M.L. (1999a). *Interpolation of Spatial Data: Some Theory for Kriging*. New York: Springer-Verlag.

[510] Stein, M.L. (1999b). Predicting random fields with increasingly dense observations. *Annals of Applied Probability*, **9**, 242–273.

[511] Stein, M. L. (2005). Space-time covariance functions. *Journal of the American Statistical Association*, **100**, 310–321.

[512] Stein, M.L., Chi, Z. and Welty, L.J. (2004) Approximating likelihoods for large spatial datasets. *Journal of the Royal Statistical Society, Series B*, **66**, 275-296.

[513] Stein, A. and Corsten, L.C.A. (1991). Universal kriging and cokriging as a regression procedure. *Biometrics*, **47**, 575–587.

[514] Stein, A., Van Eijnbergen, A.C., and Barendregt, L.G. (1991). Cokriging nonstationary data. *Mathematical Geology*, **23**, 703–719.

[515] Stern, H.S. and Cressie, N. (1999). Inference for extremes in disease mapping. In *Disease Mapping and Risk Assessment for Public Health*, eds. A. Lawson, A. Biggeri, D. Böhning, E. Lesaffre, J.-F. Viel, and R. Bertollini. Chichester: Wiley, pp. 63–84.

[516] Stigler, S. (1999). *Statistics on the Table: The History of Statistical Concepts and Methods.* Boston: Harvard University Press.

[517] Storey, J. (2002). A direct approach to false discovery rates. *Journal of the Royal Statistical Society B*, **64**, 479–498.

[518] Storey, J. (2003) The positive false discovery rate: A Bayesian interpretation and the q-value. *Annals of Statistics*, **31**, 2013–2035.

[519] Stoyan, D. (1992). Statistical estimation of model parameters of planar neyman-scott cluster processes. *Metrika*, **39**, 67–74.

[520] Strauss, D. J. (1975). A model for clustering. *Biometrika*, **62**, 467–75.

[521] Stroud, J.R., Müller, P., and Sanso, B. (2001). Dynamic models for spatio-temporal data. *J. Roy. Statist. Soc., Ser. B*, **63**, 673–689.

[522] Sun, Y., Li, B. and Genton, M.G. (2011). Geostatistics for large datasets. In *Advances And Challenges In Space-time Modelling Of Natural Events*, eds. J.M. Montero, E. Porcu, M. Schlather. Springer.

[523] Supramaniam, R., Smith, D., Coates, M., and Armstrong, B. (1998). *Survival from Cancer in New South Wales in 1980 to 1995.* Sydney: New South Wales Cancer Council.

[524] Tanner, M.A. (1996). *Tools for Statistical Inference: Methods for the Exploration of Posterior Distributions and Likelihood Functions*, 3rd ed. New York: Springer-Verlag.

[525] Tanner, M.A. and Wong, W.H. (1987). The calculation of posterior distributions by data augmentation (with discussion). *J. Amer. Statist. Assoc.*, **82**, 528–550.

[526] Theobald, D.M., Stevens, D. L., Jr., D. White, N.S. Urquhart, A.R. Olsen, and J.B. Norman. (2007). Using GIS to generate spatially balanced random survey designs for natural resource applications. *Environmental Management*, **40**, 134–146.

[527] Thomas, A.J. and Carlin, B.P. (2003). Late detection of breast and colorectal cancer in Minnesota counties: An application of spatial smoothing and clustering. *Statistics in Medicine*, **22**, 113–127.

[528] Thompson M.L., Reynolds J., Cox L.H., Guttorp P., Sampson P.D. (2001). A review of statistical methods for the meteorological adjustment of tropospheric ozone. *Atmospheric Environment*, **35**, 617–630.

[529] Tobler, W.R. (1979). Smooth pycnophylactic interpolation for geographical regions (with discussion). *J. Amer. Statist. Assoc.*, **74**, 519–536.

