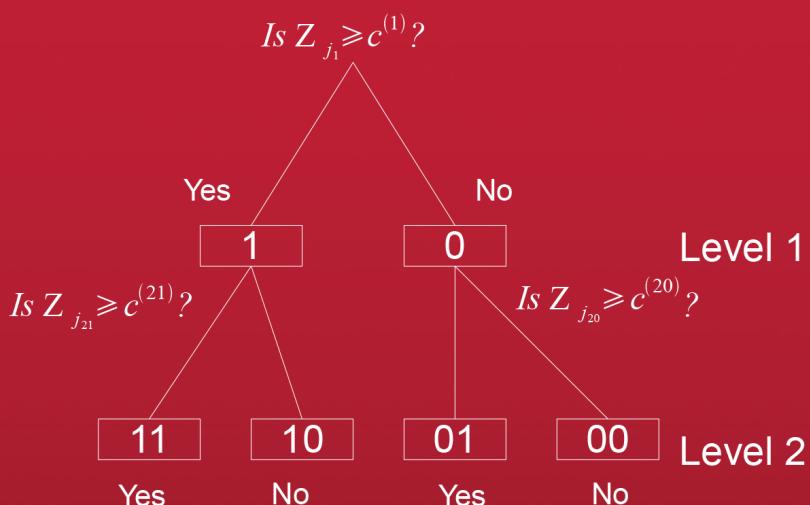


Texts in Statistical Science

Mathematical Statistics

Basic Ideas and Selected Topics

Volume II



Peter J. Bickel
Kjell A. Doksum

WITH VITALSOURCE®
EBOOK



CRC Press
Taylor & Francis Group

A CHAPMAN & HALL BOOK

Mathematical Statistics

**Basic Ideas and
Selected Topics**

Volume II

CHAPMAN & HALL/CRC

Texts in Statistical Science Series

Series Editors

Francesca Dominici, *Harvard School of Public Health, USA*

Julian J. Faraway, *University of Bath, UK*

Martin Tanner, *Northwestern University, USA*

Jim Zidek, *University of British Columbia, Canada*

Statistical Theory: A Concise Introduction

F. Abramovich and Y. Ritov

Practical Multivariate Analysis, Fifth Edition

A. Afifi, S. May, and V.A. Clark

Practical Statistics for Medical Research

D.G. Altman

Interpreting Data: A First Course in Statistics

A.J.B. Anderson

Introduction to Probability with R

K. Baclawski

Linear Algebra and Matrix Analysis for Statistics

S. Banerjee and A. Roy

Mathematical Statistics: Basic Ideas and Selected Topics, Volume I, Second Edition

P. J. Bickel and K. A. Doksum

Mathematical Statistics: Basic Ideas and Selected Topics, Volume II

P. J. Bickel and K. A. Doksum

Analysis of Categorical Data with R

C. R. Bilder and T. M. Loughin

Statistical Methods for SPC and TQM

D. Bissell

Introduction to Probability

J. K. Blitzstein and J. Hwang

Bayesian Methods for Data Analysis, Third Edition

B.P. Carlin and T.A. Louis

Second Edition

R. Caulcutt

The Analysis of Time Series: An Introduction, Sixth Edition

C. Chatfield

Introduction to Multivariate Analysis

C. Chatfield and A.J. Collins

Problem Solving: A Statistician's Guide, Second Edition

C. Chatfield

Statistics for Technology: A Course in Applied Statistics, Third Edition

C. Chatfield

Bayesian Ideas and Data Analysis: An Introduction for Scientists and Statisticians

R. Christensen, W. Johnson, A. Branscum, and T.E. Hanson

Modelling Binary Data, Second Edition

D. Collett

Modelling Survival Data in Medical Research, Third Edition

D. Collett

Introduction to Statistical Methods for Clinical Trials

T.D. Cook and D.L. DeMets

Applied Statistics: Principles and Examples

D.R. Cox and E.J. Snell

Multivariate Survival Analysis and Competing Risks

M. Crowder

Statistical Analysis of Reliability Data

M.J. Crowder, A.C. Kimber, T.J. Sweeting, and R.L. Smith

An Introduction to Generalized Linear Models, Third Edition

A.J. Dobson and A.G. Barnett

Nonlinear Time Series: Theory, Methods, and Applications with R Examples

R. Douc, E. Moulines, and D.S. Stoffer

Introduction to Optimization Methods and Their Applications in Statistics

B.S. Everitt

Extending the Linear Model with R: Generalized Linear, Mixed Effects and Nonparametric Regression Models

J.J. Faraway

Linear Models with R, Second Edition

J.J. Faraway

A Course in Large Sample Theory

T.S. Ferguson

- Multivariate Statistics: A Practical Approach**
B. Flury and H. Riedwyl
- Readings in Decision Analysis**
S. French
- Markov Chain Monte Carlo: Stochastic Simulation for Bayesian Inference, Second Edition**
D. Gamerman and H.F. Lopes
- Bayesian Data Analysis, Third Edition**
A. Gelman, J.B. Carlin, H.S. Stern, D.B. Dunson, A. Vehtari, and D.B. Rubin
- Multivariate Analysis of Variance and Repeated Measures: A Practical Approach for Behavioural Scientists**
D.J. Hand and C.C. Taylor
- Practical Longitudinal Data Analysis**
D.J. Hand and M. Crowder
- Logistic Regression Models**
J.M. Hilbe
- Richly Parameterized Linear Models: Additive, Time Series, and Spatial Models Using Random Effects**
J.S. Hodges
- Statistics for Epidemiology**
N.P. Jewell
- Stochastic Processes: An Introduction, Second Edition**
P.W. Jones and P. Smith
- The Theory of Linear Models**
B. Jørgensen
- Principles of Uncertainty**
J.B. Kadane
- Graphics for Statistics and Data Analysis with R**
K.J. Keen
- Mathematical Statistics**
K. Knight
- Introduction to Multivariate Analysis: Linear and Nonlinear Modeling**
S. Konishi
- Nonparametric Methods in Statistics with SAS Applications**
O. Korosteleva
- Modeling and Analysis of Stochastic Systems, Second Edition**
V.G. Kulkarni
- Exercises and Solutions in Biostatistical Theory**
L.L. Kupper, B.H. Neelon, and S.M. O'Brien
- Exercises and Solutions in Statistical Theory**
L.L. Kupper, B.H. Neelon, and S.M. O'Brien
- Design and Analysis of Experiments with R**
J. Lawson
- Design and Analysis of Experiments with SAS**
J. Lawson
- A Course in Categorical Data Analysis**
T. Leonard
- Statistics for Accountants**
S. Letchford
- Introduction to the Theory of Statistical Inference**
H. Liero and S. Zwanzig
- Statistical Theory, Fourth Edition**
B.W. Lindgren
- Stationary Stochastic Processes: Theory and Applications**
G. Lindgren
- Statistics for Finance**
E. Lindström, H. Madsen, and J. N. Nielsen
- The BUGS Book: A Practical Introduction to Bayesian Analysis**
D. Lunn, C. Jackson, N. Best, A. Thomas, and D. Spiegelhalter
- Introduction to General and Generalized Linear Models**
H. Madsen and P. Thyregod
- Time Series Analysis**
H. Madsen
- Pólya Urn Models**
H. Mahmoud
- Randomization, Bootstrap and Monte Carlo Methods in Biology, Third Edition**
B.F.J. Manly
- Introduction to Randomized Controlled Clinical Trials, Second Edition**
J.N.S. Matthews
- Statistical Methods in Agriculture and Experimental Biology, Second Edition**
R. Mead, R.N. Curnow, and A.M. Hasted
- Statistics in Engineering: A Practical Approach**
A.V. Metcalfe
- Statistical Inference: An Integrated Approach, Second Edition**
H. S. Migon, D. Gamerman, and F. Louzada
- Beyond ANOVA: Basics of Applied Statistics**
R.G. Miller, Jr.

A Primer on Linear Models J.F. Monahan	Decision Analysis: A Bayesian Approach J.Q. Smith
Applied Stochastic Modelling, Second Edition B.J.T. Morgan	Analysis of Failure and Survival Data P.J. Smith
Elements of Simulation B.J.T. Morgan	Applied Statistics: Handbook of GENSTAT Analyses E.J. Snell and H. Simpson
Probability: Methods and Measurement A. O'Hagan	Applied Nonparametric Statistical Methods, Fourth Edition P. Sprent and N.C. Smeeton
Introduction to Statistical Limit Theory A.M. Polansky	Data Driven Statistical Methods P. Sprent
Applied Bayesian Forecasting and Time Series Analysis A. Pole, M. West, and J. Harrison	Generalized Linear Mixed Models: Modern Concepts, Methods and Applications W. W. Stroup
Statistics in Research and Development, Time Series: Modeling, Computation, and Inference R. Prado and M. West	Survival Analysis Using S: Analysis of Time-to-Event Data M. Tableman and J.S. Kim
Introduction to Statistical Process Control P. Qiu	Applied Categorical and Count Data Analysis W. Tang, H. He, and X.M. Tu
Sampling Methodologies with Applications P.S.R.S. Rao	Elementary Applications of Probability Theory, Second Edition H.C. Tuckwell
A First Course in Linear Model Theory N. Ravishanker and D.K. Dey	Introduction to Statistical Inference and Its Applications with R M.W. Trosset
Essential Statistics, Fourth Edition D.A.G. Rees	Understanding Advanced Statistical Methods P.H. Westfall and K.S.S. Henning
Stochastic Modeling and Mathematical Statistics: A Text for Statisticians and Quantitative Scientists F.J. Samaniego	Statistical Process Control: Theory and Practice, Third Edition G.B. Wetherill and D.W. Brown
Statistical Methods for Spatial Data Analysis O. Schabenberger and C.A. Gotway	Generalized Additive Models: An Introduction with R S. Wood
Bayesian Networks: With Examples in R M. Scutari and J.-B. Denis	Epidemiology: Study Design and Data Analysis, Third Edition M. Woodward
Large Sample Methods in Statistics P.K. Sen and J. da Motta Singer	Practical Data Analysis for Designed Experiments B.S. Yandell
Spatio-Temporal Methods in Environmental Epidemiology G. Shaddick and J.V. Zidek	

Texts in Statistical Science

Mathematical Statistics

Basic Ideas and Selected Topics

Volume II

Peter J. Bickel

University of California

Berkley, California, USA

Kjell A. Doksum

University of Wisconsin

Madison, Wisconsin, 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

© 2016 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: 20150909

International Standard Book Number-13: 978-1-4987-2270-4 (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 Erich L. Lehmann

CONTENTS

PREFACE TO THE 2015 EDITION	xv
I INTRODUCTION AND EXAMPLES	1
I.0 Basic Ideas and Conventions	1
I.1 Tests of Goodness of Fit and the Brownian Bridge	5
I.2 Testing Goodness of Fit to Parametric Hypotheses	5
I.3 Regular Parameters. Minimum Distance Estimates	6
I.4 Permutation Tests	8
I.5 Estimation of Irregular Parameters	8
I.6 Stein and Empirical Bayes Estimation	10
I.7 Model Selection	11
I.8 Problems and Complements	15
I.9 Notes	20
7 TOOLS FOR ASYMPTOTIC ANALYSIS	21
7.1 Weak Convergence in Function Spaces	21
7.1.1 Stochastic Processes and Weak Convergence	21
7.1.2 Maximal Inequalities	28
7.1.3 Empirical Processes on Function Spaces	31
7.2 The Delta Method in Infinite Dimensional Space	38
7.2.1 Influence Functions. The Gâteaux and Fréchet Derivatives	38
7.2.2 The Quantile Process	47
7.3 Further Expansions	51
7.3.1 The von Mises Expansion	51
7.3.2 The Hoeffding and Analysis of Variance Expansions	54
7.4 Problems and Complements	62
7.5 Notes	72

8 DISTRIBUTION-FREE, UNBIASED, AND EQUIVARIANT PROCEDURES	73
8.1 Introduction	73
8.2 Similarity and Completeness	74
8.2.1 Testing	74
8.2.2 Testing Optimality Theory	85
8.2.3 Estimation	87
8.3 Invariance, Equivariance, and Minimax Procedures	92
8.3.1 Group Models	92
8.3.2 Group Models and Decision Theory	94
8.3.3 Characterizing Invariant Tests	96
8.3.4 Characterizing Equivariant Estimates	102
8.3.5 Minimality for Tests: Application to Group Models	104
8.3.6 Minimax Estimation, Admissibility, and Steinian Shrinkage	107
8.4 Problems and Complements	112
8.5 Notes	123
9 INFERENCE IN SEMIPARAMETRIC MODELS	125
9.1 Estimation in Semiparametric Models	125
9.1.1 Selected Examples	125
9.1.2 Regularization. Modified Maximum Likelihood	133
9.1.3 Other Modified and Approximate Likelihoods	142
9.1.4 Sieves and Regularization	145
9.2 Asymptotics. Consistency and Asymptotic Normality	151
9.2.1 A General Consistency Criterion	152
9.2.2 Asymptotics for Selected Models	153
9.3 Efficiency in Semiparametric Models	161
9.4 Tests and Empirical Process Theory	175
9.5 Asymptotic Properties of Likelihoods. Contiguity	181
9.6 Problems and Complements	193
9.7 Notes	209
10 MONTE CARLO METHODS	211
10.1 The Nature of Monte Carlo Methods	211
10.2 Three Basic Monte Carlo Methods	214
10.2.1 Simple Monte Carlo	215
10.2.2 Importance Sampling	216
10.2.3 Rejective Sampling	217

10.3 The Bootstrap	219
10.3.1 Bootstrap Samples and Bias Corrections	220
10.3.2 Bootstrap Variance and Confidence Bounds	224
10.3.3 The General i.i.d. Nonparametric Bootstrap	227
10.3.4 Asymptotic Theory for the Bootstrap	230
10.3.5 Examples Where Efron's Bootstrap Fails. The m out of n Bootstraps	235
10.4 Markov Chain Monte Carlo	237
10.4.1 The Basic MCMC Framework	237
10.4.2 Metropolis Sampling Algorithms	238
10.4.3 The Gibbs Samplers	242
10.4.4 Speed of Convergence and Efficiency of MCMC	246
10.5 Applications of MCMC to Bayesian and Frequentist Inference	250
10.6 Problems and Complements	256
10.7 Notes	263
11 NONPARAMETRIC INFERENCE FOR FUNCTIONS OF ONE VARIABLE	265
11.1 Introduction	265
11.2 Convolution Kernel Estimates on R	266
11.2.1 Uniform Local Behavior of Kernel Density Estimates	269
11.2.2 Global Behavior of Convolution Kernel Estimates	271
11.2.3 Performance and Bandwidth Choice	272
11.2.4 Discussion of Convolution Kernel Estimates	273
11.3 Minimum Contrast Estimates: Reducing Boundary Bias	274
11.4 Regularization and Nonlinear Density Estimates	280
11.4.1 Regularization and Roughness Penalties	280
11.4.2 Sieves. Machine Learning. Log Density Estimation	281
11.4.3 Nearest Neighbor Density Estimates	284
11.5 Confidence Regions	285
11.6 Nonparametric Regression for One Covariate	287
11.6.1 Estimation Principles	287
11.6.2 Asymptotic Bias and Variance Calculations	290
11.7 Problems and Complements	297
12 PREDICTION AND MACHINE LEARNING	307
12.1 Introduction	307

12.1.1	Statistical Approaches to Modeling and Analyzing Multidimensional data. Sieves	309
12.1.2	Machine Learning Approaches	313
12.1.3	Outline	315
12.2	Classification and Prediction	315
12.2.1	Multivariate Density and Regression Estimation	315
12.2.2	Bayes Rule and Nonparametric Classification	320
12.2.3	Sieve Methods	322
12.2.4	Machine Learning Approaches	324
12.3	Asymptotic Risk Criteria	333
12.3.1	Optimal Prediction in Parametric Regression Models	334
12.3.2	Optimal Rates of Convergence for Estimation and Prediction in Nonparametric Models	337
12.3.3	The Gaussian White Noise (GWN) Model	347
12.3.4	Minimax Bounds on IMSE for Subsets of the GWN Model	349
12.3.5	Sparse Submodels	350
12.4	Oracle Inequalities	352
12.4.1	Stein's Unbiased Risk Estimate	354
12.4.2	Oracle Inequality for Shrinkage Estimators	355
12.4.3	Oracle Inequality and Adaptive Minimax Rate for Truncated Estimates	357
12.4.4	An Oracle Inequality for Classification	359
12.5	Performance and Tuning via Cross Validation	361
12.5.1	Cross Validation for Tuning Parameter Choice	362
12.5.2	Cross Validation for Measuring Performance	367
12.6	Model Selection and Dimension Reduction	367
12.6.1	A Bayesian Criterion for Model Selection	368
12.6.2	Inference after Model Selection	372
12.6.3	Dimension Reduction via Principal Component Analysis	374
12.7	Topics Briefly Touched and Current Frontiers	377
12.8	Problems and Complements	381
D	APPENDIX D. SUPPLEMENTS TO TEXT	399
D.1	Probability Results	399
D.2	Supplement to Section 7.1	401
D.3	Supplement to Section 7.2	404
D.4	Supplement to Section 9.2.2	405
D.5	Supplement to Section 10.4	406

D.6 Supplement to Section 11.6	410
D.7 Supplement to Section 12.2.2	413
D.8 Problems and Complements	419
E SOLUTIONS FOR VOLUME II	423
REFERENCES	437

PREFACE TO THE 2015 EDITION

Volume II

This textbook represents our view of what a second course in mathematical statistics for graduate students with a good mathematics background should be. The mathematics background needed includes linear algebra, matrix theory and advanced calculus, but not measure theory. Probability at the level of, for instance, Grimmett and Stirzaker's *Probability and Random Processes*, is also needed. Appendix D1 in combination with Appendices A and B in Volume I give the probability that is needed. However, the treatment is abridged with few proofs.

This Volume II of the second edition presents what we think are some of the most important statistical concepts, methods, and tools developed since the first edition. Topics included are: asymptotic efficiency in semiparametric models, semiparametric maximum likelihood estimation, finite sample size optimality including Lehmann-Scheffé theory, survival analysis including Cox regression, prediction, classification, methods of inference based on sieve models, model and variable selection, Monte Carlo methods such as the bootstrap and Markov Chain Monte Carlo, nonparametric curve estimation, and machine learning including support vector machines and classification and regression trees (CART).

The basic asymptotic tools developed or presented, in part in the text and in part in appendices, are weak convergence for random processes, empirical process theory, and the functional delta method. With the tools and concepts developed in this second volume students will be ready for advanced research in modern statistics.

Volume II includes too many topics to be covered in one semester. Chapter 8 can be omitted without losing much continuity. The following outline of Volume II chapter contents can be used to select topics to be included in a one semester course. A course leaning towards basic statistical theory could include most of Chapters 7, 8, 9, and 10 plus Sections 11.6, 12.4, and 12.6, while a course leaning towards statistical learning could include most of Chapters 7, 9, 10, and 12 plus Section 11.4. A great number of other possible combinations can be constructed from the following chapter outline.

Volume II Outline

Chapter I is an introductory chapter that starts by presenting basic ideas about statistical modeling and inference, then gives a number of examples that illustrate some of the basic

ideas. Chapter 7 gives the asymptotic tools to be used in parts of the rest of the book. These tools include asymptotic empirical process theory, the delta method and derivatives on function spaces and their use in deriving influence functions, as well as von Mises and Hoeffding expansions.

Chapter 8 presents some of the classical theory of statistical optimality in a decision theoretic context. Here, we derive for fixed sample sizes procedures which are optimal over restricted classes of methods, such as distribution free tests and unbiased, invariant tests or equivariant estimates. Alternatively, we study the weak but general property of minimaxity, as well as admissibility. Some results important for Chapter 12 such as Stein's identity and estimate are also introduced. The asymptotic counterparts of these notions are introduced in Chapter 9, enabling us to present generalizations which go well beyond classical exponential family and invariant models. The methods of Chapter 7 are key in this development.

In Chapter 9, we first introduce the concepts of regularization, modified maximum likelihood, and sieves. We then develop asymptotic efficiency for semiparametric models, asymptotic optimality for tests, and Le Cam's asymptotic theory of experiments. These concepts are applied to estimation in Cox's semiparametric proportional hazard regression model, partially linear models, semiparametric linear models, biased sampling models, and models for censored data. For testing, the asymptotic optimality of Neyman's C_α and Rao's score tests are established.

Chapter 10 develops Monte Carlo methods, that is, methods based on simulation. Simulation enables us to approximate methods stochastically, which are difficult to compute analytically. Examples range from distributions of test statistics and confidence bounds to likelihood methods to estimates of risk. In this chapter we first give methods based on simple random sampling where data are generated from specified distributions, including simple posterior distributions in a Bayesian setting. Next, importance sampling and rejective sampling are introduced as methods that are able to generate random variables with a desired distribution when simple random sampling is not possible. In Chapter 10 we also present Efron's bootstrap. Here, important features of the population distribution, that is, parameters, are expressed as functionals of this unknown distribution and then the distribution is replaced by an estimate such as the empirical distribution of the experimental data. The functional evaluated at the empirical distribution is approximated by drawing Monte Carlo samples from the empirical distribution. Chapter 10 goes on to present Markov Chain Monte Carlo (MCMC) techniques. These are appropriate when direct generation of independent identically distributed variables is not possible. In this approach, a sequence of random variables are generated according to a homogenous Markov chain in such a way that variables far out in the chain are approximately distributed as a sample from a target distribution. The Metropolis algorithm and the Gibbs sampler are developed as special cases. MCMC is in particular important for Bayesian statistical inference, but also in frequentist contexts.

Chapter 11 examines nonparametric estimation of functions of one variable, including density functions and the nonparametric regression of a response on one covariate. Estimates considered are based on kernels, series expansions, roughness penalties, nearest neighbors, and local polynomials. Asymptotic properties of mean squared errors are

developed.

Chapter 12 has a number of topics related to what is known as statistical learning. Topics include prediction, classification, nonparametric estimation of multivariate densities and regression functions, penalty estimation including the least absolute shrinkage and selection operator (Lasso), classification and regression trees (CART), support vector machines, boosting, Gaussian white noise modeling, oracle inequalities, Steinian shrinkage estimation, sparsity, Bayesian model selection, regularization, sieves, cross-validation, asymptotic risk and optimality properties of predictors and classifiers, and more.

We thank Akichika Ozeki, Sangbum Choi, Sören Künzel, Joshua Cape, and many students for pointing out errors and John Kimmel and CRC Press for production support. For word processing we thank Dee Frana and especially Anne Chong who processed 95% of Volume II and helped with references, indexing and error detection. Kjell Doksum thanks the statistics departments of Harvard, Columbia and Stanford Universities for support.

Last and most important we would like to thank our wives, Nancy Kramer Bickel and Joan H. Fujimura, and our families for support, encouragement, and active participation in an enterprise that at times seemed endless, appeared gratifyingly ended in 1976 but has, with the field, taken on a new life.

Volume I

For convenience we repeat the preface to Volume I. In recent years statistics has changed enormously under the impact of several forces:

- (1) The generation of what were once unusual types of data such as images, trees (phylogenetic and other), and other types of combinatorial objects.
- (2) The generation of enormous amounts of data—terabytes (the equivalent of 10^{12} characters) for an astronomical survey over three years.
- (3) The possibility of implementing computations of a magnitude that would have once been unthinkable.

The underlying sources of these changes have been the exponential change in computing speed (Moore’s “law”) and the development of devices (computer controlled) using novel instruments and scientific techniques (e.g., NMR tomography, gene sequencing). These techniques often have a strong intrinsic computational component. Tomographic data are the result of mathematically based processing. Sequencing is done by applying computational algorithms to raw gel electrophoresis data.

As a consequence the emphasis of statistical theory has shifted away from small sample optimality results in a number of directions:

- (1) Methods for inference based on larger numbers of observations and minimal assumptions—asymptotic methods in non- and semiparametric models, models with “infinite” number of parameters.
- (2) The construction of models for time series, temporal spatial series, and other complex data structures using sophisticated probability modeling but again relying for analytical results on asymptotic approximation. Multiparameter models are the rule.

- (3) The use of methods of inference involving simulation as a key element such as the bootstrap and Markov Chain Monte Carlo.
- (4) The development of techniques not describable in “closed mathematical form” but rather through elaborate algorithms for which problems of existence of solutions are important and far from obvious.
- (5) The study of the interplay between numerical and statistical considerations. Despite advances in computing speed, some methods run quickly in real time. Others do not and some though theoretically attractive cannot be implemented in a human lifetime.
- (6) The study of the interplay between the number of observations and the number of parameters of a model and the beginnings of appropriate asymptotic theories.

There have been other important consequences such as the extensive development of graphical and other exploratory methods for which theoretical development and connection with mathematics have been minimal. These will not be dealt with in our work.

In this edition we pursue our philosophy of describing the basic concepts of mathematical statistics relating theory to practice.

Volume I Outline

This volume presents the basic classical statistical concepts at the Ph.D. level without requiring measure theory. It gives careful proofs of the major results and indicates how the theory sheds light on the properties of practical methods. The topics include estimation, prediction, testing, confidence sets, Bayesian analysis and the more general approach of decision theory.

We include from the start in Chapter 1 non- and semiparametric models, then go to parameters and parametric models stressing the role of identifiability. From the beginning we stress function-valued parameters, such as the density, and function-valued statistics, such as the empirical distribution function. We also, from the start, include examples that are important in applications, such as regression experiments. There is extensive material on Bayesian models and analysis and extended discussion of prediction and k -parameter exponential families. These objects that are the building blocks of most modern models require concepts involving moments of random vectors and convexity that are given in Appendix B.

Chapter 2 deals with estimation and includes a detailed treatment of maximum likelihood estimates (MLEs), including a complete study of MLEs in canonical k -parameter exponential families. Other novel features of this chapter include a detailed analysis, including proofs of convergence, of a standard but slow algorithm (coordinate descent) for convex optimization, applied, in particular to computing MLEs in multiparameter exponential families. We also give an introduction to the EM algorithm, one of the main ingredients of most modern algorithms for inference. Chapters 3 and 4 are on the theory of testing and confidence regions, including some optimality theory for estimation as well and elementary robustness considerations.

Chapter 5 is devoted to basic asymptotic approximations with one dimensional parameter models as examples. It includes proofs of consistency and asymptotic normality and optimality of maximum likelihood procedures in inference and a section relating Bayesian and frequentist inference via the Bernstein–von Mises theorem.

Finally, Chapter 6 is devoted to inference in multivariate (multiparameter) models. Included are asymptotic normality and optimality of maximum likelihood estimates, inference in the general linear model, Wilks theorem on the asymptotic distribution of the likelihood ratio test, the Wald and Rao statistics and associated confidence regions, and some parallels to the optimality theory and comparisons of Bayes and frequentist procedures given in the one dimensional parameter case in Chapter 5. Chapter 6 also develops the asymptotic joint normality of estimates that are solutions to estimating equations and presents Huber’s Sandwich formula for the asymptotic covariance matrix of such estimates. Generalized linear models, including binary logistic regression, are introduced as examples. Robustness from an asymptotic theory point of view appears also. This chapter uses multivariate calculus in an intrinsic way and can be viewed as an essential prerequisite for the more advanced topics of Volume II.

Volume I includes Appendix A on basic probability and a larger Appendix B, which includes more advanced topics from probability theory such as the multivariate Gaussian distribution, weak convergence in Euclidean spaces, and probability inequalities as well as more advanced topics in matrix theory and analysis. The latter include the principal axis and spectral theorems for Euclidean space and the elementary theory of convex functions on R^d as well as an elementary introduction to Hilbert space theory. As in the first edition, we do not require measure theory but assume from the start that our models are what we call “regular.” That is, we assume either a discrete probability whose support does not depend on the parameter set, or the absolutely continuous case with a density. Hilbert space theory is not needed, but for those who know this topic Appendix B points out interesting connections to prediction and linear regression analysis.

Appendix B is as self-contained as possible with proofs of most statements, problems, and references to the literature for proofs of the deepest results such as the spectral theorem. The reason for these additions are the changes in subject matter necessitated by the current areas of importance in the field.

For the first volume of the second edition we would like to add thanks to Jianging Fan, Michael Jordan, Jianhua Huang, Ying Qing Chen, and Carl Spruill and the many students who were guinea pigs in the basic theory course at Berkeley. We also thank Faye Yeager for typing, Michael Ostland and Simon Cawley for producing the graphs, Yoram Gat for proofreading that found not only typos but serious errors, and Prentice Hall for generous production support.

Peter J. Bickel
bickel@stat.berkeley.edu
Kjell Doksum
doksum@stat.wisc.edu

Chapter I

INTRODUCTION AND EXAMPLES

I.0 Basic Ideas and Conventions

Recall from Volume I that in the field of statistics we represent important data-related problems and questions in terms of questions about distributions and their parameters. Thus our goal is to use data $X \in \mathcal{X}$ to estimate or draw conclusions about aspects of the probability distribution P of X . The probability distribution P is assumed to belong to a class \mathcal{P} of distributions called the *model*. Examining what models are *useful* for answering data-related questions is an important part of statistics. In Volume I we considered three cases with a focus on the first:

- (1) P is a member of a parametric class of distributions $\{P_\theta : \theta \in \Theta\}$, $\Theta \subset R^d$, and our interest is in θ or some vector $q(\theta)$.
- (2) P is arbitrary except for regularity conditions, such as finite second moments or continuity of the distribution function, and our interest is in functionals $\nu(P)$ that may be real valued, vectors, or functions.
- (3) Our class of distributions is neither smoothly parametrizable by a Euclidean parameter nor essentially unrestricted.

In Volume I we focussed mostly⁽¹⁾ on parametric cases and on situations where the number of parameters we were dealing with was small in at least one of two ways:

- (i) The complexity of the regular model $\{P_\theta : \theta \in \Theta\}$, $\Theta \subset R^d$, as measured by the dimension d of the parametrization, was small in relation to the amount of information, as measured by the sample size n of the data. In particular, when examining the properties of statistical procedures, d does not increase with n .
- (ii) The procedures we considered, estimation of low dimensional Euclidean parameters, testing, and confidence regions, corresponded to simple (finite or low dimensional) action spaces \mathcal{A} , where \mathcal{A} is the range of the statistical decision procedure.

In this volume we will focus on inference in non- and semiparametric models. In doing so, we will not only reexamine the procedures introduced in Volume I from a more sophisticated point of view but also come to grips with new problems originating from our analysis

of estimation of functions and other complex decision procedures that appear naturally in these contexts. The mathematics needed for this work is often of a higher level than that used in Volume I. But, as before, we present what is needed in the appendices with proofs or references.

Modeling Conventions

The guiding principle of modern statistics was best formulated by George Box (see Section 1.1.):

“Models, of course, are never true, but fortunately it is only necessary that they be useful”

One implication of this statement is that the parameters we deal with are the parameters of the distribution in our model class closest to the unknown true distribution. See Sections 2.2.2, 5.4.2, and 6.2.1. For instance, a linear regression model can detect linear trends that provide useful information even if the true population relationship is not linear. See Figure 1.4.1. This leads to an interesting dilemma and accompanying research questions: The more general a class we postulate the more closely we will be able to approximate the true population distribution. However, using a very general class of models means more parameters and more variability of statistical methods. Achieving a balance leads to useful models. One approach is to use a nested sequence of “regular” parametric models (sieves) that become more general as we add parameters and then select the number of parameters by minimizing estimated prediction error (cross validation). See Chapter 12.

As in Volume I (see Section 1.1.3), except for Bayesian models, our parametric models are restricted to be *regular parametric models*. $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$, $\Theta \subset R^d$, where P_θ is either continuous or discrete, and in the discrete case $\{x : p(x; \theta) > 0\}$ does not involve θ . But see Section I.5 for a general concept of regular and irregular parameters that includes semiparametric and nonparametric models.

As in our discussion of Bayesian models in Section 1.2, conditioning of continuous variables by discrete variables and vice versa generally preserves the interpretation of the conditional p as being a continuous or discrete case density, respectively. If $\mathbf{X} = (I, Y)^T$ where I is discrete and Y is continuous, then $p(i, y)$ is a density if it satisfies $P(I = i, Y \leq y) = \int_{-\infty}^y p(i, t)dt$. Readers familiar with measure theory will recognize that all results remain meaningful when $p = dP/d\mu$, where μ is a σ -finite measure dominating all P under discussion, and conditional densities are interpreted as being with respect to the appropriate conditional measure. All of the proofs can be converted to this general case, subject only to minor technicalities. Finally, we will write $h = 0$ when $h = 0$ a.s. (almost surely). More generally, a.s. equality is denoted as equality.

As in Volume I, throughout this volume, for $x \in R^d$ we shall use $p(x)$ interchangeably for frequency functions $p(x) = P[X = x]$ and for continuous case density functions $p(x)$. We will call $p(\cdot)$ a density function in both cases. When we write $\int h(x)dP(x)$ we will mean $\sum_x h(x)P[X = x]$ or $\int h(x)p(x)dx$. Unless we indicate otherwise, statements which can be interpreted under either interpretation are valid under both, although proofs in the text will be given under one formalism or the other. That is, when we write $\int h(x)p(x)dx$, for instance, we really mean $\int h(x)dP(x)$ as interpreted above. When $d = 1$, we let $F(x) = P(-\infty, x]$ and often write $\int h(x)dF(x)$ for $\int h(x)dP(x)$.

Selected Topics

Statistical methods in Volume II include the bootstrap, Markov Chain Monte Carlo (MCMC), Steinian shrinkage, sieves, cross-validation, censored data analysis, Cox proportional hazard regression, nonparametric curve (kernel) estimation, model selection, classification, prediction, classification and regression trees (CART), penalty estimation such as the Lasso, and Bayesian procedures. The effectiveness of statistical methods is examined using classical concepts such as risk, Bayes risk, mean squared error, power, minimaxity, admissibility, invariance, and equivariance. Statistical methods that are optimal based on such criteria are obtained in a finite sample context in Chapter 8. However, most of the book is concerned with asymptotic theory including empirical process theory and efficient estimation in semiparametric models as well as the development of the asymptotic properties of the statistical methods listed above. We present in Section I.1–I.7 a few more details of some of the topics in Volume II.

Notation

$X \sim F$, X is distributed according to F

statistic, a function of observable data \mathbf{X} only

$\mathcal{L}(X)$, the distribution, or law, of X

$X_n \Rightarrow X$, $\mathcal{L}(X_n) \rightarrow \mathcal{L}(X)$ X_n converges weakly (in law) to X

df, distribution function

J , identity matrix = $\text{diag}(1, \dots, 1)$

$\bar{F} = 1 - F$, the survival function

$[t]$, greatest integer less than or equal to t

i.i.d., independent identically distributed

sample, X_1, \dots, X_n i.i.d. as $X \sim F$

\widehat{P} and \widehat{P}_n , empirical probability of a sample X_1, \dots, X_n

$\mathcal{B}(n, \theta)$, binomial distribution with parameters n and θ

$\text{Ber}(\theta)$, Bernoulli distribution = $\mathcal{B}(1, \theta)$

$\mathcal{E}(\lambda)$, exponential distribution with parameter λ (mean $1/\lambda$)

$\mathcal{H}(D, N, n)$, hypergeometric distribution with parameters D, N, n

$\mathcal{M}(n, \theta_1, \dots, \theta_q)$, multinomial distribution with parameters $n, \theta_1, \dots, \theta_q$

$\mathcal{N}(\mu, \sigma^2)$, normal (Gaussian) distribution with mean μ and variance σ^2

$\varphi, \mathcal{N}(0, 1)$ density

$\Phi, \mathcal{N}(0, 1)$ df

z_α , α th quantile of Φ : $z_\alpha = \Phi^{-1}(\alpha)$

$\mathcal{N}(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho)$, bivariate normal (Gaussian) distribution

$\mathcal{N}(\mu, \Sigma)$, multivariate normal (Gaussian) distribution

$\mathcal{P}(\lambda)$, Poisson distribution with parameter λ

$\mathcal{U}(a, b)$, uniform distribution on the interval (a, b)

d.f., degrees of freedom

χ_k^2 , chi-square distribution with k d.f.

\equiv , defined to be equal to

\perp , orthogonal to, uncorrelated

$1(\cdot)$, indicator function

The O_P , \asymp_P , and o_P Notation

The following asymptotic order in probability notation is from Section B.7. Let \mathbf{U}_n and \mathbf{V}_n be random vectors in R^d and let $|\cdot|$ denote Euclidean distance.

$$\begin{aligned}
\mathbf{U}_n = o_P(1) &\quad \text{iff } \mathbf{U}_n \xrightarrow{P} 0, \text{ that is, } \forall \epsilon > 0, P(|\mathbf{U}_n| > \epsilon) \rightarrow 0 \\
\mathbf{U}_n = O_P(1) &\quad \text{iff } \forall \epsilon > 0, \exists M < \infty \text{ such that } \forall n \quad P[|\mathbf{U}_n| \geq M] \leq \epsilon \\
\mathbf{U}_n = o_P(\mathbf{V}_n) &\quad \text{iff } \frac{|\mathbf{U}_n|}{|\mathbf{V}_n|} = o_p(1) \\
\mathbf{U}_n = O_P(\mathbf{V}_n) &\quad \text{iff } \frac{|\mathbf{U}_n|}{|\mathbf{V}_n|} = O_P(1) \\
\mathbf{U}_n \asymp_P \mathbf{V}_n &\quad \text{iff } \mathbf{U}_n = O_P(\mathbf{V}_n) \quad \text{and} \quad \mathbf{V}_n = O_P(\mathbf{U}_n) \\
\mathbf{U}_n = \Omega_P(\mathbf{V}_n) &\quad \text{iff } \mathbf{U}_n \asymp_P \mathbf{V}_n
\end{aligned}$$

Note that

$$O_P(1)o_P(1) = o_P(1), \quad O_P(1) + o_P(1) = O_P(1), \tag{I.1}$$

and $\mathbf{U}_n \xrightarrow{L} \mathbf{U} \Rightarrow \mathbf{U}_n = O_P(1)$.

Suppose $\mathbf{Z}_1, \dots, \mathbf{Z}_n$ are i.i.d. as \mathbf{Z} with $E|\mathbf{Z}| < \infty$. Set $\boldsymbol{\mu} = E(\mathbf{Z})$, then $\bar{\mathbf{Z}}_n = \boldsymbol{\mu} + o_p(1)$ by the weak law of large numbers. If $E|\mathbf{Z}|^2 < \infty$, then $\bar{\mathbf{Z}}_n = \boldsymbol{\mu} + O_p(n^{-\frac{1}{2}})$ by the central limit theorem.

I.1 Tests of Goodness of Fit and the Brownian Bridge

Let X_1, \dots, X_n be i.i.d. as X with distribution P . For one dimensional observations, the distribution function (df) $F(\cdot) = P[X \leq \cdot]$ is a natural infinite dimensional parameter to consider. In Example 4.1.5 we showed how one could use the Kolmogorov statistic $T(\hat{F}, F_0)$, where $T(\hat{F}, F) \equiv \sup_{t \in R} |\hat{F}(t) - F(t)|$ and \hat{F} is the empirical df, to construct a test of the hypothesis $H : F = F_0$. The test is designed so that one can expect it to be consistent against all alternatives, so that our viewpoint is fully nonparametric. In Example 4.4.6 we showed how to construct a simultaneous confidence band for $F(\cdot)$ using the pivot $T(\hat{F}, F)$. In both cases we noted that the critical values needed for the test and confidence band could be obtained by determining the distribution of $T(\hat{F}, \mathcal{U})$, where \mathcal{U} is the $Unif[0, 1]$ distribution function under $F = Unif[0, 1]$, and stated that these values could be determined by Monte Carlo simulation.

How does $T(\hat{F}, F)$ behave qualitatively? We will show in Section 7.1 that, although infinite dimensional, $F(\cdot)$ is a “regular” parameter. In this case, what “regular” means is that the stochastic process,

$$\mathcal{E}_n(x) \equiv \sqrt{n} (\hat{F}(x) - F(x)), \quad x \in R, \quad (I.2)$$

converges in law in a strong sense (called “weak convergence!”) to a Gaussian process $W^0(F(\cdot))$. Here $W^0(u)$, $0 \leq u \leq 1$, is a Gaussian process called the “Brownian bridge” with mean 0 and covariance structure given by,

$$\text{Cov}(W^0(u_1), W^0(u_2)) = u_1(1 - u_2), \quad u_1 \leq u_2.$$

By “Gaussian” we mean that the distribution of $W^0(u_1), \dots, W^0(u_k)$ is multivariate normal for all u_1, \dots, u_k . Note that

$$\text{Cov}(W^0(F(x_1)), W^0(F(x_2))) = \text{Cov}(\mathcal{E}_n(x_1), \mathcal{E}_n(x_2)) = F(x_1)(1 - F(x_2)), \quad x_1 \leq x_2.$$

The weak convergence of $\mathcal{E}_n(\cdot)$ to $W^0(F(\cdot))$, to be established in Section 7.1, will enable us to derive the Kolmogorov theorem, that when $F = Unif[0, 1]$, $T(\hat{F}, F)$ converges in law to $\mathcal{L}(\sup\{|W^0(u)| : 0 \leq u \leq 1\})$, which is known analytically. This approach is based on heuristics due to Doob (1949) and developed in Donsker (1952). See also Doob (1953). We will discuss the heuristics in Section 7.1 and apply them to this and other examples in Section 7.2.

These results will provide approximate size α critical values for the Kolmogorov statistics and other interesting functionals of distribution functions. The critical values yield confidence regions for distribution functions and related parameters. See Examples 4.4.6, 4.4.7 and Problems 4.4.17–4.4.19, 4.5.14–4.5.16. \square

I.2 Testing Goodness of Fit to Parametric Hypotheses

In Examples 4.1.6 and 4.4.6 we considered the important problem of testing goodness-of-fit to a Gaussian distribution $H : F(\cdot) = \Phi(\frac{\cdot - \mu}{\sigma})$ for some μ, σ . We deduced that the

goodness-of-fit statistic

$$\sup_x |\hat{G}(x) - \Phi(x)|,$$

where \hat{G} is the empirical distribution function of (Z_1, \dots, Z_n) with $Z_i = (X_i - \bar{X})/\hat{\sigma}$, has a null distribution which does not depend on μ and σ , so that critical values can be calculated by simulating from $\mathcal{N}(0, 1)$. In Section 8.2, we will consider other classes of hypothesis models which admit reasonable tests whose critical values can be specified without knowledge of which particular hypothesized distribution is true. However, when we consider the general problem of testing $H : X \sim P \in \mathcal{P}$ where $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$ is a regular parametric model, we quickly come to situations where the methods of Chapter 8 will not apply. For instance, suppose that in the Gaussian goodness-of-fit problem above, our observations X_1, \dots, X_n which are i.i.d. as $F(x) = \Phi([x - \mu]/\sigma)$ are truncated at 0, that is, we assume that we observe Y_1, \dots, Y_n i.i.d. distributed as $X|X \geq 0$ where $X \sim \mathcal{N}(\mu, \sigma^2)$. Then, $P[0, \infty) = 1$ and H is

$$\begin{aligned} P[Y \leq t] \equiv G_{\mu, \sigma^2}(t) &= \frac{P(0 \leq X \leq t)}{P(X \geq 0)} = 1 - \left(\Phi\left(\frac{\mu-t}{\sigma}\right) / \Phi\left(\frac{\mu}{\sigma}\right) \right), & t \geq 0 \\ &= 0, & t < 0. \end{aligned}$$

The only promising approach here is to estimate μ and σ^2 consistently using, for instance, maximum likelihood or the method of moments (Problem I.2.1) by $\hat{\mu}$ and $\hat{\sigma}^2$ and estimate the null distribution of

$$T_n \equiv \sup_{t \geq 0} \sqrt{n} |\hat{F}(t) - G_{\hat{\mu}, \hat{\sigma}^2}(t)|$$

by simulating samples of size n from $\mathcal{N}(\hat{\mu}, \hat{\sigma}^2)$ truncated at 0, i.e., keep as observations only the nonnegative ones. But can this method, called the “parametric bootstrap,” be justified? To answer this question we need to consider asymptotics: It turns out that, under H , T_n converges in law to a limit. This helps us little in approximating the null distribution of T_n since an analytic form for its limiting distribution is not available. But, as we show in Section 9.4, such results are essential in justifying the parametric bootstrap.

The more important “nonparametric bootstrap” and other methods for simulating or approximately simulating observations from complicated distributions, often dependent on the data, such as Markov Chain Monte Carlo, are developed in Chapter 10.

I.3 Regular Parameters. Minimum Distance Estimates

We have seen in Chapters 5 and 6 how to establish asymptotic normality and approximate linearity of estimates that are solutions to estimation equations (M estimates) and then used these results to establish efficiency of the MLE under suitable conditions.

There are many types of estimates which cannot be characterized as solutions of estimating equations. Examples we have discussed are the linear combinations of order statistics, such as the trimmed mean introduced in Section 3.5,

$$\bar{X}_\alpha = (n - 2[n\alpha])^{-1} \sum_{i=[n\alpha]+1}^{n-[n\alpha]} X_{(i)}$$

where $X_{(1)} \leq \dots \leq X_{(n)}$ are the ordered X_i .

Here is another important class of such estimates. Suppose the data $X \in \mathcal{X}$ have probability distribution P . Let $\{P_\theta : \theta \in \Theta, \Theta \subset R^q\}$ be a regular parametric model. There may be a unique θ such that $P_\theta = P$, or if not, we choose the θ that makes P_θ closest to P in some metric d on the space of probability distributions on \mathcal{X} . For instance, if $\mathcal{X} = R$, examples of such metrics are

$$d_\infty(P, Q) = \sup_t |P(-\infty, t] - Q(-\infty, t]| \quad (\text{I.3})$$

and

$$d_2^2(P, Q) = \int_{-\infty}^{\infty} (P(-\infty, t] - Q(-\infty, t])^2 \psi(t) dt \quad (\text{I.4})$$

where $\psi(\cdot)$ is a nonnegative weight function of our choice with $\int_{-\infty}^{\infty} \psi(t) dt < \infty$.

A *minimum distance* estimate $\theta(\hat{P})$ is obtained by plugging the empirical distribution distribution \hat{P} into the parameter

$$\theta(P) = \arg \min \{d(P, P_\theta) : \theta \in \Theta\}$$

where we assume that $d(Q, P_\theta)$ is well defined for Q in \mathcal{M} , a general class of distributions containing all distributions with finite support. Thus \mathcal{M} contains the probability distribution P generating X and the empirical probability \hat{P} . See Problems 7.2.10 and 7.2.18 for examples and properties of minimum distance estimates $\theta(\hat{P})$. These problems show \sqrt{n} consistency of $\theta(\hat{P})$. They also show that $\theta(\hat{P})$ may not have a linear approximation in the sense of Section 7.2.1, and they may not be asymptotically normally distributed. Note that the minimum contrast estimates of Section 2.1 are of this form but, in this case, $d(Q, P_\theta) = \int \rho(x, \theta) dQ(x)$, which is not a metric, but is linear in Q , whereas metrics are not.

Can minimum distance estimates $\theta(\hat{P})$ be linearized in the sense of (6.2.3), and are they asymptotically Gaussian as we have shown M estimates to be in Section 6.2.1 and 6.2.2? When this is true asymptotic inference is simple as we have seen in Section 6.3. We have effectively studied this question for \mathcal{X} finite in Theorem 5.4.1. To do the general case, we need to extend the notion of Taylor expansion to function spaces, and apply so called maximal inequalities discussed in Section 7.1. In fact, we shall go further and examine function valued estimates such as the quantile function and study conditions under which these can be linearized and shown to be asymptotically Gaussian in the sense of weak convergence which will be rigorously defined in Section 7.1. Moreover, we want to conclude that asymptotic Gaussianity holds uniformly in a suitable sense. This is an important issue which we did not focus on in Volume I. As the Hodges Example 5.4.2 shows, it is possible to have estimates whose asymptotic behavior is not a good guide to the finite n case because of lack of uniformity of convergence when we vary the underlying distribution.

In Section 9.3 we will be concerned with regular parameters $\theta(P)$, ones whose plug in estimates $\theta(\hat{P})$ converge to $\theta(P)$ at rate $n^{-\frac{1}{2}}$ uniformly over a suitable subset \mathcal{M}_0 of a nonparametric family \mathcal{M} of probability distributions.

Definition I.1. $\theta(\hat{P})$ converges to $\theta(P)$ at rate δ_n over $\mathcal{M}_0 \subset \mathcal{M}$ iff for all $\varepsilon > 0$ there exists $c < \infty$ such that

$$\sup\{P[|\theta(\hat{P}) - \theta(P)| \geq c\delta_n] : P \in \mathcal{M}_0\} \leq \varepsilon.$$

There is another important set of questions having to do with the extension of the notion of efficient estimation from parametric to non- and semiparametric models. For instance, consider the parameter corresponding to a minimum contrast estimate

$$\nu(P) = \operatorname{argmin} \left\{ \int \rho(x, \boldsymbol{\theta}) dP(x) : \boldsymbol{\theta} \in \Theta \right\}.$$

Suppose that, as usual, we assume ν is defined for all $P \in \mathcal{M}$, a nonparametric model. Is there any estimate of $\nu(P)$ which behaves regularly and yet is able to achieve smaller asymptotic variance than the minimum contrast estimate $\nu(\hat{P})$ at some P in \mathcal{M} ? Note that this is not the same question as asking whether $\nu(\hat{P})$ is improvable by another such estimator $\eta(\hat{P})$ such that $\eta(P_{\boldsymbol{\theta}}) = \nu(P_{\boldsymbol{\theta}}) = \boldsymbol{\theta}$ on a parametric model $\{P_{\boldsymbol{\theta}} : \boldsymbol{\theta} \in \Theta\}$. Intuitively, $\nu(\hat{P})$ is to first order the only regular estimate of $\nu(P)$ on all of \mathcal{M} and should not be and indeed, is not improvable. Regularity is essential here to exclude the Hodges Example phenomena. We develop this theme further in the context of semiparametric as well as nonparametric models in Section 9.3.

I.4 Permutation Tests

In Section 4.9.3 we considered the Gaussian two-sample problem. We want to compare two samples X_1, \dots, X_{n_1} (control) and Y_1, \dots, Y_{n_2} (treatment) from distributions F and G and, in particular, want to test the hypothesis $H : F = G$ of no treatment effect. We studied the classical case in which F and G were $\mathcal{N}(\mu, \sigma^2)$, $\mathcal{N}(\mu + \Delta, \sigma^2)$, respectively. Then H becomes $\Delta = 0$ and we arrived at the classical two-sample t-test. In Example 5.3.7 we showed that, even if $F = G$ was not Gaussian, if $\int x^2 dF(x) < \infty$, the asymptotic level of the test is preserved as $n_1, n_2 \rightarrow \infty$. This can be thought of as a result for testing the hypothesis that the semiparametric model $\{(F, G) : F = G\}$ holds within the full nonparametric model $\{(F, G) : F, G \text{ arbitrary}, \int x^2 dF(x) < \infty, \int y^2 dG(y) < \infty\}$. But is it possible to construct tests with reasonable properties which have level $0 < \alpha < 1$ for fixed n_1, n_2 and all F, G ? The answer is yes. We shall study such *permutation tests* and their simple special subclass, the *rank tests*, in Sections 8.2 and 8.3 in the context of classes of composite semiparametric and parametric hypotheses $H : P \in \mathcal{P}_0 \subset \mathcal{P}$ which allow the construction of tests whose null distribution does not depend on where we are in \mathcal{P}_0 . We have, in fact, already seen examples of such parametric hypotheses in the Gaussian one- and two-sample problems with unknown variance.

I.5 Estimation of Irregular Parameters

We now show that the phenomena we encounter with irregular parameters are not simply a function of infinite dimension but rather manifest themselves as soon as the parameter

space complexity p , as measured naively by parameter space dimension, is comparable to the information in the data as measured naively by the sample size n . Although the distribution function is a reasonable object to study for $\mathcal{X} = R$, a much more visually informative parameter, even for this dimension and certainly for $\mathcal{X} = R^d$ with $d > 1$, is the density function, assuming that we postulate that only P which have continuous case densities p (in the usual sense) are considered. If $P \in \mathcal{P} = \{P_\theta : \theta \in \Theta\}$ is a regular parametric model and P_θ has density $p(\cdot, \theta)$, we can estimate the density by plugging in, say, $\hat{p} = p(\cdot, \hat{\theta})$ where $\hat{\theta}$ is the MLE. If \mathcal{P} is essentially nonparametric, there is no natural extension of the function valued parameter $\nu(P) \equiv p(\cdot)$ to all of the nonparametric class \mathcal{M} since, if \hat{P} denotes the empirical probability (2.1.15), $\nu(\hat{P})$, the density of \hat{P} , has no meaning. This leads to a strategy of “regularization” by which we first approximate $\nu(P)$ on \mathcal{P} by $\nu_n(P)$, which extends smoothly to $\tilde{\nu}_n$ on \mathcal{M} , i.e., for which $\tilde{\nu}_n(\hat{P})$ makes sense and yet $\nu_n(P) - \nu(P) \rightarrow 0$ in some uniform sense as $n \rightarrow \infty$. “Regularization” refers precisely to the change of estimation from the “irregular” ν to the “regular” ν_n . In essence, there is now a natural decomposition of the estimation error,

$$\tilde{\nu}_n(\hat{P}) - \nu(P) = (\tilde{\nu}_n(\hat{P}) - \nu_n(P)) + (\nu_n(P) - \nu(P)). \quad (\text{I.5})$$

The first term in parenthesis can be interpreted as the source of random variability, and the second as that of deterministic “bias.” We will loosely refer to this as the “bias-variance decomposition.” For instance, consider the usual *histogram estimate* of a one dimensional density $p(\cdot)$,

$$\hat{p}_h(t) = \hat{P}[I_j(t)]/h$$

where $I_j = (jh, (j+1)h]$, $-\infty < j < \infty$, and $I_j(t)$ is the unique I_j which contains t . Now $\hat{p}_h(t)$ is the plug-in estimate for the parameter,

$$p_h(t) \equiv P[I_j(t)]/h.$$

Of course, $p_h \neq p$ for $h > 0$ but if $h = h_n \downarrow 0$, then $p_h(t) \rightarrow p(t)$ for all t and $p_{h_n}(\cdot)$ is a sequence of approximating parameters. Now,

$$E(\hat{p}_h(t) - p_h(t)) = 0,$$

$$\text{Var } \hat{p}_h(t) = p_h(t)(1 - hp_h(t))/hn.$$

Thus, the “variance” part of the decomposition tends to 0 only if $h \rightarrow 0$ slower than n^{-1} . On the other hand, the rate of convergence of the “bias” part to 0

$$\text{BIAS}(h) \equiv \frac{1}{h} \int_{jh}^{(j+1)h} (p(s) - p(t))ds$$

is fastest when $h \rightarrow 0$ fastest. In fact, typically, at best $h^{-1}\text{BIAS}(h) \rightarrow c \neq 0$ (Problem I.5.1). So we see a tension present here in choosing the rate at which $h \rightarrow 0$, which is a new phenomenon whose consequences we shall investigate in Chapter 11 and 12.

Irregular parameters play a critical role in nonparametric regression, classification, and prediction which we shall also study in Chapters 11 and 12, as well as even in some aspects

of regular parameter estimation in semiparametric models. As mentioned in Section I.3, the distinction between regular and irregular is loose, roughly corresponding to the distinction between parameters which can at least asymptotically be estimated unbiasedly at rate $n^{-\frac{1}{2}}$ in the sense that $\text{MSE}(\hat{\theta}) \asymp n^{-\frac{1}{2}}$ for some $\hat{\theta}$, and for which the usual asymptotic Gaussian-based methods of inference apply straightforwardly; and those which cannot be treated in this way.

I.6 Stein and Empirical Bayes Estimation

For conceptual reasons, we consider the analysis of variance p -sample Gaussian model (Example 6.1.3), specified by $\mathbf{X} = \{X_{ij} : 1 \leq i \leq n, 1 \leq j \leq p\}$ where the X_{ij} are independent, $\mathcal{N}(\mu_j, \sigma_0^2)$, with $\boldsymbol{\mu}_p = (\mu_1, \dots, \mu_p)^T$ unknown. We write $\mathcal{P}(n, p)$ for this class of distributions for \mathbf{X} . The MLE of $\boldsymbol{\mu}_p$ is

$$\bar{\mathbf{X}}_p \equiv (X_{.1}, \dots, X_{.p})^T$$

where $X_{.j} \equiv n^{-1} \sum_{i=1}^n X_{ij}$. Evidently, $\bar{\mathbf{X}}_p \sim \mathcal{N}(\boldsymbol{\mu}_p, (\sigma_0^2/n)J)$ where $J_{p \times p}$ is the identity. Let our loss function be $l(\boldsymbol{\mu}_p, \mathbf{d}) = |\boldsymbol{\mu}_p - \mathbf{d}|^2/p$, where $|\mathbf{t}|$ is the Euclidean norm of the vector \mathbf{t} . Then, the MSE is,

$$R(\boldsymbol{\mu}_p, \bar{\mathbf{X}}_p) = \frac{1}{p} \sum_{j=1}^p E(X_{.j} - \mu_j)^2 = \frac{\sigma_0^2}{n}. \quad (\text{I.6})$$

We can show $\bar{\mathbf{X}}_p$ is minimax (Problem I.6.1) for each n and p and asymptotically efficient as $n \rightarrow \infty$ for p fixed. But, even if p is only ≥ 3 , a remarkable phenomenon discovered by Stein (1956(b)) occurs: $\bar{\mathbf{X}}_p$ is not admissible, whatever be n . That is, there exist minimax estimates $\delta^*(\bar{\mathbf{X}}_p)$ such that $R(\boldsymbol{\mu}_p, \delta^*(\bar{\mathbf{X}}_p)) < \sigma_0^2/n$ for all $\boldsymbol{\mu}_p$. A famous simple and intuitively reasonable estimate is *Stein's positive part estimate*,

$$\delta^*(\bar{\mathbf{X}}_p) = \left(1 - \frac{p-2}{|\bar{\mathbf{X}}_p|^2} \right)_+ \bar{\mathbf{X}}_p \quad (\text{I.7})$$

where if y is a scalar, $y_+ \equiv \max(y, 0)$. This estimate shrinks $\bar{\mathbf{X}}_p$ towards $\mathbf{0}$ and if the distance of $\bar{\mathbf{X}}_p$ from $\mathbf{0}$ is smaller than $\sqrt{p-2}$ declares the estimate to be $\mathbf{0}$.

This result can be made more plausible by considering why $\bar{\mathbf{X}}_p$ becomes a poor estimate as $p \rightarrow \infty$ for fixed n . Suppose n is fixed. Because $\bar{\mathbf{X}}_p$ is sufficient and normally distributed with independent components we can without loss of generality set $n = 1$. In this case we write X_1, \dots, X_p for the data. Put a prior distribution Π on R^p according to which the μ_i are i.i.d. with density π_0 on R . If π_0 is known, the posterior mean of (μ_1, \dots, μ_p) given (X_1, \dots, X_p) , which minimizes the Bayes risk, is

$$\delta_0(X_1, \dots, X_p) = (\delta_0(X_1), \dots, \delta_0(X_p))$$

where

$$\delta_0(x) = \frac{\int_{-\infty}^{\infty} \mu \phi\left(\frac{x-\mu}{\sigma_0}\right) \pi_0(\mu) d\mu}{\int_{-\infty}^{\infty} \phi\left(\frac{x-\mu}{\sigma_0}\right) \pi_0(\mu) d\mu} = x + \sigma_0^2 \frac{f'_0(x)}{f_0(x)} \quad (\text{I.8})$$

and

$$f_0(x) = \frac{1}{\sigma_0} \int_{-\infty}^{\infty} \phi\left(\frac{x-\mu}{\sigma_0}\right) \pi_0(\mu) d\mu,$$

the marginal density of X_1 (Problem I.6.2). The Bayes risk of this estimate is just

$$r(\pi_0, \delta_0) \equiv \sigma_0^2 [1 - \sigma_0^2 I(f_0)] < \sigma_0^2 \quad (\text{I.9})$$

where $I(f_0) = \int \{[f'_0(x)]^2 / f_0(x)\} dx$, the *Fisher information for location of f_0* (Problem I.6.4).

Suppose now that π_0 is unknown. We shall show, in Chapter 12, following Robbins (1956), that we can construct estimates \hat{f}'_0/\hat{f}_0 using X_1, \dots, X_p which, when plugged in to (I.8), yield a purely data dependent estimate $\hat{\delta}_0$, such that if $r(\pi_0, \delta)$ denotes Bayes risk (see Section 3.3), then $r(\pi_0, \hat{\delta}_0) \rightarrow r(\pi_0, \delta_0)$ as $p \rightarrow \infty$ for all π_0 . This strictly improves the performance of $\bar{\mathbf{X}}_p$ for $n = 1$, for all π_0 (but not uniformly). Here $\hat{\delta}_0$ is an example on an *empirical Bayes* estimate.

What if both p and n tend to ∞ ? There is no change in our conclusion that δ_0 is optimal if π_0 is fixed and known. An alternative and more informative analysis leads us to consider priors π_{0n} such that $\sqrt{n}\mu/\sigma_0$, the signal to noise ratio, stabilizes. The extent to which we can or want to try to estimate π_{0n} now depends on the family of priors. We shall consider this as well as related questions in the context of the so called Gaussian white noise model in Chapter 12.

The next section looks at this situation from a different point of view.

I.7 Model Selection

How complex should our model be? Usually this question can be reduced to asking how many parameters should be included in the model. We continue to consider the model $\mathcal{P}(n, p)$ of Section I.6. We identify the model with its parameter space R^p for (μ_1, \dots, μ_p) . Consider nested submodels $\mathcal{P}_t(n, p)$, $0 \leq t \leq p$, specified by $\omega_t = \{\boldsymbol{\mu}_p : \mu_{t+1} = \dots = \mu_p = 0\}$. Here t is unknown and is to be selected on the basis of the data \mathbf{X} . Consider the problem of estimating $\boldsymbol{\mu}_p$ as a vector with quadratic loss $l(\boldsymbol{\mu}_p, d)$. If we knew t and, in fact, that $\boldsymbol{\mu}_p \in \omega_t$, $t < p$ we could use $\delta_t(\mathbf{X}) = (X_{.1}, \dots, X_{.t}, 0, \dots, 0)^T$ and obtain

$$R(\boldsymbol{\mu}_p, \delta_t) = \frac{t\sigma_0^2}{n} < \frac{p\sigma_0^2}{n} = R(\boldsymbol{\mu}_p, \delta_p), \quad \boldsymbol{\mu}_p \in \omega_t$$

where R is the risk function

$$R(\boldsymbol{\mu}, \delta) = E|\delta(\mathbf{X}) - \boldsymbol{\mu}|^2$$

and $|\cdot|$ denotes Euclidean distance.

This seemingly artificial appearing situation is, in fact, a canonical model for a general linear model replicated n times:

$$Y_{ki} = \sum_{j=1}^p z_{kj} \beta_j + e_{ki}, \quad 1 \leq k \leq p, \quad 1 \leq i \leq n$$

where the e_{ki} are i.i.d. $\mathcal{N}(0, \sigma_0^2)$ and the z_{kj} are distinct covariate (predictor) values. Then, if $\hat{\beta}_p$ denotes the MLE,

$$\hat{\beta}_p \equiv (\hat{\beta}_1, \dots, \hat{\beta}_p)^T \sim \mathcal{N}_p(\beta_p, \frac{\sigma_0^2}{n} [\mathbf{Z}_p^T \mathbf{Z}_p]^{-1})$$

where $\beta_p = (\beta_1, \dots, \beta_p)^T$, $\mathbf{Z}_p = \|z_{kj}\|_{p \times p}$, and our model $\mathcal{P}(n, p)$ is the special case $\mathbf{Z}_p^T \mathbf{Z}_p = J_{p \times p}$. Our submodels $\{\beta_p : \beta_{t+1} = \dots = \beta_p = 0\}$ are natural if we think the p predictors or covariates Z_1, \dots, Z_p whose influence on the distribution of Y is governed by the β_j can be ordered in terms of importance and we expect $(p - t)$ of them to be independent of the response Y . In the context of the model $\mathcal{P}(n, p)$, the questions we address briefly now and more extensively in Chapter 12 are

- (1) Suppose $p = \infty$ (the possible number of predictors we can measure is “very large”) and we believe that $\mu \in \omega_t$ for some $t < \infty$. That is, $\mu \in \{\mu : \mu_j = 0, j > t\}$.

Can we, without knowledge of t , estimate t by \hat{t} so that, as $n \rightarrow \infty$,

$$E|\delta_{\hat{t}}(\mathbf{X}) - \mu|^2 = \frac{t\sigma_0^2}{n}(1 + o(1)) \quad (\text{I.10})$$

where $|\mathbf{a}|^2 = \sum_{j=1}^{\infty} a_j^2$ for $\mathbf{a} = (a_1, a_2, \dots)^T$, that is, can we asymptotically do as well not knowing t as knowing it? The optimal procedure for the case where t is known is called the *oracle* solution.

- (2) Suppose that all μ_j can be nonzero but $\sum_{j=1}^{\infty} \mu_j^2 < \infty$. Then, we can write the risk as

$$V_n(t, \mu) \equiv E|\delta_t(\mathbf{X}) - \mu|^2 = \frac{t\sigma_0^2}{n} + \sum_{j=t+1}^{\infty} \mu_j^2 \quad (\text{I.11})$$

There is clearly a best $t(\mu, n)$, one such that

$$t(\mu, n) = \arg \min_t V_n(t, \mu).$$

Note that $t(\mu, n) = \infty$, that is, estimating each μ_j by $X_{.j}$ is always a bad idea! Putting $t = 0$ will always do better. Can we select $\hat{t}(n)$ such that

$$\frac{E|\delta_{\hat{t}(n)}(\mathbf{X}) - \mu|^2}{V_n(t(\mu, n), \mu)} \rightarrow 1 \quad (\text{I.12})$$

as $n \rightarrow \infty$?

For question (1) we want (I.10) and (I.12) to hold uniformly over “moderately large” sets of μ . It is natural to consider \hat{t} of the form: \hat{t} is the largest k such that for a suitable decreasing sequence $\{c_n\}$ of positive numbers, $|X_{.j}| \leq c_n$ for all j such that $k+1 \leq j \leq n$. Suppose $p \leq n$. Finding the c_n such that this \hat{t} solves (I.10) then turns out to lead to a special case of the solution to Schwarz’s Bayes criterion (SBC) (1978) which also is called

the “Bayes information criterion” (BIC). Let P_{μ} denote computation under μ ; then take c_n such that

$$P_0[\max\{|X_{.j}| : 1 \leq j \leq n\} \geq c_n] \rightarrow 0 \quad (\text{I.13})$$

and

$$P_{\mu}[|X_{.j}| \leq c_n] \rightarrow 0. \quad (\text{I.14})$$

Since, for $j \geq t+1$, the $X_{.j}$ are i.i.d. $\mathcal{N}(0, \sigma_0^2/n)$ it is easy to see (Problem I.7.1) that

$$c_n = \sigma_0 \sqrt{(2 \log n)/n}$$

will achieve (I.10) without knowledge of t . To see this compute for $\mu \in \omega_t$,

$$\begin{aligned} P_{\mu}[\hat{t} \neq t] &= P_{\mu}[\hat{t} < t] + P_{\mu}[\hat{t} > t] \\ &\leq P_{\mu}[|X_t| \leq c_n] + P_{\mu}[\max\{|X_{.j}| : t+1 \leq j \leq n\} > c_n] \end{aligned} \quad (\text{I.15})$$

(Problem I.7.2). So, (I.13), (I.14), and (I.15) establish our answer to question (1).

The solution to question (2) is subtler and somewhat different and was first proposed in a time series context by Akaike (1969) and in the regression context by Mallows (1973). We will show that $\sum_{j=t+1}^n X_{.j}^2$ can be used to obtain an unbiased estimate of $V_n(t, \mu)$. Evidently

$$E_{\mu} \left(\sum_{j=t+1}^n X_{.j}^2 \right) = (n-t) \frac{\sigma_0^2}{n} + \sum_{j=t+1}^n \mu_j^2.$$

Therefore,

$$E_{\mu} \left(\sum_{j=t+1}^n X_{.j}^2 + 2t \frac{\sigma_0^2}{n} - \sigma_0^2 \right) = V_n(t, \mu),$$

and it seems plausible since σ_0^2 doesn't depend on t that minimizing the contrast

$$\rho_n(\mathbf{X}, k) \equiv \sum_{j=k+1}^n X_{.j}^2 + 2k \frac{\sigma_0^2}{n}, \quad 1 \leq k \leq n$$

will yield \hat{t} which achieves (I.12). This can be shown under suitable conditions (Shibata (1981)). However, it is interesting to note that this solution, which is called Mallows' C_p , is quite different from the SBC solution. Because

$$\rho_n(\mathbf{X}, j) - \rho_n(\mathbf{X}, j-1) = -X_{.j}^2 + 2 \frac{\sigma_0^2}{n}$$

then $\rho_n(X, j) \leq \rho_n(X, j-1)$ iff $(nX_{.j}^2/\sigma_0^2) \geq 2$. That is, we prefer j parameters to $j-1$ iff $\sqrt{n}|X_{.j}| \geq \sqrt{2}$. Thus, Mallows' C_p essentially uses a threshold of $\sqrt{2}$ on $\sqrt{n}|X_{.j}|/\sigma_0$

rather than the corresponding threshold $\sqrt{2 \log n} \rightarrow \infty$ for SBC. These methods and others will be discussed and contrasted with oracle solutions in Chapter 12. \square

Remark I.7.1. Mallows' C_p is usually discussed in the context of squared prediction error rather than squared estimation error. These two criteria are equivalent under general conditions; see Problem I.7.5.

Remark I.7.2. More generally we can consider risks other than those based on squared error. We shall do so in Chapter 12.

Remark I.7.3. This introductory chapter has barely touched on some of the main topics of Chapter 12. These topics include what is referred to as "Statistical Learning." See Section 12.1 for an introduction to this topic.

Summary We have in this chapter introduced, through examples, some of the main issues, topics, and procedures to be considered in this volume. One issue is the level of accuracy possible in the estimation of a parameter. In Section I.3, we illustrated *regular parameters*, which are parameters that are relatively easy to estimate in the sense that \sqrt{n} times the estimation error is well behaved as $n \rightarrow \infty$. We also discussed, in Section I.1, *Doob's heuristic*, which suggests how to approximate the distribution of a functional of a sample-based stochastic process by the distribution of the functional of the limiting stochastic process. Other methods for obtaining approximations, *Monte Carlo Methods* and *the bootstrap*, are illustrated in Section I.2. We also gave, in Section I.3, important parameters that are regular but nevertheless do not fall under the framework of Vol. I and discussed efficient estimation of parameters for a given model. In Section I.4 we discussed situations where we can construct tests, called *permutations tests*, whose significance level α remain the same for every member of a general class of distributions \mathcal{P}_0 . An important subclass of the permutation tests is the *rank tests*. In Section I.5 we illustrated *irregular parameters*, and the notion of plugging into approximations of irregular parameters such as densities. In this context we introduced the bias variance tradeoff. In Section I.6 we introduced Stein estimation, which, in our illustration with p Gaussian population means, consists of providing vector estimates with smaller average mean squared error than the vector of p sample means. We then connected this approach to the *empirical Bayes* approach where we solve the Bayes estimation problem with p Gaussian means assuming the prior is known and then use the data to estimate the unknowns in the Bayes procedure. Finally, in Section I.7 we considered model selection where the question is how many parameters should be included in a model used to analyze the results of an experiment. A complex model with many parameters may represent the true distribution of the data better than a simpler model with fewer parameters. However, we illustrated that trying to estimate too many parameters for the given sample size may result in worse inference than fitting a simpler model providing a less adequate approximation to the truth. This is another reflection of the bias-variance tradeoff discussed in Section I.5. We used a simple p -sample framework to outline the derivations of two procedures that give the most accurate result in the sense of minimizing average mean squared error for the problem of estimating a vector of means. The first method, which applies to the situation where all but a finite number of parameters are zero, turns out to be a special case of the method generated by *Schwarz's Bayes criterion* (also called BIC), while the

second method, which allows for an infinite sequence of positive means, yields a special case of *Mallows'* C_p .

I.8 PROBLEMS AND COMPLEMENTS

Problems for Section I.1

1. Let X_1, \dots, X_n be i.i.d. as $X \sim F$, let $\hat{F}(x) = n^{-1} \sum_{i=1}^n 1[X_i \leq x]$ be the empirical distribution function, and let $\mathcal{E}_n(x) = \sqrt{n}[\hat{F}(x) - F(x)]$, $x \in R$.

- (a) Show that for fixed x_0 , $\mathcal{E}_n(x_0) \xrightarrow{\mathcal{L}} \mathcal{N}(0, F(x_0)[1 - F(x_0)])$
- (b) Find $\text{Cov}(\mathcal{E}_n(x_1), \mathcal{E}_n(x_2))$ for $x_1 \leq x_2$.
- (c) Find the limiting law of $(\mathcal{E}_n(x_1), \mathcal{E}_n(x_2))$ for $x_1 \leq x_2$ using the bivariate central limit theorem.

2. Let X_1, \dots, X_m be i.i.d. as $X \sim F$, Y_1, \dots, Y_n i.i.d. as $Y \sim G$, and let the X 's and Y 's be independent (see Example 1.1.3). Let $\hat{F}(\cdot)$ and $\hat{G}(\cdot)$ be the empirical *df*'s, define $N = m + n$, and

$$\mathcal{E}_N(t) = \sqrt{\frac{mn}{N}} \{ \hat{G}(t) - \hat{F}(t) - [G(t) - F(t)] \}$$

Assume that $\frac{m}{N} \rightarrow \lambda$ as $N \rightarrow \infty$ with $0 < \lambda < 1$.

- (a) Show that for fixed t_0 , as $N \rightarrow \infty$,

$$\mathcal{E}_N(t_0) \xrightarrow{\mathcal{L}} \sqrt{\lambda} Z_1(t_0) + \sqrt{(1-\lambda)} Z_2(t_0)$$

where $Z_1(t_0)$ and $Z_2(t_0)$ are independent with $Z_1(t_0) \sim \mathcal{N}(0, G(t_0)[1 - G(t_0)])$ and $Z_2(t_0) \sim \mathcal{N}(0, F(t_0)[1 - F(t_0)])$.

- (b) Describe the weak limit of $\mathcal{E}_N(t)$.

Problems for Section I.2

1. Let X_1, \dots, X_n be i.i.d. as $X \sim F(x) = \Phi([x - \mu]/\sigma)$. Suppose we observe Y_1, \dots, Y_m i.i.d. as $Y \sim (X | X \geq 0)$. Let $\mu = E(X)$, $\sigma^2 = \text{Var}(X)$, and $\theta = (\mu, \sigma^2)^T$.

- (a) Express $E(Y)$ and $\text{Var}(Y)$ as functions of θ . By replacing $E(Y)$ and $\text{Var}(Y)$ by \bar{Y} and $\hat{\sigma}_Y^2$ in these two equations and solving them for θ numerically, we have method of moment estimates of μ and σ^2 .
- (b) Use Proposition 5.2.1 to outline an argument for the consistency of the estimates of μ and σ^2 in part (a).

- (c) Use Theorem 5.2.3 to outline an argument for the consistency of the MLE of θ .

Problems for Section I.5

1. Suppose p is positive and Lipschitz continuous over I_j , that is, for some $\gamma > 0$ and all $x, z \in I_j$, $|p(x) - p(z)| \leq \gamma|x - z|$. Show that

- (a) $\text{BIAS } \hat{p}_h(t) = O(h)$, all $t \in I_j$.
- (b) $\text{Var } \hat{p}_h(t) = O((nh)^{-1})$, all $t \in I_j$.
- (c) If $p(x)$ is not constant on I_j , then for some constant $c > 0$,

$$c < \inf_{h>0} h^{-1} |\text{BIAS } \hat{p}_h(t)| \leq \sup h^{-1} |\text{BIAS } \hat{p}_h(t)| \leq c^{-1} .$$

(d) Let t be as in (c) above. Assume that p' exists and that $0 < |p'(x)| \leq M < \infty$ for x in a neighbourhood of t . Show that

$$\inf_h \text{MSE } \hat{p}_h(t) = \text{Var } \hat{p}_h(t) + [\text{BIAS } \hat{p}_h(t)]^2 \asymp (n^{-\frac{2}{3}}) ,$$

where $A_n \asymp B_n$ iff $A_n = O(B_n)$, $B_n = O(A_n)$.

Hint (a) and (b). By the mean value theorem, for some $x_0 \in I_j$,

$$P[I_j(t)] = \int_{I_j} p(x)d(x) = hp(x_0) .$$

Now show that

$$|\text{BIAS } \hat{p}_h(t)| \leq \gamma|x_0 - t| \leq \gamma h, \quad \text{Var } \hat{p}_h(t) \leq \frac{p(x_0)}{nh} .$$

Hint (d). Show that $\text{MSE } \hat{p}_n(t) = b(nh)^{-1} + ch^2$ plus smaller order terms for some constants b and c . Now minimize $A(h) \equiv b(nh)^{-1} + ch^2$ with respect to h .

Problems for Section I.6

1. Show that $\bar{\mathbf{X}}_P$ is the minimax estimate for the model $\mathcal{P}(n, p)$ and the risk (I.6).

Hint. See Example 3.3.3.

2. Derive formula (I.8).

Hint. The first equality is from (3.2.6).

3. Show that the Bayes risk for estimating $q(\theta)$ with quadratic loss when θ has prior π is

$$E\text{Var}(q(\theta|\mathbf{X})) = \int [q(\theta) - \delta^*(\mathbf{x})]^2 \pi(\theta) p(\mathbf{x}, \theta) d\theta d\mathbf{x} ,$$

where $\delta^*(\mathbf{x})$ is the Bayes estimate of $q(\theta)$.

4. Derive formula (I.9).

Hint. Use (I.8), (5.4.32), and Problem I.6.3.

Problems for Section I.7

1. Show that if X_1, \dots, X_n are i.i.d. $\mathcal{N}(0, 1)$, then

$$P[\max(X_1, \dots, X_n) \leq \sqrt{2 \log n}] \rightarrow 1.$$

Hint. Use the following refinement of (7.1.9) (Feller (1968), p. 175)

$$(x^{-1} - x^{-3})\phi(x) \leq (1 - \Phi(x)) \leq x^{-1}\phi(x) \quad \text{for all } x > 0.$$

2. Verify (I.15).

Hint. If Z_1, \dots, Z_n are i.i.d. $\mathcal{N}(0, 1)$, $P[\max\{|Z_i| : 1 \leq i \leq n\} \geq z] \leq n P[|Z_1| \geq z]$. Use $1 - \Phi(z) \leq \phi(z)/z$ for all $z > 0$.

3. *Selecting the model to minimize MSE. The t-test revisited.* Consider the simple linear regression model (e.g. Example 2.2.2)

$$Y_i = \beta_0 + \beta_1 z_i + \varepsilon_i, \quad i = 1, \dots, n \tag{I.8.1}$$

where $\varepsilon_1, \dots, \varepsilon_n$ are i.i.d. as $\varepsilon \sim \mathcal{N}(0, \sigma^2)$ and the z_i 's are not all equal. Without loss of generality we assume $\sum z_i = 0$ and $\beta_0 = E(\bar{Y})$. We are interested in estimating

$$\mu_i \equiv \mu(z_i) \equiv E(Y_i) = \beta_0 + \beta_1 z_i, \quad i = 1, \dots, n.$$

In this problem we assume that (I.8.1) is the true model. Even so, it may be that the MLE $\hat{\mu}_{10} = \hat{\beta}_0 = \bar{Y}$ based on the submodel with $\beta_1 = 0$ has smaller risk than the MLE $\hat{\mu}_{11} \equiv \bar{Y} + \hat{\beta}_1 z_i$ based on the full model. For any estimate $\hat{\mu} = (\hat{\mu}_1, \dots, \hat{\mu}_n)^T$ we define the risk for estimating $\mu = (\mu_1, \dots, \mu_n)^T$ as

$$R(\mu, \hat{\mu}) = \frac{1}{n} E|\hat{\mu} - \mu|^2 = \frac{1}{n} \sum_{i=1}^n E|\hat{\mu}_i - \mu_i|^2$$

where the expectation is computed for the full model (I.8.1).

(a) Show that for $\hat{\mu}_j = (\hat{\mu}_{j1}, \dots, \hat{\mu}_{jn})^T$,

$$R(\mu, \hat{\mu}_j) = \frac{(1+j)\sigma^2}{n} + \frac{(1-j)\beta_1^2 \sum_{i=1}^n z_i^2}{n}, \quad j = 0, 1 \tag{I.8.2}$$

Hint. See Example 6.1.2, pages 381–382.

(b) Show that unbiased estimates of $R(\mu, \hat{\mu}_j)$, $j = 0, 1$, are

$$\begin{aligned} \hat{R}(\mu, \hat{\mu}_0) &\equiv n^{-1}[RSS_0 - RSS_1] = n^{-1}|\hat{\mu}_1 - \hat{\mu}_0|^2 = \hat{\beta}_1^2 \left(\frac{\sum_{i=1}^n z_i^2}{n} \right) \\ \hat{R}(\mu, \hat{\mu}_1) &= 2s^2/n \end{aligned}$$

where $RSS_j = \sum_{i=1}^n [Y_i - \hat{\mu}_{ji}]^2$, $j = 0, 1$, and $s^2 = RSS_1/(n - 2)$.

Hint. See Section 6.1.3.

- (c) Show that the model selection rule that decides to keep β_1 in the model iff $\widehat{R}(\boldsymbol{\mu}, \widehat{\boldsymbol{\mu}}_1) < \widehat{R}(\boldsymbol{\mu}, \widehat{\boldsymbol{\mu}}_0)$ is equivalent to a likelihood ratio test of $H : \beta_1 = 0$ vs $K : \beta_1 \neq 0$. Show that the rule and the test can be written as

“Keep β_1 in the model iff $t^2 > 2$ ”

where $t = \widehat{\beta}_1 / (s/\sqrt{n}\widehat{\sigma}_z)$ with $\widehat{\sigma}_z^2 = n^{-1} \sum_{i=1}^n z_i^2$ is the t -statistic for regression.

Hint. See Example 6.1.2, pages 381–382.

- (d) Show that the limit as $n \rightarrow \infty$ of the significance level of the test in (c) equals $P(|Z| > \sqrt{2}) = 0.16$ where $Z \sim \mathcal{N}(0, 1)$. That is, we decide to keep β_1 if the p -value is less than 0.16. Show this result without the normality assumption. Instead assume that $\varepsilon_1, \dots, \varepsilon_n$ are i.i.d. as ε with $E(\varepsilon) = 0$, $E(\varepsilon^2) = \sigma^2$, $E(\varepsilon^4) < \infty$. Also assume $\max_i \{z_i^2 / \Sigma z_i^2\} \rightarrow 0$.

Hint. Use Lindeberg-Feller and Slutsky’s theorems. See Example 5.3.3.

- (e) Using (a), show that $R(\boldsymbol{\mu}, \widehat{\boldsymbol{\mu}}_1) < R(\boldsymbol{\mu}, \widehat{\boldsymbol{\mu}}_0)$ iff $\tau^2 > 1$ where $\tau = \beta_1 / (\sigma/\sqrt{n}\widehat{\sigma}_z)$ is the noncentrality parameter of the distribution of the t -statistic.