[530] Tolbert, P., Mulholland, J., MacIntosh, D., Xu, F., Daniels, D., Devine, O., Carlin, B.P., Klein, M., Dorley, J., Butler, A., Nordenberg, D., Frumkin, H., Ryan, P.B., and White, M. (2000). Air pollution and pediatric emergency room visits for asthma in Atlanta. *Amer. J. Epidemiology*, **151:8**, 798–810.

[531] Tonellato, S. (1997). Bayesian dynamic linear models for spatial time series. Technical report (Rapporto di riceria 5/1997), Dipartimento di Statistica, Universita CaFoscari di Venezia, Venice, Italy.

[532] Ugarte, M.D., Goicoa, T., and Militino, A.F. (2010). Spatio-temporal modeling of mortality risks using penalized splines. *Environmetrics*, **21**, 270–289.

[533] United States Department of Health and Human Services (1989). *International Classification of Diseases*, 9th revision. Washington, D.C.: DHHS, U.S. Public Health Service.

[534] Vargas-Guzmán, J.A., Warrick, A.W., and Myers, D.E. (2002). Coregionalization by linear combination of nonorthogonal components. *Mathematical Geology*, **34**, 405–419.

[535] Vaupel, J.W., Manton, K.G., and Stallard, E. (1979). The impact of heterogeneity in individual frailty on the dynamics of mortality. *Demography*, **16**, 439–454.

[536] Vecchia, A.V. (1988). Estimation and model identification for continuous spatial processes. *Journal of the Royal Statistical Society Series B*, **50**, 297–312.

[537] Ver Hoef, J.M. and Barry, R.P. (1998). Constructing and fitting models for cokriging and multivariable spatial prediction. *Journal of Statistical Planning and Inference*, **69**, 275–294.

[538] Ver Hoef, J.M. and Cressie, N.A.C. (1993). Multivariable spatial prediction: *Mathematical Geology*, **25**, 219-240.

[539] Ver Hoef, J.M., Cressie, N.A.C. and Barry, R.P. (2004). Flexible spatial models for Kriging and Cokriging using moving averages and the Fast Fourier Transform (FFT). *Journal of Computational and Graphical Statistics*, **13**, 265–282.

[540] Waagepetersen, R. (2007). An estimating function approach to inference for inhomogeneous Neyman-Scott processes. *Biometrics*, **63**, 252–258.

[541] Waagepetersen, R. (2008). Estimating functions for inhomogeneous spatial point processes with incomplete covariate data. *Biometrika*, **95**, 351–363..

[542] Waagepetersen, R. and Guan, Y. (2009). Two-step estimation for inhomogeneous spatial point processes. *Journal of the Royal Statistical Society, Ser. B.*, **71**, 685–702.

[543] Wackernagel, H. (1994). Cokriging versus kriging in regionalized multivariate data analysis. *Geoderma*, **62**, 83–92.

[544] Wackernagel, H. (2003). *Multivariate Geostatistics: An Introduction with Applications*, 3rd ed. New York: Springer-Verlag.

[545] Wakefield, J. (2001). A critique of ecological studies. *Biostatistics*, **1**, 1–20.

[546] Wakefield, J. (2003). Sensitivity analyses for ecological regression. *Biometrics*, **59**, 9–17.

[547] Wakefield, J. (2005). Ecological inference for 2×2 tables. To appear (with discussion) *J. Roy. Statist. Soc., Ser. B.*

[548] Wakefield, J. and Morris, S. (2001). The Bayesian modeling of disease risk in relation to a point source. *J. Amer. Statist. Assoc.*, **96**, 77–91.

[549] Wakefield, J. and Salway, R. (2001). A statistical framework for ecological and aggregate studies. *J. Roy. Statist. Soc., Ser. A*, **164**, 119–137.

[550] Wall, M.M. (2004). A close look at the spatial structure implied by the CAR and SAR models. *J. Statist. Plann. Inf.*, **121**, 311–324.

[551] Waller, L.A. and Gotway, C.A. (2004). *Applied Spatial Statistics for Public Health Data*. New York: Wiley.

[552] Waller, L.A., Carlin, B.P., and Xia, H. (1997). Structuring correlation within hierarchical spatio-temporal models for disease rates. In *Modelling Longitudinal and Spatially Correlated Data*, eds. T.G. Gregoire, D.R. Brillinger, P.J. Diggle, E. Russek-Cohen, W.G. Warren, and R.D. Wolfinger, New York: Springer-Verlag, pp. 308–319.

[553] Waller, L.A., Carlin, B.P., Xia, H., and Gelfand, A.E. (1997). Hierarchical spatio-temporal mapping of disease rates. *J. Amer. Statist. Assoc.*, **92**, 607–617.

[554] Waller, L.A., Turnbull, B.W., Clark, L.C., and Nasca, P. (1994). Spatial pattern analyses to detect rare disease clusters. In *Case Studies in Biometry*, eds. N. Lange, L. Ryan, L. Billard, D. Brillinger, L. Conquest, and J. Greenhouse. New York: Wiley, pp. 3–23.

[555] Ward, G., Hastie, T., Barry, S.C., Elith, J. and Leathwick, J.R. (2009). Presence-only data and the EM algorithm. *Biometrics*, **65**, 554–563.

[556] Warton, D.I. and Shepherd, L.C. (2010). Poisson point process models solve the pseudo-absence problem for presence-only data in ecology. *Annals of Applied Statistics*, **4(3)**, 1383–1402.

[557] Wei, W.W.S. (1990). *Time Series Analysis: Univariate and Multivariate Methods.* Menlo Park, CA: Addison-Wesley.

[558] Weiss, R.E. (1996). Bayesian model checking with applications to hierarchical models. Technical report, Department of Biostatistics, UCLA School of Public Health. Available online at `rem.ph.ucla.edu/~rob/papers/index.html`.

[559] West, M. and Harrison, P.J. (1997). *Bayesian Forecasting and Dynamic Models*, 2nd ed. New York: Springer-Verlag.

[560] Whittaker, J. (1990). *Graphical Models in Applied Multivariate Statistics.* Chichester: Wiley.

[561] Whittle, P. (1954). On stationary processes in the plane. *Biometrika*, **41**, 434–449.

[562] Wikle, C.K. and Berliner, L.M. (2005). Combining information across spatial scales. *Technometrics*, **47**, 80–91.

[563] Wikle, C.K. , Millif, R.F., Nychka, D. and Berliner, L.M. (2001). Spatiotemporal hierarchical Bayesian modeling: Tropical ocean surface winds. *Journal of the American Statistical Association*, **96**, 382–397.

[564] Wikle, C. and Cressie, N. (1999). A dimension reduced approach to space-time Kalman filtering. *Biometrika*, **86**, 815–829.

[565] Wikle, C.K. and Hooten, M.B. (2006). Hierarchical Bayesian Spatio-Temporal Models for Population Spread. In *Applications of Computational Statistics in the Environmental Sciences: Hierarchical Bayes and MCMC Methods*, J.S. Clark and A. Gelfand (eds). Oxford: Oxford University Press, pp. 145–169.

[566] Wikle, C.K. and M.B. Hooten, (2010). A general science-based framework for spatiotemporal dynamical models (with discussion). *Test*, **19**, 417–451.

[567] Wikle, C.K., Milliff, R.F., Nychka, D., and Berliner, L.M. (2001). Spatiotemporal hierarchical Bayesian modeling: Tropical ocean surface winds. *J. Amer. Statist. Assoc.*, **96**, 382–397.

[568] Wilkinson, J.H. (1965). *The Algebraic Eigenvalue Problem.* Oxford: Clarendon Press.

[569] Wolpert, R.L. and Ickstadt, K. (1998). Poisson/gamma random field models for spatial statistics. *Biometrika*, **85**, 251–269.

[570] Woodbury, A. (1989). Bayesian updating revisited. *Math. Geology*, **21**, 285–308.

[571] Xia, H. and Carlin, B.P. (1998). Spatio-temporal models with errors in covariates: mapping Ohio lung cancer mortality. *Statistics in Medicine*, **17**, 2025–2043.