Hint. See page 260, Section 4.9.2.

- (f) Show that $\widehat{\tau}^2 \equiv ct^2 - 1$ with $c = (n-4)/(n-2)$ is an unbiased estimate of τ^2 . Using $\widehat{\tau}^2$, deciding in favor of keeping β_1 is now equivalent to $t^2 > 2(n-2)/(n-4)$ for $n \geq 5$.

Hint. $t \stackrel{D}{=} (Z + \tau)/\sqrt{V/(n-2)}$ where $Z \sim \mathcal{N}(0, 1)$, $V \sim \chi_{n-2}^2$, and Z and V are independent. Use Problem B.2.4.

- (g) Show that $\widehat{R}(\boldsymbol{\mu}, \widehat{\boldsymbol{\mu}}_1) < \widehat{R}(\boldsymbol{\mu}, \widehat{\boldsymbol{\mu}}_0)$ is also equivalent to $r^2 > 2/n$, where r^2 is the sample correlation coefficient and we assume $n \geq 3$. Also show $t^2 \geq 2(n-2)/(n-4)$ iff $r^2 \geq 2/(n-2)$.

Hint. $t^2 = n\widehat{\beta}_1^2 \widehat{\sigma}_z^2 / RSS_1$ and $r^2 = \widehat{\beta}_1^2 \widehat{\sigma}_z^2 / \widehat{\sigma}_Y^2$ where $\widehat{\sigma}_Y^2 = n^{-1} \sum_{i=1}^n (Y_i - \bar{Y})^2$. Now use the ANOVA decomposition $n\widehat{\sigma}_Y^2 = RSS_1 + \widehat{\beta}_1^2 n\widehat{\sigma}_z^2$ (see Section 6.1.3).

4. *Selecting the model to minimize MSE. The F-test revisited.* Consider the linear model (see 6.1.26)

$$\mathbf{Y} = \mathbf{Z}_1 \boldsymbol{\beta}_1 + \mathbf{Z}_2 \boldsymbol{\beta}_2 + \boldsymbol{\varepsilon} \quad (I.8.3)$$

where \mathbf{Z}_1 is $n \times q$ and \mathbf{Z}_2 is $n \times (p-q)$, $\boldsymbol{\beta}_1$ is $q \times 1$, $\boldsymbol{\beta}_2$ is $(p-q) \times 1$, and $\boldsymbol{\varepsilon} = (\varepsilon_1, \dots, \varepsilon_n)^T$ with $\varepsilon_1, \dots, \varepsilon_n$ i.i.d. as $\varepsilon \sim \mathcal{N}(0, \sigma^2)$. The entries of \mathbf{Z}_1 and \mathbf{Z}_2 are constants. Assume that $\mathbf{Z} \equiv (\mathbf{Z}_1, \mathbf{Z}_2)$ has rank p . We are interested in estimating the effect of the covariates on the response mean $E(Y)$. Thus our parameters are

$$\mu_i \equiv \mu(\mathbf{z}_i) \equiv E(Y_i) = \mathbf{z}_i^{(1)} \boldsymbol{\beta}_1 + \mathbf{z}_i^{(2)} \boldsymbol{\beta}_2, \quad i = 1, \dots, n$$

where $\mathbf{z}_i^{(j)}$ is the i th row of \mathbf{Z}_j , $j = 1, 2$. It may be that the coefficients in $\boldsymbol{\beta}_2$ are so small that in the bias-variance tradeoff we get more efficient estimates of the μ_i ’s if we use the

model with $\beta_2 = \mathbf{0}$. Thus we want to compare the risks of the two estimates $\hat{\mu}_0 = H_1 \mathbf{Y}$ and $\hat{\mu} = H \mathbf{Y}$ where H_1 and H are the hat matrices (e.g., $H_1 = \mathbf{Z}_1 (\mathbf{Z}_1^T \mathbf{Z}_1)^{-1} \mathbf{Z}_1^T$) for the models with $\beta_2 = \mathbf{0}$ and general $\beta = (\beta_1^T, \beta_2^T)^T$, respectively. The risk for estimating $\mu = (\mu_1, \dots, \mu_n)^T$ is

$$R(\mu, \hat{\mu}) = n^{-1} E|\hat{\mu} - \mu|^2 = n^{-1} \sum_{i=1}^n E[\hat{\mu}_i - \mu_i]^2$$

where the expectation is computed for the full model.

- (a) Set $\mu_0 = H_1 \mu$. Show that

$$R_q \equiv R_q(\mu, \hat{\mu}_0) = \frac{q\sigma^2}{n} + \frac{|\mu - \mu_0|^2}{n}, \quad 1 \leq q \leq p.$$

Note that when $q = p$, then $\hat{\mu}_0 = \hat{\mu}$, and $R_p(\mu, \hat{\mu}) = p\sigma^2/n$.

Hint. See (6.1.15).

- (b) (i) Let $s^2 = [n - (p + 1)]^{-1} \sum [Y_i - \hat{Y}_i]^2$ where \hat{Y}_i is the predicted value of Y_i in the full model (I.8.3). Show that an unbiased estimate of $R_p - R_q$ for model (I.8.3) is

$$\hat{R}_p - \hat{R}_q = \frac{2(p - q)s^2}{n} - \frac{|\hat{\mu} - \hat{\mu}_0|^2}{n}.$$

- (ii) Show that deciding to keep $Z_2 \beta_2$ in the model when $\hat{R}_p < \hat{R}_q$ is equivalent to deciding $\beta_2 = \mathbf{0}$ when $F > 2(p - q)$ where “ $F > 2(p - q)$ ” is a likelihood ratio test of $H : \beta_2 = \mathbf{0}$ vs $K : \beta_2 \neq \mathbf{0}$.

Hint. See Example 6.1.2.

- (c) (i) Use (a) to show that $R_p < R_q$ is equivalent to $\theta^2 > p - q$, where $\theta^2 = \sigma^{-2} |\mu - \mu_0|^2$ is the noncentrality parameter of the distribution of the F -statistic of Example 6.1.2.
- (ii) Show that $\hat{\theta}^2 = (n - p)^{-1} (p - q)(n - p - 2)F - (p - q)$ is an unbiased estimate of θ^2 . Thus, using $\hat{\theta}^2 > (p - q)$, we select the full model iff

$$F > \frac{(p - q + 1)(n - p)}{(p - q)(n - p - 2)}, \quad n > p + 2.$$

Hint. By Problem 8.3.13, $F = [(p - q)^{-1} (Z_1 + \theta)^2 + \sum_{i=1}^{p-q} Z_i^2 (n - p)^{-1} V^{-1}]$ where $Z_i \sim \mathcal{N}(0, 1)$, $V \sim \mathcal{X}_{n-p}^2$, and Z_1, \dots, Z_{p-q} , V are independent.

- 5. Prediction and estimation are equivalent for squared error.** Assume that $E(Y_i) = \mu_i$ and $\text{Var}(Y_i) = \sigma_i^2$ where μ_i depends on a vector \mathbf{z}_i of covariates while σ_i^2 does not, $i = 1, \dots, n$. Let $\mathbf{Y} = (Y_1, \dots, Y_n)^T$, $\mu = (\mu_1, \dots, \mu_n)^T$, and let $\hat{\mu}$ be any estimate of μ based on \mathbf{Y} with $0 < \text{Var}\hat{\mu}_i < \infty$ for $i = 1, \dots, n$. Let Y_1^0, \dots, Y_n^0 be variables to be predicted using Y_1, \dots, Y_n . We assume that $\mathbf{Y}^0 = (Y_1^0, \dots, Y_n^0)^T$ is independent of \mathbf{Y} ,

and that Y_i^0 has the same mean and variance as Y_i , $i = 1, \dots, n$. We will use $\hat{\mu}$ to predict \mathbf{Y}^0 and define the mean squared prediction error as

$$\text{MSPE} = n^{-1} E|\hat{\mu} - \mathbf{Y}^0|^2.$$

Show that selecting the model (covariates, as in Problem 4 preceding) that minimizes MSPE is equivalent to selecting the model (covariates) that minimizes the mean squared error MSE = $n^{-1} E|\hat{\mu} - \mu|^2$.

Hint. $\hat{\mu}_i - Y_i^0 = [\mu_i - Y_i^0] + [\hat{\mu}_i - \mu_i]$. Complete the square keeping the square brackets intact. Because Y_i^0 is independent of $\hat{\mu}_i$, we have

$$E[\hat{\mu}_i - Y_i^0]^2 = \sigma_i^2 + E[\hat{\mu}_i - \mu_i]^2. \quad (\text{I.8.4})$$

Note that the equivalence fails if $\sum_{i=1}^n \sigma_i^2$ depends on the covariates.

I.9 Notes

Notes for Section I.1.

- (1) Important exceptions are Sections 1.4 and 6.6.

Chapter 7

TOOLS FOR ASYMPTOTIC ANALYSIS

In this chapter we will present the basic tools needed for most of our subsequent chapters. These tools include empirical process theory and maximal inequalities, the delta method in infinite dimensional space, influence functions, functional derivatives, and Taylor type expansions in function space including the von Mises and Hoeffding expansions. These tools are essential for developing asymptotic inference for semiparametric models. Some of the proofs and references will be deferred to Appendix D.

7.1 Weak Convergence in Function Spaces

7.1.1 Stochastic Processes and Weak Convergence

As we discussed in Section I.1, when our inference involves estimates of functions such as the distribution F and functionals $\nu(F)$ such as $\nu(F) = \sup_x |F(x) - F_0(x)|$, we are faced with problems involving the convergence of stochastic processes such as

$$\{\sqrt{n}[\hat{F}(x) - F_0(x)] : x \in R\}.$$

We now turn to this subject. Our treatment largely parallels that of van der Vaart and Wellner (1996). Our story has to do with sequences of collections of random variables $\{Z(t) : t \in T\}$, also written as $\{Z(\cdot)\}$ or $\{Z(t; w) : t \in T, w \in \Omega\}$, defined on a suitable probability space (Ω, \mathcal{A}, P) , which have a type of approximability property called *separability* by Doob (1953). Specifically, $\{Z(t) : t \in T\}$ is *separable* if there exists a countable set $C \equiv \{t_j \in T : j \geq 1\}$ such that for any open interval $(a, b) \subset R$ and any $S \subset T$,

$$\{\omega \in \Omega_0 : Z(t; w) \in (a, b), \text{ all } t \in S\} = \{\omega \in \Omega_0 : Z(t; w) \in (a, b), \text{ all } t \in S \cap C\} \quad (7.1.1)$$

where $\Omega_0 = \Omega - N$ with N a set in \mathcal{A} with $P(N) = 0$. See Doob (1953) or van der Vaart and Wellner (1996) for more details.

The reason we need to introduce separability is that if T is uncountable, the distribution of functionals such as $\sup\{Z(t) : t \in T\}$ is not determined by the finite dimensional distributions. For instance, suppose $Z(t) = c > 0$ for all $0 \leq t \leq 1$. Let τ be uniform on $T = [0, 1]$ independent of $Z(\cdot)$ and let

$$\begin{aligned} Z^*(t) &= Z(t), \quad t \neq \tau \\ Z^*(\tau) &= 2c. \end{aligned}$$

Then $Z^*(t) = c$ with probability 1, but $\sup\{Z^*(t) : t \in T\}$.

Doob essentially shows that for certain collections of finite dimensional distributions $\mathcal{L}_{t_1, \dots, t_k}$, $t_j \in T$, $1 \leq j \leq k$, there is, on a suitable probability space, a separable stochastic process $\{Z(t) : t \in T\}$ with the same finite dimensional distributions — see Doob (1953) for an extensive discussion.

What (7.1.1) means in addition to guaranteeing that we can rigorously talk about the probability of the left hand side of (7.1.1), is that if, for all $\{t_{i_1}, \dots, t_{i_m}\} \subset S \cap C$, all $m \geq 1$,

$$P[Z(t_{i_j}) \in (a, b), 1 \leq j \leq m] \geq 1 - \epsilon,$$

then

$$P[Z(t) \in (a, b) \text{ for all } t \in S] \geq 1 - \epsilon.$$

We shall call a collection $\{Z(t) : t \in T\}$ satisfying (7.1.1) a *stochastic process*. An important example of a sequence of processes which we have already encountered in Section I.1 is the *classical empirical process*,

$$\mathcal{E}_n(t) \equiv \sqrt{n} (\widehat{F}(t) - F(t)), \quad t \in R,$$

where \widehat{F} is the empirical df of X_1, \dots, X_n , i.i.d. F . Note, however, that this does not mean that we restrict T to be discrete or Euclidean⁽¹⁾.

Definition 7.1.1. *Finite Dimensional Convergence.* A sequence of stochastic processes $Z_n(\cdot)$ converge *FIDI* to the stochastic process $Z(\cdot)$ ($Z_n(\cdot) \xrightarrow{\text{FIDI}} Z(\cdot)$) iff

$$\mathcal{L}(Z_n(t_1), \dots, Z_n(t_k)) \rightarrow \mathcal{L}(Z(t_1), \dots, Z(t_k)) \tag{7.1.2}$$

as $n \rightarrow \infty$ for all $\{t_1, \dots, t_k\} \in T$, all $k < \infty$.

In Definition 7.1.1 $Z(\cdot)$ is not uniquely defined by (7.1.2) but we assume that a particular $Z(\cdot)$ is given.

As we noted in Section I.1, $\mathcal{E}_n(\cdot) \xrightarrow{\text{FIDI}} W^0(F(\cdot))$, where $\{W^0(u) : 0 \leq u \leq 1\}$ is the Brownian bridge defined in Section I.1.

As the notation suggests, it is possible and important to think of any $Z(\cdot)$ as an abstract valued random quantity, at the very least $Z(\cdot) \in \mathcal{F}(T)$, the set of all real valued functions on T . More precisely, if points of the set Ω , on which $Z(\cdot)$ are defined, are denoted by ω , then, for each $\omega \in \Omega$, $Z(\cdot; \omega)$ is a function on T . If T is Euclidean, it is customary to speak of a realization of $Z(\cdot)$ as a *sample function*. It is clear that $\mathcal{E}_n(\cdot)$ has bounded sample

functions. It is far less obvious but true (Problem D.2.3) that $W^0(\cdot)$ has sample functions which are not only bounded but continuous.

$$P[\omega : W^0(\cdot; \omega) \text{ is continuous}] = 1. \quad (7.1.3)$$

Much more is known about $W^0(\cdot)$ including the distribution of random variables such as $\sup\{|W^0(u)| : 0 \leq u \leq 1\}$, $\int_0^1 [W^0]^2(u) du$, and $\int_0^1 W^0(u) a(u) du$ (Problem 7.1.11). A discussion of some of its properties and the closely related *Brownian motion* or *Wiener process* $W(\cdot)$ defined in Section 7.1.2 is given in Billingsley (1968) and Shorack and Wellner (1986).

Here is an example which is important for asymptotic inference.

Example 7.1.1. The log likelihood process. Let X_1, \dots, X_n be i.i.d. P , where $P \in \mathcal{P} = \{P_\theta : \theta \in R\}$, a regular 1 dimensional model satisfying conditions A0–A4 and A6 of Section 5.4.2 applied to $\psi = \partial l / \partial \theta$. Define as usual,

$$l_n(\theta) = \sum_{i=1}^n l(X_i, \theta),$$

where $l \equiv \log p$ and $p = p(\cdot, \theta)$ is the density of X under P_θ . Define for all $t \in R$,

$$Z_n(t) = l_n(\theta_0 + \frac{t}{\sqrt{n}}) - l_n(\theta_0)$$

with P_{θ_0} being the true underlying probability. Then, for all $t \in R$, by Problem 5.4.5, as $n \rightarrow \infty$,

$$Z_n(t) = t Z_n^0 - \frac{t^2}{2} I(\theta_0) + o_{P_{\theta_0}}(1) \quad (7.1.4)$$

where $I(\theta_0)$ is Fisher information and $Z_n^0 = n^{-\frac{1}{2}} \sum_{i=1}^n l'(X_i, \theta_0)$ with $l' = \partial l / \partial \theta$. By the central limit theorem and Slutsky's theorem, for all t_1, \dots, t_k ,

$$(Z_n(t_1), \dots, Z_n(t_k)) \xrightarrow{\mathcal{L}} (Z(t_1), \dots, Z(t_k))$$

where

$$Z(t) \equiv t Z^0 - \frac{t^2}{2} I(\theta_0) \quad (7.1.5)$$

and $Z^0 \sim \mathcal{N}(0, I(\theta_0))$. Therefore, $Z_n(\cdot) \xrightarrow{FIDI} Z(\cdot)$ for the index set $T = R$.

The result may readily be generalized to the case $\Theta \subset R^p$, Θ open with the model satisfying A0–A4 and A6 of Section 6.2.1. We take $T = \Theta$ and define, for all $\mathbf{t} \in \Theta$, n sufficiently large,

$$Z_n(\mathbf{t}) \equiv l_n(\boldsymbol{\theta}_0 + \mathbf{t} n^{-\frac{1}{2}}) - l_n(\boldsymbol{\theta}_0).$$

It follows (Problem 7.1.2) that

$$Z_n(\mathbf{t}) = \mathbf{t}^T \mathbf{Z}_n^0 - \frac{1}{2} \mathbf{t}^T I(\boldsymbol{\theta}_0) \mathbf{t} + o_{P_{\boldsymbol{\theta}_0}}(1) \quad (7.1.6)$$

where $I(\boldsymbol{\theta}_0)$ is the Fisher information matrix, and

$$\mathbf{Z}_n^0 \equiv n^{-\frac{1}{2}} \sum_{i=1}^n \dot{\mathbf{l}}(X_i, \boldsymbol{\theta}_0)$$

with $\dot{\mathbf{l}}$ the gradient vector $(\partial l / \partial \theta_1, \dots, \partial l / \partial \theta_p)^T$. Generalizing (7.1.5), $Z_n(\cdot) \xrightarrow{FIDI} Z(\cdot)$ where

$$Z(\mathbf{t}) = \mathbf{t}^T \mathbf{Z}^0 - \frac{1}{2} \mathbf{t}^T I(\boldsymbol{\theta}_0) \mathbf{t}$$

and $\mathbf{Z}^0 \sim \mathcal{N}_p(\mathbf{0}, I(\boldsymbol{\theta}_0))$. \square

We return to the general case where the index set T may not be Euclidean but for instance be a function space. Let $|h|_\infty \equiv \sup\{|h(t)| : t \in T\}$, where $|\cdot|$ is the Euclidean norm in R^d .

Definition 7.1.2. For the index set T , $l_\infty(T)$ denotes the class of *sample functions* $Z(\cdot)$ on T consisting of the set of all bounded real valued functions h on T endowed with the sup norm $|h|_\infty$.

Note that $\mathcal{E}_n(\cdot) \in l_\infty(R)$ and $W^0(\cdot) \in l_\infty([0, 1])$. Doob's (1949) heuristics which started the fields of weak convergence and empirical processes were of the following type: We know from the multivariate central limit theorem that $\mathcal{E}_n(\cdot) \xrightarrow{FIDI} W^0(F(\cdot))$. This does not imply that the law of functions of $\mathcal{E}_n(\cdot)$ such as $\sup_t |\mathcal{E}_n(t)|$ whose value depends on more than a fixed finite number of t 's is necessarily close to that of the corresponding function of $W^0(\cdot)$, but it seems plausible that further analysis will yield $\mathcal{L}(\sup_t |\mathcal{E}_n(t)|) \rightarrow \mathcal{L}(\sup_t |W^0(t)|)$. This heuristic was shown to be correct in this and related cases by Donsker (1952), but the theory has taken on its most satisfactory form only during the last two decades.

What are needed are conditions on processes $Z_n(\cdot) \xrightarrow{FIDI} Z(\cdot)$ and on functions $q : \mathcal{F}(T) \rightarrow R$ such that $\mathcal{L}(q(Z_n(\cdot))) \rightarrow \mathcal{L}(q(Z(\cdot)))$. This strong and useful notion is called *weak convergence*.

Definition 7.1.3. Suppose that $Z(\cdot)$ and $Z_n(\cdot)$, $1 \leq i \leq n$, are in $l_\infty(T)$. $Z_n(\cdot)$ converges weakly to $Z(\cdot)$, written $\mathcal{L}(Z_n(\cdot)) \rightarrow \mathcal{L}(Z(\cdot))$, or more commonly, $Z_n \Rightarrow Z$, iff

$$(i) \quad Z_n \xrightarrow{FIDI} Z.$$

$$(ii) \quad \text{For every continuous function } q(\cdot) : l_\infty(T) \rightarrow R \text{ such that } q(Z_n(\cdot)) \text{ is a random variable,}$$

$$\mathcal{L}(q(Z_n(\cdot))) \rightarrow \mathcal{L}(q(Z(\cdot))).$$

- (iii) For each m , there exists a partition of T into a finite number m of sets and $Z^{(m)}(\cdot) \in l_\infty(T)$, each constant on the members of the partition, such that $|Z^{(m)} - Z|_\infty \xrightarrow{P} 0$, as $m \rightarrow \infty$.

Continuity of q here means that if $h_n, h \in l_\infty(T)$ and $|h_n - h|_\infty \rightarrow 0$, then $q(h_n) \rightarrow q(h)$. Note that $q(h) = |h|_\infty$ is certainly continuous, in fact, *uniformly continuous*, that is, for all $\epsilon > 0$, there exists $\delta(\epsilon) > 0$ such that if $|h_1 - h_2|_\infty \leq \delta(\epsilon)$, then $|q(h_1) - q(h_2)| \leq \epsilon$. Even more, this q is *Lipschitz continuous*, that is,

$$|q(h_1) - q(h_2)| \leq M|h_1 - h_2|_\infty,$$

for all h_1, h_2 for some constant $M > 0$.

Property (iii), finite approximation, is called *tightness* and any $Z(\cdot)$ with this property is called *tight*. It implies separability.

We relate our definition of weak convergence to classical definitions such as the one in Billingsley (1968) or more modern ones in terms of outer measures such as the one in van der Vaart and Wellner (1996) by example as follows: Let $C(T)$ be all real valued continuous functions on T . If $P(Z(\cdot) \in C(T)) = 1$ and $T = [0, 1]$, then (iii) is satisfied by taking $t_{mj} = j/m$, $0 \leq j \leq m$, and $Z^{(m)}$ the piecewise constant interpolation of $Z(\cdot)$ based on $\{Z(t_{mj}), 0 \leq j \leq m\}$. More generally, (iii) essentially says that with high probability, $Z(\cdot)$ is arbitrarily close to a compact subset of $l_\infty(T)$ such as an equicontinuous uniformly bounded subset of $l_\infty(T)$. See Theorem 7.1.1 below. See also van der Vaart and Wellner (1996) for further discussion.

We can now state the basic theorem giving sufficient (and essentially necessary) checkable conditions for weak convergence.

Theorem 7.1.1. Suppose $\{Z_n(\cdot)\}, Z(\cdot) \in l_\infty(T)$, and

- (1) $Z_n(\cdot) \xrightarrow{\text{FIDI}} Z(\cdot)$.
- (2) There exist subsets T_{m1}, \dots, T_{mk_m} of T with $\cup_{j=1}^{k_m} T_{mj} = T$ and $\delta_m, \varepsilon_m \rightarrow 0$ as $m \rightarrow \infty$ such that, for all m ,

$$\limsup_{n \rightarrow \infty} P[\max\{\sup[|Z_n(s) - Z_n(t)| : s, t \in T_{mj}] : 1 \leq j \leq k_m\} > \varepsilon_m] \leq \delta_m.$$

Then, $Z_n \Rightarrow Z$.

A proof of a weaker form of this theorem is given in Appendix D.2. The full proof may be found in van der Vaart and Wellner (1996), for instance. The basic idea is that for processes Z_n of the form $Z_n(t) = \sum_{j=1}^k Z_n(t_j)1(t \in T_{mj})$ where $t_j \in T_{mj}$, FIDI and weak convergence coincide. Condition (2) links uniform approximation of $\{Z_n(\cdot)\}$ by processes $\{Z_{nm}(\cdot)\}$ in a way which makes the weak limit of the Z_{nm} , call them $Z^{(m)}$, approximate Z in the sense of condition (iii) of Definition 7.1.3. Here is a development of Example 7.1.1.

Example 7.1.2.(Example 7.1.1 continued) *Weak convergence of likelihood processes.* Note that, for $\theta \in R$, conditions A0–A4 and A6 of Section 5.4.2 applied to $\psi = \partial l / \partial \theta$ imply

that, under P_{θ_0} ,

$$\sup\{|Z_n(t) - tZ_n^0 + \frac{t^2}{2}I(\theta_0)| : |t| \leq M\} = o_{P_{\theta_0}}(1).$$

We claim that weak convergence of $Z_n(\cdot)$ to $Z(\cdot)$ holds on all intervals $T = [-M, M]$. To see this note first that $Z_n(\cdot)$ is separable on all such T since $\theta \rightarrow l_n(\theta)$ is continuous. FIDI convergence was shown in Example 7.1.1. Next, to show condition (2) of Theorem 7.1.1, let $T_{mj} = M(\frac{j-1}{m}, \frac{j}{m}]$, $-m+1 \leq j \leq m$. Then,

$$\sup\{|(s-t)Z_n^0 - \left(\frac{s^2}{2} - \frac{t^2}{2}\right)I(\theta_0)| : s, t \in T_{mj}\} \leq \left[\frac{|Z_n^0|}{m} + \frac{M}{m}I(\theta_0)\right]M.$$

It follows that,

$$\begin{aligned} & P\left[\max\{\sup\{|Z_n(s) - Z_n(t)| : s, t \in T_{mj}\}, : -m+1 \leq j \leq m\} > \varepsilon\right] \\ & \leq \sum_{j=-m+1}^m P\left[\max\{\sup\{|(s-t)Z_n^0 - (\frac{s^2}{2} - \frac{t^2}{2})I(\theta_0)| : s, t \in T_{mj}\}\} > \frac{\varepsilon}{2}\right] \\ & \quad + P\left[\sup\{|Z_n(s) - sZ_n^0 + \frac{s^2}{2}I(\theta_0)| : |s| \leq M\} > \frac{\varepsilon}{2}\right] \\ & \leq 2m P\left[|Z_n^0| + MI(\theta_0) > \frac{\varepsilon m}{2M}\right] + o(1). \end{aligned} \tag{7.1.7}$$

Choose $\varepsilon = \varepsilon_m \rightarrow 0$ so slowly that

$$P\left[|Z^0| + MI(\theta_0) > \frac{\varepsilon_m m}{2M}\right] \leq \frac{1}{2m^2}.$$

Since Z^0 is Gaussian, using Problem D.2.10 or (7.1.9) of the next subsection, $\varepsilon_m = (\log m)/m$ will do. Now take $\delta_m = 1/m$ and weak convergence follows from Theorem 7.1.1.

This example also illustrates condition (iii) of Definition 7.1.3 since it is evident that we can approximate $Z(t)$ on the partition member T_{mj} by

$$Z^{(m)}(t) = M\frac{j}{m}Z_0 - \left(M\frac{j}{m}\right)^2 \frac{1}{2}I(\theta_0).$$

□

Theorem 7.1.1 has an easy corollary (Problem 7.1.3). Let

$$p_m(\varepsilon) = \limsup_n \max\{P[\sup\{|Z_n(s) - Z_n(t)| : s, t \in T_{mj}\} \geq \varepsilon] : 1 \leq j \leq k_m\}. \tag{7.1.8}$$

Corollary 7.1.1. *If $k_m p_m(\varepsilon) \rightarrow 0$ for all $\varepsilon > 0$ and $Z_n \xrightarrow{\text{FIDI}} Z$, then $Z_n \Rightarrow Z$.*

Here is a continuous mapping theorem for processes:

Proposition 7.1.1. *If $Z_n \Rightarrow Z \in l_\infty(T)$ and g is a continuous map from $l_\infty(T)$ to $l_\infty(S)$ for some set S , then*

$$g(Z_n) \Rightarrow g(Z) \in l_\infty(S).$$

Proof. We show that for any function q that maps $l_\infty(S)$ to R continuously, $q(g(Z_n)) \rightarrow q(g(Z))$ on $l_\infty(S)$. Because the continuity of q and g implies the continuity of the composition $q \cdot g$, (ii) of Definition 7.1.3 follows from the weak convergence of Z_n . We leave (i) and (iii) of Definition 7.1.3 for Problem 7.1.6. \square

The previous concepts extend readily from processes $Z_n(t) \in l_\infty(T)$ to processes $(Z_n(t), U_n(s)) \in l_\infty(T) \times l_\infty(S)$. For such processes a useful result is *Slutsky's theorem for processes*.

Corollary 7.1.2. For sets T , S , and V , suppose $g : l_\infty(T) \times l_\infty(S) \rightarrow l_\infty(V)$ continuously and,

$$(i) \quad Z_n \Rightarrow Z$$

$$(ii) \quad U_n \xrightarrow{P} u$$

where $Z_n \in l_\infty(T)$, $U_n, u \in l_\infty(S)$, u is a nonrandom function, and \xrightarrow{P} convergence is in $l_\infty(S)$. Then,

$$g(Z_n, U_n) \Rightarrow g(Z, u).$$

Proof. It is equivalent to show that $(Z_n, U_n) \Rightarrow (Z, u)$. We establish the result for Z_n, U_n obeying the conditions of Theorem 7.1.1. Let $\{T_{mj}\}$ be the sets for $\{Z_n\}$, $\{S_{mj}\}$ for $\{U_n\}$. Take $\{Q_{ml}\} = \{S_{mj} \times T_{mk} : \text{all } j, k\}$. Evidently $\{Q_{ml}\}$ satisfy (2) of Theorem 7.1.1 for (Z_n, U_n) . On the other hand,

$$(Z_n(t_1), \dots, Z_n(t_a), U_n(s_1), \dots, U_n(s_b)) \xrightarrow{FIDI} (Z(t_1), \dots, Z(t_a), u(s_1), \dots, u(s_b))$$

by the “classical” Slutsky theorem (Theorem B.7.2). Thus, by Theorem 7.1.1, $(Z_n, U_n) \Rightarrow (Z, u)$ and the result follows.

Example 7.1.3. *The standardized empirical process.* Let X_1, \dots, X_n be i.i.d. F on R with F continuous. Assume Donsker’s theorem 7.1.4,

$$\mathcal{E}_n(\cdot) \Rightarrow W^0(\cdot).$$

For t such that $\widehat{F}(t)[1 - \widehat{F}(t)] > 0$, define

$$Z_n(t) = \sqrt{n} \frac{(\widehat{F}(t) - F(t))}{\sqrt{\widehat{F}(t)(1 - \widehat{F}(t))}},$$

and set $Z_n(t) = 0$, otherwise. Let $I \equiv [F^{-1}(\varepsilon), F^{-1}(1 - \varepsilon)]$ for $\varepsilon > 0$. Then

$$Z_n(\cdot) \implies \frac{W^0(F(\cdot))}{\sqrt{F(\cdot)(1 - F(\cdot))}}$$

on I . This follows by representing X_i as $X_i = F^{-1}(U_i)$, $i = 1, \dots, n$, where U_i are i.i.d. uniform $(0, 1)$ so that

$$\sqrt{n}(\hat{F}(t) - F(t)) = \mathcal{E}_n(F(t)),$$

and using Corollary 7.1.2 — see Problem 7.1.9.

Remark 7.1.1. Weak convergence in this general sense also has many of the important properties of FIDI convergence, mainly Theorems B7.1, B7.2, B7.4, B7.5 and generalizations of these to functions $g : l_\infty(T) \rightarrow l_\infty(S)$. We shall use these results as needed in this section and subsequently.

Summary. We introduced the concept of a *stochastic process* as a collection $\{Z(t) : t \in T\}$ of random variables that are *separable* in the sense that the probability that $Z(t)$ will stay in the interval (a, b) for all t in a set $S \subset T$ can be computed by restricting t to the set $S \cap C$ for some countable set C . We consider stochastic processes that take values in the class $l_\infty(T)$ of all bounded real valued functions h on T with the sup norm $|h|_\infty$. A sequence $\{Z_n(t) : t \in T\}$ of stochastic processes *converges weakly* to a stochastic process $Z(t)$, $t \in T$, if each finite collection $Z_n(t_1), \dots, Z_n(t_k)$ converges in law to $Z(t_1), \dots, Z(t_k)$, if $q(Z_n(\cdot))$ converges in law to $q(Z(\cdot))$ for all continuous real valued functions q on $l_\infty(T)$, and if $Z(\cdot)$ is *tight* in the sense that $|Z^{(m)} - Z|_\infty \xrightarrow{P} 0$ as $m \rightarrow \infty$ for some $\{Z^{(m)}(\cdot)\} \in l_\infty(T)$, where each $Z^{(m)}(\cdot)$ takes on only a finite set of values. We give conditions on fluctuations of $Z_n(\cdot)$ over small sets that imply weak convergence in Theorem 7.1.1 and Corollary 7.1.1, and verify that they hold for the likelihood process which is defined to be the increments of the log likelihood over intervals $(\theta_0, \theta_0 + t/\sqrt{n}]$ for a smooth parametric model for i.i.d. X_1, \dots, X_n .

7.1.2 Maximal Inequalities

Corollary 7.1.1 suggests that the essential tools we need are bounds on $P[\sup\{|Z_n(s) - Z_n(t)| : s, t \in S\} \geq \lambda]$ for subsets S of T of the form T_{mj} described in Theorem 7.1.1. There is a standard way of bounding tail probabilities for random variables (see (A.15.4)):

$$P[|W| \geq \lambda] \leq \frac{Eh(W)}{h(\lambda)} \text{ for } h \geq 0, h \text{ non-decreasing.}$$

By studying $h(t) = \exp(\gamma t)$ for appropriate $\gamma > 0$, one can derive Bernstein's and Hoeffding's inequalities (B.9.5), (B.9.6). If W is $\mathcal{N}(0, 1)$, an important bound we shall use (Problem D.2.10) is, for $\lambda \geq 1$,

$$P[|W| \geq \lambda] \leq 2 \frac{\phi(\lambda)}{\lambda} \leq 2 \exp\left\{-\frac{\lambda^2}{2}\right\}. \quad (7.1.9)$$

An important refinement of Bernstein's inequality (B.9.5) — see Hoeffding (1963) or Shorack and Wellner (1986), is the following *Hoeffding's inequality*. Let Y_1, \dots, Y_n be independent, $|Y_i| \leq M$ for some constant $M > 0$, $E(Y_i) = 0$ for $i = 1, \dots, n$, and $\sigma_n^2 = \text{Var}(Y_1 + \dots + Y_n)$. Then for each $\lambda > 0$,

$$P\left[\frac{|Y_1 + \dots + Y_n|}{\sigma_n} > \lambda\right] \leq 2 \exp\left\{-\frac{1}{2}\lambda^2\left(1 + \frac{M\lambda}{3\sigma_n}\right)^{-1}\right\}. \quad (7.1.10)$$

The $M\lambda/3\sigma_n$ term makes this inequality weaker than (7.1.9) for the normal case.

If, however, we have a collection of random variables, $\{Z(t) : t \in T\}$, tail bounds $P(|Z(t)| > \lambda)$ on the individual $Z(t)$ are not translatable into useful maximal inequalities without additional conditions. To take a trivial example, suppose T is the rational numbers and $Z(t)$ are independent $\mathcal{N}(0, 1)$. Then, no matter how small $\delta > 0$ is, for each $\lambda > 0$

$$P[\sup\{|Z(t)| : |t| \leq \delta\} \geq \lambda] = 1.$$

Yet, as we shall see in this section and Appendix D.2, it is quite possible to have inequalities of the form (7.1.9) and (7.1.10) improved to maximal inequalities of the same form for stochastic processes. Essential ingredients are appropriate tail bounds on $|Z(s) - Z(t)|$ and the structure of T . To develop intuition we begin with $T = [0, 1]$. The following criterion, due to Kolmogorov, has been elaborated and extended in Billingsley ((1968), p.89).

Proposition 7.1.2 Suppose that the separable stochastic process $Z(\cdot)$ on $[0, 1]$ satisfies

$$P[|Z(t) - Z(s)| \geq \varepsilon] \leq M(\varepsilon)\delta^{1+\gamma} \quad (7.1.11)$$

for all $s, t, |s - t| \leq \delta$, all $\delta > 0$, and some $\gamma > 0$. Then, if the length of $S \subset [0, 1]$ is $\leq \delta_0$,

$$P[\sup\{|Z(t) - Z(s)| : s, t \in S\} \geq \varepsilon] \leq KM(\varepsilon)\delta_0^{1+\gamma} \quad (7.1.12)$$

for a universal constant $K = K(\gamma)$.

Note that by (A.15.4), (7.1.11) follows if, for some $c > 0, r > 0$ and $\gamma > 0$,

$$E|Z(t) - Z(s)|^r \leq c|t - s|^{1+\gamma}. \quad (7.1.13)$$

A remarkable feature of the proposition is that (7.1.11) is a condition which can be checked using bivariate distributions.

In the following application, we will need

Definition 7.1.4. The Wiener process or Brownian motion on $[0, 1]$ is a separable Gaussian process $W(t)$, $0 \leq t \leq 1$, with $EW(t) = 0$, $\text{Cov}(W(s), W(t)) = s \wedge t$.

Example 7.1.4. The partial sum process. The following process arises in statistics primarily in the context of sequential analysis. Let X_1, \dots, X_n be i.i.d. as X , $E(X) = 0$, $\text{Var}(X) = \sigma^2 < \infty$, $E|X|^{2+\gamma} < \infty$, $\gamma > 0$. Define $S_k = \sum_{j=1}^k X_j$,

$$W_n(t) = n^{-\frac{1}{2}} \frac{S_{[nt]}}{\sigma}, \quad 0 \leq t \leq 1 \quad (7.1.14)$$

where $[x]$ is the largest integer $\leq x$.

Proposition 7.1.3. *If $W(\cdot)$ is the Wiener process on $[0,1]$, then*

$$W_n \Rightarrow W.$$

Proof: Note that

$$W_n(t) = \left(\frac{[nt]}{n} \right)^{\frac{1}{2}} \frac{S_{[nt]}}{\sigma [nt]^{\frac{1}{2}}}, \quad n^{-1} \leq t \leq 1.$$

Let $t_1 < \dots < t_k$. Then (Problem 7.1.16), by the k variate central limit theorem,

$$(W_n(t_1), W_n(t_2) - W_n(t_1), \dots, W_n(t_k) - W_n(t_{k-1})) \Rightarrow (U_1, \dots, U_k), \quad (7.1.15)$$

where U_1, \dots, U_k are independent $U_1 \sim \mathcal{N}(0, t_1)$, $U_j \sim \mathcal{N}(0, t_j - t_{j-1})$, $2 \leq j \leq k$. It is easy to see that

$$(U_1, U_1 + U_2, \dots, U_1 + \dots + U_k) \sim (W(t_1), \dots, W(t_k)).$$

Now we show that convergence is not just FIDI but weak by using Proposition 7.1.1 and Corollary 7.1.1. By Markov's inequality

$$\begin{aligned} P[|W_n(t) - W_n(s)| \geq \lambda] &\leq \lambda^{-(2+\gamma)} E|W_n(t) - W_n(s)|^{2+\gamma} \\ &= \lambda^{-(2+\gamma)} E \left| \sum_{j=[ns]+1}^{[nt]} \frac{X_j}{\sigma \sqrt{n}} \right|^{2+\gamma}. \end{aligned} \quad (7.1.16)$$

Without loss of generality, take $\sigma = 1$. By an extension of the proof of (5.3.3) to all $\gamma > 0$ based on the Marcinkiewicz-Zygmund's inequality (1939),

$$E \left| \sum_{j=[ns]+1}^{[nt]} X_j \right|^{2+\gamma} \leq ([nt] - [ns])^{1+\frac{\gamma}{2}} E|X|^{2+\gamma},$$

and thus for $M = E|X|^{2+\gamma}$,

$$E|W_n(t) - W_n(s)|^{2+\gamma} \leq M|s - t|^{1+\frac{\gamma}{2}}. \quad (7.1.17)$$

We conclude from Proposition 7.1.2 that, for all fixed t_0 ,

$$P[\sup\{|W_n(t_0) - W_n(s)| : t_0 \leq s \leq t_0 + \delta\} \geq \lambda] \leq \lambda^{-(2+\gamma)} M \delta^{(1+\frac{\gamma}{2})}$$

and

$$\begin{aligned} P[\sup\{|W_n(t) - W_n(s)| : t_0 \leq s, t \leq t_0 + \delta\} \geq \lambda] \\ \leq 2P[\sup\{|W_n(s) - W_n(t_0)| : t_0 \leq s \leq t_0 + \delta\} \geq \frac{\lambda}{2}] \\ \leq \lambda^{-(2+\gamma)} 2^{\gamma+3} M \delta^{1+\frac{\gamma}{2}}. \end{aligned} \quad (7.1.18)$$

Take $T_{mj} = \left(\frac{j}{m}, \frac{j+1}{m}\right]$, $0 \leq j \leq m-1$. Then, $\delta = m^{-1}$, and by (7.1.18), $p_m(\varepsilon)$ of Corollary 7.1.1 is bounded above by $cm^{-(1+\frac{\gamma}{2})}$. Thus because $k_m = m$, Corollary 7.1.1 establishes weak convergence. \square

Remark 7.1.2.

- 1) The structure of the processes $W_n(\cdot)$ appears only in a weak way, through the FIDI convergence and the inequality (7.1.12). Thus, results such as Proposition 7.1.2 can be proved for dependent X_j as well.
- 2) The condition $E|X|^{2+\gamma} < \infty$ in Example 7.1.3 is superfluous. Only second moments are needed—see Billingsley(1968).
- 3) The process $W(\cdot)$ has continuous sample functions (Problem D.2.3).
- 4) The property (7.1.15) reveals the fundamental property of the Wiener process. It has stationary independent Gaussian increments. \square

Summary. We gave a result which gives bounds on the probability that the maximum of the fluctuations of a stochastic process over a set S exceeds ε . We introduced the Wiener process and the partial sum process and showed how the result can be used to establish the weak convergence of the partial sum process to a Wiener process.

7.1.3 Empirical Processes on Function Spaces

We could at this stage prove weak convergence results for the classical empirical process of Section I.1. We choose instead to introduce and develop the general theory of weak convergence of empirical processes which we use in Chapters 9–12.

We specialize in the rest of this section to a subset T of the set of functions h from \mathcal{X} to R , in the context of $X_1, \dots, X_n \in \mathcal{X}$ i.i.d as $X \sim P$, and $E_Ph^2(X) < \infty$. We define the *empirical process on T* by

$$\mathcal{E}_n(h) = \frac{1}{\sqrt{n}} \sum_{i=1}^n (h(X_i) - E_Ph(X_i)) \quad (7.1.19)$$

for $h \in T$. The empirical process on the line that we defined earlier is a special case with $T = \{h : h(\cdot) = h_x(\cdot), x \in R\}$ where $h_x(\cdot) = 1(-\infty, x](\cdot)$. Note that we have made a notational change by now writing $\mathcal{E}_n(h_x)$ for what was denoted by $\mathcal{E}_n(x)$ in Section I.1 and Section 7.1.1. Weak convergence of \mathcal{E}_n for this T will be what we need for Donsker's theorem and, as we noted, maximal inequalities for \mathcal{E}_n on suitable subsets of T are what we'll need for the proof.

Another \mathcal{E}_n on a suitable T appeared in Examples 7.1.1 and 7.1.2 with

$$T = \{\log p(\cdot, \theta) : \theta \in \Theta\}.$$

If we define, generally,

$$\|\mathcal{E}_n\|_T \equiv \sup\{|\mathcal{E}_n(h)| : h \in T\}$$

we will find that remainder terms in the proofs of asymptotic results in this book will be naturally bounded by $\|\mathcal{E}_n\|_T$ for suitable T .

To obtain useful maximal inequalities for $\mathcal{E}_n(\cdot)$, we need essentially to consider T which are not too big. If $T = \{\text{all bounded functions}\}$ for instance, we do not get useful bounds (Problem 7.1.7). There are many ways of restricting T . We limit ourselves here to bounding the *bracketing number*. Note in passing that by specializing to \mathcal{E}_n , we have completely specified the joint behavior of $\mathcal{E}_n(h_1), \dots, \mathcal{E}_n(h_k)$ and hence FIDI limits of the $\mathcal{E}_n(\cdot)$. The *bracketing numbers* take advantage of this structure.

Definition 7.1.5. A bracket, (\underline{f}, \bar{f}) , is a set of functions h on \mathcal{X} such that $h \in (\underline{f}, \bar{f})$, that is $\underline{f}(x) \leq h(x) \leq \bar{f}(x)$ for all x . We assume both \underline{f} and \bar{f} are in $L_2(P)$, but \underline{f} and \bar{f} don't have to belong to the bracket.

Definition 7.1.6. The δ bracketing number $N_{[]}(\delta, T, L_2(P))$ for a subset T of functions on \mathcal{X} is the smallest number of brackets $(\underline{f}_i, \bar{f}_i)$ with $\underline{f}_i, \bar{f}_i \in T$ such that

- (a) for every $h \in T$, there exist $\underline{f}_i, \bar{f}_i$ such that $h \in (\underline{f}_i, \bar{f}_i)$,
- (b) $E_P(\bar{f}_i - \underline{f}_i)^2(X) \leq \delta^2$

Definition 7.1.7. The *envelope function* for sets T in Definition 7.1.6 is $s(T) = \sup\{|h(\mathcal{X})| : h \in T\}$.

We can get upper bounds on $N_{[]}$ by finding the number of brackets for some collection of brackets satisfying (b). Here is an example.

Example 7.1.5. *The classical empirical process.* Suppose $\mathcal{X} = R$, $F(x) = P(-\infty, x]$ is continuous and strictly increasing, and $T = \{1(-\infty, x) : x \in R\}$. For a given $\delta \in (0, 1/2)$, set $k = [1/\delta^2]$ where $[t]$ is the greatest integer $\leq t$ and let $x(\delta^2), x(2\delta^2), \dots, x(k\delta^2)$ be the unique $\delta^2, 2\delta^2, \dots, k\delta^2$ quantiles of P . We set $x(0) = -\infty$ and $x((k+1)\delta^2) = \infty$. Next, for $i = 0, \dots, k$, define the brackets

$$(\underline{f}_i, \bar{f}_i) = [1(-\infty, x(i\delta^2)], 1(-\infty, x((i+1)\delta^2))] .$$

Then, for each $h \in T$, there is $i \in \{0, \dots, k\}$ such that $h \in (\underline{f}_i, \bar{f}_i)$, and

$$E_P(\bar{f}_i - \underline{f}_i)^2(X) = P(x(i\delta^2) < X \leq x((i+1)\delta^2)) = \delta^2 .$$

For this collection of brackets, the number of brackets is $k + 1 = [1/\delta^2] + 1$. Thus, $N_{[]}(\delta, T, L_2(P)) \leq 1 + [1/\delta^2]$. This upper bound is, in fact, sharp — but only upper bounds are needed. For this \mathcal{X} and T , the envelope function is $s(T) \equiv 1$. \square

We continue by stating and proving a maximal inequality not for the empirical process but for its Gaussian limit, the “Brownian bridge.” We then state the appropriate analogue for the empirical process and discuss the new difficulties introduced but only give proofs of the theorem in Appendix D.2.

Let T be a set of functions as above.

Definition 7.1.8. The *Brownian bridge* on T (with respect to P) is the Gaussian process $W_P^0(f)$, $f \in T$ with

$$EW_P^0(f) = 0, \quad \text{Cov}(W_P^0(f), W_P^0(g)) = \text{Cov}_P(f(X), g(X)) .$$

Note that by the central limit theorem, for $h_j \in T$, $\mathcal{E}_n(h_1), \dots, \mathcal{E}_n(h_k)$ will converge in law to $W_P^0(h_1), \dots, W_P^0(h_k)$, which incidentally establishes that W_P^0 is definable as a Gaussian process as well as that $\mathcal{E}_n(\cdot)$ converges FIDI to $W_P^0(\cdot)$.

The critical connection between $W_P^0(f)$ and bracketing is that if (\underline{f}, \bar{f}) is a bracket with bracketing number δ and if $f, g \in (\underline{f}, \bar{f})$, then

$$E(W_P^0(g) - W_P^0(f))^2 \leq E_P(g - f)^2(X) \leq E(\bar{f} - \underline{f})^2(X) \leq \delta^2.$$

We state an analogue of Theorem 2.14.16 of van der Vaart and Wellner (1996), which is also a crude version of a result of Talagrand (1994).

Theorem 7.1.2. Suppose that for all $x \in \mathcal{X}$, $f \in T$

- (i) For some fixed enveloped function $s \in T$, $|(f(x)| \leq s(x)$.
- (ii) For some positive constants c and d , $N_{[]}(\delta, T, L_2(P)) \leq c\delta^{-d}$ for all $0 < \delta < 1$.
- (iii) For some positive constant γ , $E_P f^2(X) \leq \gamma^2$.

Then,

$$P[\sup\{|W_P^0(f)| : f \in T\} \geq \lambda] \leq C \left(1 + \frac{\lambda}{\gamma}\right)^{2d(1+\epsilon)} \exp\left\{-\frac{\lambda^2}{2\gamma^2}\right\}, \quad (7.1.20)$$

for all $\epsilon > 0$ and C a constant.

Note that (ii) implies (i) (Problem D.2.9).

The proof we shall sketch in Appendix D.2 is essentially that of van der Vaart and Wellner (1996) following Pollard (1984). A much better result due to Talagrand (1994) shows that $2d(1+\epsilon)$ can be replaced by $d-1$ with C replaced by $C\gamma^{-2d}$ — see Proposition A.2.7 in van der Vaart and Wellner. This is the best possible (Problem D.2.7).

In the future, we shall refer to the condition (ii) above as the “*polynomial bracketing*” condition.

In the discussion following the proof of Theorem 7.1.2 in Appendix D.2 it is argued that the result follows from a “chaining” argument which involves showing that we can write

$$W_P^0(f) = W_P^0(g_{m_0}) + \sum_{m=m_0}^{\infty} [W_P^0(g_{m+1}) - W_P^0(g_m)] \quad (7.1.21)$$

where the $\{g_m\}$ are representative members of the bracket sets $(\underline{f}_{m_j}, \bar{f}_{m_j})$. It is also argued that the order of magnitude of $P(\{\sup|W_P^0(f)| : f \in T\} \geq \lambda)$ is of the same order $\exp\{-\lambda^2/2\gamma^2\}$ as that of a single $W_P^0(f)$. This result enables us to show that $W_P^0(f)$ is continuous in $|\cdot|_\infty$ on T — see below.

Remarkably, a result almost as good is available for \mathcal{E}_n , independent of n . This is a slightly cruder form of Theorem 2.14.16 of van der Vaart and Wellner (1996).

Theorem 7.1.3. Suppose T satisfies the conditions of Theorem 7.1.2 and the envelope function $s(\cdot)$ in condition (i) of Theorem 7.1.2 is uniformly bounded by $M < \infty$. Then,

$$P[\|\mathcal{E}_n\|_T \geq M\lambda] \leq C\gamma^{-2d} \left(1 + \frac{\lambda}{\gamma}\right)^{4d} \exp\left\{-\frac{1}{2}[\lambda^2(\gamma^2 + (3+\lambda)n^{-\frac{1}{2}})^{-1}]\right\}. \quad (7.1.22)$$

The argument for this theorem, which we discuss somewhat further in Appendix D.2, replaces the use of the Gaussian tail inequality (7.1.9) by Hoeffding's inequality (7.1.10) and the change in the bound reflects this. The chaining method described in Appendix D.2 relies on bounds for the probabilities of deviation from 0 of increments with small variance. For \mathcal{E}_n , this behaviour is intrinsically worse than for W_P^0 . For instance, the classical empirical process sample functions, $\sqrt{n}(\hat{F}(\cdot) - F(\cdot))$, have jumps while $W_P^0(F(\cdot))$ has continuous sample functions. This leads to the need for arguing that although individual $g_{m+1} - g_m$ may be unusually large on their scale, having more than one such element of a chain deviate is sufficiently unlikely. Furthermore, the requirement that $s(\cdot)$ is bounded has to be imposed since, for an arbitrary $F(\cdot)$, a bound of the type given in (7.1.22) couldn't hold even if T were a single function.

We now apply Theorem 7.1.2 to a first example.

Corollary 7.1.3. Continuity of $W_P^0(f)$. Suppose T satisfies the conditions of Theorem 7.1.2. Let $\|f\|_2 \equiv E_P^{\frac{1}{2}} f^2(X)$. Then, $W_P^0(\cdot)$ is continuous in probability with respect to $\|\cdot\|_2$, in the sense that, for $f_0 \in T$,

$$P[\sup\{|W_P^0(f) - W_P^0(f_0)| : \|f - f_0\|_2 \leq \delta, f \in T\} \geq \epsilon] \rightarrow 0 \quad (7.1.23)$$

for each $\epsilon > 0$ as $\delta \rightarrow 0$.

Proof. Set $W_0(f) = W_P^0(f) - W_P^0(f_0)$. Then $W_0(f)$ is a Gaussian process with mean zero and covariance $E[W_0(f)W_0(g)] = E[(f - f_0)(X)(g - f_0)(X)]$. That is, $W_0(f)$ has the same distribution as $W_P^0(f - f_0)$. The result follows from the bound of Theorem 7.1.2. \square

Note that (7.1.23) is weaker than the more natural definition of *continuity with probability one*: On a suitable probability space (Ω, \mathcal{A}, P) , we can define $W_P^0(\cdot, w)$ such that

$$P[\omega : \sup\{|W_P^0(f, \omega) - W_P^0(f_0, \omega)| : \|f - f_0\|_2 \leq \delta, f \in T\} \rightarrow 0 \text{ as } \delta \rightarrow 0] = 1. \quad (7.1.24)$$

But this can also be derived. See Problem D.2.3.

As an application of Corollary 7.1.3, take $T = \{1[0, u] : u \in [0, 1]\}$ and P the uniform distribution on $[0, 1]$. Identify $1[0, u]$ with u . Then $W^0(u)$ is the usual Brownian bridge of Section I.1. We deduce from Corollary 7.1.3 that, for $u_0 \in [0, 1]$,

$$P[\sup\{|W^0(u) - W^0(u_0)| : |u - u_0| \leq \delta, u \in [0, 1]\} \geq \epsilon] \rightarrow 0$$

as $\delta \rightarrow 0$ for all $\epsilon > 0$. To see this, note that if $v > u$,

$$E\{1(X \in [0, v]) - 1(X \in [0, u])\}^2 = E\{1(X \in (u, v])\}^2 = v - u,$$

and use Example 7.1.4 and Theorem 7.1.2. The statement (7.1.24) also holds in this example.

We continue with the identification of T with $\{1[0, u] : u \in [0, 1]\}$ and write $\mathcal{E}_n(u)$ and $W_P^0(u)$ for $\mathcal{E}_n(1[0, u])$ and $W_P^0(1[0, u])$ as before.

Theorem 7.1.4 (Donsker (1952)). *If $T = \{1[0, u] : u \in [0, 1]\}$ and $P = \mathcal{U}[0, 1]$, then for \mathcal{E}_n and W^0 as above*

$$\mathcal{E}_n \Longrightarrow W^0.$$

Proof. The structure of this argument parallels that of Proposition 7.1.2. We need only check the conditions of Corollary 7.1.1. FIDI convergence, as we noted previously, follows from the multivariate central limit theorem. Let $T_{mj} = (j\delta_m, (j+1)\delta_m]$, $0 \leq j < [1/\delta_m] - 1$ with $\delta = \delta_m$ to be chosen.

Now $\mathcal{E}_n(v) - \mathcal{E}_n(j\delta_m) = \mathcal{E}_n(1(j\delta_m, v])$. The set of all $\{1(j\delta_m, v] : j\delta_m < v \leq (j+1)\delta_m\}$ is easily seen to obey the conditions of Theorem 7.1.2 with $d = 1$, $\gamma^2 = \delta_m(1 - \delta_m)$ by arguing as in Example 7.1.4. Thus using (7.1.22),

$$\begin{aligned} P \left[\sup \{|\mathcal{E}_n(v) - \mathcal{E}_n(j\delta_m)| : j\delta_m < v \leq (j+1)\delta_m\} > \frac{\lambda}{2} \right] \\ \leq C\delta_m^{-1}(1 + \lambda\delta_m^{-\frac{1}{2}})^4 \exp \left\{ -\frac{1}{2} \left(\frac{\lambda^2}{2} (\delta_m + (3 + \lambda)n^{-\frac{1}{2}})^{-1} \right) \right\}. \end{aligned}$$

Put $\lambda = 2/m$, $\delta_m = 1/m^3$ (say), and the condition of Corollary 7.1.1 holds. The theorem follows. \square

More generally, we state (see Problem 7.1.17)

Theorem 7.1.5. *Suppose T can be partitioned into sets $\{T_{mj}\}$, $j = 1, \dots, k_m$, such that each T_{mj} satisfies the conditions of Theorem 7.1.3, polynomial $\{c, d\}$ bracketing, and uniformly bounded envelope $s(\cdot) \leq M$. Suppose c , d , and M are independent of m , and suppose*

$$\sup \{E_P(f - g)^2(X) : f, g \in T_{mj}\} \leq \gamma_m^2$$

for all m , where $\gamma_m^2 = o(1/\log k_m)$. Then, if X_1, \dots, X_n are i.i.d. P and $f \in T$,

$$\mathcal{E}_n(f) \Longrightarrow W_P^0(f).$$

Note, for future reference, that the only properties specific to \mathcal{E}_n that are used are the maximal inequalities. Thus weak convergence results can be expected to hold for much wider classes of processes, for instance, empirical processes for weakly dependent stationary sequences — see Doukhan (1995) for instance.

We continue with

Example 7.1.6. *The Kolmogorov statistic.* We immediately deduce from Donsker's theorem that if $F_0 = \mathcal{U}(0, 1)$, under $H : F = F_0$,

$$P[\sqrt{n} \sup_u |\widehat{F}(u) - F_0(u)| > \lambda] \rightarrow P[\sup_u |W^0(u)| > \lambda] = \sum_{j=1}^{\infty} (-1)^{j+1} e^{-2j^2\lambda^2} \quad (7.1.25)$$

where the second equality is from Shorack and Wellner (1986). From Proposition 4.1.1, we know that the left hand side of (7.1.25) in fact has the same distribution under F_0 for any F_0 continuous. What if F_0 is not continuous? It follows from Theorem 7.1.5 that, under $H : F = F_0$, for $T = \{1(-\infty, u] : u \in R\}$, with $1(-\infty, u]$ identified with u ,

$$\mathcal{E}_n(\cdot) \implies W_{F_0}^0(\cdot) \equiv W^0(F_0(\cdot)).$$

Although the distribution of $\sup_x |W_{F_0}^0(x)|$ depends on F_0 if F_0 is not continuous, it is still true that the distribution of $\sqrt{n} \sup_x |\hat{F}(x) - F_0(x)|$ converges weakly to that of $\sup_x |W_{F_0}^0(x)|$. We will use this remark and the bootstrap in Chapter 10 to set confidence bounds for F based on Kolmogorov statistic which are narrower than ones using the tabled values based on the right hand side of (7.1.25). See Example 10.3.4. \square

Example 7.1.7. The *Cramér-von Mises statistic* of Problem 4.1.11 is

$$C_n \equiv n \int_{-\infty}^{\infty} (\hat{F}(x) - F_0(x))^2 dF_0(x).$$

If F_0 is continuous, the distribution of this statistic under H does not depend on F_0 . Take $F_0 = \mathcal{U}(0, 1)$. Since the map $f \rightarrow \int_0^1 f^2(u) du$ is continuous from $l_\infty[0, 1]$ to R^+ , we immediately deduce from Donsker's theorem that

$$\mathcal{L}_{F_0}(C_n) \rightarrow \mathcal{L}\left(\int_0^1 [W^0]^2(u) du\right).$$

Although the density function of this limit does not have a nice analytic form (see Example 7.3.1), its characteristic function does.

$$E \exp\{it \int_0^1 [W^0]^2(u) du\} = \prod_{j=1}^{\infty} (1 - 2\lambda_j it)^{-\frac{1}{2}}$$

with $\lambda_j = (j\pi)^{-2}$ from Shorack and Wellner (1986), p.215. \square

Classes of functions T for which $\mathcal{E}_n(\cdot) \implies W_P^0(\cdot)$ are called *Donsker classes*. They include indicators of quadrants and hyperplanes in R^d and much else. Validation of weak convergence involves, as we have seen, being able to find partitions of the sets of functions T such that the oscillations of \mathcal{E}_n over each member of the partition tend to be small. This requires using maximal inequalities in the way we have. An important, but in general overly crude application of the maximal inequalities we have, establishes generalizations of the law of large numbers. We obtain from Theorem 7.1.3

Proposition 7.1.4. *If T obeys the conditions of Theorem 7.1.3,*

$$\sup\left\{\frac{1}{n} \sum_{i=1}^n |h(X_i) - E_P h(X)| : h \in T\right\} \xrightarrow{P} 0. \quad (7.1.26)$$

Proof. The result follows because Theorem 7.1.3 tells us that

$$\sup\{|\mathcal{E}_n(h)| : h \in T\} = O_P(1)$$

which implies (7.1.26) on multiplying both sides by $n^{-\frac{1}{2}}$. \square

In fact, more can be shown: The convergence in (7.1.26) is almost sure (Problem 7.1.10). A famous special case is the *Glivenko-Cantelli theorem*: With probability 1

$$\sup_x |\hat{F}(x) - F(x)| \rightarrow 0 \quad (7.1.27)$$

We will apply the weak convergence results further in Sections 7.2 and 9.2. The maximal inequalities will be directly useful both in Section 7.2 and in Chapters 9 and 11.

We close the discussion in this section noting that we do not enter further into the other beautiful ideas underlying maximal inequalities, such as symmetrization, viewing the sample obtained as a sample without replacement from a larger sample and Vapnik-Chervonenkis theory, among others. Our focus in this section and the following sections has been and will be the application of these ideas to develop the asymptotic tools needed for statistical theory.

Summary. We introduced *separable stochastic processes* $\{Z(t) : t \in T\}$ on a general space T which typically is a space of functions and defined the *weak convergence* of a sequence of stochastic processes $Z_n(\cdot)$ to a *tight* stochastic process $Z(\cdot)$. Weak convergence allows us to conclude that the distribution of a l_∞ continuous function q of $Z_n(\cdot)$ converges to the distribution of $q(Z(\cdot))$. We gave sufficient conditions for weak convergence in terms of probability bounds on oscillations of $Z_n(\cdot)$. This led to the consideration of *maximal inequalities* that provide such bounds. We applied this framework and results to the *likelihood and partial sum processes* and established weak convergence of the partial sum process to the Wiener process. We devoted Section 7.1.3 to the *empirical process*

$$\mathcal{E}_n(h) = n^{\frac{1}{2}} \int [h(x) - Eh(X)] d\hat{P}(x)$$

for h in some space T of functions, where \hat{P} is the empirical probability of X_1, \dots, X_n i.i.d. as $X \sim P$. $\mathcal{E}_n(\cdot)$ converges weakly to a Brownian bridge $W_P^0(\cdot)$ on T , which is defined to be a zero mean Gaussian process with

$$\text{Cov}(W_P^0(g), W_P^0(h)) = \text{Cov}_P(g(X), h(X)).$$

For such processes, we introduced the bracketing entropy (number) and established maximal inequalities for $W_P^0(\cdot)$ and $\mathcal{E}_n(h)$ under *envelope* bounds on $h(x)$ and polynomial bounds on the bracketing number (the *polynomial bracketing* condition). We established *Donsker's theorem* which states that for uniform P the classical empirical process converges weakly to the Brownian bridge W^0 and we generalized this result to give conditions under which $\{\mathcal{E}_n(h) : h \in T\}$ converges weakly to $\{W_P^0(h) : h \in T\}$. A class of functions T such that $\{\mathcal{E}_n(h) : h \in T\}$ converges weakly to $\{W_P^0(h) : h \in T\}$ is called a *Donsker class*. Finally, we used the results to derive the asymptotic distributions of the Kolmogorov and Cramér-von Mises statistics, and to give a proof of the Glivenko-Cantelli theorem.

7.2 The Delta Method in Infinite Dimensional Space

We have applied the stochastic delta method successfully to M estimates in Sections 5.4 and 6.2 by Taylor expanding the estimating equations defining the parameter $\nu(P)$. There are many examples, such as the minimum distance estimates and trimmed mean of Section I.3, where it is unclear what a linear approximation to $\hat{\nu}_n = \nu(\hat{P})$ means. The first and perhaps the most valuable part of this section is the development of a heuristic method which yields the “correct” linear approximation if one exists. In this section, as in most of the text, we are limited to the i.i.d. case.

7.2.1 Influence Functions. The Gâteaux and Fréchet Derivatives

Let \hat{P} denote the empirical distribution of X_1, \dots, X_n i.i.d. as P . What is a linear approximation to a statistic $\hat{\nu}_n = \nu(\hat{P})$ which is a consistent estimate of a parameter $\nu(P)$? In analogy with the one and d dimensional approximations (5.3.23), (5.4.22), and (6.2.3), we define it to be an expression of the form

$$\nu(P) + \frac{1}{n} \sum_{i=1}^n \psi(X_i, P) = \nu(P) + \int \psi(x, P) d\hat{P}(x) \quad (7.2.1)$$

where $E_P \psi(X, P) = 0$ and $E_P \psi^2(X, P) < \infty$. Since

$$\frac{1}{n} \sum_{i=1}^n \psi(X_i, P) = O_P(n^{-\frac{1}{2}})$$

we shall say that (7.2.1) is a *valid linear approximation* to $\hat{\nu}_n$ if

$$\hat{\nu}_n - \nu(P) - \int \psi(x, P) d\hat{P}(x) = o_P(n^{-\frac{1}{2}}). \quad (7.2.2)$$

Thus $\hat{\nu}_n - \nu(P)$ is asymptotically an average of i.i.d. variables and $\sqrt{n}(\hat{\nu}_n - \nu(P))$ will be asymptotically normal. For instance, if $\hat{\nu}_n = \nu(\hat{P}) = h(\bar{X})$ for some differentiable function h , then $\psi(x, P) = h'(\mu)(x - \mu)$ and

$$\sqrt{n}[h(\bar{X}) - \mu] = \sqrt{n}[h'(\mu)(\bar{X} - u)] + o_P(1).$$

See the proof of (A.14.17).

The function $\psi(\cdot, P)$ is unique if it exists (Problem 7.2.19) and we shall call it the *influence function* of $\nu(\hat{P})$.

The terminology and the basic idea of the approximation come from Section 5.4.1. Let

$$\nu : \mathcal{M} \rightarrow R$$

be a parameter where

- (i) $\mathcal{M} \supset \mathcal{P}$, with \mathcal{P} the model of primary interest.
- (ii) $\mathcal{M} \supset$ all distributions with finite support $\subset \mathcal{X}$, i.e., \mathcal{M} includes the empirical distribution.
- (iii) \mathcal{M} is convex. If P and Q belong to \mathcal{M} , then

$$(1 - \varepsilon)P + \varepsilon Q \in \mathcal{M} \text{ for all } 0 \leq \varepsilon \leq 1 .$$

Define, in accord with Huber (1981) the *Gâteaux derivative* of ν at P in the direction of Q by

$$\psi_0(Q, P) = \frac{\partial}{\partial \varepsilon} \nu((1 - \varepsilon)P + \varepsilon Q)|_{\varepsilon=0} \quad (7.2.3)$$

where the one-sided derivative is assumed to exist. If the influence function as defined above exists, then under general conditions (e.g., Huber (1981), Bickel, Klassen, Ritov and Wellner (1993,1998)), it can be computed as a function of $x \in \mathcal{X}$ as,

$$\psi(x, P) \equiv \psi_0(\delta_x, P) \quad (7.2.4)$$

where δ_x is point mass at x . Under appropriate conditions (see Example 7.2.1 and Problem 7.2.3)

$$\psi_0(Q, P) = \int \psi(x, P) dQ(x) \quad (7.2.5)$$

for all $Q \in \mathcal{M}$. In particular, since $\psi_0(P, P) = 0$, we have a fundamental property of the influence function,

$$E_P \psi(X, P) = 0 . \quad (7.2.6)$$

Why should we expect (7.2.2) and (7.2.5) to hold? Here is an example where they are valid. If $\mathcal{X} = \{x_1, \dots, x_k\}$, then we are in the multinomial case we studied in Section 5.4.1.

Example 7.2.1. *The multinomial case.* We can identify P with $(p_1, \dots, p_k)^T$ where $p_j = P\{X = x_j\}$, $j = 1, \dots, k$, and we can write the parameter $\nu(P)$ as $h(p_1, \dots, p_k)$ for some $h : R^k \rightarrow R$. In this case we know from Theorem 5.3.2 and Section 5.4.1 that the influence function $\psi(x, P)$ is obtained from the total differential. We take $\mathcal{M} = \mathcal{P} = \{\mathbf{q} : 0 < q_i < 1, \sum_{i=1}^k q_i = 1\}$. Because of the restriction $\sum_{i=1}^k q_i = 1$, the total differential evaluated at $x = x_j$ under the conditions of Section 5.4.1 is (see (B.8.14))

$$\begin{aligned} \psi(x_j, P) &= \lim_{\varepsilon \downarrow 0} \frac{1}{\varepsilon} (h((1 - \varepsilon)p_1, \dots, (1 - \varepsilon)p_{j-1}, (1 - \varepsilon)p_j + \varepsilon, \\ &\quad (1 - \varepsilon)p_{j+1}, \dots, (1 - \varepsilon)p_k) - h(p_1, \dots, p_k)) \\ &= \frac{\partial h}{\partial p_j}(\mathbf{p}) - \sum_{t=1}^k \frac{\partial h}{\partial p_t}(\mathbf{p}) p_t = \sum_{t=1}^k \frac{\partial h}{\partial p_j}(\mathbf{p}) [1(x_t = x_j) - p_t] \end{aligned} \quad (7.2.7)$$

and if $Q \leftrightarrow (q_1, \dots, q_k)$,

$$\begin{aligned}\psi_0(Q, P) &= \lim_{\varepsilon \downarrow 0} \frac{1}{\varepsilon} (h((1 - \varepsilon)p_1 + \varepsilon q_1, \dots, (1 - \varepsilon)p_k + \varepsilon q_k) - h(p_1, \dots, p_k)) \\ &= \sum_{j=1}^k \frac{\partial h}{\partial p_j}(\mathbf{p})(q_j - p_j) = \sum_{j=1}^k \psi(x_j, P) q_j\end{aligned}\quad (7.2.8)$$

which is just (7.2.5). We see that

$$\psi(X, P) = \sum_{j=1}^k \frac{\partial h}{\partial p_j}(\mathbf{p})(1(X = x_j) - p_j)$$

is the function appearing in (5.4.10) and that (5.4.10) exactly gives (7.2.2) with $\widehat{\nu}_n = \nu(\widehat{P})$ where \widehat{P} is the empirical probability, that is,

$$\nu(\widehat{P}) = h\left(\frac{N_1}{n}, \dots, \frac{N_k}{n}\right) = \nu(P) + \frac{1}{n} \sum_{i=1}^n \psi(X_i, P) + o_P(n^{-\frac{1}{2}}) \quad (7.2.9)$$

Equation (7.2.9) just says that (7.2.1) is a valid linear approximation to $\nu(\widehat{P})$ in this case. \square

In the multinomial case, continuity of $\frac{\partial h}{\partial p_j}(\cdot)$ in a neighborhood of \mathbf{p} is enough to legitimize the existence of a total differential for h at \mathbf{p} , which is what (7.2.7) corresponds to. Also, all expressions such as $\int \Psi(x, P) dQ(x)$ naturally exist and are finite. There are different generalizations of the total differential to infinite dimensional spaces. At this point, we simply operate formally, derive the influence function, and show, by example, the validity of (7.2.5), and, in some cases, (7.2.2) for a number of examples. We also relate the influence function to the sensitivity curve of Section 3.5 (see Remark 3.5.1 and Problem 7.2.2).

Here is an extension of the multinomial example.

Example 7.2.2. Moments and functions of moments. Suppose \mathcal{M} is as in (i), (ii), and (iii),

$$\nu(P) = h\left(\int \mathbf{g}(x) dP\right)$$

where $\mathbf{g}(\cdot) = (g_1, \dots, g_k)^T$ is a vector of functions from \mathcal{X} to R such that

$$\int |\mathbf{g}|^2(x) dP(x) < \infty \quad \text{for all } P \in \mathcal{M}$$

and $h : R^k \rightarrow R$ has a total differential at $\mu_{\mathbf{g}}(P) \equiv \int \mathbf{g}(x) dP(x)$. Then, by definition, for $P, Q \in \mathcal{M}$,

$$\nu((1 - \varepsilon)P + \varepsilon Q) = h\left(\int \mathbf{g}(x) dP(x) + \varepsilon \left[\int \mathbf{g}(x) d(Q - P)(x)\right]\right)$$

and, by definition,

$$\begin{aligned}\psi_0(\delta_x, P) &= \frac{\partial h}{\partial \varepsilon}(\boldsymbol{\mu}_g(P) + \varepsilon(g(x) - \boldsymbol{\mu}_g(P)))|_{\varepsilon=0} \\ &= \sum_{j=1}^k \frac{\partial h}{\partial \mu_{gj}}(\boldsymbol{\mu}_g(P))(g_j(x) - \mu_{gj}(P)) \\ &= \nabla^T h(\boldsymbol{\mu}_g(P))(g(x) - \boldsymbol{\mu}_g(P)).\end{aligned}\quad (7.2.10)$$

It can be shown (Problem 7.2.3) that (7.2.5) holds. Moreover, if we set $\psi(x, P) = \psi_0(\delta_x, P)$, then by (5.3.23),

$$\nu(\hat{P}) = \nu(P) + \frac{1}{n} \sum_{i=1}^n \psi(X_i, P) + o_P(n^{-\frac{1}{2}}) \quad (7.2.11)$$

and (7.2.2) holds. We have already seen in Chapters 5 and 6 how to apply (7.2.10) to get expressions for the asymptotic behaviour of moments and cumulants. \square

Recall that a minimum contrast (MC) estimate is defined in Section 5.4.2 as $\hat{\theta} = \arg \min \int \rho(x, \theta) d\hat{F}(x)$ for some contrast function ρ , while a M estimate is defined as the solution to $\int \psi(x, \theta) d\hat{F}(x) = 0$ for some function ψ . An important special case of M estimates is obtained by choosing ψ as the derivative of a contrast function.

Example 7.2.3. Quantiles. The median is a MC estimate which can be shown to be consistent and asymptotically normal (Problem 5.4.1), yet does not satisfy the conditions of Theorems 5.4.2. Specifically, suppose $X \in R$ has distribution function F with density f and uniquely defined median $\nu(F) = F^{-1}(.5)$, such that $f(\nu(F)) > 0$. Then, from Problem 5.4.1, if $\hat{F}^{-1}(.5)$ is a sample median,

$$\mathcal{L}(\sqrt{n}(\hat{F}^{-1}(.5) - \nu(F))) \rightarrow \mathcal{N}(0, \sigma^2(F))$$

where, $\sigma^2(F) = 1/4f^2(\nu(F))$.

The mean and median are informative of the center of a population. Sometimes quantiles are also of interest. For instance in addition to median income we may want to examine the 10th or the 99th percentile of income. We next formally work out the influence function of the α quantile $\nu_\alpha(F)$, $0 < \alpha < 1$, defined by

$$F(\nu_\alpha(F)) = \alpha. \quad (7.2.12)$$

We assume that equation (7.2.12) has a unique solution for $F \leftrightarrow P \in \mathcal{M}$. In general, $\nu_\alpha(F)$ corresponds (Problem 7.2.7) to the MC estimate, with

$$\begin{aligned}\rho(x, \theta) &= (1 - \alpha)|x - \theta|, \quad x < \theta \\ &= \alpha|x - \theta|, \quad x \geq \theta.\end{aligned}$$

If F is continuous and increasing, $\nu_\alpha(F)$ is the parameter (corresponding to an M estimate) which solves $\int \phi_\alpha(x, \theta) dF(x) = 0$ with

$$\begin{aligned}\phi_\alpha(x, \theta) &= -(1 - \alpha), \quad x < \theta \\ &= \alpha, \quad x \geq \theta.\end{aligned}\quad (7.2.13)$$

The median indeed corresponds to $\alpha = .5$.

If $F = \widehat{F}$, as we know for the median, $\nu_\alpha(\widehat{F})$ is not uniquely defined in general. For convenience, we use the *sample quantile* (see (2.1.17))

$$\begin{aligned}\widehat{x}_\alpha &= \frac{1}{2}[\widehat{F}^{-1}(\alpha) + \widehat{F}_U^{-1}(\alpha)] = X_{([n\alpha]+1)}, \text{ if } n\alpha \notin \{1, \dots, n\} \\ &= \frac{1}{2}[X_{([n\alpha])} + X_{([n\alpha]+1)}], \text{ otherwise}\end{aligned}\quad (7.2.14)$$

where $\widehat{F}_U^{-1}(\alpha) = \sup\{x : \widehat{F}(x) \leq \alpha\}$, $X_{(n+1)} = X_{(n)}$, and $X_{(1)} \leq \dots \leq X_{(n)}$ are the ordered X_1, \dots, X_n .

We now proceed formally from (7.2.12) assuming $\nu_\alpha(F)$ is uniquely defined on \mathcal{M} , and that we are evaluating the influence function at $F \leftrightarrow P$ such that $f(\nu_\alpha(F))$ is defined and > 0 . Then, we can differentiate (7.2.12) in the form

$$((1 - \varepsilon)F + \varepsilon G)(\nu_\alpha((1 - \varepsilon)F + \varepsilon G)) = \alpha$$

with respect to ε , evaluated at $\varepsilon = 0$, to get

$$(G - F)(\nu_\alpha(F)) + f(\nu_\alpha(F)) \frac{\partial}{\partial \varepsilon} \nu_\alpha(F + \varepsilon(G - F))|_{\varepsilon=0} = 0$$

which, by (7.2.4), yields, by substituting $G(y) = 1(y \geq x)$ corresponding to δ_x , the influence function,

$$\psi_\alpha(x, F) = \frac{-(1(\nu_\alpha(F) \geq x) - \alpha)}{f(\nu_\alpha(F))}. \quad (7.2.15)$$

Expressions (7.2.15) and (7.2.2) lead to the expectation that we have, for $x_\alpha = F^{-1}(\alpha)$,

$$\widehat{x}_\alpha = x_\alpha + \frac{1}{n} \sum_{i=1}^n \frac{-(1(X_i \leq F^{-1}(\alpha)) - \alpha)}{f(F^{-1}(\alpha))} + o_P(n^{-\frac{1}{2}}) \quad (7.2.16)$$

and hence, by the central limit theorem,

$$\sqrt{n}(\widehat{x}_\alpha - x_\alpha) \implies \mathcal{N}(0, \sigma_\alpha^2(F))$$

where $\sigma_\alpha^2(F) = \alpha(1 - \alpha)/f^2(F^{-1}(\alpha))$. That (7.2.16) is, in fact, correct is shown in Problem 7.2.8. \square

The Fréchet derivative

To obtain a general theorem yielding (7.2.2) we need stronger notions of differentiation than that due to Gâteaux. These necessarily rely on some definition of (metric) topology on \mathcal{M} , a convex subset of probabilities on \mathcal{X} containing \mathcal{P} , and all discrete distributions. We can regard the Gâteaux derivative as a partial derivative in the direction $(1 - \varepsilon)P + \varepsilon Q$. We next turn to a stronger differentiability that corresponds to the existence of a total differential.

Definition 7.2.1. d is a *suitable metric* on \mathcal{M} , if it satisfies the usual a)–d) below, and e).

- a) $d \geq 0$
- b) $d(P, Q) = d(Q, P)$
- c) $d(P, Q) = 0$ iff $P = Q$
- d) $d(P, Q) \leq d(P, R) + d(R, Q)$ all $P, R, Q \in \mathcal{M}$
- e) $d((1 - \varepsilon)P + \varepsilon Q, P) \leq \varepsilon d(P, Q)$ all $P, Q \in \mathcal{M}$

Examples of suitable metrics are

$$d(P, Q) = \sup\left\{\left|\int f(dP - dQ)\right| : f \in \mathcal{C}\right\}$$

where \mathcal{C} is a Donsker class satisfying: $\int f dP = \int f dQ$ for all $f \in \mathcal{C} \implies P = Q$. For instance, if $\mathcal{X} = \mathbb{R}$ and $\mathcal{C} = \{1_z(\cdot) : z \in \mathbb{R}\}$ where $1_z(t) = 1(t \leq z)$, then d is the Kolmogorov (sup) metric.

The strongest (often too strong) definition of differentiability of a function $\nu : \mathcal{M} \rightarrow \mathbb{R}$ is F (for Fréchet) differentiability. In the following we write “Rem” for the remainder of the approximation $\int \psi(x, P) dQ(x)$ to $\nu(Q) - \nu(P)$.

Definition 7.2.2. ν is F differentiable at P iff there exists $\psi(\cdot, P)$ such that

- (a) $\int \psi(x, P) dP(x) = 0$.
- (b) For all $\varepsilon > 0$, there exists $\delta(\varepsilon) > 0$ such that if $d(P, Q) \leq \delta(\varepsilon)$, then

$$\text{Rem}(P, Q) \equiv |\nu(Q) - \nu(P) - \int \psi(x, P) dQ(x)| \leq \varepsilon d(P, Q).$$

Note that $\int \psi(x, P) dQ(x) = \int \psi(x, P) d(Q - P)$. Equivalently, if Q_m is any sequence such that $d(Q_m, P) \rightarrow 0$, and if (a) holds, then (b) is equivalent to

$$\text{Rem}(P, Q_m) = o(d(P, Q_m)).$$

Write \widehat{P}_n for the empirical probability \widehat{P} to clearly indicate the dependence on sample size.

Theorem 7.2.1. If ν is F differentiable at P with derivative ψ with respect to a suitable metric d and

- (i) $d(\widehat{P}_n, P) = O_P(n^{-\frac{1}{2}})$
- (ii) $\psi(\cdot, P) \in L_2(P)$

then

$$\nu(\widehat{P}_n) = \nu(P) + \frac{1}{n} \sum_{i=1}^n \psi(X_i, P) + o_P(n^{-\frac{1}{2}}).$$

That is, $\widehat{\nu}_n = \nu(\widehat{P}_n)$ has influence function ψ .