[572] Yaglom, A.M. (2004). *An Introduction to the Theory of Stationary Random Functions.* New York: Dover Publications.

[573] Yaglom, A.M. (1987). *Correlation Theory of Stationary and Related Random Functions: Volume I: Basic Results*, London: Springer.

[574] Yakovlev, A.Y. and Tsodikov, A.D. (1996). *Stochastic Models of Tumor Latency and their Biostatistical Applications.* New Jersey: World Scientific.

[575] Zhang, H. (2004). Inconsistent estimation and asymptotically equal interpolations in model-based geostatistics. *Journal of the American Statistical Association*, **99**, 250–261.

[576] Zhang, H. and Zimmerman, D.L. (2005). Towards reconciling two asymptotic frameworks in spatial statistics. *Biometrika*, **92**, 921–936.

[577] Zhang, Y., Hodges, J.S. and Banerjee, S. (2009). Smoothed ANOVA with spatial effects as a competitor to MCAR in multivariate spatial smoothing. *Annals of Applied Statistics*, **3**, 1805–1830.

[578] Zhu, J., Lahiri, S.N., and Cressie, N. (2005). Asymptotic inference for spatial CDFs over time. To appear *Statistica Sinica*.

[579] Zhu, L. and Carlin, B.P. (2000). Comparing hierarchical models for spatio-temporally misaligned data using the Deviance Information Criterion. *Statistics in Medicine*, **19**, 2265–2278.

[580] Zhu, L., Carlin, B.P., English, P. and Scalf, R. (2000). Hierarchical modeling of spatio-temporally misaligned data: Relating traffic density to pediatric asthma hospitalizations. *Environmetrics*, **11**, 43–61.

[581] Zhu, L., Carlin, B.P., and Gelfand, A.E. (2003). Hierarchical regression with misaligned spatial data: Relating ambient ozone and pediatric asthma ER visits in Atlanta. *Environmetrics*, **14**, 537–557.

[582] Zimmerman, D.L. (1993). Another look at anisotropy in geostatistics. *Mathematical Geology*, **25**, 453–470.

[583] Zimmerman D.L. (2008). Estimating the intensity of a spatial point process from locations coarsened by incomplete geocoding. *Biometrics* **64**, 262-70.

[584] Zimmerman, D.L. and Cressie, N. (1992). Mean squared prediction error in the spatial linear model with estimated covariance parameters. *Ann. Inst. Stastist. Math.*, **44**, 27–43.

In the ten years since the publication of the first edition, the statistical landscape has substantially changed for analyzing space and space-time data. More than twice the size of its predecessor, **Hierarchical Modeling and Analysis for Spatial Data, Second Edition** reflects the major growth in spatial statistics as both a research area and an area of application.

New to the Second Edition

- New chapter on spatial point patterns developed primarily from a modeling perspective
- New chapter on big data that shows how the predictive process handles reasonably large datasets
- New chapter on spatial and spatiotemporal gradient modeling that incorporates recent developments in spatial boundary analysis and wombling
- New chapter on the theoretical aspects of geostatistical (point-referenced) modeling
- Greatly expanded chapters on methods for multivariate and spatiotemporal modeling
- New special topics sections on data fusion/assimilation and spatial analysis for data on extremes
- Double the number of exercises
- Many more color figures integrated throughout the text
- Updated computational aspects, including the latest version of WinBUGS, the new flexible spBayes software, and assorted R packages

This second edition continues to provide a complete treatment of the theory, methods, and application of hierarchical modeling for spatial and spatiotemporal data. It tackles current challenges in handling this type of data, with increased emphasis on observational data, big data, and the upsurge of associated software tools. The authors also explore important application domains, including environmental science, forestry, public health, and real estate.

CRC Press
Taylor & Francis Group
an informa business
www.crcpress.com

6000 Broken Sound Parkway, NW
Suite 300, Boca Raton, FL 33487
711 Third Avenue
New York, NY 10017
2 Park Square, Milton Park
Abingdon, Oxon OX14 4RN, UK

K11011
ISBN: 978-1-4398-1917-3
90000
9 781439 819173