Proof. By hypothesis, for $\delta(\varepsilon)$ as in Definition 7.2.2,

$$P[d(\hat{P}_n, P) \leq \delta(\varepsilon)] \rightarrow 1 \text{ for every } \varepsilon > 0.$$

Therefore, by F differentiability,

$$P[|\nu(\hat{P}_n) - \nu(P) - \int \psi(x, P)d(\hat{P}_n - P)(x)| \leq \varepsilon' d(\hat{P}_n, P)] \rightarrow 1 \text{ for every } \varepsilon' > 0.$$

By (i), for all $\varepsilon > 0$, there exists M such that

$$P[d(\hat{P}_n, P) \leq Mn^{-\frac{1}{2}}] \geq 1 - \varepsilon.$$

Take $\varepsilon' = \frac{\varepsilon}{M}$, then, for every $\varepsilon > 0$,

$$P[|\nu(\hat{P}_n) - \nu(P) - \int \psi(x, P)d(\hat{P}_n - P)(x)| \leq \varepsilon n^{-\frac{1}{2}}] \rightarrow 1.$$

□

Example 7.2.4. *Cramér-von Mises (C-vM) Statistic.* We want to see how the C-vM statistic behaves when $H : P = P_0$ is false and $\mathcal{X} = R$. The C-vM statistic of Example 7.1.6 is $\nu(\hat{F})$, where

$$\nu(F) = \int (F(y) - F_0(y))^2 dF_0(y).$$

By a standard calculation (Problem 7.2.9), the influence function is given by

$$\psi(x, P) = 2 \int_{-\infty}^{\infty} (1(y \leq x) - F(y))(F(y) - F_0(y))dF_0(y). \quad (7.2.17)$$

It can be shown (Problem 7.2.9) that, if we take d as the Kolmogorov metric

$$d(P, Q) = \sup_y |P(-\infty, y] - Q(-\infty, y)|$$

then,

$$\text{Rem}(P, Q) = o(d(P, Q)).$$

Thus, if $P \neq P_0$, the Cramér-von Mises statistic is asymptotically normal with mean $\int_{-\infty}^{\infty} (F - F_0)^2(y) dF_0(y)$ and variance $n^{-1} \int_{-\infty}^{\infty} \psi^2(x, P) dF(x)$ where ψ is given by (7.2.17). This is quite different from its behaviour under the hypothesis $H : F = F_0$ where its asymptotic behaviour is non-Gaussian as we will see in Example 7.3.1. □

There is a weaker but more broadly applicable differentiability notion due to Hadamard. We refer the reader to van der Vaart (1998), Chapter 20, for a clear discussion of this notion and its utility.

We shall give some more applications of Theorem 7.2.1 in the problems. Unfortunately, it is hard to apply, even when weakened to the Hadamard form. The problem is

that many parameters are defined implicitly and what is most serious — the argument P appears as dP , as in the median which minimizes $\int |x - \theta| dP(x)$. Convergence in a suitable metric having the \sqrt{n} consistency property, $d(\hat{P}, P) = O_P(n^{-\frac{1}{2}})$, is not sufficient to have continuity of such $\nu(P)$ in general, much less differentiability. It is often the case that, after formal calculation of the influence function, proving that it provides a suitable approximation is more easily done by ad hoc methods. We illustrate this in the following Theorem 7.2.2.

In this result, we also illustrate that the influence function gives us all relevant first order asymptotic information about $\nu(\hat{P})$ if we have a vector parameter

$$\boldsymbol{\nu}(P) = (\nu_1(P), \dots, \nu_d(P))^T.$$

If the expansion (7.2.2) is valid for $\nu_j(P)$, $j = 1, \dots, d$, and $\psi_j(x, P)$ is the influence function corresponding to $\nu_j(P)$, then

$$\sqrt{n}(\boldsymbol{\nu}(\hat{P}) - \boldsymbol{\nu}(P)) = \frac{1}{\sqrt{n}} \sum_{i=1}^n \boldsymbol{\psi}(X_i, P) + o_P(1) \quad (7.2.18)$$

where $\boldsymbol{\psi}(x, P) \equiv (\psi_1(x, P), \dots, \psi_d(x, P))^T$ and $E_P \boldsymbol{\psi}(X, P) = \mathbf{0}$, $E_P |\boldsymbol{\psi}|^2(X, P) < \infty$. Here $\boldsymbol{\psi}$ is the natural extension of the concept of an influence function from one dimensional parameters to d dimensional parameters. By the multivariate central limit theorem, we conclude that

$$\sqrt{n}(\boldsymbol{\nu}(\hat{P}) - \boldsymbol{\nu}(P)) \implies \mathcal{N}_d(\mathbf{0}, E_P \boldsymbol{\psi} \boldsymbol{\psi}^T(x)). \quad (7.2.19)$$

We now present a method of Huber(1967) for establishing influence function approximations for M estimates which covers situations like the median or its analogues in more than 1 dimension. It is an extension of Theorem 6.2.2.

Theorem 7.2.2. Suppose $\phi : \mathcal{X} \times R^p \rightarrow R^p$ satisfies

A0': $\hat{\theta}_n$ is such that

$$\int \phi(x, \hat{\theta}_n) d\hat{P}_n(x) = o_P(n^{-\frac{1}{2}}). \quad (7.2.20)$$

A1: $\boldsymbol{\theta}(P)$ uniquely satisfies $\int \phi(x, \boldsymbol{\theta}(P)) dP(x) = 0$ for all $P \subset \mathcal{Q}$.

A2: $E_P |\phi(X, \boldsymbol{\theta}(P))|^2 < \infty$ for all $P \in \mathcal{Q}$.

A5': $\hat{\theta}_n$ is consistent for $\boldsymbol{\theta}(P)$ for $P \in \mathcal{Q}$.

A4': The set $T_\varepsilon = \{\phi(\cdot, \boldsymbol{\theta}) - \phi(\cdot, \boldsymbol{\theta}(P)) : |\boldsymbol{\theta} - \boldsymbol{\theta}(P)| \leq \varepsilon\}$ satisfies the conditions of Theorem 7.1.3 with envelope function $s(\cdot, \varepsilon)$ such that $E_P s^2(X, \varepsilon) = o(1)$ as $\varepsilon \rightarrow 0$, where statements about the vector ϕ refer to each of its components ϕ_j .

A3': The map $\boldsymbol{\theta} \rightarrow \int \phi(x, \boldsymbol{\theta}) dP(x)$ has a total differential $H(P)$ at $\boldsymbol{\theta}(P)$ which is non-singular;

$$H(P) = \left[\frac{\partial}{\partial \theta_j} E_P \phi_i(X, \boldsymbol{\theta}(P)) \right]_{1 \leq i, j \leq p}$$

where $[\cdot]$ denotes a matrix. Then,

$$\widehat{\boldsymbol{\theta}} = \boldsymbol{\theta}(P) + \frac{1}{n} \sum_{i=1}^n [-H]^{-1}(P) \boldsymbol{\phi}(X_i, \boldsymbol{\theta}(P)) + o_P(n^{-\frac{1}{2}}). \quad (7.2.21)$$

Note (Problem 7.2.11) that formal calculation of the influence function for $\boldsymbol{\theta}(P)$ as a solution of (6.2.2) yields $H^{-1}\boldsymbol{\phi}$.

Proof: By A4' and Theorem 7.1.3,

$$\sup\{|\mathcal{E}_n(\boldsymbol{\phi}(\cdot, \boldsymbol{\theta}) - \boldsymbol{\phi}(\cdot, \boldsymbol{\theta}(P)))| : |\boldsymbol{\theta} - \boldsymbol{\theta}(P)| \leq \varepsilon_n\} = o_P(1) \quad (7.2.22)$$

if $\varepsilon_n \rightarrow 0$. By A5', it follows that

$$\mathcal{E}_n(\boldsymbol{\phi}(\cdot, \widehat{\boldsymbol{\theta}}_n) - \boldsymbol{\phi}(\cdot, \boldsymbol{\theta}(P))) = o_p(1). \quad (7.2.23)$$

Translating (7.2.23) by using (7.1.19) and using A0' and A1, we obtain that

$$-n^{\frac{1}{2}} E_P \boldsymbol{\phi}(X, \widehat{\boldsymbol{\theta}}_n) = n^{-\frac{1}{2}} \sum_{i=1}^n \boldsymbol{\phi}(X_i, \boldsymbol{\theta}(P)) + o_P(1). \quad (7.2.24)$$

But, by A3' and A1, since $n^{-\frac{1}{2}} \sum_{i=1}^n \boldsymbol{\phi}(X_i, \boldsymbol{\theta}(P)) = O_P(1)$ by the central limit theorem,

$$E_P \boldsymbol{\phi}(X, \widehat{\boldsymbol{\theta}}_n) = H(P)(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}(P)) + o_P(|\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}(P)|) \quad (7.2.25)$$

since $E_p \boldsymbol{\phi}(X, \boldsymbol{\theta}(P)) = \mathbf{0}$. Combining (7.2.24) and (7.2.25) we obtain the result. \square

Here is an example.

Example 7.2.5. *Multivariate “medians.”* Define for P on R^d .

$$\boldsymbol{\theta}(P) \equiv \arg \min \int |\mathbf{x} - \boldsymbol{\theta}|^r dP(\mathbf{x})$$

where $|\mathbf{x}|$ is the Euclidean norm and $1 \leq r \leq 2$. It is not hard to show (Problem 7.2.12) that $\boldsymbol{\theta}(P)$ is defined if $\int |\mathbf{x}|^r dP(\mathbf{x}) < \infty$ and is unique for $r > 1$. It is defined when $r = 1$, which corresponds to the median for $d = 1$, but then $\boldsymbol{\theta}(P)$ may not be unique. If P has a density $p > 0$, then $\boldsymbol{\theta}(P)$ is unique even if $r = 1$ (Problem 7.2.12). Suppose P is such that

- (i) P has a positive density p .
- (ii) $\int |\mathbf{x}|^{2(r-1)} p(\mathbf{x}) d\mathbf{x} < \infty$.

Then,

$$\boldsymbol{\theta}(\widehat{P}_n) = \boldsymbol{\theta}(P) + \frac{1}{n} \sum_{i=1}^n \boldsymbol{\psi}(X_i, P) + o_P(n^{-\frac{1}{2}})$$

where

$$\boldsymbol{\psi}(\mathbf{x}, P) = M(P)(\mathbf{x} - \boldsymbol{\theta}(P))|\mathbf{x} - \boldsymbol{\theta}(P)|^{r-2}$$

and, if $\mathbf{X} \equiv (X_1, \dots, X_d)^T$, $\delta_{ij} = 1[i = j]$,

$$M^{-1}(P) = E_P \left\{ \frac{(X_i - \theta_i(P))(X_j - \theta_j(P))}{|\mathbf{X} - \boldsymbol{\theta}(P)|^{4-r}} (r-2) + \delta_{ij} |\mathbf{X} - \boldsymbol{\theta}(P)|^{r-2} \right\}. \quad (7.2.26)$$

The matrix $M(P)$ is well defined for $r > 1$ and corresponds to $[-H]$ in (7.2.21). For $r = 1$, expression (7.2.26) becomes $\infty - \infty$, but passage to the limit as r decreases to 1 yields $\psi(\mathbf{x}, P) = \text{sgn}(\mathbf{x} - \boldsymbol{\theta}(P))/2p(\theta(P))$ as in Problem 5.4.1(e). If $r > 1$, $E_P |\mathbf{X}|^{2r-2} < \infty$ makes (7.2.26) well defined — see Problem 7.2.13. Checking the conditions of Theorem 7.2.2 otherwise requires some arguments — see Huber (1967) and Problems 7.2.11 and 7.2.12. \square

7.2.2 The Quantile Process

We now give another approach to justifying the quantile representation (7.2.13) and deducing weak convergence for an infinite dimensional parameter, the *quantile process* defined by

$$Q_n(t) \equiv \sqrt{n}(\hat{F}^{-1}(t) - F^{-1}(t)), \quad 0 < t < 1 \quad (7.2.27)$$

for suitable F . The approach is due to Shorack (1982); see also Shorack and Wellner (1986). Let $W^0(u)$ denote the Brownian bridge on $[0, 1]$. Note that $-W_0(u)$ is also a Brownian bridge. We state the result as

Theorem 7.2.3. Suppose P on R has distribution F with continuous positive density $f = F'$. Then, $F(F^{-1}(t)) = t$ for $0 < t < 1$ and

- (i) For all $0 < \varepsilon \leq \frac{1}{2}$, $Q_n(t)$ can be approximated by $-\mathcal{E}_n(F^{-1}(t))/f(F^{-1}(t))$ in the sense that

$$\sup\{|\hat{F}^{-1}(t) - F^{-1}(t) + \frac{(\hat{F}(F^{-1}(t)) - t)}{f(F^{-1}(t))}| : \varepsilon \leq t \leq 1 - \varepsilon\} = o_P(n^{-\frac{1}{2}}). \quad (7.2.28)$$

- (ii) If $T = [\varepsilon, 1 - \varepsilon]$, and $Q_n(t)$ defined in (7.2.27) is viewed as a stochastic process on T , then

$$Q_n \implies W^0(\cdot)/f(F^{-1}(\cdot)) \quad (7.2.29)$$

where, by definition, $W^0(\cdot)/f(F^{-1}(\cdot))$ is a Gaussian process with mean 0 on $(0, 1)$,

$$\text{Cov} \left(\frac{W^0(s)}{f(F^{-1}(s))}, \frac{W^0(t)}{f(F^{-1}(t))} \right) = \frac{s(1-t)}{f(F^{-1}(s))f(F^{-1}(t))}, \quad s \leq t,$$

which has continuous sample functions since $W^0(\cdot)$ does.

- (iii) If $f(F^{-1}(t)) \geq \delta$, $0 \leq t \leq 1$, then ε can be taken as 0 in (i) and (ii).

(iv) If $P = \mathcal{U}(0, 1)$ and $\mathcal{E}_n(\cdot)$ is the classical empirical process, we have

$$\|Q_n + \mathcal{E}_n\|_\infty = o_P(1) \quad (7.2.30)$$

and $Q_n(\cdot)$ converges weakly to the Brownian bridge $-W^0(\cdot)$.

Proof. Write

$$\begin{aligned} \widehat{F}^{-1}(t) - F^{-1}(t) &= \frac{(\widehat{F}^{-1}(t) - F^{-1}(t))}{F(\widehat{F}^{-1}(t)) - t} \cdot (F(\widehat{F}^{-1}(t)) - t) \\ &= \frac{(\widehat{F}^{-1}(t) - F^{-1}(t))}{F(\widehat{F}^{-1}(t)) - F(F^{-1}(t))} \{F(\widehat{F}^{-1}(t)) - \widehat{F}(\widehat{F}^{-1}(t)) + \Delta_{n1}(t)\} \\ &\equiv I(t) \times II(t) \end{aligned} \quad (7.2.31)$$

where $\Delta_{n1}(t) = \widehat{F}(\widehat{F}^{-1}(t)) - t$ satisfies $|\Delta_{n1}(t)| \leq 1/n$ by the definition of $\widehat{F}^{-1}(t)$ since F is continuous.

We prove (iv) first. If $F(x) = x$, $0 \leq x \leq 1$,

$$\|\widehat{F}^{-1} - F^{-1}\|_\infty = o_P(1) \quad (7.2.32)$$

by the Glivenko-Cantelli Theorem (Problem 7.2.16). Now, in the $\mathcal{U}(0, 1)$ case,

$$I(t) \equiv 1.$$

On the other hand,

$$\begin{aligned} II(t) &= -(\mathcal{E}_n(\widehat{F}^{-1}(t)) + O(n^{-\frac{1}{2}}))n^{-\frac{1}{2}} \\ &= -(\mathcal{E}_n(t) + \Delta_{n2}(t) + O(n^{-\frac{1}{2}}))n^{-\frac{1}{2}} \end{aligned} \quad (7.2.33)$$

where

$$|\Delta_{n2}(t)| \leq \|\Delta_{n2}\|_\infty \leq \sup\{|\mathcal{E}_n(u) - \mathcal{E}_n(v)| : |u - v| \leq \|\widehat{F}^{-1} - F^{-1}\|_\infty\}.$$

Then

$$[\|\Delta_{n2}\|_\infty > \varepsilon] \subset [\sup\{|\mathcal{E}_n(u) - \mathcal{E}_n(v)| : |u - v| \leq \delta\} > \varepsilon] + [\|\widehat{F}^{-1} - F^{-1}\| > \delta]$$

by intersecting the event on the left hand side with $[\|\widehat{F}^{-1} - F^{-1}\|_\infty \leq \delta]$ and its complement. Thus,

$$\begin{aligned} \limsup_n P[\|\Delta_{n2}\|_\infty > \varepsilon] &\leq \limsup_n P\left[\sup\{|\mathcal{E}_n(s) - \mathcal{E}_n(t)| : |s - t| \leq \delta\} > \varepsilon\right] \\ &\quad + \limsup_n P[\|\widehat{F}^{-1} - F^{-1}\|_\infty > \delta] \end{aligned} \quad (7.2.34)$$

for all $\delta, \varepsilon > 0$. But the first term in (7.2.34) is just

$$P\left[\sup\{\|W^0(s) - W^0(t)\| : |s - t| \leq \delta\} > \varepsilon\right]$$

and the second is 0 for all $\delta > 0$ by the Glivenko-Cantelli theorem (Problem 7.2.16). Then (7.2.30) for $P = \mathcal{U}(0, 1)$ follows from (7.2.33) and (7.2.34).

For the (i) case, let U_1, \dots, U_n be i.i.d. $\mathcal{U}(0, 1)$ and \widehat{G} be their empirical df. Then, if $X_i \equiv F^{-1}(U_i)$, $1 \leq i \leq n$, X_1, \dots, X_n are i.i.d. with df F . Taking X_i represented in this way as our sample from F , we can write

$$\widehat{F}^{-1}(t) - F^{-1}(t) = F^{-1}(\widehat{G}^{-1}(t)) - F^{-1}(G^{-1}(t)) \quad (7.2.35)$$

and

$$\widehat{F}(F^{-1}(t)) - t = \widehat{G}(t) - t. \quad (7.2.36)$$

Set $t_n = \widehat{G}^{-1}(t)$. Then $t_n \xrightarrow{P} G^{-1}(t) = t$ uniformly on $[\varepsilon, 1 - \varepsilon]$ by (7.2.32). Thus, uniformly on $[\varepsilon, 1 - \varepsilon]$,

$$\frac{F^{-1}(t_n) - F^{-1}(t)}{t_n - t} \xrightarrow{P} \frac{\partial}{\partial t} F^{-1}(t) = \frac{1}{f(F^{-1}(t))}.$$

More precisely, by (7.2.30),

$$\frac{F^{-1}(\widehat{G}^{-1}(t)) - F^{-1}(G^{-1}(t))}{\widehat{G}^{-1}(t) - G^{-1}(t)} = \frac{1}{f(F^{-1}(t))}(1 + \Delta_n(t))$$

where

$$\sup\{|\Delta_n(t)| : \varepsilon \leq t \leq 1 - \varepsilon\} \xrightarrow{P} 0$$

because $f(F^{-1}(t))$ is continuous and bounded away from 0 on $[\varepsilon, 1 - \varepsilon]$ if $\varepsilon > 0$, by assumption. Thus (i) follows. The case (iii) also follows. Finally, (ii) follows from Donsker's theorem (Theorem 7.1.4) and Slutsky's theorem generalized to weak convergence in function spaces (Corollary 7.1.2). \square

Here is a first application of this result.

Example 7.2.6. *Linear combinations of order statistics.* A class of parameters which includes the trimmed means (3.5.3) are

$$\nu(P) = \int_0^1 F^{-1}(t)d\Lambda(t) \quad (7.2.37)$$

where $\Lambda(t)$ is the df of a probability distribution on $[0, 1]$. If

$$\begin{aligned} \Lambda(t) &= 0, & t < \varepsilon \\ &= 1, & t > 1 - \varepsilon \end{aligned}$$

or equivalently $\Lambda[\varepsilon, 1 - \varepsilon] = 1$, the parameter $\nu(P)$ is defined for all F . A special case is Λ_α with density

$$\begin{aligned} \lambda_\alpha(t) &= (1 - 2\alpha)^{-1}, & \alpha \leq t \leq 1 - \alpha \\ &= 0 & \text{otherwise} \end{aligned} \quad (7.2.38)$$

for $0 < \alpha < \frac{1}{2}$. Then $\nu(\widehat{P})$ is just the α trimmed mean. From Theorem 7.2.3 (iii), since

$$h \rightarrow \int_{\varepsilon}^{1-\varepsilon} h(u) d\Lambda(u)$$

is continuous in $\|\cdot\|_\infty$ we have that

$$\int_0^1 \sqrt{n}(\widehat{F}^{-1}(t) - F^{-1}(t))d\Lambda(t) \implies \int_0^1 [f(F^{-1}(t))]^{-1}W^0(t)d\Lambda(t). \quad (7.2.39)$$

The limit is $\mathcal{N}(0, \sigma^2(\Lambda, f))$ (Problem 7.2.16) with

$$\sigma^2(\Lambda, f) = \int_0^1 \int_0^1 [f(F^{-1}(s))f(F^{-1}(t))]^{-1}(s \wedge t - st)d\Lambda(s)d\Lambda(t). \quad (7.2.40)$$

We can get a simpler expression for $\sigma^2(\Lambda, f)$ and more insight by using part (ii). We obtain, if $\lambda(s) = 0$, $s \leq \varepsilon$ or $s \geq 1 - \varepsilon$, $\varepsilon > 0$,

$$\int_0^1 (\widehat{F}^{-1}(s) - F^{-1}(s))\lambda(s)ds = - \int_0^1 \frac{(\widehat{F}(F^{-1}(s)) - s)}{f(F^{-1}(s))}\lambda(s)ds + o_P(n^{-\frac{1}{2}}).$$

So the influence function of $\int_0^1 \widehat{F}^{-1}(s)\lambda(s)ds$ is

$$\begin{aligned} \psi(x, P) &= - \int_0^1 \frac{(1(x \leq F^{-1}(s)) - s)}{f(F^{-1}(s))}\lambda(s)ds \\ &= - \int_0^1 (1(x \leq F^{-1}(s)) - s)\lambda(s)dF^{-1}(s) \\ &= - \int (1(x \leq y) - F(y))\lambda(F(y))dy \\ &= - \left(\int_x^\infty \lambda(F(y))dy - E_P \left(\int_X^\infty \lambda(F(y))dy \right) \right). \end{aligned} \quad (7.2.41)$$

Specializing to the trimmed mean, with $\lambda = \lambda_\alpha$ given by (7.2.38), we get, as influence function,

$$\begin{aligned} \psi_\alpha(x, P) &= \frac{F^{-1}(\alpha)}{1-2\alpha} - \mu_\alpha(P), \quad x \leq F^{-1}(\alpha) \\ &= \frac{x}{1-2\alpha} - \mu_\alpha(P), \quad F^{-1}(\alpha) \leq x \leq F^{-1}(1-\alpha) \\ &= \frac{F^{-1}(1-\alpha)}{1-2\alpha} - \mu_\alpha(P), \quad x \geq F^{-1}(1-\alpha) \end{aligned} \quad (7.2.42)$$

where

$$\mu_\alpha(P) = \frac{\alpha(F^{-1}(\alpha) + F^{-1}(1-\alpha))}{1-2\alpha} + (1-2\alpha)^{-1} \int_{F^{-1}(\alpha)}^{F^{-1}(1-\alpha)} x dF(x).$$

From (7.2.41) the asymptotic variance of $\sqrt{n}\nu_\alpha(\hat{P})$ is

$$\begin{aligned}\sigma^2(\Lambda, f) &= (1 - 2\alpha)^{-2} \{(F^{-1}(\alpha) - \mu_\alpha(P))^2 + (F^{-1}(1 - \alpha) - \mu_\alpha(P))^2 \\ &\quad + \int_{F^{-1}(\alpha)}^{F^{-1}(1-\alpha)} (x - \mu_\alpha(P))^2 dF(x)\}.\end{aligned}\tag{7.2.43}$$

Note that if f is symmetric, $f(a + x) = f(a - x)$ for some a , then

$$\nu_\alpha(P) = (1 - 2\alpha)^{-1} \int_{F^{-1}(\alpha)}^{F^{-1}(1-\alpha)} x dF(x) = a.$$

Note the agreement between this formula and the sensitivity curve of Figure 3.5.2. \square

Remark 7.2.1. Finally, note that (7.2.39) makes sense even if $\lambda > 0$ on all of $(0,1)$ and is correct for the mean itself, $\lambda \equiv 1$. In fact (7.2.39) and (7.2.41) are valid much more generally; see Problems 7.2.21–7.2.24 and results in Bickel (1967), Chernoff, Gastwirth and Johns (1967), and Stigler (1969). \square

Summary. We introduced a generalization of the stochastic delta method from functions of i.i.d. vectors of dimension d to functions $\nu(\hat{P})$ of the empirical distribution. We did this by introducing Gâteaux and Fréchet derivatives of functions defined on ∞ dimensional linear spaces. These derivatives correspond, respectively, to partial differentiation in a fixed direction and the existence of a total differential. They lead to the important notion of the *influence function* which tells us what is the “right” approximation of the form $n^{-1} \sum_{i=1}^n \psi(X_i, P)$ to $\nu(\hat{P})$ we should aim for. We derive influence functions for functions of multinomial proportions, functions of sample moments, sample quantiles, the Cramér-von Mises statistic, multivariate medians, and linear combination of order statistics. The Fréchet derivative is used to obtain the validity of influence function approximation for statistics which are not simply functions of vector means or characterizable as M estimates. Empirical process theory is used to justify more general results for M estimates than those of Chapter 5, and in Section 7.2.2 yields weak convergence of the *quantile process*.

7.3 Further Expansions

7.3.1 The von Mises Expansion

The influence function gives us the analogue of the first derivative of a parameter. The von Mises expansion is the formal analogue of a Taylor series and tells us, for instance, what we can expect if the “first derivative” vanishes. Write, given $\nu : \mathcal{M} \rightarrow \mathbb{R}$,

$$\begin{aligned}\nu(P + \varepsilon(Q - P)) &= \nu(P) + \sum_{k=1}^m \frac{\nu^{(k)}(P - Q)}{k!} \varepsilon^k + o(\varepsilon^{m+1}) \\ \text{where } \nu^{(k)}(Q - P) &= \frac{\partial^k}{\partial \varepsilon^k} \nu(P + \varepsilon(Q - P))|_{\varepsilon=0}\end{aligned}\tag{7.3.1}$$

the usual Taylor expansion. Suppose, as can easily be shown to hold under mild conditions in case \mathcal{X} is finite (Problem 7.3.1), that

$$\frac{\partial^k}{\partial \varepsilon^k} \nu(P + \varepsilon(Q - P))|_{\varepsilon=0} = \int \dots \int \psi(x_1, \dots, x_k, P) \Pi_{j=1}^k dQ(x_j) \quad (7.3.2)$$

where

- (i) ψ is symmetric in x_1, \dots, x_k .
- (ii) $\int \psi(x, X_2, \dots, X_k, P) dP(x) = 0$ with P probability 1 or, equivalently,

$$E_P(\psi(X_1, \dots, X_k, P) | X_2, \dots, X_k) = 0. \quad (7.3.3)$$

From (ii) it follows that

$$\int \dots \int \psi(x_1, \dots, x_k, P) \Pi_{j=1}^k dQ(x_j) = \int \dots \int \psi(x_1, \dots, x_k, P) \Pi_{j=1}^k d(Q-P)(x_j). \quad (7.3.4)$$

Further, ψ may be computed formally as follows. For $Q_1, \dots, Q_k \in \mathcal{M}$, define

$$\psi_0(Q_1, \dots, Q_k, P) \equiv \frac{\partial^k}{\partial \varepsilon_1 \dots \partial \varepsilon_k} \nu(P + \sum_{j=1}^k \varepsilon_j (Q_j - P))|_{\varepsilon_1=\dots=\varepsilon_k=0} \quad (7.3.5)$$

where $\varepsilon_j \geq 0$, $\sum_{j=1}^k \varepsilon_j < 1$. Then, extending (7.2.4),

$$\psi(x_1, \dots, x_k, P) = \psi_0(\delta_{x_1}, \dots, \delta_{x_k}, P). \quad (7.3.6)$$

It is easy to see that (7.3.5) and (7.3.2) imply (7.3.6) and (7.3.3) (Problem 7.3.2). Formally, (7.3.2) and (7.3.4) suggest that

$$\nu(\widehat{P}) = \nu(P) + n^{-\frac{1}{2}} \int \psi(x, P) d\mathcal{E}_n(x) + \frac{n^{-1}}{2} \int \psi(x, y, P) d\mathcal{E}_n(x) d\mathcal{E}_n(y) + \dots \quad (7.3.7)$$

Still formally, if $\psi(\cdot, P) \neq 0$ (7.3.7) suggests that $n^{\frac{1}{2}}(\nu(\widehat{P}) - \nu(P))$ is of order 1 and its behaviour is the same as that of $\int \psi(x, P) d\mathcal{E}_n(x)$. This is just the linear approximation (7.2.2).

Interpretation of von Mises expansion terms. Stochastic integrals

As we have already noted, by the multivariate central limit theorem, for $T = L_2(P)$, $\mathcal{E}_n \xrightarrow{\text{FID}} W_P^0$. Now, $\mathcal{E}_n(f) = \int f(x) d\mathcal{E}_n(x)$ by definition. Can we similarly interpret $W_P^0(f)$ as $\int f(x) dW_P^0(x)$ where $W_P^0(x)$ is the (transformed) Brownian bridge? This cannot be done in the Riemann-Stieltjes or Lebesgue-Stieltjes way since the sample functions

of W^0 are not of bounded variation (Breiman (1968), pp. 261–263). However, if f is piecewise constant,

$$f(x) = \sum_{j=1}^J a_j 1(c_j < x \leq c_{j+1}),$$

then, if we naturally define $\int f(x)dW_P^0(x) \equiv \sum_{j=1}^J a_j (W_P^0(c_{j+1}) - W_P^0(c_j))$, it is easy to see that, indeed, $\int f(x)dW_P^0(x)$ has a $\mathcal{N}(0, \text{Var}_P f(X))$ distribution and can be identified with $W_P^0(f)$.

More generally, if f_1, \dots, f_k are k such functions the joint distribution of

$$\int f_j(x)dW_P^0(x), \quad 1 \leq j \leq k$$

coincides with that of $\{W_P^0(f_j) : 1 \leq j \leq k\}$. It may be shown, since any $f \in L_2(P)$ can be approximated in the $L_2(P)$ sense by such functions, we can define $\int f dW_P^0$ for all $f \in L_2(P)$ with the correct FIDI's for any finite set of such variables. It turns out that, for $k \geq 2$, if $\int b^2(x)dP(x) < \infty$, it is, unfortunately, possible to define

$$\int \dots \int b(x_1, \dots, x_k) dW_P^0(x_1) \dots dW_P^0(x_k)$$

in two distinct ways as the *Ito* or *Stratonovich* stochastic integral. However, if $b(\mathbf{x}) = 0$ whenever $\mathbf{x} = (x_1, \dots, x_k)$ has at least two coordinates equal, then the two integrals coincide and our heuristics apply.

Our heuristics suggest that if $\psi(\cdot, P) = 0$, then the expansion (7.3.7) is of order 2 and

$$2n(\nu(\widehat{P}) - \nu(P)) \implies \mathcal{L} \left(\int \int \psi(x_1, x_2, P) dW_P^0(x_1) dW_P^0(x_2) \right). \quad (7.3.8)$$

If $P[X_1 = X_2] = 0$, the right hand side is uniquely defined. What makes (7.3.8) valuable, if it holds, is that the distribution of $\int \int \psi(x_1, x_2, P) dW_P^0(x_1) dW_P^0(x_2)$ is always of the form $\sum_{k=1}^{\infty} \lambda_k Z_k^2$, where the λ_k are the eigenvalues of the integral operator which sends $h \in L_2(P)$ into $\int \psi(x_1, x_2, P) h(x_1) dP(x_1)$ and the Z_j are i.i.d. $\mathcal{N}(0, 1)$. Here is an example.

Example 7.3.1. (Example 7.2.4. continued). *Null distribution of the Cramér-von Mises statistic.* Consider $\nu(F)$ given in that example. By (7.2.17), $\psi(x, P_0) = 0$ if $F_0 \leftrightarrow P_0$. However, we can compute, if F_0 is continuous and G_j is the df of the probability Q_j ,

$$\frac{\partial \nu}{\partial \varepsilon_1 \partial \varepsilon_2} (F_0 + \varepsilon_1(G_1 - F_0) + \varepsilon_2(G_2 - F_0))|_{\varepsilon_1=\varepsilon_2=0} = 2 \int (G_1 - F_0)(G_2 - F_0)(z) dF_0(z).$$

Thus,

$$\begin{aligned} \psi(x, y, P_0) &= 2 \int (1(z \geq x) - F_0(z))(1(z \geq y) - F_0(z)) dF_0(z) \\ &= 2F_0(x \wedge y) - (F_0^2(y) + F_0^2(x)) + \frac{2}{3}. \end{aligned}$$

Without loss of generality (Problem 4.1.11), we can take $F_0(t) = t$, $0 \leq t \leq 1$, and obtain the required eigenvalues as $\lambda_j = 1/j^2 n^2$. It may be shown (Shorack and Wellner (1986), Ch. 5) that, indeed, $\int_0^1 [W^0]^2(u) du = \int_0^1 \int_0^1 (2(x \wedge y) - (x^2 + y^2) + \frac{2}{3}) dW^0(x) dW^0(y)$. \square

Unfortunately, higher order terms in the von Mises expansion do not have closed form limiting distributions. However, the limit theorems we can prove do suggest how to properly estimate the limit distribution of $n^{\frac{m}{2}}(\nu(\hat{P}) - \nu(P))$ if $\psi(x_1, \dots, x_j, P) = 0$ for all $j < m$ via the m out of n bootstrap discussed in Chapter 10.

7.3.2 The Hoeffding and Analysis of Variance Expansions

Here is another expansion of a symmetric statistic. Its terms are harder to compute but have important orthogonality properties which ease asymptotic analysis. Let \hat{P} be the empirical df of the sample X_1, \dots, X_n and think of $\nu(\hat{P})$ as a function $q(X_1, \dots, X_n)$ where q is symmetric, and represent q as

$$q(x_1, \dots, x_n) = \int q(\mathbf{y}) dH_{x_1}(y_1) \dots dH_{x_n}(y_n)$$

where $H_x(y) = 1(y \geq x)$ the distribution function of δ_x , point mass at x . On the other hand, consider

$$E_P q(X_1, \dots, X_n) = \int q(\mathbf{y}) dP(y_1) \dots dP(y_n).$$

Note the expansion to n terms for the product $\prod_{i=1}^n y_i$ around $\mathbf{y} = \mathbf{x}$

$$\prod_{i=1}^n y_i = \prod_{i=1}^n x_i + \sum_{m=1}^n \sum_{J_m} \{ \prod_{i \in J_m} x_i : i \in \bar{J}_m \} \{ \prod_{i \in J_m} (y_i - x_i) : i \in J_m \}$$

where J_m ranges over all distinct subsets of m integers out of $\{1, \dots, n\}$ and \bar{J}_m is the complement of J_m . We get, after some algebra,

$$q(X_1, \dots, X_n) = E_P q(X_1, \dots, X_n) + \quad (7.3.9)$$

$$\sum_{m=1}^n \sum_{J_m} \left\{ \int \dots \int q(y_1, \dots, y_n) \right\} \prod_{i \in \bar{J}_m} dP(y_i) \prod_{i \in J_m} d(H_{X_i} - P)(y_i).$$

Call the sums in the series $M_q^{(1)}, \dots, M_q^{(n)}$ and let $M_q^{(0)} = E_P q(X_1, \dots, X_n)$. Write

$$M_q^{(m)} = \sum_{J_m} M_q(J_m),$$

where $M_q(J_m)$ is the integral in (7.3.9) for a fixed J_m . Note that $M_q(J_m)$ is a function of $\{X_i : i \in J_m\}$ only. Thus,

$$M_q(\{i\}) = E_P \{ q(X_1, \dots, X_n) | X_i \} - M_q^{(0)}. \quad (7.3.10)$$

So we may write, abusing notation, $M_q(\{i\})$ as $M_q(X_i)$, and similarly, $M_q(\{i, j\})$ as $M_q(X_i, X_j)$ given by,

$$M_q(\{i, j\}) = E_P\{q(X_1, \dots, X_n) | X_i, X_j\} - M_q(\{i\}) - M_q(\{j\}) + M_q^{(0)}. \quad (7.3.11)$$

If S is a subset of $\{1, \dots, n\}$ with $|S| = m + 1$ where $|S|$ denotes cardinality, then, more generally, expanding in (7.3.9) we get the recursion

$$\begin{aligned} M_q(S) &= E_P\{q(X_1, \dots, X_n) | \{X_i : i \in S\}\} \\ &\quad - \sum\{M_q(T) : T \subset S, |T| = m\} \\ &\quad + \sum\{M_q(T) : T \subset S, |T| = m - 1\} \\ &\quad - \sum\{M_q(T) : T \subset S, |T| = m - 2\} \\ &\quad + \cdots (-1)^{m+1} M_q^{(0)}. \end{aligned} \quad (7.3.12)$$

This is the Moebius expansion of $M_q(S)$ — see Stanley (1986), p.116.

Suppose we drop the assumption of symmetry on q and simply take X_i independent with $X_i \sim P_i$, $i = 1, \dots, n$. Then, our development remains valid once we replace P by P_i appropriately in (7.3.9) and, in particular,

$$q(X_1, \dots, X_n) = \sum_{j=0}^n M_q^{(j)} = \sum_{j=0}^n \sum\{M_q(S) : |S| = j\} \quad (7.3.13)$$

with $M_q(S)$ defined by $M_q^{(0)} = E q(X_1, \dots, X_n)$, and the recursion (7.3.12) (provided that $E|q(X_1, \dots, X_n)| < \infty$).

The expression simplifies considerably if q is symmetric and X_1, \dots, X_n are i.i.d. In that case, $M_q(T) = u_q(X_i : i \in T)$ with $|T| = m$, and the integral in (7.3.9) is

$$\begin{aligned} w_q(x_1, \dots, x_m) &\equiv \int \dots \int E q(y_1, \dots, y_m, X_{m+1}, \dots, X_n) \pi_{i=1}^m d(H_{x_i} - P)(y_i) \\ &= \sum_{r=0}^m (-1)^{n-r} \sum\{v_q(x_{i_1}, \dots, x_{i_r}) : \{i_1, \dots, i_r\} \subset \{1, \dots, m\}\} \end{aligned} \quad (7.3.14)$$

where

$$v_q(x_1, \dots, x_r) = E q(x_1, \dots, x_r, X_{r+1}, \dots, X_n).$$

The importance of the expansion (7.3.9) comes from the following interpretation (Theorem 7.3.1) of the $M_q^{(j)}$ which does not make use of either the identical distribution of the X_i or the symmetry of q . Suppose X_1, \dots, X_n are independent. Let Λ_0 be the space of all constants, let Λ_1 be the linear subspace of $L_2(P)$ of all random variables of the form $\sum_{i=1}^n u_i(X_i)$, Λ_2 that of all random variables of the form, $\sum_{1 \leq i < j \leq n} u_{ij}(X_i, X_j)$, and so on. Let $V_j = \Lambda_j \cap \Lambda_{j-1}^\perp$, the linear subspace of all members of Λ_j orthogonal to all members of Λ_{j-1} , and let $\pi(q|V_j)$ denote the $L_2(P)$ projection of q on V_j in $L_2(P)$. We write $U \perp V$ iff $E(UV) = 0$. See Section B.10.3.2. We state our interpretation as

Theorem 7.3.1. *With the definitions and assumptions of the previous paragraph, if $Eq^2(X_1, \dots, X_n) < \infty$, then*

$$M_q^{(j)} = \pi(q(X_1, \dots, X_n) | V_j) \quad (7.3.15)$$

and the representation,

$$q(X_1, \dots, X_n) = \sum_{j=0}^n M_q^{(j)}, \quad (7.3.16)$$

is a decomposition of $q(X_1, \dots, X_n)$ into mutually orthogonal functions of \mathbf{X} . In fact,

$$M_q(S) \perp M_q(T) \quad (7.3.17)$$

unless $S = T$ and so each of the $M_q^{(j)}$ have an orthogonal representation as

$$M_q^{(j)} = \sum \{M_q(S) : |S| = j\} \quad (7.3.18)$$

Proof. The key observation is that, if $Eh^2(X_i : i \in S) < \infty$, and $S \cap T \neq S$, then

$$E \left\{ \int h(y_i : i \in S) \Pi \{d(H_{X_i} - P_i)(y_i) : i \in S\} | X_k : k \in T \right\} = 0. \quad (7.3.19)$$

The reason is that the conditional expectation form passes through the integral and product, since the variables are independent, and

$$\begin{aligned} & E[\{\Pi \{d(H_{X_i} - P_i)(y_i) : i \in S\} | X_k : k \in T\}] \\ &= \Pi \{E[d(H_{X_i} - P_i)(y_i)] | X_k : k \in T\} : i \in S \} = 0 \end{aligned} \quad (7.3.20)$$

since one of the terms in the product must vanish given $S \cap T \neq S$. Statements like (7.3.20) are evidently rigorous if the P_i are discrete, so that

$$d(H_{X_i} - P_i)(y_i) = 1(X_i = y_i) - P[X_i = y_i].$$

The general case can be obtained by induction on the cardinality of S (Problem 7.3.7). Claim (7.3.17) follows since $S \neq T$ implies that either $S \cap T \neq S$ or $S \cap T \neq T$ or both and

$$\begin{aligned} EM_q(S)M_q(T) &= EM_q(S)E\{M_q(T) | X_i : i \in S\} \\ &= EM_q(T)E\{M_q(S) | X_i : i \in T\}. \end{aligned}$$

Then, (7.3.19) implies that one or the other conditional expectation is 0. Finally, we claim that

$$M_q^{(j)} \in V_j, \quad 0 \leq j \leq m. \quad (7.3.21)$$

It is enough to argue that, if $|S| = j$, $M_q(S) \in \Lambda_j$, and that

$$M_q(S) \perp \sum \{h_T\{X_i : i \in T\} : |T| = j - 1\}.$$

That $M_q(S) \in \Lambda_j$ follows by definition. On the other hand,

$$\begin{aligned} & EM_q(S)h_T\{X_i : i \in T\} \\ &= E[E\{M_q(S)|X_i : i \in T\}]h_T(X_i : i \in T) = 0 \end{aligned}$$

since $|S| = j$, $|T| = j - 1$ implies that $S \neq T$. Claim (7.3.21) follows. Write

$$\begin{aligned} \pi(q|V_j) &= M_q^{(j)} + \pi(\sum_{i \neq j} M_q^{(i)}|V_j) \\ &= M_q^{(j)} + \sum_{i \neq j} \pi(M_q^{(i)}|V_j). \end{aligned}$$

The second term is 0 since by construction $M_q^{(i)} \perp V_j$ for $i < j$ since $V_j \subset \Lambda_{j-1}^\perp$ and $M_q^{(i)} \in \Lambda_{j-1}$ while, if $i > j$, $M_q^{(i)} \perp V_j$ since $V_j \subset \Lambda_j \subset \Lambda_{i-1}$. Claim (7.3.15) and the theorem follows. \square

Theorem 7.3.1 has a number of interesting consequences: Let

$$\tau^2(S) = \text{Var } M_q(S).$$

Then,

$$\text{Var } q(X_1, \dots, X_n) = \sum_{m=1}^n \sum \{\tau^2(S) : |S| = m\} = \sum_{m=1}^n \text{Var } M_q^{(m)}. \quad (7.3.22)$$

If we specialize to q symmetric, X_1, \dots, X_n i.i.d., (7.3.22) becomes

$$\text{Var } q(X_1, \dots, X_n) = \sum_{r=1}^n \binom{n}{r} \text{Var } v_q(X_1, \dots, X_r). \quad (7.3.23)$$

A number of important inequalities and other results follow from (7.3.22) and (7.3.23); see Serfling (1980) and van Zwet (1984) as well as the classic paper of Hoeffding (1948).

The first term $M_q^{(1)}$ of the Hoeffding expansion of a statistic of the form $\nu(\widehat{P})$ is, by (7.3.10), a sum of i.i.d. variables, and is not surprisingly closely related to the first term of the von Mises expansion — the influence function approximation to ν . Unfortunately, $E\{\nu(\widehat{P})|X_1 = x\}$ is not, in general, easy to compute and unlike $\psi(x, P)$ depends on n . However, showing that it is a good approximation can be easier. Here is a simple result.

Theorem 7.3.2. Suppose X_1, \dots, X_n are i.i.d., $q_n(X_1, \dots, X_n)$ is a sequence of symmetric statistics with $0 < \text{Var}(q_n(\mathbf{X})) < \infty$, and

$$u_n(x) \equiv E(q_n(\mathbf{X})|X_1 = x).$$

Write q_n for $q_n(\mathbf{X})$. Suppose

$$\text{Var}(q_n) = n \text{Var } u_n(X_1) + o(1).$$

Then,

$$q_n = Eq_n + \sum_{i=1}^n u_n(X_i) + o_P(1). \quad (7.3.24)$$

If further $\mathcal{L}\left\{n^{-\frac{1}{2}}\left[\sum_{i=1}^n u_n(X_i) - E(u_n(X_i))\right]\right\} \rightarrow \mathcal{N}(0, \sigma^2)$, then

$$\mathcal{L}\left\{n^{-\frac{1}{2}}[q_n - E(q_n)]\right\} \rightarrow \mathcal{N}(0, \sigma^2). \quad (7.3.25)$$

Proof. Define $M_q^{(0)} = Eq_n(\mathbf{X})$, and $M_q^{(1)}$ via

$$M_q^{(1)} = \sum_{i=1}^n (u_n(X_i) - Eu_n(X_i)).$$

The result follows because

$$E[(q_n - E(q_n)) - \sum_{i=1}^n u_n(X_i)]^2 = \text{Var}(q_n) - \text{Var}(\sum_{i=1}^n u_n(X_i))$$

since

$$(q_n - E(q_n)) - \sum_{i=1}^n u_n(X_i) \perp \sum_{i=1}^n u_n(X_i)$$

by the definition of u_n and Theorem 7.3.1. So (7.3.24) follows and then so does (7.3.25) by Slutsky's theorem. \square

We illustrate with,

Example 7.3.2. U statistics. Our model is X_1, \dots, X_n i.i.d. as $X \sim P \in \mathcal{P}$. Given a function $\phi : \mathcal{X} \rightarrow R$ which is symmetric, a U statistic of order m is given by

$$U_n \equiv \frac{1}{\binom{n}{m}} \sum \{\phi(X_{i_1}, \dots, X_{i_m}) : 1 \leq i_1 < \dots < i_m \leq n\}. \quad (7.3.26)$$

If $\mathcal{P} = \{P : E_P U_n < \infty\}$, then it may be shown that U_n is the UMVU estimate of $\theta_n(P) \equiv E_P \phi(X_1, \dots, X_m)$ (Problem 7.3.7). A U statistic of order 1 is just a mean of i.i.d. variables such as the sample mean.

An example of a U statistic of order 2, for $\mathcal{X} = R$, is the one-sample Wilcoxon rank statistic for testing symmetry of P ,

$$W = \frac{1}{\binom{n}{2}} \sum_{i < j} 1(X_i + X_j > 0).$$

To apply Theorem 7.3.2, set $q_n = \sum_{i < j} 1(x_i + x_j > 0)$, and note

$$E(1(X_1 + X_2 > 0 | X_1 = x)) = P(x + X_2 > 0) = 1 - F(-x).$$

It follows that the projection of q_n is

$$\hat{q}_n \equiv \sum_{i=1}^n u_n(X_i) = \sum_{i=1}^n [1 - F(-X_i)].$$

Consider the hypothesis $H : F$ is symmetric about zero. Then

$$\text{Var}_H(q_n) = 4(n-2)\text{Var}[1 - F(-X_1)]$$

and the assumptions of Theorem 7.3.2 are satisfied. Because $1 - F(-X_1) = F(X_1) \sim \mathcal{U}(0, 1)$ under H , $\text{Var}(q_n) = (n-2)/3$, and $n^{-\frac{1}{2}}(q_n - \frac{1}{2}) \rightarrow \mathcal{N}(0, \sqrt{1/3})$.

For $\mathcal{X} = R^2$, $X = (Z, V)$, Kendall's τ statistic for testing independence of Z and V ,

$$\tau = \frac{1}{\binom{n}{2}} \sum_{i < j} 1((Z_i - Z_j)(V_i - V_j) > 0),$$

is also a U statistic of order 2. We will study these two statistics further in Problems 7.3.11 and 7.3.16.

Next we analyze U statistics generally using the Hoeffding and von Mises approaches. Note first that for a U statistic of order m with $E\phi^2(X_1, \dots, X_m) < \infty$, the Hoeffding expansion is truncated, since $U_n \in \Lambda_m$,

$$U_n = \sum_{j=0}^m M_{T_n}^{(j)}.$$

Therefore, from (7.3.22) and (7.3.23), for

$$v_q(x_1, \dots, x_j) \equiv E(U_n(x_1, \dots, x_j, X_{j+1}, \dots, X_n)),$$

$$\text{Var}(U_n) = \sum_{j=1}^m \binom{n}{j} \text{Var} v_q(X_1, \dots, X_j).$$

By symmetry,

$$\begin{aligned} v_q(x) &= E\left[\frac{1}{\binom{n}{m}} \sum \{\phi(X_{i_1}, \dots, X_{i_m}) : 1 \leq i_1 \leq \dots \leq i_m | X_1 = x\}\right] \\ &= \frac{\binom{m-1}{m-1}}{\binom{n}{m}} \phi(x, X_2, \dots, X_m) = \frac{m}{n} E\phi(x, X_2, \dots, X_m). \end{aligned} \tag{7.3.27}$$

More generally, for $1 \leq r \leq m$,

$$v_q(x_1, \dots, x_r) = \frac{\binom{n-r}{m-r}}{\binom{n}{m}} E\phi(x_1, \dots, x_r, X_{r+1}, \dots, X_m)$$

so that

$$\text{Var}(U_n) = \sum_{r=1}^m \frac{\binom{m}{r} \binom{n-r}{m-r}}{\binom{n}{m}} \xi_r, \quad (7.3.28)$$

where

$$\xi_r \equiv \text{Var}E[\phi(X_1, \dots, X_m) | X_1, \dots, X_r]. \quad (7.3.29)$$

For the one-sample Wilcoxon statistic we have $E(\phi(X_1, X_2) | X_1 = x) = 1 - F(-x)$ and $E(\phi(X_1, X_2) | X_1, X_2) = 1(X_1 + X_2 > 0)$. It follows that $\xi_1 = \text{Var}(1 - F(-X_1))$, $\xi_2 = \text{Var}(1(X_1 + X_2 > 0))$, and

$$\text{Var}(W) = 2[2(n-2)\xi_1 + \xi_2(1 - \xi_2)]/n(n-1).$$

□

Note that closely related to U_n is the von Mises statistic,

$$\begin{aligned} V_n &\equiv \int \dots \int \phi(x_1, \dots, x_m) d\hat{P}(x_1) \dots d\hat{P}(x_m) \\ &= \frac{1}{n^m} \sum \{\phi(X_{i_1}, \dots, X_{i_m}) : 1 \leq i_1, \dots, i_m \leq n\}. \end{aligned}$$

So,

$$V_n = \frac{\binom{n}{m}}{n^m} U_n + \int_S \dots \int \phi(x_1, \dots, x_m) d\hat{P}(x_1) \dots d\hat{P}(x_m) \quad (7.3.30)$$

where $S = \{(x_1, \dots, x_m) : x_i = x_j \text{ for some } i \neq j\}$. The influence function calculation readily yields for V_n ,

$$\psi(x, P) = m \left(\int \dots \int \phi(x, x_2, \dots, x_m) dP(x_2) \dots dP(x_m) - E_P \phi(X_1, \dots, X_m) \right).$$

However, application of Theorem 7.3.2 for V_n is difficult since F appears as dF . On the other hand, consider U_n .

Proposition 7.3.1. *For U_n as (7.3.26), if $0 < \text{Var}(v_q(X_1)) < \infty$, then*

$$n^{\frac{1}{2}} (U_n - E_P \phi(X_1, \dots, X_m)) \implies \mathcal{N}(0, E_P \psi^2(X_1, P))$$

where

$$\psi(X, P) = m E_P [\phi(X_1, \dots, X_m) | X_1].$$

Proof. By (7.3.27) and (7.3.23), if $q_n = n^{\frac{1}{2}}U_n$,

$$v_q(X_1) = n^{-\frac{1}{2}}\psi(X_1, P)$$

and from (7.3.28),

$$\text{Var}(n^{\frac{1}{2}}U_n) = m\xi_1 + O_P(n^{-1}), \quad (7.3.31)$$

since $\binom{n}{r}^{-1} = O(n^{-r})$. Now apply Theorem 7.3.2 to q_n above. Then,

$$\sum_{i=1}^n u_n(X_i) = n^{-\frac{1}{2}} \sum_{i=1}^n \psi(X_i, P)$$

so that, since $\xi_1 = \text{Var } \psi(X_1, P)$, the conditions of the theorem are satisfied and Proposition 7.3.1 follows. \square

Remark 7.3.1. If $\psi(\cdot, P) \equiv 0$ the limit of $2n[U_n - E(U_n)]$ is of the form

$$\int \int \psi(x, y, P) dW^0(F(x)) dW^0(F(y)).$$

See (7.3.8). This is also true for V_n . That is, the second term in (7.3.30) is negligible if

$$E\phi^2(\mathbf{X}^{(i_1)}, \mathbf{X}^{(i_2)}, \dots, \mathbf{X}^{(i_r)}) < \infty$$

for all $i_1 + \dots + i_r = m$ and $\mathbf{X}^{(i_j)} = (X_j, X_j, \dots, X_j)$ with the dimension of $\mathbf{X}^{(i_j)}$ being i_j . We leave this to Problem 7.3.9.

The von Mises and Hoeffding expansions can be compared quite precisely for U statistics if we assume that $\phi(x_1, \dots, x_m) = 0$ whenever $x_i = x_j$ for some $i \neq j$. In that case,

$$V_n = n^{-m} \binom{n}{m} U_n$$

and further, for $\psi(x_1, \dots, x_j, P)$ and w_q as defined in (7.3.14) and (7.3.6),

$$w_0(x_1, \dots, x_j) = \psi(x_1, \dots, x_j, P) n^{-j} \quad (7.3.32)$$

for $j = 1, \dots, m$. We leave this result to Problem 7.3.9. We again refer to Serfling (1980) for an extensive treatment of U statistics. \square

The analysis of variance expansion

Suppose Z_1, \dots, Z_k are independent predictor variables, Z_j taking on values in $\mathcal{Z}_j \equiv \{z_{j_1}, \dots, z_{j_{n_j}}\}$ with equal probabilities n_j^{-1} , for $1 \leq j \leq k$. Then if Y is a “response” variable we are interested in the regression surface,

$$\mu(z_1, \dots, z_k) \equiv E(Y|Z_1 = z_1, \dots, Z_k = z_k).$$

Here $\mu(Z_1, \dots, Z_k)$ is a general function of (Z_1, \dots, Z_k) and thus can be expanded in the Hoeffding sense as

$$\begin{aligned}\mu(Z_1, \dots, Z_k) &= E\mu(Z_1, \dots, Z_k) + \sum_{j=1}^k \{E(\mu(Z_1, \dots, Z_k)|Z_j) - E\mu(Z_1, \dots, Z_k)\} \\ &\quad + \dots + U_\mu^{(k)}.\end{aligned}$$

Here the first term is the mean response which is

$$E\mu(Z_1, \dots, Z_k) \equiv \mu(\cdot, \dots, \cdot) = \frac{1}{n_1 \dots n_k} \sum \{\mu(z_1, \dots, z_k) : z_j \in \mathcal{Z}_j, 1 \leq j \leq k\}.$$

The next term is

$$E(\mu(Z_1, \dots, Z_k)|Z_1 = z) - E\mu(Z_1, \dots, Z_k) = [\mu(z, \cdot, \dots, \cdot) - \mu(\cdot, \dots, \cdot)]$$

where we use dots in the usual ANOVA sense for averaging. But in the case where the Z 's are treatments, this is just the effect of treatment 1 at level z . Continuing, $U^{(2)}(\{1, 2\})(z_1, z_2)$ is just the first order interaction between Z_1 and Z_2 at levels z_1 and z_2 . So in this context, the Hoeffding expansion is the classical ANOVA expansion mentioned in terms of the corresponding variance decomposition in Example 6.1.3 and discussed extensively in books such as Box and Hunter (1978).

Summary. In Section 7.3.1 we consider a generalization of Taylor expansion due to von Mises. The limiting behaviour of statistics whose first order Gâteaux derivative vanishes is related to the Brownian bridge. Finally in Section 7.3.2 we introduce another type of expansion due to Hoeffding and closely related to the analysis of variance. This is an expansion into orthogonal terms and as such enables us to develop powerful bounds used in establishing asymptotic approximations. The theory is applied to U statistics.

7.4 PROBLEMS AND COMPLEMENTS

Problems for Section 7.1

1. Establish (7.1.4) using Problem 5.4.5.
2. Establish (7.1.6) using the multivariate Taylor expansion and the multivariate CLT.
3. Prove Corollary 7.1.1 from Theorem 7.1.1.

Hint. Let A_{mj} be the event where

$$\sup \{|Z_n(s) - Z_n(t)| : s, t \in T_{mj}\} > \varepsilon.$$

Then $P(U_j A_{mj}) \leq \Sigma_j P(A_{mj})$.

4. Suppose T is a metric space and $Z_n \Rightarrow Z$ on $l_\infty(T)$ and $P[Z \text{ is continuous}] = 1$. Let $t_n \xrightarrow{P} t$ on T , t_n possibly random but t a constant. Show that then

$$Z_n(t_n) \Rightarrow Z(t).$$

Hint. The map $M : l_\infty(T) \times T \rightarrow R$ given by $M(ft) = f(t)$ is continuous at all (f_0, t_0) such that f_0 is continuous at t_0 . Apply Corollary 7.1.2.

5. Assuming the existence of the Wiener process $W(\cdot)$ on $[0, 1]$, show that if $U(t) \equiv W(t) - tW(1)$, then $U(\cdot) \sim W^0(\cdot)$ where $W^0(\cdot)$ is the Brownian bridge.

6. Establish FIDI convergence and tightness in Proposition 7.1.1.

Hint. $[nt]/n - [ns]/n = (t - s) + O(1/n)$.

7. Show that, if $\mathcal{X} = R$, $T = \{\text{All } f \text{ with } |f|_\infty \leq 1\}$ and P is continuous, then

$$\sup\{|\mathcal{E}_n(f)| : f \in T\} = n^{\frac{1}{2}}.$$

8. Show that in Definition 7.1.3, (ii) implies (i).

9. Prove Theorem 7.1.5.

10. Show that the convergence in (7.1.27) is almost sure.

Hint. Use the Borel-Cantelli Lemma; see Appendix D.1.

11. Let $W^0(u)$, $0 \leq u \leq 1$, be the Brownian bridge.

(a) Show that $\int_0^1 W^0(u)a(u)du$ has a $\mathcal{N}(0, \sigma^2(a))$ distribution where

$$\sigma^2(a) = 2 \int_0^1 \int_0^1 \int_{s \leq t} s(1-t)a(s)a(t)dsdt.$$

(b) If $A(v) \equiv \int_0^v a(s)ds$, $\int_0^1 a(s)ds = 0$, show that

$$\sigma^2(a) = \int_0^1 A^2(s)ds.$$

Hint. From Problem D.2.3, W^0 is continuous.

12. For the Section I.2 example show that the process $t \rightarrow \sqrt{n}(\widehat{F}(t) - G_{\widehat{\mu}, \widehat{\sigma}^2}(t))$ converges weakly to a Gaussian process $Z(\cdot; \mu, \sigma^2)$ if $\widehat{\mu} \xrightarrow{P} \mu$, $\widehat{\sigma}^2 \xrightarrow{P} \sigma^2$, whenever the data are generated from the conditional distribution of Z_n given $Z_n > 0$ where $Z_n \sim \mathcal{N}(\mu_n, \sigma_n^2)$ and $\mu_n \rightarrow \mu$, $\sigma_n^2 \rightarrow \sigma^2$.

13. Suppose U, V, W_1, W_2, \dots are i.i.d. uniform $[0, 1]$ and set

$$Z(t) = U + Vt, \quad Z_n(t) = Z(t) + n^{-1} \sum_{i=1}^{r_n} 1[W_i \leq t], \quad t \in [0, 1]$$

where the integer r_n satisfies $(r_n/n) \rightarrow 0$ as $n \rightarrow \infty$. Show that

(a) $Z_n \xrightarrow{FIDI} Z$.

Hint. Compute $|Z_n - Z|$ and show that $Z_n - Z \xrightarrow{FIDI} 0$. Apply Slutsky's theorem.

(b) $Z_n \Rightarrow Z$.

Hint.

$$Z_n(s) - Z_n(t) = (s-t)V + n^{-1} \sum_{i=1}^{r_n} \{1[W_i \leq s] - 1[W_i \leq t]\}.$$

Introduce

$$T_{mj} = \left[\frac{j-1}{k_m}, \frac{j}{k_m} \right], \quad j = 1, \dots, k_m$$

and show that $|Z_n(s) - Z_n(t)|$ is bounded on T_{mj} by $k_m^{-1} + r_n/n$.

14. Let U, V, W_1, W_2, \dots be i.i.d. uniform $[0, 1]$, let Φ be the $\mathcal{N}(0, 1)df$, and set

$$Z(t) = \Phi\left(\frac{t-U}{V+1}\right), \quad Z_n(t) = Z(t) + n^{-1} \sum_{i=1}^{r_n} \Phi\left(\frac{t-W_i}{V+1}\right).$$

Show that if $r_n/n \rightarrow 0$ as $n \rightarrow \infty$, then

(a) $Z_n(\cdot) \xrightarrow{FIDI} Z(\cdot)$, **(b)** $Z_n(\cdot) \Rightarrow Z(\cdot)$.

15. If X_1, \dots, X_n are i.i.d. with continuous df F and empirical df \hat{F} , then, in $l_\infty(R)$,

$$\sqrt{n}(\hat{F}(\cdot) - F(\cdot)) \Longrightarrow W^0 F(\cdot).$$

Hint. Show that for any df $F(\cdot)$ if U_1, \dots, U_n are i.i.d. $\mathcal{U}(0, 1)$ then $F^{-1}(U_1), \dots, F^{-1}(U_n)$ are i.i.d. F , where $F^{-1}(t) = \inf\{x : F(x) \geq t\}$.

16. Establish (7.1.15).

17. Establish Theorem 7.1.5 by extending the proof of Theorem 7.1.4.

18. (a) Show there exists a constant $\alpha > 0$ ($\alpha \approx 6.308$), such that for every real random variable X and every $h > 0$,

$$P(|X| \geq h^{-1}) \leq \frac{\alpha}{2h} \int_{-h}^h \{1 - \phi_X(t)\} dt,$$

where $\phi_X(t) = E\{\exp(itX)\}$ denotes the characteristic function for X .

Remark: In this problem, we do not have the assumption $E(|X|) < \infty$.

Hint. You can use (without proving) the inequality, $\frac{\sin(x)}{x} \leq \sin(1)$, which holds for any $|x| \geq 1$.

(b) Let X_1, X_2, \dots be a sequence of real random variables. Assume that ϕ_{X_n} converges pointwise to ϕ , as $n \rightarrow \infty$, where ϕ is continuous (but not necessarily a characteristic function). Show that $X_n = O_P(1)$.

19. Show the following

(a) For $\delta > 0$,

$$\frac{\delta e^{-\delta^2/2}}{1+\delta^2} \leq \int_{\delta}^{\infty} e^{-x^2/2} dx \leq \frac{e^{-\delta^2/2}}{\delta}.$$

Hint. You may find integration by parts helpful for the first inequality.

(b) For X_1, \dots, X_n independent and identically distributed $N(0, 1)$ random variables,

- (i) $P\left\{\max_{1 \leq i \leq n} |X_i| > \sqrt{2 \log(n)}\right\} \rightarrow 0$, as $n \rightarrow \infty$,
- (ii) $\sqrt{\log(n)} P\left\{\max_{1 \leq i \leq n} |X_i| > \sqrt{2 \log(n)}\right\} \rightarrow \frac{1}{\sqrt{\pi}}$, as $n \rightarrow \infty$,
- (iii) $n^{c/2-1} \sqrt{\log(n)} P\left\{\max_{1 \leq i \leq n} |X_i| > \sqrt{c \log(n)}\right\} \rightarrow \frac{\sqrt{2}}{\sqrt{c\pi}}$, as $n \rightarrow \infty$.

20. Let $\{Z_n\}$ be a sequence of independent and identically distributed $N(0, 1)$ random variables. Prove that there is a random variable X such that $P(X < \infty) = 1$ and

$$|Z_n| \leq X \sqrt{\log(n)}, \text{ for all } n \geq 2.$$

21. Let $\varepsilon_1, \dots, \varepsilon_n$ be independent and identically distributed random variables with $P(\varepsilon_1 = 1) = P(\varepsilon_1 = -1) = 1/2$.

(a) Let a_1, \dots, a_n be real numbers and let $\|a\|^2 = \sum_{i=1}^n a_i^2$. Prove that for any $x > 0$,

$$P\left(\left|\sum_{i=1}^n \varepsilon_i a_i\right| > x\right) \leq 2 \exp\left(-\frac{x^2}{2\|a\|^2}\right).$$

Hint: Show that for any real λ , $E \exp(\lambda \varepsilon_1) \leq \exp(\lambda^2/2)$.

(b) Let Z_1, \dots, Z_n be independent and identically distributed random variables, independent of $\varepsilon_1, \dots, \varepsilon_n$ such that $0 < EZ_1^2 = \tau < \infty$. Prove that for any $x > 0$,

$$\limsup_{n \rightarrow \infty} P\left(\left|\sum_{i=1}^n \varepsilon_i Z_i\right| > \sqrt{n}x \text{ given } Z_1, \dots, Z_n\right) \leq 2 \exp\left(-\frac{x^2}{2\tau}\right)$$

with probability one.

Problems for Section 7.2

1. Show that (7.2.5) holds if the map, $Q \rightarrow \psi(Q, P)$, is continuous for weak convergence, i.e. $Q_m \rightharpoonup Q$ implies that $\psi(Q_m, P) \rightarrow \psi(Q, P)$.

Hint. Use the result of Section 5.4.1.

2. The sensitivity curve $\nu(\hat{P})$ is defined by $SC_n(x) = n(\nu(\hat{P}_{n-1,x}) - \nu(\hat{P}_{n-1}))$ where \hat{P}_{n-1} is the empirical distribution of X_1, \dots, X_{n-1} , and

$$\hat{P}_{n-1,x} = \frac{n-1}{n} \hat{P}_{n-1} + \frac{1}{n} \delta_x.$$

Suppose that the Gâteaux derivative of $\nu(\cdot)$ is well defined in every direction and (7.2.2) holds at all $P \in \mathcal{M}$.

(a) Show that if $\widehat{\nu} = \nu(\widehat{P})$ has influence function ψ , then

$$SC_n(x : \widehat{\nu}) = n \int_0^{\frac{1}{n}} [\psi(x, (1-t)\widehat{P}_{n-1} + t\delta_x) + \psi(\widehat{P}_{n-1}, (1-t)\widehat{P}_{n-1} + t\delta_x)] dt.$$

(b) Deduce that if $P_m \xrightarrow{\mathcal{L}} P$ implies that $\psi(x, P_m) \rightarrow \psi(x, P)$, then, as $n \rightarrow \infty$,

$$SC_n(x, \widehat{\nu}) \xrightarrow{P} \psi(x, P).$$

3. Show that $\psi(x, P)$ as given by (7.2.10) satisfies (7.2.5).

4. Let $(X_1, Y_1), \dots, (X_n, Y_n)$ be i.i.d. as $(X, Y) \sim P$, where $0 < E(X^4), E(Y^4) < \infty$. Let $\nu_1(P) = \text{Cov}(X, Y)$ and let $\nu_2(P) = \nu_1^2(P)/\text{Var}(X)\text{Var}(Y)$ be the squared correlation coefficient. Compute the influence function of (a) $\nu_1(P)$, (b) $\nu_2(P)$.

Hint. Use (7.2.10) and Example 5.3.6.

5. In the Hardy-Weinberg model where three categories have frequencies

$$p_1 = \theta^2, p_2 = 2\theta(1-\theta), p_3 = (1-\theta)^2, 0 < \theta < 1$$

(Examples 2.1.4, 2.2.6) we can write the parameter θ as $\nu_1(P) = \sqrt{p_1}$, $\nu_2(P) = 1 - \sqrt{p_3}$, or $\nu_3(P) = p_1 + \frac{1}{2}p_2$. Use Example 7.2.1 to compute the influence functions ψ_1, ψ_2 , and ψ_3 of $\nu_1(\widehat{P}), \nu_2(\widehat{P})$, and $\nu_3(\widehat{P})$. Which ψ_j has the smaller value of the asymptotic variance $E_P \psi_j^2(X, P)$?

6.(a) Show that if ψ_α is given by (7.2.13), then the definition of $\nu(\widehat{F})$ in Problem 5.4(a) is equivalent to (7.2.14).

Hint. $E\psi_\alpha(X, \theta) = \alpha[1 - \widehat{F}^-(\theta)] - (1 - \alpha)\widehat{F}^-(\theta)$.

(b) Show that for ρ and \widehat{x}_α as defined in Example 7.2.3, $\widehat{x}_\alpha = \arg \min_\theta \int \rho(x, \theta) d\widehat{F}(x)$.

7. Show that $\nu_\alpha(F)$ minimizes $\int \rho(x, \theta) dF(x)$ for ρ and $\nu_\alpha(F)$ as in Example 7.2.3.

8. Establish (7.2.16) under appropriate assumptions.

Hint. Use the assumptions and method of Problem 5.4.1 to show that

$$(\sqrt{n}(\widehat{x}_\alpha - F^{-1}(\alpha)), n^{-\frac{1}{2}} \sum_{i=1}^n \psi_\alpha(X_i, F))$$

have an asymptotically *joint* Gaussian distribution concentrating in the plane on the diagonal $\{(u, v) \in R^2 : u = v\}$, which establishes the claim.

9. (a) Establish (7.2.17).

(b) Give the details needed to establish the asymptotic normality of the Cramér-von Mises statistics if the true $F \neq F_0$ using Theorem 7.2.1.

10. *Consistency for minimum distance estimates.* Let d be a suitable metric and $\{P_\theta : \theta \in \Theta \text{ compact } \subset R^p\}$ be a regular parametric model such that θ is identifiable and $\theta \rightarrow d(P, P_\theta)$ is continuous for fixed P . Let $\widehat{\theta} = \arg \min d(\widehat{P}, P_\theta)$.

(a) Show that $\hat{\theta}$ is consistent.

Hint. If θ_0 is true, $d(\hat{P}, P_{\theta_0}) = O_P(n^{-\frac{1}{2}})$. But

$$d(P_{\theta_0}, P_{\hat{\theta}}) \leq d(P_{\theta_0}, \hat{P}) + d(P_{\hat{\theta}}, \hat{P}) \leq 2d(P_{\theta_0}, \hat{P})$$

and 1-1 continuous maps on compacts have continuous inverses.

(b) Show that if the conditions of Theorem 7.2.1 are satisfied, then $\hat{\theta}$ is \sqrt{n} consistent in the sense that $\sqrt{n}(\hat{\theta} - \theta) = O_P(1)$.

11. In Theorem 7.2.2, show that if $\theta(Q)$ is defined as the unique solution of

$$\int \phi(x, \theta(Q)) dQ(x) = \mathbf{0}$$

for all Q in a convex neighbourhood of P , and A3' holds, then the influence function of $\theta(P)$ is indeed

$$\psi(x, P) = [H^{-1}(P)]\phi(x, \theta(P)).$$

12. (a) In Example 7.2.5 show that if $\int |\mathbf{x}|^r dP(\mathbf{x}) < \infty$, then $\int |\mathbf{x} - \theta|^r dP(\mathbf{x})$ takes its infimum over θ for $r \geq 1$ and the minimizing $\theta(P)$ is unique if $r > 1$.

(b) Show that $\theta(P) = \arg \min \int (|\mathbf{x} - \theta|^r - |\mathbf{x}|^r) dP(\mathbf{x})$ which is defined iff $\int |\mathbf{x} - \theta|^{r-1} dP(\mathbf{x}) < \infty$.

Hint. If $|\theta| \rightarrow \infty$, $\int |\mathbf{x} - \theta|^r dP(\mathbf{x}) \rightarrow \infty$ and $\theta \rightarrow \int |\mathbf{x} - \theta|^r dP(\mathbf{x})$ is strictly convex for $r > 1$.

(c) Show that $\theta(P)$ is uniquely defined even if $r = 1$ if P has a positive density.

Hint. (i) Prove the result for $d = 1$ where $\theta(P)$ is just the median.

(ii) If the minimizer is not unique, then by convexity, there exists $\theta_0, \Delta \neq 0$ such that all $\theta_0 + \varepsilon \Delta$ are minimizers $0 \leq \varepsilon \leq 1$. Rotate and scale so that $\Delta = (1, 0, \dots, 0)$.

13. (a) Show that $\Psi(\cdot, P) \in L_2(P)$ and $M(P)$ given by (7.2.26) is well defined if $r > 1$ and $E_P |\mathbf{X}|^{2r-2} < \infty$ or $r = 1$ and P has a positive density.

Hint. For $r = 1$, use spherical polar coordinates.

(b) Check the conditions of Theorem 7.2.2 for the multivariate medians.

14. Show that A4' in Theorem 7.2.2 is implied by:

A4'': $\sup\{ |\psi(x, \theta) - \psi(x, \theta_0)| : |\theta - \theta_0| \leq \gamma \} \leq V(x, \theta_0) \gamma^\alpha$, $\alpha > 0$
for all $\theta_0, \gamma > 0$ where $E_P V^2(X, \theta) < \infty$ for all $|\theta - \theta(P)| \leq \varepsilon$.

Hint. Consider a box $\{\theta : |\theta - \theta(P)|_\infty \leq \varepsilon\}$ and break it up into a grid of $(\varepsilon/\delta)^p$ boxes of side δ . If $\theta^{(1)}, \dots, \theta^{(K)}$, where $K = 2^p$ are the vertices of such a box B , then for θ in this box, $\psi(x, \theta) \geq \underline{f}$ where \underline{f} is defined as $\min\{\psi_1(x, \theta) - \psi(x, \theta(P)) : \theta \text{ in } B\}$ and similarly for \bar{f} .

15. Apply Problem 7.2.14 to Example 7.2.4 to justify A4 for $\psi(\mathbf{x}, \theta) \equiv (\mathbf{x} - \theta) |\mathbf{x} - \theta|^{r-1}$ for $1 \leq r \leq 2$, $p \geq 2$, by showing that

$$|\psi(\mathbf{x}, \theta) - \psi(\mathbf{x}, \theta_0)| \leq |\theta - \theta_0| |\mathbf{x} - \theta|^{r-2}.$$

16. (a) Establish (7.2.32).

Hint. $F^{-1}(F(t)) = t$, $t \in R$.

(b) Show that (7.2.40) follows from (7.2.39).

17. Derive (7.2.43) from (7.2.41).

18. Let $F_\theta = F_0(\cdot - \theta)$ be a location parameter family. Let $\hat{\theta}$ be the minimum distance estimate

$$\hat{\theta} = \arg \min_{\theta} \int_{-\infty}^{\infty} (\hat{F}(x) - F_\theta(x))^2 dF_0(x).$$

(a) Compute formally the influence function of $\hat{\theta}$.

(b) Show that the expansion (7.2.1) is valid if $\int_{-\infty}^{\infty} |f'(x)| dx < \infty$.

Hint. Without loss of generality, take $\theta_0 = 0$. Note that $\theta(F)$ solves

$$\int (F_0(x - \theta) - F(x)) f_0(x - \theta) dF_0(x) = 0.$$

(c) Set $d(F, F_\theta) = \int_{-\infty}^{\infty} [F(x) - F_0(x - \theta)]^2 dF_0(x)$. Show that $\hat{\theta}$ is consistent using Problem 7.2.10 and

$$\frac{\partial d}{\partial \theta}(F_0, F_0) = 0, \quad \frac{\partial^2 d}{\partial \theta^2}(F_0, F_0) \neq 0.$$

(d) Let

$$\hat{\theta} = \arg \min_{\theta} \sup_x |\hat{F}(x) - F_\theta(x)|.$$

Show that $\sqrt{n}(\hat{\theta} - \theta)$ is not asymptotically normal and hence does not have an influence function.

Hint. Consider θ with $|\theta| = O(n^{-\frac{1}{2}})$ and $Z_n(x; \theta) = \sqrt{n}[\hat{F}(x) - F_\theta(x)] = \sqrt{n}[\hat{F}(x) - F_0(x)] + f_0(x)\sqrt{n}\theta + O_p(1)$. Apply Donsker's theorem.

19 (a) Show that in (7.2.1), $n^{-1} \sum_{i=1}^n \psi(X_i, P) = O_P(n^{-\frac{1}{2}})$.

(b) Show that ψ is unique if the approximation is valid.

That is, show that if

$$\begin{aligned} \nu(\hat{P}) &= \nu(P) + \frac{1}{n} \sum_{i=1}^n \psi_1(X_i, F) + o_p(n^{-\frac{1}{2}}) \\ &= \nu(P) + \frac{1}{n} \sum_{i=1}^n \psi_2(X_i, F) + o_p(n^{-\frac{1}{2}}) \end{aligned}$$

with $E_P \psi_j(X) = 0$, $E_P \psi_j^2(X) < \infty$, $j = 1, 2$, then $P[\psi_1(X) = \psi_2(X)] = 1$.

Hint. Use the CLT and $n^{-\frac{1}{2}} \sum_{i=1}^n \Delta(X_i) = o_P(1)$ for $\Delta \in L_2(P)$ iff $\Delta = 0$.

20. Let $X_{(1)} < \dots < X_{(n)}$ be the order statistics of X_1, \dots, X_n i.i.d. as $X \sim F$ where F is continuous. Show that if $E|X|^\varepsilon < \infty$ for $\varepsilon > 0$, then for all $r > 0$ there exist $C_r > 0$ and $n(r, k)$ such that for $n \geq n(r, k)$

$$E|X_{(k)}|^r \leq C_r \quad \text{if } \alpha \leq \frac{k}{n} \leq 1 - \alpha, \quad 0 < \alpha < \frac{1}{2}.$$

- 21.** Show that if $f = F' > 0$ and continuous, then, if $\frac{k}{n} \rightarrow s$, $\frac{l}{n} \rightarrow t$,

$$n \operatorname{Cov}(X_{(k)}, X_{(l)}) \rightarrow \frac{s(1-t)}{f(F^{-1}(s))f(F^{-1}(t))}; \quad s \leq t$$

uniformly for $\alpha \leq \frac{k}{n}, \frac{l}{n} \leq 1 - \alpha$.

Hint. You may use the fact that if $U_n \implies U$ and $E|U_n|^{r+1} \leq C$, all n , then $EU_n^r \rightarrow EU^r$ uniformly. If $U_n(\cdot) \implies U(\cdot)$, $\sup_{t,n} E|U_n(t)|^{r+1} \leq C$, then $\sup_t |EU_n^r(t) - EU^r(t)| \rightarrow 0$. Use Problem 7.1.1.4 and the quantile process convergence.

- 22.** If (U, V) are such that $E(U^2 + V^2) < \infty$ and $E(V|U = u)$ is nondecreasing in u , then $\operatorname{Cov}(U, V) \geq 0$.

Hint. $\operatorname{Cov}(U, V) = \operatorname{Cov}(U, E(V|U)) = \frac{1}{2}E(U - U')(E(V|U) - E(V'|U'))$ where (U, V) and (U', V') are i.i.d.. (See below A.11.14.)

- 23.** Show that under the conditions of Problem 21, we always have

$$\operatorname{Cov}(X_{(k)}, X_{(l)}) \geq 0.$$

- 24.** Suppose that $\lambda = \Lambda$ is as in Example 7.2.5, save that we no longer require $\lambda(t) = 0$ for $t < \varepsilon$ and $t > 1 - \varepsilon$ but only that $\lambda(t) \leq C < \infty$ and $EX^2 < \infty$. Show that the conclusion (7.2.39) still holds. (Bickel (1967)).

Hint. Argue that

$$(i) \int_{\varepsilon}^{1-\varepsilon} \sqrt{n}(\widehat{F}^{-1}(t) - F^{-1}(t))\lambda(t)dt \implies \mathcal{N}(0, \sigma_{\varepsilon}^2(\Lambda)).$$

$$(ii) \int_{1-\varepsilon}^1 \sqrt{n}(\widehat{F}^{-1}(s) - F^{-1}(s))\lambda(s)ds^2 \rightarrow 0 \text{ as } \varepsilon \rightarrow 0 \text{ if } \lambda \equiv 1$$

and Problem 7.2.20 ensures that we can pass to the general case.

- 25.** Consider the nonparametric regression model $Y = \mu(X) + e$, where $\mu(\cdot)$ is unknown, $E(e) = 0$, $\operatorname{Var}(e) = \sigma^2$, and X and e are independent. The *nonparametric correlation* is defined as $\eta^2 = \operatorname{CORR}^2(\mu(X), Y)$. Assume η^2 exists.

$$(a) \text{ Show that } \eta^2 = \frac{E(\mu^2(X)) - \mu_Y^2}{\operatorname{Var}(Y)}.$$

- (b) Let $\nu(P) = E[\mu^2(X)] = \int \mu^2(x)f(x)dx$ where $X \sim F$, $f = F'$ is assumed to exist. Informally, show that the influence function is $\mu^2(x) + 2\mu(x)e$ by computing the Gâteaux derivative and evaluating it at $Q = \delta_{(x,y)}$.

Note: For the model $(1 - \varepsilon)P + \varepsilon Q$, both $f(x)$ and $\mu(x)$ depend on ε .

- (c) Find the influence function of η^2 .

- (d) Evaluate your answer to (c) when $\mu(X) = \alpha + bX$.

- (e) Compare your answer to (d) with the influence function of the Pearson squared correlation coefficient ρ^2 when $\mu(X) = \alpha + \beta X$.

Hint. See Problem 9.2.4.

Problems for Section 7.3

1. Verify (7.3.2) when \mathcal{X} is finite.
2. Show that (7.3.2) and (7.3.5) imply (7.3.3) and (7.3.6) and that ψ is symmetric in x_1, \dots, x_n .
3. (a) Verify that if h in Theorem 5.4.1 has continuous derivatives order of k , then $\theta(P) \equiv h(\mathbf{p})$ can be represented as in (7.3.1).
(b) Identify the $\psi(\cdot, \dots, P)$ functions in terms of the derivatives of h and expectations of these.
(c) Verify (7.3.2) in this case.
4. A U statistics of order 2 is called *degenerate* iff

$$U = \frac{1}{\binom{n}{2}} \sum_{i < j} u(X_i, X_j)$$

with $E\{u(X_1, X_2)|X_1\} = 0$ and hence $Eu(X_1, X_2) = 0$. Suppose $u(x, x) = 0$, $Eu^2(X_1, X_2) < \infty$. Show that

- (a) $nU = \int u(x, y)d\mathcal{E}_n(x)d\mathcal{E}_n(y)$.
- (b) If $\mathcal{L}(nU) \rightarrow \mathcal{L}_0$ for some probability law \mathcal{L}_0 , what form would you expect \mathcal{L}_0 to have?
- (c) Is the result you conjecture a consequence of weak convergence of $\mathcal{E}_n(\cdot)$?

5. *Establishing weak convergence for degenerate U statistics.* Suppose $\phi(x, y)$ in (7.3.2) is symmetric. Then there exist real eigenvalues λ_j , $j \geq 1$, and eigenfunctions $\psi_j(x)$ of the kernel $K(x, y) = u(x, y)$, such that

$$\int \psi_j(y)K(x, y)dF(y) = \lambda_j \psi_j(x), K(x, y) = \sum_{j=1}^{\infty} \lambda_j \psi_j(x)\psi_j(y)$$

and

$$\int \psi_i(x)\psi_j(x)dF(x) = \delta_{ij} \equiv 1[i = j].$$

- (a) Assuming the preceding, show that

$$\begin{aligned} \int u(x, y)d\mathcal{E}_n(x)d\mathcal{E}_n(y) &= \sum_{j=1}^{\infty} \lambda_j \int \psi_j(x)\psi_j(y)d\mathcal{E}_n(x)d\mathcal{E}_n(y) \\ &= \sum_{j=1}^{\infty} \lambda_j \underbrace{\left(\int \psi_j(x)d\mathcal{E}_n(x) \right)^2}_{Z_{jn}}. \end{aligned}$$

(b) Show that $\text{Cov}(Z_{in} Z_{jn}) = \delta_{ij}$ and hence,

$$(Z_{1n}, \dots, Z_{kn}) \implies (Z_1, \dots, Z_k), Z_j \text{ i.i.d. } \mathcal{N}(0, 1).$$

(c) If $\sum_{j=1}^{\infty} |\lambda_j| < \infty$, then $nU \implies \sum_{j=1}^{\infty} \lambda_j Z_j^2$.

Hint. $\int \sum_{j=1}^k \lambda_j \psi_j(x) \psi_j(y) d\mathcal{E}_n(x) d\mathcal{E}_n(y) \implies \sum_{j=1}^k \lambda_j Z_j^2$ and

$$\begin{aligned} & E \left(\int (U(x, y) - \sum_{j=1}^k \lambda_j \psi_j(x) \psi_j(y)) d\mathcal{E}_n(x) d\mathcal{E}_n(y) \right)^2 \\ &= E \left(\sum_{j=k+1}^{\infty} Z_{jn}^2 \lambda_j \right)^2 = \text{Var} \left(\sum_{j=k+1}^{\infty} Z_{jn}^2 \lambda_j \right) + \left(\sum_{j=k+1}^{\infty} \lambda_j \right)^2. \end{aligned}$$

6. Let $\mathbf{U} \sim \mathcal{N}_p(\mathbf{0}, 1)$ and let $M_{p \times p}$ be symmetric. Show that

(a) $\mathbf{U}^T M \mathbf{U} \sim \sum_{j=1}^p \lambda_j Z_j^2$ where the Z_j are i.i.d. $\mathcal{N}(0, 1)$ and $\lambda_1, \dots, \lambda_k$ are the eigenvalues of M . That is,

$$M = P \Lambda P^T \quad (7.4.1)$$

where $\Lambda = \text{diag}(\lambda_1, \dots, \lambda_p)$, and P is orthonormal. (You may assume (7.4.1), a generalization of the principal axis theorem. (B.10.1.1))

(b) Show that (7.4.1) is equivalent to the existence of unique $\mathbf{v}_1, \dots, \mathbf{v}_p$ orthonormal such that

$$M \mathbf{v}_j = \lambda_j \mathbf{v}_j, \quad \mathbf{v}_j^T M = \lambda_j \mathbf{v}_j^T, \quad 1 \leq j \leq p.$$

7. Show that U_n defined in (7.3.26) is the UMVU estimate of $E_P \phi(X_1, \dots, X_m)$.

8. Establish (7.3.20) rigorously.

9. Establish the conclusion of Remark 7.3.1 for V_n .

10. Establish (7.3.32).

11. (a) Show that the one-sample Wilcoxon statistic W is asymptotically normal with mean $\int (1 - F(-x)) dF(x)$ and variance $\sigma^2(F)/n$, where

$$\sigma^2(F) = \int (1 - F(-x))^2 dF(x) - \mu^2(F)$$

and $F(x) = P[X \leq x]$.

(b) Show that if $F(x)$ is continuous and $F(x) = 1 - F(-x)$ for all x , then $\mu(F) = 1/2$ and $\sigma^2(F) = 1/12$.

12. Suppose $\widetilde{W} \equiv n^{-2} \sum_{i,j} 1(X_i + X_j > 0)$.

(a) Show that \widetilde{W} is a von Mises statistic and that under the assumption of Problem 11 (b), $\widetilde{W} = W + O_p(n^{-1/2})$, where W is the one-sample Wilcoxon statistic.

(b) Show that the identity of (a) fails if F is not continuous.

13. Develop a multivariate central limit theorem for U statistics. We state it for U statistics of order 2 but the generalization is obvious. Let $\varphi = (\varphi_1, \dots, \varphi_k)$ be a function from $\mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}^k$ such that $\int \varphi_j^2(x, y) dF(x) dF(y) < \infty$, for all j , and φ_j are symmetric, $1 \leq j \leq k$.

(a) Show that $U_n = \binom{n}{2} \sum_{i < j} \varphi(X_i, X_j)$ is asymptotically normal with expectation $E\varphi(X_1, X_2)$ and variance $2\text{Var}(E[\varphi(X_1, X_2)|X_1])$.

(b) Extend this result to U statistics of order m_1, m_2, \dots, m_k .

14. Show that even if $P[X_i = X_j] > 0$ for $i \neq j$ nevertheless the von Mises statistic

$$n^{-2} \sum_{i,j} \varphi(X_i, X_j)$$

is asymptotically normal if $E\varphi^2(X_1, X_2) < \infty$ and $E\varphi^2(X_1, X_1) < \infty$.

Hint. Use Problem 13.

15. Apply Problem 14 to derive the asymptotic distribution of \widetilde{W} in Problem 12 generally.

16. Derive the results parallel to Problems 11, 12, and 15 for Kendall's τ .

17. Establish (7.3.31).

7.5 Notes

Notes for Section 7.1

(1) It may be shown that if T is Euclidean, given the FIDIS of $\{X(t) : t \in T\}$, it is always possible to choose $X(t)$ separable and $X(\cdot, \cdot)$ defined by $X(u, v) = X(v) - X(u)$ separable — see Doob (1953), pp. 46–47. Unfortunately the empirical processes in general, can be defined on sets of functions T , which may not be identified with a subset of Euclidean, or even a separable metric space. The only recourse is then to go to so called “outer probabilities” — see van der Vaart and Wellner (1996) and Pollard (1990). However, in all cases of interest that we treat, separability of $X(\cdot)$ and $X(\cdot, \cdot)$ (and measurability) are more or less evident, so we shall ignore this technicality.

Chapter 8

DISTRIBUTION-FREE, UNBIASED, AND EQUIVARIANT PROCEDURES

8.1 Introduction

In this chapter we leave asymptotic theory and elaborate on some general inference principles initiated in Chapter 1, Sections 1.2 and 1.3 and Chapter 3, Sections 3.3 and 3.4. The results will guide the construction of procedures with optimal asymptotic properties for models much more general than those of this chapter. Recent work relating to the topics in this chapter deals with optimal inference after data based model selection. See Fithian, Sun and Taylor (2014).

One motivating question is, when can we construct a test of composite hypotheses whose power function, probability of Type I error, is constant on the hypothesis and, if so, how can we do it? In examining answers to this question we will establish results important for answering other decision theoretic questions having to do with estimation as well.

The first answer, elaborated on in Section 8.2, is based on the concept of a complete sufficient statistic when the null hypothesis is our model. Such a statistic, which, if it exists, must be minimal sufficient, can be conditioned on, and the resulting tests with conditional Type I error probability equal to α have the desired property. An important example discussed is permutation tests. It turns out that the notion of completeness is also what's needed to construct uniformly minimum variance unbiased (UMVU) estimates in situations where the information bound for unbiased estimates is not achievable. This is a much more powerful method than the one discussed in Section 3.4.

The second answer is based on the notion that some models exhibit symmetries which, for loss functions having corresponding symmetries, lead us to consider decision procedures which are also symmetric. That is, there is a group acting on the model which is isomorphic to a group acting on the action space \mathcal{A} which in turn is isomorphic to a group acting on decision procedures. For instance in testing $H : \theta = 0$ vs $K : \theta \neq 0$ when we observe X_1, \dots, X_n i.i.d. with density $f(x - \theta)$ and f is symmetric about 0 it seems reasonable to restrict consideration to tests δ such that $\delta(X_1, \dots, X_n) = \delta(-X_1, \dots, -X_n)$. The argument, much elaborated in Section 8.3, is that if H holds for X_1, \dots, X_n it does so for $-X_1, \dots, -X_n$ as well and if it doesn't then $-X_1, \dots, -X_n$ are symmetrically distributed about $-\theta \neq 0$. So, it seems reasonable that X_1, \dots, X_n and $-X_1, \dots, -X_n$

should lead to the same action. Note, however, that, implicitly, we are postulating a problem of a malevolent nature which might flip the signs on the X_i we see to confuse us. This suggests via Section 3.3 a connection to the minimax principle which we shall explore.

How do symmetries relate to our initial question? Symmetry of the procedure leads to critical regions with constant probability of Type I error if $\theta = 0$. If in the problem we just discussed we specify that $f(x) = (1/\sigma)f_0(x/\sigma)$, with f_0 known and σ unknown, then symmetry arguments lead us to require $\delta(aX_1, \dots, aX_n) = \delta(X_1, \dots, X_n)$, $a \neq 0$. We pursue this argument to show that such tests, which are called *invariant* with respect to multiplication, have constant probability of Type I error in σ if $\theta = 0$.

These notions can be extended to requiring properties for estimators δ of θ , such as

$$\delta(X_1 + c, \dots, X_n + c) = \delta(X_1, \dots, X_n) + c. \quad (8.1.1)$$

Examples of estimators satisfying (8.1.1) are the mean and the median. See Problem 3.5.6. Such estimates are called *equivariant* with respect to addition. Such notions turn out to be closely linked to minimaxity so that we can find minimax procedures by restricting to such procedures to begin with.

Although these ideas are initially applicable only in restricted classes of models, their applicability where the X_i are i.i.d. $\mathcal{N}(\mu, \Sigma_0)$ with Σ_0 known can be used to establish asymptotic versions of these results for all smooth, regular i.i.d. parametric models.

8.2 Similarity and Completeness

We have seen that in the case of the t test under the $\mathcal{N}(\mu, \sigma^2)$ model, the t -statistic has the same t -distribution for all values of σ^2 . In this section, we shall study a key property that a number of models have which permits such a reduction. This reduction is based on a set of ideas introduced and developed by Lehmann and Scheffé (1950, 1955) that have another interesting application we shall discuss. Using these notions, it is possible to essentially characterize situations in which we can expect a UMVU estimate to exist and give a constructive method to obtain these procedures. As we shall see, this approach is far more powerful than that of Section 3.4. We develop the concept of completeness for a model we denote by \mathcal{P}_0 . In the testing context, \mathcal{P}_0 is the class of probabilities specified by the hypothesis. In the estimation context, it is typically the class \mathcal{P} of all probabilities under consideration.

8.2.1 Testing

We consider the general testing problem in the Neyman–Pearson framework: we observe $X \in \mathcal{X}$, $X \sim P \in \mathcal{P}$, with hypothesis $H : P \in \mathcal{P}_0$, $\mathcal{P}_0 \subset \mathcal{P}$ and alternative $K : P \notin \mathcal{P}_0$.

Recall that a test ϕ has level α if $E_P(\phi(X)) \leq \alpha$ for all $P \in \mathcal{P}_0$ and ϕ has size α if $\sup\{E_P(\phi(X)) : P \in \mathcal{P}_0\} = \alpha$.

Definition 8.2.1. A (randomized) level α test ϕ is *similar*⁽¹⁾ *level* α if $E_P\phi(X) = \alpha$ for all $P \in \mathcal{P}_0$.

Evidently, a similar test has size α as well as level α . A statistic $T(X)$ is *distribution-free* if $P(T(X) \leq t)$ is the same for all $P \in \mathcal{P}_0$. For such $T(X)$, tests of the form $\varphi(X) = 1(T(X) \in A)$ are similar. Similar tests other than $\phi(X) \equiv \alpha$ may not exist, and may exist only for some α (Problem 8.2.19). There is one very important general situation in which they may readily be constructed: when a complete, sufficient statistic exists.

Similarity is a property of a test and a model \mathcal{P}_0 . Completeness is a property of a statistic $T(X)$ and the model obtained when $X \sim P$, $P \in \mathcal{P}_0$. We consider $T(X)$, and the model $\mathcal{Q}_0 = \{PT^{-1}; P \in \mathcal{P}_0\}$, the set of distributions $P(T(X) \leq t)$ of $T(X)$ as P ranges over \mathcal{P}_0 .

Definition 8.2.2. The model $\mathcal{Q}_0 = \{PT^{-1}; P \in \mathcal{P}_0\}$ and the statistic T are *complete* for \mathcal{Q}_0 and \mathcal{P}_0 , respectively, if, for any function $v(T)$, the system of equations,

$$E_Q v(T) = 0, \text{ for all } Q \in \mathcal{Q}_0, \quad (8.2.1)$$

or, equivalently,

$$E_P v(T(X)) = 0, \text{ for all } P \in \mathcal{P}_0,$$

implies that

$$Q[v(T) = 0] = P[v(T(X)) = 0] = 1$$

for all $Q \in \mathcal{Q}_0$, respectively, $P \in \mathcal{P}_0$.

That is, the system (8.2.1) has only the trivial solution $v = 0$. The main utility of completeness in testing comes from the following theorem.

Theorem 8.2.1. Let $T(X)$ be a complete, sufficient statistic for \mathcal{P}_0 . Let ϕ be a randomized test of H . Then, ϕ is similar level α iff

$$E_P(\phi(X)|T(X)) = \alpha \quad (8.2.2)$$

with P probability 1 for all $P \in \mathcal{P}_0$.

Property (8.2.2) is sometimes called *Neyman structure*.

Proof. Evidently, by the iterated expectation theorem, B.1.20, (8.2.2) implies similarity. On the other hand, suppose that, for $P \in \mathcal{P}_0$,

$$E_P \phi(X) = E_P(E_P(\phi(X)|T(X))) = \alpha \quad (8.2.3)$$

where, by sufficiency, $E_P(\phi(X)|T(X))$ is a function of $T(X)$ not depending on P so that the subscript P on $E_P(\phi(X)|T(X))$ can be dropped. That is, $E(\phi(X)|T(X))$ is a statistic. Rewrite (8.2.3) as

$$E_P(E(\phi(X)|T(X)) - \alpha) = 0 \quad (8.2.4)$$

for all $P \in \mathcal{P}_0$. Completeness leads to (8.2.2). \square

The following simple but useful result follows from (8.2.3).

Proposition 8.2.1. If T is sufficient for \mathcal{P}_0 and if $E_P(\phi(X)|T) \leq \alpha$ for $P \in \mathcal{P}_0$, then ϕ is level α .

Completeness and sufficiency are properties of a statistic that push in opposite directions. For instance, the trivial sufficient statistic $T(X) = X$ is not complete unless \mathcal{P}_0 contains all point masses. On the other hand, if $T(X)$ is a constant, then $T(X)$ is complete but sufficient only if $\mathcal{P}_0 = \{P_0\}$. To see if the concept of completeness is useful, we need nontrivial examples of complete, sufficient statistics.

Example 8.2.1. *Bernoulli trials.* Suppose X_1, \dots, X_n are i.i.d. as X where $P_\theta[X = 1] = \theta = 1 - P_\theta[X = 0]$, $0 < \theta < 1$. Then, $T \equiv \sum_{i=1}^n X_i$ is sufficient for θ . We claim that T is complete. To see this, write (8.2.1) as

$$\sum_{k=0}^n v(k) \binom{n}{k} \theta^k (1-\theta)^{n-k} = 0, \quad 0 < \theta < 1.$$

or, equivalently, with $\rho = \theta/(1-\theta)$,

$$\sum_{k=0}^n v(k) \binom{n}{k} \rho^k = 0, \tag{8.2.5}$$

for all $\rho > 0$. But (8.2.5) evidently requires $v(k) = 0$, $0 \leq k \leq n$, since any polynomial of degree n has at most n roots. \square

Not surprisingly, Example 8.2.1 is a special case of a general result for exponential families.

Theorem 8.2.2. Suppose $p(\mathbf{x}; \boldsymbol{\theta})$, $\boldsymbol{\theta} \in \Theta \subset R^k$, is an exponential family of rank k with sufficient statistic $\mathbf{T}(\mathbf{X}) \equiv (T_1(\mathbf{X}), \dots, T_k(\mathbf{X}))^T$,

$$p(\mathbf{x}; \boldsymbol{\theta}) = \exp \left\{ \sum_{j=1}^k \eta_j(\boldsymbol{\theta}) T_j(\mathbf{x}) - A(\boldsymbol{\eta}(\boldsymbol{\theta})) \right\} h(\mathbf{x}),$$

and $\boldsymbol{\eta}(\Theta)$ has a nonempty interior. Then, $\mathbf{T}(\mathbf{X})$ is complete as well as sufficient.

Proof. The system of equations (8.2.1) implies that

$$\int_{R^k} h(\mathbf{t}) \exp \left\{ \sum_{j=1}^k \eta_j t_j - A(\boldsymbol{\eta}) \right\} v(\mathbf{t}) d\mathbf{t} = 0$$

for all $\boldsymbol{\eta}$ belonging to an open set, or, equivalently,

$$\int_{R^k} h(\mathbf{t}) \exp \left\{ \sum_{j=1}^k \eta_j t_j \right\} v(\mathbf{t}) d\mathbf{t} = 0 \tag{8.2.6}$$

for all $\boldsymbol{\eta}$ in an open set. Now (8.2.6) implies that $h(\mathbf{t})v(\mathbf{t}) = 0$ by a classical theorem on multivariate Laplace transforms (Widder (1941)). \square

We gave the proof of this result for the continuous case but it applies equally well in the discrete case or for any exponential family dominated by some measure μ .

Here are some immediate consequences of Theorem 8.2.2.

- (i) If X_1, \dots, X_n are i.i.d. $\mathcal{N}(\mu, \sigma^2)$ with μ, σ^2 both unknown, then $(\bar{X}, \hat{\sigma}^2)^T$ is complete and sufficient. If σ^2 is known, \bar{X} is complete, sufficient. If μ is known, $\sum(X_i - \mu)^2$ is complete, sufficient.
- (ii) If X_1, \dots, X_n are i.i.d. Poisson (λ) , $\lambda > 0$, $\sum_{i=1}^n X_i$ is complete, sufficient.

Completeness is not limited to exponential families.

Example 8.2.2. *Completeness in the $\mathcal{U}(0, \theta)$ family, $\theta > 0$.* If X_1, \dots, X_n are i.i.d. $\mathcal{U}(0, \theta)$, then $M_n \equiv \max_i\{X_i\}$ is sufficient. Note that $P(M_n \leq t) = P(\text{all } X_i \leq t) = (t/\theta)^n$ for all $0 \leq t < \theta$ so that (8.2.1) becomes

$$\frac{1}{\theta} \int_0^\theta v(t)n\left(\frac{t}{\theta}\right)^{n-1} dt = n\theta^{-n} \int_0^\theta v(t)t^{n-1} dt = 0$$

or, equivalently, $\int_0^\theta v(t)t^{n-1} dt = 0$ for all $\theta > 0$. Differentiating with respect to θ we obtain $v(\theta)\theta^{n-1} = 0 \implies v(\theta) = 0$ and completeness is proved. \square

Example 8.2.3. *Completeness of the order statistics.* Here is an example which will prove important for both testing and estimation. Let X_1, \dots, X_n be i.i.d. $f \in \mathcal{F}$ where \mathcal{F} is the set of all continuous case densities. Then, $X_{(1)} < \dots < X_{(n)}$, the order statistics, are sufficient (Problem 1.5.8). We next argue that they are complete. Now, $v(x_1, \dots, x_n)$ defined for $x_1 < x_2 < \dots < x_n$ can be extended to all of R^n by symmetry $v(x_{i_1}, \dots, x_{i_n}) \equiv h(x_1, \dots, x_n)$ for all permutations (i_1, \dots, i_n) . So (8.2.1) becomes

$$\int \dots \int v(x_1, \dots, x_n) f(x_1) \dots f(x_n) dx_1, \dots, dx_n = 0 \quad (8.2.7)$$

for all $f \in \mathcal{F}$. Let g_1, \dots, g_n be arbitrary densities in \mathcal{F} so that $\sum_{j=1}^n \alpha_j g_j \in \mathcal{F}$ if $\alpha_j \geq 0$, for all j , $\sum_{j=1}^n \alpha_j = 1$. Then (8.2.7) implies

$$G(\alpha_1, \dots, \alpha_n, g_1, \dots, g_n) \equiv \int \dots \int v(x_1, \dots, x_n) \prod_{i=1}^n \left(\sum_{j=1}^n \alpha_j g_j(x_i) \right) dx_i = 0$$

for all α , $(g_1, \dots, g_n)^T$ as above. Write $\alpha_j = w_j / \sum_{k=1}^n w_k$ where $w_k \geq 0$, $1 \leq k \leq n$. Simplifying further,

$$G(w_1, \dots, w_n, g_1, \dots, g_n) = \int \dots \int v(x_1, \dots, x_n) \prod_{i=1}^n \left(\sum_{j=1}^n w_j g_j(x_i) \right) dx_i = 0. \quad (8.2.8)$$

The coefficient of $\prod_{j=1}^n w_j$ is

$$\frac{\partial^n G}{\partial w_1 \dots \partial w_n}(w_1, \dots, w_n, g_1, \dots, g_n) \Big|_{\mathbf{w}=0} = \int \dots \int v(x_1, \dots, x_n) \prod_{i=1}^n g_i(x_i) dx_i. \quad (8.2.9)$$

Put $g_j(x) = 1(x \in A_j)/|A_j|$, the uniform density on A_j , where $|A|$ is the length of A and A_1, \dots, A_n are disjoint arbitrary intervals. Then (8.2.9) becomes

$$\int_{A_1}, \dots, \int_{A_n} v(x_1, \dots, x_n) dx_1, \dots, dx_n = 0 \quad (8.2.10)$$

for all A_1, \dots, A_n and (8.2.10) implies that $v = 0$. \square

The order statistics are also complete for the case where $X_i \sim F$ with F continuous. See Bell, Blackwell and Breiman (1960).

Let us now apply completeness to obtain distribution-free tests in a number of important examples.

Testing hypotheses in k parameter exponential families

Suppose $\mathbf{X} \sim \{P_{\boldsymbol{\theta}} : \boldsymbol{\theta} \in \Theta \subset R^k\}$, a canonical k parameter exponential family, with $k \geq 2$ and Θ open, with density function,

$$p(\mathbf{x}; \boldsymbol{\theta}) = \exp \left\{ \sum_{j=1}^k \theta_j T_j(\mathbf{x}) - A(\boldsymbol{\theta}) \right\} h(\mathbf{x}).$$

We want to test $H : \theta_1 = \theta_{10}$ where θ_{10} is a specified value, that is, $\boldsymbol{\theta} \in \Theta_0 = \{\boldsymbol{\theta} : \theta_1 = \theta_{10}, \boldsymbol{\theta} \in \Theta\}$. Thus, H is $P_{\boldsymbol{\theta}} \in \mathcal{P}_0$ where \mathcal{P}_0 is the $k-1$ parameter exponential family with densities

$$q(\mathbf{x}; \theta_2, \dots, \theta_k) = \exp \left\{ \sum_{j=2}^k \theta_j T_j(\mathbf{x}) - B(\theta_2, \dots, \theta_k) \right\} h_0(\mathbf{x})$$

where $h_0(\mathbf{x}) = \exp[\theta_{10} T_1(\mathbf{x})] h(\mathbf{x})$. By Theorem 8.2.2, $(T_2(\mathbf{X}), \dots, T_k(\mathbf{X}))^T$ is complete and sufficient for \mathcal{P}_0 . Thus, if $\phi(\mathbf{T})$, where $\mathbf{T} = (T_1(\mathbf{X}), \dots, T_k(\mathbf{X}))^T$, is any test such that $E_{\boldsymbol{\theta}} \phi(\mathbf{T}) = \alpha$, $\boldsymbol{\theta} \in \Theta_0$, we must have

$$E_{\boldsymbol{\theta}} (\phi(\mathbf{T}) | (T_2, \dots, T_k)) = \alpha. \quad (8.2.11)$$

Computing the conditional expectation in (8.2.11) is straightforward since the conditional distribution of \mathbf{T} given $T_j = t_j$, $j \geq 2$ is determined by that of T_1 given $T_j = t_j$, $j \geq 2$. By the k dimensional version of Theorem 1.6.1, (T_1, \dots, T_k) has density of the form

$$p(t_1, \dots, t_k; \boldsymbol{\theta}) = \exp\{\boldsymbol{\theta}^T \mathbf{t} - A_1(\boldsymbol{\theta})\} h_1(\mathbf{t}).$$

Here (Problem 8.2.20), the conditional distribution of T_1 given $T_{[2,k]} \equiv (T_2, \dots, T_k)^T$ has density of the form

$$p(t) = \exp\{\theta_{10}t - B_0(\theta_{10}, t_2, \dots, t_k)\}h_0(t_2, \dots, t_k) \quad (8.2.12)$$

so that (8.2.11) is just

$$\int \phi(t, t_2, \dots, t_k) p(t) dt = \alpha. \quad (8.2.13)$$

The usual analogous formulae apply for the discrete case.

Conversely, suppose we start with any test statistic $S(\mathbf{T})$ and

$$\begin{aligned} \phi(\mathbf{T}) &= 1 \text{ if } S(\mathbf{T}) > s \\ &= 0 \text{ if } S(\mathbf{T}) < s. \end{aligned}$$

In general, we can not choose s so that $E_{\boldsymbol{\theta}}\phi(\mathbf{T}) \leq \alpha$ for all $\boldsymbol{\theta} \in \Theta_0$. However, consider the randomized test,

$$\begin{aligned} \phi^*(\mathbf{T}) &= 1, \quad S(\mathbf{T}) > s(T_2, \dots, T_k) \\ &= \gamma, \quad S(\mathbf{T}) = s(T_2, \dots, T_k) \\ &= 0, \quad S(\mathbf{T}) < s(T_2, \dots, T_k), \end{aligned}$$

where $s(t_2, \dots, t_k)$ and $\gamma(t_2, \dots, t_k)$ are determined by

$$E_{\boldsymbol{\theta}}\{\phi^*(\mathbf{T})|T_2, \dots, T_k\} = \alpha$$

under H . This is a useful test since the conditional distribution of $S(\mathbf{T})$ given $T_2 = t_2, \dots, T_k = t_k$ does not depend on $\theta_2, \dots, \theta_k$ and it has level α by Theorem 8.2.1. By Proposition 8.2.1, if we are satisfied with $E_{\boldsymbol{\theta}}\phi(\mathbf{T}) \leq \alpha$ for all $\boldsymbol{\theta} \in \Theta_0$, then we can dispense with randomization in $\phi^*(\mathbf{T})$. In the continuous case $\phi^*(\mathbf{T})$ is similar without randomization. Note that $\phi^*(\mathbf{T})$ may not be the same as the test $\phi(\mathbf{T})$ we started with, but asymptotically they are usually equivalent to first order.

Remark 8.2.1 We can start with any statistic $S(\mathbf{T})$ to construct a similar test $\phi^*(\mathbf{T})$. What makes the test we have just obtained appropriate? The test $\phi^*(\mathbf{T})$ is “aimed” at one-sided alternatives. We shall see later in this section that the tests we have introduced do have optimality properties. \square

Example 8.2.4. *Another look at testing location and scale for the Gaussian distribution.* Suppose X_1, \dots, X_n are i.i.d. $\mathcal{N}(\mu, \sigma^2)$. This is an exponential family with $T_1(\mathbf{X}) = \sum_{i=1}^n X_i$, $T_2(\mathbf{X}) = \sum_{i=1}^n X_i^2$, $\theta_1 = \mu/\sigma^2$, $\theta_2 = -1/2\sigma^2$. We seek a test with guaranteed level α of $H : \mu = 0$ vs $K : \mu > 0$. This is the appropriate test if the X_i are case control differences. If $\sigma^2 = \sigma_0^2$ is known, the UMP test is

$$\phi(T_1) = 1 \text{ iff } \sum_{i=1}^n X_i \geq \sqrt{n}\sigma_0\Phi^{-1}(1 - \alpha).$$

If we are wrong about σ_0 , this test can have arbitrarily large level. We can apply the principles of this section to conclude that the test

$$\phi(T_1, T_2) = 1 \text{ iff } T_1 \geq c(T_2, \alpha)$$

with $T_1 = \sum X_i$ and

$$P_{\boldsymbol{\theta}}[T_1 \geq c(t_2, \alpha) | T_2 = t_2] = \alpha \quad (8.2.14)$$

for all t_2 has Neyman structure. In our case, the distribution of $T_1 | T_2 = t_2$ can be determined from

$$(T_1, T_2) = \left(\sum X_i, \sum (X_i - \bar{X})^2 + \frac{(\sum X_i)^2}{n} \right)$$

where, by Theorem B.3.3, $\sum X_i, \sum (X_i - \bar{X})^2$ are independent $\mathcal{N}(0, \sigma^2/n)$, $\sigma^2 \chi_{n-1}^2$, respectively. Here $V \sim \sigma^2 \chi^2$ means that $(V/\sigma^2) \sim \chi^2$. Consider the t -statistic

$$t = \frac{\sqrt{n}\bar{X}}{s} = \frac{\frac{1}{\sqrt{n}}T_1}{\{\frac{1}{n-1}[T_2 - T_1^2/n]\}^{\frac{1}{2}}}.$$

For fixed $T_2 = t_2$, t is an increasing function of T_1 . Thus for each $s(t_2)$ and $T_2 = t_2$, $T_1 \geq s(t_2)$ is equivalent to $t \geq d(t_2)$ for some $d(t_2)$, and it is enough to find $d(t_2)$ such that

$$P(T_1 \geq s(t_2) | T_2 = t_2) = P_H(t \geq d(T_2) | T_2 = t_2) = \alpha. \quad (8.2.15)$$

Here the distribution of t does not depend on σ ; thus, if b_α is such that

$$P_H(t \geq b_\alpha) = \alpha$$

then by the Neyman structure result (8.2.2), for all $\sigma > 0$,

$$P_H(t \geq b_\alpha | T_2 = t_2) = \alpha \quad (8.2.16)$$

and we can determine $s(t_2)$ from (8.2.15) and (8.2.16). Note that (8.2.16) holding for all $\alpha \in (0, 1)$ is equivalent to independence of t and T_2 under H . This example illustrates the computation of the conditional critical value $s(t_2)$, and it shows connections to earlier normal model procedures. In practice, we would carry out the test as in Vol. I using the t -distribution of t .

We can extend the argument leading to (8.2.13) for testing the hypothesis $H : \theta_1 = \theta_{10}$ to $H : \boldsymbol{\theta}_S = \boldsymbol{\theta}_{0S}$ where $\boldsymbol{\theta}_S \equiv \{\theta_i : i \in S\}$, $S \subset (1, \dots, k)$. Here is a simple application.

Example 8.2.5. Gaussian goodness-of-fit. Suppose we wish to test $H : F \in \{\mathcal{N}(\mu, \sigma^2) : \mu \in R, \sigma^2 > 0\}$ versus a particular alternative with specified density f which is not

Gaussian. If we knew $\mu = \mu_0$ and $\sigma^2 = \sigma_0^2$, the Neyman-Pearson Lemma implies that the most powerful statistic would be

$$S(\mathbf{X}) = \sum_{i=1}^n \log f(X_i) + \frac{1}{2\sigma_0^2} \sum X_i^2 - \frac{\mu_0}{\sigma_0^2} \sum X_i + \frac{n\mu_0^2}{2\sigma_0^2}.$$

Using our $\phi^*(\mathbf{T})$ construction the similar size α test based on S does not depend on μ_0 and σ_0^2 . It is just, if $n \geq 3$,

$$\begin{aligned} \phi^*(\mathbf{X}) &= 1 \quad \text{iff} \quad \sum_{i=1}^n \log f(X_i) \geq c\left(\bar{X}, \sum_{i=1}^n (X_i - \bar{X})^2\right) \\ &= 0 \quad \text{otherwise} \end{aligned}$$

where

$$P\left[\sum_{i=1}^n \log f(X_i) \geq c\left(\bar{X}, \sum_{i=1}^n (X_i - \bar{X})^2\right) \mid \sum X_i, \sum X_i^2\right] = \alpha.$$

Since, given \bar{X} and $\sum_{i=1}^n (X_i - \bar{X})^2$, $(X_1, \dots, X_n)^T$ is uniformly distributed on the surface of the n sphere with center $(\bar{X}, \dots, \bar{X})$ and radius $\sqrt{\sum_{i=1}^n (X_i - \bar{X})^2}$, computing involves evaluating the area of the intersection of the surface of the sphere with

$$\left\{(x_1, \dots, x_n)^T : \sum_{i=1}^n \log f(x_i) \geq c\right\}$$

which is, unfortunately, not trivial in general.

Note that unlike the test “Reject iff $\sum_{i=1}^n \log f(X_i)$ is large,” $\phi^*(\mathbf{X})$ has probability of Type I error α whatever be μ and σ^2 . \square

Here is an important and typical example where we can, in general, not obtain similarity without randomization but can guarantee level α .

Example 8.2.6. Independence in contingency tables. Suppose $(X_i, Y_i), 1 \leq i \leq n$ are, as in Section 6.4, i.i.d. with X_i taking values x_1, \dots, x_r and Y_i values y_1, \dots, y_s , with arbitrary probabilities, $\theta_{ab} = P[X = x_a, Y = y_b]$, $0 < \theta_{ab} < 1$ for all a, b . We can immediately reduce by sufficiency to $\mathbf{N} = \{N_{ab} : 1 \leq a \leq r, 1 \leq b \leq s\}$ with $N_{ab} \equiv \sum 1(X_i = x_a, Y_i = y_b)$ which has a multinomial $\mathcal{M}(n, \{\theta_{ab}\})$ distribution. The hypothesis of independence of X and Y is $H : \theta_{ab} = \theta_{a+} \theta_{+b}$, for all a, b where $+$ indicates summation over the index. The Wald test of Section 6.3.2 for this hypothesis is based on the Pearson χ^2 statistic,

$$\chi^2 = \sum_{a=1}^r \sum_{b=1}^s \frac{n(N_{ab} - \frac{N_{a+} N_{+b}}{n})^2}{N_{a+} N_{+b}},$$

see (6.4.9). Note that we can write the density function of \mathbf{N} under the hypothesis as

$$p(\mathbf{N} : \boldsymbol{\eta}) = \exp \left\{ \sum_{a,b} N_{ab} (\eta_{a1} + \eta_{b2}) \right\}$$

where $\eta_{a1} = \log \theta_{a+}$, $\eta_{b2} = \log \theta_{+b}$, an exponential family in canonical form with canonical sufficient statistics, $N_{a+} \equiv \sum_{b=1}^s N_{ab}$, $1 \leq a \leq r$, and $N_{+b} = \sum_{a=1}^r N_{ab}$, $1 \leq b \leq s$. Let $\mathbf{N}_{r+} = (N_{1+}, \dots, N_{r+})$ and $\mathbf{N}_{+s} = (N_{+1}, \dots, N_{+s})$. Consider the test which rejects H iff

$$\chi^2 \geq c(\mathbf{N}_{r+}, \mathbf{N}_{+s}),$$

and let c be chosen so that

$$P_H [\chi^2 \geq c(n_{a+}, n_{+b}, \mathbf{N}_{r+}, \mathbf{N}_{+s}) | N_{a+} = n_{a+}, N_{+b} = n_{+b}, \mathbf{N}_{r+}, \mathbf{N}_{+s}] \leq \alpha$$

and thus the test has level α under H . Note that (see Problem 8.2.23), given $N_{a+} = n_{a+}$, $N_{+b} = n_{+b}$, for all a, b ,

$$\chi^2 = n \left\{ \sum_{a,b} \frac{N_{ab}^2}{n_{a+} n_{+b}} - 1 \right\}. \quad (8.2.17)$$

To implement the test, we need to compute, under H , the conditional distribution of $\{\mathbf{N}_{ab}\}$ given $\mathbf{N}_{r+} = \mathbf{n}_{r+}$, $\mathbf{N}_{+s} = \mathbf{n}_{+s}$, which is multiple hypergeometric. In fact, computing the conditional null distribution of this statistic is computationally difficult if r and s are large. Exact results for low values of r and s are given in the package STATEXACT. More generally, Markov chain Monte Carlo methods such as that of Diaconis and Stumpfels (1998) can be used.

A particularly interesting example is Fisher's exact test, where $r = s = 2$. We note that the conditional χ^2 test given $N_{1+} = n_{1+}$, $N_{+1} = n_{+1}$ can be based on the distribution of N_{11} given N_{1+} , N_{+1} which, under H , is hypergeometric (n, n_{1+}, n_{+1}) . The one-sided version of the test, rejecting H for N_{11} large, has optimality properties. See Lehmann and Romano (2005). \square

Example 8.2.7. Two-sample permutation tests. Consider the nonparametric two-sample problem. We observe X_1, \dots, X_m i.i.d. F and Y_1, \dots, Y_n i.i.d. G with the X 's independent of the Y 's, and want to test $H : F = G$ against the alternative that G tends to give higher values than F , which can be formally expressed by $K : \bar{G}(x) \geq \bar{F}(x)$ for all x , which we refer to as G is *stochastically larger* than F — see Problems 1.1.4 and 8.3.11(d). If F and G are $\mathcal{N}(\mu, \sigma^2)$ under H and $\mathcal{N}(\mu, \sigma^2)$, $\mathcal{N}(\mu + \Delta, \sigma^2)$ under K , we have the Gaussian two-sample problem discussed in Section 4.9.5. The test for that parametric model is to reject iff

$$T \equiv \frac{\sqrt{\frac{mn}{m+n}}(\bar{Y} - \bar{X})}{s} \geq t_{m+n-2}(1 - \alpha)$$

where $s^2 = (m+n-2)^{-1} [\sum_{i=1}^m (X_i - \hat{\mu})^2 + \sum_{i=1}^n (Y_i - \hat{\mu})^2]$ and $\hat{\mu} = \lambda \bar{X} + (1-\lambda) \bar{Y}$, $\lambda \equiv m/(m+n)$. This test is not similar for general $H : F = G$. What is the similar version for this nonparametric hypothesis when F and G are continuous? Let Z_1, \dots, Z_N be $X_1, \dots, X_m, Y_1, \dots, Y_n$, $N = m+n$, and let $Z_{(1)}, \dots, Z_{(N)}$ be the ordered values. Since $(Z_{(1)}, \dots, Z_{(N)})$ is a complete sufficient statistic under H by Example 8.2.3, the

similar test is

$$\psi(\mathbf{X}, \mathbf{Y}) = \begin{cases} 1 & \text{if } T \geq c(Z_{(1)}, \dots, Z_{(N)}) \\ \gamma(Z_{(1)}, \dots, Z_{(N)}) & \text{if } T = c(Z_{(1)}, \dots, Z_{(N)}) \\ 0 & \text{otherwise} \end{cases}$$

where c and γ are chosen so that

$$E(\psi(\mathbf{X}, \mathbf{Y}) | Z_{(1)}, \dots, Z_{(N)}) = \alpha.$$

We simplify by noting that

$$\sum_{i=1}^m X_i + \sum_{i=1}^n Y_i = \sum_{i=1}^N Z_{(i)}, \quad \sum_{i=1}^m X_i^2 + \sum_{i=1}^n Y_i^2 = \sum_{i=1}^N Z_{(i)}^2.$$

Then

$$T = \sqrt{\frac{mn}{N+1}} \left(\frac{n^{-1} \sum_{i=1}^n Y_i - (\sum_{i=1}^N Z_{(i)} - \sum_{i=1}^n Y_i) m^{-1}}{\frac{1}{m+n-2} \left\{ \sum_{i=1}^N Z_{(i)}^2 - \frac{(\sum Z_{(i)})^2}{N} \right\}} \right),$$

a monotone function of $\sum_{i=1}^n Y_i$ given $Z_{(1)}, \dots, Z_{(N)}$ and hence given $\sum_{i=1}^N Z_{(i)}$ and $\sum_{i=1}^N Z_{(i)}^2$. Our test is therefore equivalent to

$$\psi(\mathbf{X}, \mathbf{Y}) = \begin{cases} 1 & \text{if } \sum_{i=1}^n Y_i > c(Z_{(1)}, \dots, Z_{(N)}) \\ \gamma(Z_{(1)}, \dots, Z_{(N)}) & \text{if } \sum_{i=1}^n Y_i = c(Z_{(1)}, \dots, Z_{(N)}) \\ 0 & \text{otherwise} \end{cases}$$

How do we find c ? Note that, given $(Z_{(1)}, \dots, Z_{(N)})$, the values of Y_1, \dots, Y_n are known, but not their order. Each ordering is equally likely under H ; thus, the conditional distribution of Y_1, \dots, Y_n given $(Z_{(1)}, \dots, Z_{(N)})$ is

$$P[Y_j = Z_{(i_j)}, 1 \leq j \leq n] = \binom{N}{n}^{-1}.$$

The test is thus equivalent to

- (i) Form $s_{(1)} < s_{(2)} < \dots < s_{(\frac{N}{n})}$, where the $s_{(k)}$ are the ordered $s = \sum_{i=1}^n Z_{(i)}$ as (i_1, \dots, i_n) ranges over distinct permutations.
- (ii) “Reject H if $\sum_{j=1}^n Y_j$ is among the top $\lceil \binom{N}{n} \alpha \rceil + 1$ of the $s_{(k)}$ and reject with probability $\gamma(Z_{(1)}, \dots, Z_{(N)})$ if $\sum_{i=1}^n Y_i$ is the $\lceil \binom{N}{n} \alpha \rceil$ th s_k .”

This test is distribution-free in the sense that it does not require the knowledge of the distribution $F = G$ under H . It is the so called size α *permutation t-test* introduced by R.A. Fisher. In practice, randomization is not used and one settles for a level α test with size $\leq \alpha$.

The main downside with permutation tests is that the calculation and ordering of the s_j become prohibitive for m, n large. However, there is a simple alternative, the *Monte Carlo*

permutation test: Resample $Z_{1b}^*, \dots, Z_{nb}^*$, $1 \leq b \leq B$, from $\{Z_{(1)}, \dots, Z_{(N)}\}$ with equal probability

$$P[Z_{11}^* = z_{(i_1)}, \dots, Z_{n1}^* = z_{(i_n)} | Z_{(i)} = z_{(i)}, 1 \leq i \leq N] = \binom{N}{n}^{-1}.$$

That is, draw n times without replacement from $\{Z_{(1)}, \dots, Z_{(N)}\}$ to get $(Z_{11}^*, \dots, Z_{n1}^*)$, then repeat B times. Order the resulting B sums $\sum_{i=1}^n Z_{ib}^*$, $1 \leq b \leq B$, and reject if $\sum_{i=1}^n Y_i$ is among the top $[B\alpha] + 1$. This test has level α . \square

Basu's theorem

We noted that in Example 8.2.4, we were able to eliminate the need for computing a conditional distribution since the t -statistic and $\sum_{i=1}^n X_i^2$ are independent under H . The existence of such statistics is closely linked to *ancillarity*.

Definition 8.2.3. A statistic $S(X)$ is said to be *ancillary* for $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$ if $\mathcal{L}_\theta S(X)$ is the same for all $\theta \in \Theta$.

Theorem 8.2.3. (Basu) Suppose $T(X)$ is sufficient and complete for \mathcal{P} . If $S(X)$ is ancillary for \mathcal{P} , then $T(X)$ and $S(X)$ are statistically independent for all $P \in \mathcal{P}$. If the support of P is the same for all $P \in \mathcal{P}$, then a converse holds: If $T(X)$ and $S(X)$ are independent for all P , then $S(X)$ is ancillary.

Proof. For any set A in the range of S write

$$P[S(X) \in A] = E_\theta [P[S(X) \in A | T]] . \quad (8.2.18)$$

Note that we need not put a θ subscript on the left hand side since S is ancillary, nor on the inside term of the right hand side since T is sufficient. Now (8.2.18) becomes

$$E_\theta [P[S(X) \in A | T] - P[S(X) \in A]] = 0$$

for all θ . By completeness $P[S(X) \in A | T] = P[S(X) \in A]$ and the independence of S and T follows. We leave the converse for Problem 8.2.21. \square

Example 8.2.4 (Continued). Here $T_2 = \sum X_i^2$ is complete, sufficient, and $t = \sqrt{n}\bar{X}/s$ is ancillary under $H : \mu = 0$. Thus T_2 and t are independent under H .

Example 8.2.7 (Continued). *Rank tests.* Let R_i be the rank of Y_i among $Z_{(1)}, \dots, Z_{(N)}$, that is, $Y_i = Z_{(R_i)}$. Then $\mathbf{R} = (R_1, \dots, R_n)$ is, under H , necessarily uniform on all combinations $\{\mathbf{r}\}$ of size n from $\{1, \dots, N\}$. That is, $P(\mathbf{R} = \mathbf{r}) = 1/\binom{N}{n}$. So \mathbf{R} is ancillary and tests based on it are distribution-free. An example of such a test is the two-sample Wilcoxon test which is based on the statistic $W = \sum_{i=1}^n R_i$, or equivalently, $U = \sum_{j=1}^m \sum_{i=1}^n 1(X_i < Y_j)$. Most software provide p -values for the Wilcoxon test when $\min\{m, n\} \leq 10$. For $\min\{m, n\} \geq 10$, $(U - \frac{1}{2}mn)/\sigma_W$ is close to $\mathcal{N}(0, 1)$, where $\sigma_W^2 = mn(m+n+1)/12$. \square

8.2.2 Testing Optimality Theory

We begin with

Definition 8.2.4. A test ϕ for $H : \theta \in \Theta_0$ vs. $K : \theta \in \Theta_1$ is *unbiased level α* iff $E_\theta \phi(X) \leq \alpha$, $\theta \in \Theta_0$, and $E_\theta \phi(X) \geq \alpha$, $\theta \in \Theta_1$.

Thus an unbiased test does at least as well as the trivial test $\phi \equiv \alpha$ at all θ . As with unbiasedness of estimates, being biased in testing is not necessarily bad. For instance, in testing $H : \mu = 0$ vs $K : \mu \neq 0$ on the basis of n observations from $\mathcal{N}(\mu, 1)$, the one-sided tests, “Reject iff $\sqrt{n}\bar{X} \geq z(1 - \alpha)$ ” and “Reject iff $\sqrt{n}\bar{X} \leq z(\alpha)$ ” are both biased yet reasonable. Nevertheless, unbiasedness often leads to reasonable procedures. Recall that UMP stands for “uniformly most powerful.”

Definition 8.2.5. The test ϕ^* is a *UMP unbiased level α test* if it is unbiased and $E_\theta \phi^*(X) \geq E_\theta \phi(X)$ for all $\theta \in \Theta_1$ if ϕ is also unbiased level α .

Let $\omega \equiv \partial\Theta_0 \cap \partial\Theta_1$ denote the intersection of the boundaries of Θ_0 and Θ_1 . We assume that ω is not empty. Note that if ϕ is unbiased level α and $\theta \rightarrow E_\theta \phi(x)$ is continuous, then ϕ is similar on ω , that is, $E_\theta \phi(x) = \alpha$ for $\theta \in \omega$.

Lemma 8.2.1. Lehmann. If $\theta \rightarrow E_\theta \phi(X)$ is continuous for each test ϕ , and if ϕ^* is level α and UMP among the class of test ϕ satisfying $E_\theta \phi(X) = \alpha$ for $\theta \in \omega$, then ϕ^* is UMP unbiased α .

Proof. The class of tests that satisfies $E_\theta \phi(X) = \alpha$ for $\theta \in \omega$ contains the class of tests that are unbiased level α . □

Remark 8.2.2. If $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$, $\Theta \subset R^d$ is an exponential family of distributions, then $\theta \rightarrow E_\theta \phi(X)$ is continuous, for every test function ϕ .

The connection between unbiasedness and completeness comes via

Theorem 8.2.4. Suppose $\theta \rightarrow E_\theta \phi(X)$ is continuous for all ϕ and $T(X)$ is sufficient and complete for $\mathcal{P}_0 = \{P_\theta : \theta \in \omega\}$. Suppose $\theta_1 \in \Theta_1$. Then the UMP unbiased level α test of $H : \theta \in \Theta_0$ vs $K : \theta = \theta_1$ is

$$\phi^*(X) = \begin{cases} 1 & \text{if } \frac{P_{\theta_1}(X|T(X))}{p_w(X|T(X))} > c(T(X), \theta_1) \\ 0 & \quad = c(T(X), \theta_1) \\ & \quad < c(T(X), \theta_1) \end{cases}$$

with c, γ determined by $E_\theta [\phi^*(X)|T(X)] = \alpha$, $\theta \in \omega$.

Proof. Optimality among all unbiased level α tests follows from:

- (i) Every unbiased level α test is similar on $\omega = \partial\Theta_0 \cap \partial\Theta_1$.
- (ii) Every test similar on ω has Neyman structure on ω .
- (iii) The conditional power, $E_{\theta_1}(\phi(X)|T(X) = t)$, is maximized for each t by ϕ^* . Here the conditional probability defining the Neyman-Pearson likelihood ratio is

$$p_{\theta_1}(x|T(X) = t) = \frac{p_{\theta_1}(X = x, T(X) = t)}{P_{\theta_1}(T(X) = t)} = \frac{p(x; \theta_1) \mathbf{1}(T(x) = t)}{P_{\theta_1}(T(X) = t)}.$$

By the iterated expectation theorem, (B.1.20), ϕ^* is also unconditionally most powerful. \square

A simple application of this result is

Theorem 8.2.5. Suppose X is distributed according to $P \in \mathcal{P}$, a canonical exponential family with $\log p(x, \theta) = \theta^T \mathbf{T}(x) - A(\theta) + b(x)$. Write $\mathbf{T} = (T_1, \mathbf{T}_2^T)^T$ and $\theta = (\theta_1, \theta_2^T)^T$. Then, the test

$$\begin{aligned}\phi^*(x) &= 1 && \text{if } T_1(x) > c(\mathbf{T}_2(x)) \\ &= \gamma(\mathbf{T}_2(x)) && \text{if } T_1(x) = c(\mathbf{T}_2(x)) \\ &= 0 && \text{otherwise}\end{aligned}\quad (8.2.19)$$

is UMP unbiased level α for $H : \theta_1 \leq \theta_{10}$ vs $K : \theta_1 > \theta_{10}$, where c and γ are determined by $E_\theta(\phi^*(X)|T(X)) = \alpha$, $\theta \in \omega$. Similarly, $1 - \phi^*$ is UMP unbiased for $H : \theta_1 \geq \theta_{10}$ vs $K : \theta_1 < \theta_{10}$.

Proof. The family of conditional distributions of X given $\mathbf{T}_2(X) = \mathbf{t}$ is a one parameter exponential family of the form

$$q(x; \mathbf{t}) = \exp\{\theta_1 T_1(x) - A(\theta_1, \mathbf{t})\} h(x, \mathbf{t}) dx.$$

Thus, we can apply Theorem 8.2.4 to conclude that the test (8.2.19) is most powerful among all unbiased level α tests. The test ϕ^* is UMP for $K : \theta_1 > \theta_{10}$ because it does not depend on θ_1 .

Remark 8.2.3. It is possible to extend this theory to show that procedures such as the two-sided t -test are UMP unbiased. See Lehmann and Romano (2005).

Example 8.2.6. (Continued). 2×2 contingency tables. Suppose that in Example 8.2.6, we have $r = s = 2$ and we are interested in testing H vs $K : \theta_{ab} > \theta_{a+}\theta_{+b}$ for all a, b . Write

$$\begin{aligned}p &= \theta_{11}^{N_{11}} \theta_{01}^{N_{1+}-N_{11}} \theta_{10}^{N_{10}} \theta_{00}^{N_{++}-N_{10}} \\ &= \left(\frac{\theta_{11}}{\theta_{01}}\right)^{N_{11}} \left(\frac{\theta_{10}}{\theta_{00}}\right)^{N_{10}} \theta_{01}^{N_{1+}} \theta_{10}^{N_{++}} \\ &= \gamma^{N_{11}} \left(\frac{\theta_{10}}{\theta_{00}}\right)^{N_{1+}} \theta_{01}^{N_{1+}} \theta_{10}^{n-N_{1+}} \\ &= \gamma^{N_{11}} \lambda^{N_{1+}} \eta^{N_{++}} \theta_{01}^n\end{aligned}\quad (8.2.20)$$

$$\text{where } \gamma = \left(\frac{\theta_{11}}{\theta_{01}}\right) / \left(\frac{\theta_{10}}{\theta_{00}}\right) = \frac{P[Y=1|X=1]P[Y=0|X=0]}{P[Y=1|X=0]P[Y=0|X=1]}$$

and λ and η are appropriately defined from (8.2.20). Note that $\gamma > 0 \iff \theta_{ab} > \theta_{a+}\theta_{+b}$ for all a, b , and thus $\gamma = 0$ corresponds to independence. Therefore, by Theorem 8.2.4, the test which rejects H if $N_{11} > c(N_{1+}, N_{++})$ randomizes on the boundary and accepts otherwise, is UMP unbiased. By Remark 8.2.3, there is also a two-sided UMP unbiased test ϕ^* of the form, Reject if $N_{11} > c_2(N_{1+}, N_{++})$ or $N_{11} < c_1(N_{1+}, N_{++})$ with randomization γ_1, γ_2 at both c_1 and c_2 , determined by

$$(i) E_H(\phi^*(\mathbf{N})|N_{1+}, N_{+1}) = \alpha \text{ and}$$

$$(ii) E_H(N_{11}\phi^*(\mathbf{N})|N_{1+}, N_{+1}) = \alpha E_H(N_{11}|N_{1+}, N_{+1}) = \alpha \frac{N_{1+}N_{+1}}{n}.$$

The condition (ii) comes from the requirement that the conditional power function of the test has derivative 0 at $\gamma = 0$ which is necessary for unbiasedness. Unfortunately, this test doesn't coincide with the conditional χ^2 test discussed earlier nor with the naive two-tailed test: Reject iff $N_{11} < c_1^*(N_{1+}, N_{+1})$ or $N_{11} > c_2^*(N_{1+}, N_{+1})$ with randomization on the boundary, all chosen so that, if \mathbf{N}_1^+ denotes (N_{1+}, N_{+1}) ,

$$P_H[N_{11} > c_1^*(\mathbf{N}_1^+)|\mathbf{N}_1^+] + \gamma_1(\mathbf{N}_1^+)P[N_{11} = c_1^*(\mathbf{N}_1^+)|\mathbf{N}_1^+] = \frac{\alpha}{2}$$

with a similar identity for c_2, γ_2 . \square

Finally, consider

Example 8.2.7. (Continued). Permutation t -test. Consider an alternative with $F = \mathcal{N}(\mu, \sigma^2)$, and $G = \mathcal{N}(\mu + \Delta, \sigma^2)$, $\Delta > 0$. Then the test of Theorem 8.2.4 can be written as

$$\text{Reject } H \text{ if } \frac{p(\mathbf{X}, \mathbf{Y}; \boldsymbol{\theta}_1)}{p(\mathbf{X}, \mathbf{Y}; \boldsymbol{\theta}_0)} \mathbb{1}(T(\mathbf{X}, \mathbf{Y}) = T) > c(Z_{(1)}, \dots, Z_{(N)}) \quad (8.2.21)$$

where $\boldsymbol{\theta}_1 = (\mu, \mu + \Delta, \sigma^2)$ and $\boldsymbol{\theta}_0 = (\mu, \mu, \sigma^2)$. This holds, since by sufficiency of $T(\mathbf{X}, \mathbf{Y}) = (Z_{(1)}, \dots, Z_{(N)})^T$ on $w = \Theta_0 = \{\boldsymbol{\theta}_0 : \mu \in R, \sigma^2 > 0\}$ and the general principles of conditioning,

$$p_{\boldsymbol{\theta}_0}[\mathbf{X}, \mathbf{Y}|\mathbf{T}(\mathbf{X}, \mathbf{Y}) = \mathbf{t}] = p(\mathbf{X}, \mathbf{Y}; \boldsymbol{\theta}_0) \mathbb{1}[\mathbf{T}(\mathbf{X}, \mathbf{Y}) = \mathbf{t}] / P_{\boldsymbol{\theta}_0}(\mathbf{T}(\mathbf{X}, \mathbf{Y}) = \mathbf{t})$$

for any $\boldsymbol{\theta}_0 \in \Theta_0$ corresponding to H . But by Section 4.9.3 the ratio in (8.2.21) is a monotone increasing function of $\sum_{i=1}^n Y_i$ given $Z_{(1)}, \dots, Z_{(N)}$. Since $\sum_{i=1}^n Y_i$ and $Z_{(1)}, \dots, Z_{(N)}$ do not involve any parameters, we conclude that the permutation t test is UMP unbiased against the usual $\{\mathcal{N}(\mu, \sigma^2), \mathcal{N}(\mu + \Delta, \sigma^2) : \Delta > 0\}$ alternatives.

Remark 8.2.4. Recently the Lehmann-Scheffé theory of this section has been used to derive most powerful unbiased selective tests for inference in exponential family models after data based model selection. See Fithian, Sun and Taylor (2014).

8.2.3 Estimation

We will now show how completeness provides the most powerful tool that we have for the construction of uniformly minimum variance unbiased (UMVU) estimates. Recall from Section 3.4.1 that given a model $\mathcal{P} \equiv \{P_\theta : \theta \in \Theta\}$, an estimate δ of a parameter $q(\theta) \in R$ is unbiased iff

$$E_\theta \delta(X) = q(\theta) \quad \text{for all } \theta.$$

A UMVU estimate δ^* is an unbiased estimate such that, for all θ and all unbiased δ ,

$$\text{Var}_\theta \delta^*(X) \leq \text{Var}_\theta \delta(X).$$

Given any unbiased estimate $\delta(X)$ and sufficient statistic $T(X)$ we now show how to construct an unbiased Rao-Blackwell (RB) estimate depending on T only which is at least as good.

Theorem 8.2.6. (Rao-Blackwell) Suppose δ is an estimate of $q(\theta)$ and $E_\theta \delta^2(X) < \infty$ for all θ . Let $T(X)$ be a sufficient statistic and denote the Rao-Blackwell estimate of $q(\theta)$ by

$$\delta_{RB}(T(X)) = E\{\delta(X)|T(X)\}. \quad (8.2.22)$$

Then, for all $\theta \in \Theta$,

- a) $\text{Bias}_\theta(\delta_{RB}) = \text{Bias}_\theta(\delta)$.
- b) $\text{MSE}_\theta(\delta_{RB}(T)) \leq \text{MSE}_\theta(\delta(X))$. (8.2.23)

Equality holds in (8.2.23) iff $\delta(X)$ is a function of $T(X)$ (with probability 1).

Note. As in the case of conditional tests, sufficiency enables us to conclude that δ_{RB} does not depend on θ so that it is a bona fide estimate.

Proof. Part a) is a consequence of the iterated expectation theorem,

$$E_\theta \delta_{RB}(T(X)) = E_\theta E(\delta(X)|T(X)) = E_\theta \delta(X). \quad (8.2.24)$$

Further,

$$\begin{aligned} \text{MSE}_\theta(\delta(X)) &= E_\theta(E(\delta_{RB}(X) - q(\theta))^2 | T(X)) \\ &= E_\theta(E(\delta(X)|T(X)) - q(\theta))^2 \\ &= E_\theta(E(\delta(X) - q(\theta)|T(X))^2 \\ &\leq E_\theta E(\delta(X) - q(\theta))^2 | T(X)) = E_\theta(\delta(X) - q(\theta))^2 \end{aligned}$$

by applying the inequality $(EU)^2 \leq EU^2$ conditionally on $T(X)$. By A.11.9, equality can hold iff $\delta(X)$ is constant given $T(X)$. The theorem follows. \square

Note. The theorem is a special case of Problem 3.4.2 which shows that δ_{RB} improves on $\delta(X)$ for convex rather than just quadratic loss functions.

The following Lehmann-Scheffé theorem (1950, 1955) completes the story of complete sufficient statistics and UMVU estimates.

Corollary 8.2.2. (Lehmann–Scheffé) Suppose $T(X)$ is complete as well as sufficient for \mathcal{P} and $\delta(X)$ is unbiased for $q(\theta)$. Then $\delta_{RB}(T)$ is the unique UMVU estimate of $q(\theta)$.

Proof. By part a) of the theorem, $\delta_{RB}(T)$ is unbiased. Suppose $\tilde{\delta}$, unbiased, is better than $\delta_{RB}(T)$, i.e. $E_\theta(\tilde{\delta}(X) - q(\theta))^2 \leq E_\theta(\delta_{RB}(X) - q(\theta))^2$ with strict inequality for some θ . Then $\tilde{\delta}_{RB}(T)$, the Rao-Blackwell version of $\tilde{\delta}$ has

$$E_\theta(\tilde{\delta}_{RB}(T) - q(\theta))^2 \leq E_\theta(\delta_{RB}(T) - q(\theta))^2. \quad (8.2.25)$$

On the other hand,

$$E_\theta \delta_{RB}(T(X)) - E_\theta (\tilde{\delta}_{RB}(T(X))) = E_\theta (\delta_{RB}(T) - \tilde{\delta}_{RB}(T)) = 0$$

for all θ . The result follows because completeness of T implies that

$$\delta_{RB}(T) - \tilde{\delta}_{RB}(T) = 0.$$

□

Constructing UMVU estimates when a complete, sufficient T is available.

Method 1. Given $q(\theta)$, find δ such that $E_\theta \delta(T(X)) = q(\theta)$ for all θ . Then $\delta(T(X))$ is UMVU.

Method 2. Given $q(\theta)$, find $\delta(X)$ such that $E_\theta \delta(X) = q(\theta)$ for all θ . Then $\delta^* \equiv E(\delta(X)|T(X))$ is UMVU.

Example 8.2.8. *Exponential families.* $T(\mathbf{X})$ is UMVU for $E_\theta T(\mathbf{X}) \equiv \dot{A}(\theta)$. Recall two special cases:

- a) \mathbf{N}/n is UMVU for \mathbf{p} if $\mathbf{N} \sim \mathcal{M}(n, \mathbf{p})$. All $\mathbf{a}^T \mathbf{N}$ are UMVU estimates of $\mathbf{a}^T \mathbf{p}$.
- b) $\bar{\mathbf{X}}$ is UMVU for $\boldsymbol{\mu}$ if X_1, \dots, X_n are i.i.d. $\mathcal{N}_p(\boldsymbol{\mu}, \Sigma)$.

This example is also covered by Theorem 3.4.3. □

More examples will be given in the problems. However, consider the closely related

Example 8.2.9. *Estimating the variance covariance matrix Σ when $\mathbf{X}_1, \dots, \mathbf{X}_n$ are i.i.d. $\mathcal{N}_p(\boldsymbol{\mu}, \Sigma)$ with both $\boldsymbol{\mu}$, Σ unknown for $n \geq 2$.* By Theorem 8.2.2, a complete sufficient statistic here is $\mathbf{T}(\mathbf{X}) = (n^{-1} \sum_{i=1}^n \mathbf{X}_i \mathbf{X}^T)$ where $\mathbf{X} = (X_{ij})$, $1 \leq i \leq n$, $1 \leq j \leq d$. We try Method 2. The MLE $\hat{\Sigma} \equiv n^{-1} \sum_{i=1}^n (\mathbf{X}_i - \bar{\mathbf{X}})(\mathbf{X}_i - \bar{\mathbf{X}})^T$ has

$$\begin{aligned} E(\hat{\Sigma}) &= E \left\{ \frac{1}{n} \sum_{i=1}^n (\mathbf{X}_i - \boldsymbol{\mu})(\mathbf{X}_i - \boldsymbol{\mu})^T - (\bar{\mathbf{X}} - \boldsymbol{\mu})(\bar{\mathbf{X}} - \boldsymbol{\mu})^T \right\} \\ &= \Sigma \left(1 - \frac{1}{n} \right) = \frac{n-1}{n}. \end{aligned} \tag{8.2.26}$$

Therefore $[n/(n-1)]\hat{\Sigma}$ is an unbiased estimate of Σ and since it is a function of \mathbf{T} only, it is UMVU. Note that $\hat{\Sigma}$ is *not* a linear function of \mathbf{T} so that the theory of Section 3.4 doesn't apply. □

Remark 8.2.5. We could not establish in Section 3.4 that the unbiased estimate s^2 of σ^2 in the Gaussian linear model is UMVU because $\text{Var}(s^2)$ does not achieve the information lower bound. Now we have shown that s^2 is UMVU.

Here is another example.

Example 8.2.10. *Estimation of p for geometric observations.* For illustration, we apply both methods in this example. Suppose X_1, \dots, X_n are i.i.d. geometric (p)

$$P[X_1 = k] = q^k p, \quad 0 < p < 1, \quad k = 0, 1, \dots.$$

Then, $T \equiv \sum_{i=1}^n X_i$ is complete and sufficient and

$$P[T = k] = \binom{n+k-1}{n-1} q^k p^n, \quad k = 0, 1, \dots.$$

We want an unbiased estimate of p .

Method 1. We are supposed to find an estimate δ such that

$$p = \sum_{k=0}^{\infty} \delta(k) \binom{n+k-1}{n-1} q^k p^n,$$

or, equivalently,

$$1 = \sum_{k=0}^{\infty} \delta(k) \binom{n+k-1}{n-1} q^k p^{n-1}.$$

Then, for $n \geq 2$, by the unicity of power series and the fact that the negative binomial distribution with parameters n, k , and p assigns probability 1 to the nonnegative integers,

$$\delta(k) \binom{n+k-1}{n-1} = \binom{n+k-2}{n-2}.$$

Thus the solution is

$$\delta(k) = \frac{\binom{n+k-2}{n-2}}{\binom{n+k-1}{n-1}} = 1 - \frac{k}{n+k-1} = \frac{n-1}{n+k-1}.$$

Then, $\delta^*(T) = (n-1)/(n+T-1)$, $n \geq 2$, is UMVU.

Note that the MLE of \hat{p} solves the likelihood equation,

$$E_{\hat{p}}(\bar{X}) = \bar{X},$$

or, equivalently,

$$\frac{1}{\hat{p}} - 1 = \bar{X}$$

yielding $\hat{p} = n/(n+T)$, which is close to but not the same as the UMVUE.

Method 2. We begin with a simple unbiased estimate, $\delta(X_1) \equiv 1(X_1 = 0)$. Then,

$$\begin{aligned} E(\delta(X_1)|T = k) &= P[X_1 = j|T = k] \\ &= P\left[X_1 = j, \sum_{i=2}^n X_i = k-j\right] / P\left[\sum_{i=2}^n X_i = k\right] \\ &= \binom{n+k-j-2}{n-2} / \binom{n+k-1}{n-1}. \end{aligned}$$

By Corollary 8.2.2, $\delta^*(k) = P[X_1 = 0 | T = k]$ is the UMVU estimate which we already derived using Method 1. \square

Here is a nonparametric example.

Example 8.2.11. Estimation of F . Suppose X is univariate, F is the df of X which is assumed completely unknown. We assume \mathcal{P} is the family of all distributions with continuous case densities. Here $X_{(1)} < \dots < X_{(n)}$ are complete, sufficient by Example 8.2.3. Since

$$\widehat{F}(x) \equiv \frac{1}{n} \sum_{i=1}^n 1(X_i \leq x) = \frac{1}{n} \sum_{i=1}^n 1(X_{(i)} \leq x)$$

and \widehat{F} is unbiased, we conclude that $\widehat{F}(x)$ is an UMVU estimate of $F(x)$ for all x . More generally, every function $q(\widehat{F})$ of $\widehat{F}(\cdot)$ is a UMVU estimate of its own expectation. \square

Remark 8.2.6. The basis of our treatment and many more examples may be found in Lehmann and Casella (1998) and the problems.

Summary. We introduce the concept of completeness for a class of probability distributions \mathcal{P} and show how it can be used in the construction of tests and estimates with desirable properties. In essence, \mathcal{P} is complete if it admits a sufficient statistic T such that the only function $v(T)$ that is an unbiased estimate of zero for all $P \in \mathcal{P}$ is the function $v(T) = 0$. For this case T is also called *complete*. We call a test $\phi(x)$ *similar* or *distribution-free* with respect to a hypothesized class of probabilities \mathcal{P}_0 if $E_P \phi(X) = \alpha$ for all $P \in \mathcal{P}_0$ and show that if $T(X)$ is a complete, sufficient statistic for \mathcal{P}_0 , then a randomized test ϕ is similar level α iff $E_P(\phi(X)|T(X)) = \alpha$ with P probability one for all $P \in \mathcal{P}_0$. A test ϕ satisfying $E_P(\phi(X)|T(X)) = \alpha$, $P \in \mathcal{P}_0$, for a complete, sufficient statistic T is said to have *Neyman structure*. Important examples of complete, sufficient statistics are the natural sufficient statistic in k -dimensional exponential family models and the vector of order statistics or the empirical distribution. It is shown that when a complete, sufficient statistic exists, it can be used to construct similar tests by conditioning on a statistic which is complete and sufficient for the null hypothesis class of probabilities \mathcal{P}_0 . Important special cases are testing problems in exponential family models, Gaussian models, contingency tables, and the two-sample models where conditioning leads to permutation tests and rank tests. We define $S(\mathbf{X})$ to be an *ancillary statistic* with respect to \mathcal{P} if $\mathcal{L}_P(S(\mathbf{X}))$ does not depend on P for $P \in \mathcal{P}$, and show Basu's theorem stating that if $S(\mathbf{X})$ is ancillary for \mathcal{P} and $T(\mathbf{X})$ is complete, sufficient for \mathcal{P} , then $S(\mathbf{X})$ and $T(\mathbf{X})$ are independent when $\mathbf{X} \sim P \in \mathcal{P}$. For testing $H : \theta \in \Theta_0$ vs $k : \theta \in \Theta_1$, we define a test ϕ to be *unbiased level α* if $E_\theta \phi(\mathbf{X}) \leq \alpha$ for all $\theta \in \Theta_0$ and $E_\theta \phi(\mathbf{X}) \geq \alpha$ for all $\theta \in \Theta_1$; that is, ϕ is no more likely to accept H when K is true than when H is true. We define ϕ to be *UMP unbiased level α* if it is UMP among all unbiased level α tests and show how such tests can be found for some models when a simple statistic which is complete, sufficient with respect to $\{P_\theta : \theta \in \partial\Theta_0 \cap \partial\Theta_1\}$ is available. We give the Rao-Blackwell theorem which establishes that if S is any estimate of $q(\theta)$, then the conditional expected value $E(S|T)$ of S given a complete, sufficient statistic T has mean squared error at least as small as that of S . The Lehmann-Scheffé theorem establishes that if S is unbiased, $E(S|T)$ is UMVU. We give

two methods for finding UMVU estimates and show, for example, that in the multivariate normal model $\bar{\mathbf{X}} = n^{-1} \sum \mathbf{X}_i$ and $(n-1)^{-1} \sum (\mathbf{X}_i - \bar{\mathbf{X}})(\mathbf{X}_i - \bar{\mathbf{X}})^T$ are UMVU for μ and Σ .

8.3 Invariance, Equivariance, and Minimax Procedures

The topics we present in this section are a small subset of the extensive discussion given in Lehmann's classics, Theory of Point Estimation (TPE) and Testing Statistical Hypotheses (TSH), now Lehmann and Casella (1986) and Lehmann and Romano (2005).

8.3.1 Group Models

In Section 8.2 we have seen how to reduce composite hypotheses to simple ones by conditioning on a sufficient statistic. Basu's theorem further enabled us to identify situations where ancillary statistics could be used to turn conditional tests into unconditional ones. In this section we will begin by identifying an important structural property which is typical of such situations and which has further important implications.

We recall the definition of a group⁽¹⁾ \mathcal{G} of transformations g of a set \mathcal{L} . Recall that all $g \in \mathcal{G}$ map \mathcal{L} onto itself in a $1 - 1$ fashion. Group multiplication is composition, and \mathcal{G} is closed under composition and inversion. That is,

- i) If $g_1, g_2 \in \mathcal{G}$ so does $g_1 \circ g_2$ given by $g_1 \circ g_2(x) \equiv g_1(g_2(x))$.
- ii) If $g \in \mathcal{G}$ so does g^{-1} defined by $g \circ g^{-1} = g^{-1} \circ g = j$ where $j(x) = x$, all $x \in \mathcal{L}$.

Here is an important example of such a group.

The affine group on \mathbf{R}^d

Here $\mathcal{L} = \mathbf{R}^d$ and \mathcal{G} can be parametrized by $\theta \equiv (A, \mathbf{b})$ where $A_{d \times d}$ ranges over all nonsingular matrices and \mathbf{b} over \mathbf{R}^d . Then, $g_{(A, \mathbf{b})}(\mathbf{x}) \equiv A\mathbf{x} + \mathbf{b}$. It is easy to see that \mathcal{G} is a group and that the group \mathcal{G} is isomorphic to the group $\overline{\mathcal{G}}$ on $\mathcal{A} \times \mathbf{R}^d$ where \mathcal{A} is the set of all nonsingular matrices via the correspondence

$$\begin{aligned} g_{(A_1, \mathbf{b}_1)} \circ g_{(A_2, \mathbf{b}_2)} &= g_{(A_1 A_2, A_1 \mathbf{b}_2 + \mathbf{b}_1)} \\ g_{(A, \mathbf{b})}^{-1} &= g_{(A^{-1}, -A^{-1}\mathbf{b})}. \end{aligned}$$

As we know \mathcal{G} is not commutative. Most of the groups we consider will be subgroups, some commutative, of the affine group.

Groups on the sample space \mathcal{X} induce models. The simplest situation is when we start with a single P_0 and define

$$\mathcal{P} = \{P_0 g^{-1} : g \in \mathcal{G}\} \tag{8.3.1}$$

or, equivalently, if $X \sim P_0$, $g(X) \sim P_0 g^{-1} \in \mathcal{P}$. Note that \mathcal{G} parametrizes \mathcal{P} . If the parametrization is identifiable, $P_0 g_0^{-1} = P_0 g_1^{-1} \Rightarrow g_0 = g_1$, then \mathcal{G} induces a group $\overline{\mathcal{G}}$

on \mathcal{P} defined by $\bar{g}(P) \equiv Pg^{-1} \in \mathcal{P}$. More generally, if $\theta \rightarrow P_\theta, \theta \in \Theta$ is an identifiable parametrization of \mathcal{P} , then \mathcal{G} induces a group $\bar{\mathcal{G}}$ on Θ via

$$P_{\bar{\theta}} = P_\theta g^{-1}. \quad (8.3.2)$$

It is easy to establish (Problem 8.3.1) that (8.3.2) defines \bar{g} , a transformation of Θ uniquely, and that $\bar{\mathcal{G}} = \{\bar{g} \leftrightarrow g \in \mathcal{G}\}$ is a group isomorphic to \mathcal{G} . That is, the map $g \rightarrow \bar{g}$ has the properties $\bar{g}_1 \circ \bar{g}_2 = \bar{g}_1 \circ \bar{g}_2, g^{-1} = \bar{g}^{-1}$. We shall call models such as (8.3.1) *transformation models* induced by P_0, \mathcal{G} . Here is a basic example.

Example 8.3.1. *The multivariate Gaussian model.* Take $P_0 = \mathcal{N}_d(\mathbf{0}, J)$, where J is the $d \times d$ identity matrix. That is, if $\varepsilon \equiv (\varepsilon_1, \dots, \varepsilon_d) \sim P_0$, the $\varepsilon_1, \dots, \varepsilon_d$ are i.i.d. $\mathcal{N}(0, 1)$. The transformation model induced by P_0 and the affine group is just $\mathcal{P} = \{\mathcal{N}_d(\mu, \Sigma); \mu \in R^d, \Sigma \text{ nonsingular}\}$ (Problem 8.3.2). Note that the parametrization $P_{A, b}$ corresponding to $g_{(A, b)}(\varepsilon) = A\varepsilon + b$ is not identifiable—see Example 9.1.5. \square

More generally, we define a *group model* \mathcal{P} by the property that if $P \in \mathcal{P}$, $X \sim P$, and $g \in \mathcal{G}$, then $P_{g^{-1}} \in \mathcal{P}$. Here is a simple example.

Example 8.3.2. *The shift model.* Fix a density f_0 on R^d and let \mathcal{G}_0 be the group of all translations on R^d , $g_c(x) = x + c$, and \mathcal{P}_0 the corresponding transformation model. \mathcal{G}_0 is clearly a subgroup of the affine group and if f_0 is Gaussian $\mathcal{P}_0 \subset \mathcal{P}$ of Example 8.3.1. Note that the parametrization is identifiable. The model can be written structurally,

$$\mathbf{X} = \mathbf{c} + \varepsilon \quad (8.3.3)$$

where $\varepsilon \sim f_0$. For Example 8.3.1, $\mathbf{X} = A\varepsilon + b$, where A is nonsingular and $\varepsilon \sim \mathcal{N}(\mathbf{0}, J)$.

Now consider the set of all probability distributions corresponding to (8.3.3) when we take f_0 to be arbitrary. This remains a group model under the shift group but is not a transformation model since we cannot transform one density shape, say Gaussian, into another, say Cauchy, by simply shifting. Equivalently the parametrization $(\mathbf{c}, f) \rightarrow f(\cdot - \mathbf{c})$ is not identifiable. \square

The major group we consider in addition to the affine group, and some of its subgroups, is the group of increasing transformations:

Example 8.3.3. *The group \mathcal{G} of all continuous strictly increasing functions from R onto R .* Let F_0 be the cdf of P_0 where $F_0 \in \mathcal{G}$. Then the model \mathcal{P} generated by P_0 and \mathcal{G} is the set of all P on R with continuous strictly increasing df. If we consider the subgroup of g which have positive derivatives and P_0 has a density which is positive we generate the submodel of all P having positive densities (Problem 8.3.3). \square

Remarks 8.3.1.

- (a) Any given point x_0 generates an *orbit* $O_x \equiv \{y : y = gx \text{ for some } g \in \mathcal{G}\}$. Belonging to the same orbit is an equivalence relation on \mathcal{X} (Problem 8.3.4).
- (b) The groups we have defined so far allow only for samples of size 1. However, if a group \mathcal{G} is defined on \mathcal{X} and we have a sample (X_1, \dots, X_n) taking values in \mathcal{X}^n then we

naturally define

$$g^{(n)}(x_1, \dots, x_n)^T = (g(x_1), \dots, g(x_n)) \quad (8.3.4)$$

the tensor group with the natural inherited group operations and properties.

(c) For $n = 1$ all the groups we have considered are *transitive*, that is, there is only one orbit, \mathcal{X} itself. That is, given $x_0, x_1 \in \mathcal{X}$ there exists $g \in \mathcal{G}$ such that $gx_0 = x_1$. For $n > 1$, following the definition in (8.3.4) this is no longer the case. For instance, for the shift group with $d = 1$, orbits are indexed by R^{n-1} . Given $(x_1^0, \dots, x_{n-1}^0)^T$ the orbit is the hyperplane, $(x_1^0, \dots, x_{n-1}^0, 0)^T + c(1, \dots, 1)^T$ (Problem 8.3.5). We re-encounter such quantities subsequently.

8.3.2 Group Models and Decision Theory

The basic connection between group models and inference is made via a group \mathcal{G}^* on the action space \mathcal{A} and the loss function relation,

$$l(Pg^{-1}, g^*a) = l(P, a) \quad (8.3.5)$$

for all $P \in \mathcal{P}, g \in \mathcal{G}, g^* \in \mathcal{G}^*, a \in \mathcal{A}$. In other words, the cost of taking action a when you observe $X \sim P$ is the same as that of taking action g^*a when you observe $gX \sim Pg^{-1}$. Consistency with (8.3.5) for a decision rule δ is defined by

$$\delta(gx) = g^*\delta(x) \quad (8.3.6)$$

for all $x \in \mathcal{X}$. Such rules are called *equivariant*. In the case of tests they are called *invariant* for reasons that will become apparent. In other words again, equivariance means that if action a is “right” when you see x then g^*a is “right” if you see gx . Requiring decision procedures to have this property has the same nature as requiring unbiasedness. This time rules in the class obey certain symmetries. What \mathcal{G}^* appear?

Here are examples showing what happens in estimation and testing.

Example 8.3.4. *Testing in the Gaussian two-sample problem.* Suppose that as in Section 4.9.3, we observe a sample X_1, \dots, X_{n_1} of i.i.d. $\mathcal{N}(\mu, \sigma^2)$ observations and an independent sample Y_1, \dots, Y_{n_2} of $\mathcal{N}(\mu + \Delta, \sigma^2)$ observations. We wish to test $H : \Delta = 0$, and postulate the usual 0 – 1 loss function. The model \mathcal{P} can be viewed as a group model with respect to the shift group applied to $(X_1, \dots, X_{n_1}, \dots, Y_1, \dots, Y_{n_2})$. If

$$X_i \rightarrow X_i + c, 1 \leq i \leq n_1, Y_j \rightarrow Y_j + c, 1 \leq j \leq n_2$$

under g_c , and we parametrize P by $\theta \equiv (\mu, \Delta, \sigma^2)$ in the standard way, then $\bar{g}_c\theta = (\mu + c, \Delta + c, \sigma^2)$. Note that $\Theta_0 \equiv \{(\mu, \Delta, \sigma^2) : \Delta = 0\}$ is an orbit of $\bar{\mathcal{G}}$ as is $\Theta_1 = \{(\mu, \Delta, \sigma^2) : \Delta \neq 0\}$. \square

Following Lehmann and Romano (2005), we shall in general call a testing problem $H : \theta \in \Theta_0$ vs $K : \theta \in \Theta_1$ *invariant* if \mathcal{P} is a group model and Θ_0 and Θ_1 are disjoint

unions of $\overline{\mathcal{G}}$ orbits. Consider now the usual 0 – 1 loss. If Θ_0 is an orbit, if $\theta_0 \in \Theta_0$, $l(\overline{g}\theta, 0) = 1 - l(\overline{g}\theta, 1)$ for all \overline{g} and similarly for Θ_1 . It is then natural to take $\mathcal{G}^* = \{\text{Identity}\}$, the trivial group, and see that (8.3.6) defines an *invariant test*

$$\delta(gx) = \delta(x) \quad (8.3.7)$$

for all g, x . We take (8.3.7) as our general definition of an invariant test.

Example 8.3.4 (continued). In this example invariance means

$$\delta(x_1 + c, \dots, x_{n_1} + c, y_1 + c, \dots, y_{n_2} + c) = \delta(x_1, \dots, x_{n_1}, y_1, \dots, y_{n_2}) \quad (8.3.8)$$

for all $\mathbf{x}, \mathbf{y}, c$. For $n_1 = n_2 = 1$, this model is invariant under the location-scale group,

$$g_{(a,b)}(x_1, y_1) = (ax_1 + b, ay_1 + b) \quad a \neq 0, b \in R.$$

Now

$$\overline{g}_{(a,b)}(\mu, \Delta, \sigma^2) = (a\mu + b, a\mu + b + a\Delta, a^2\sigma^2)$$

and Θ_0 and Θ_1 are again orbits. This invariance is preserved for general n_1, n_2 if g is extended to $R^{n_1} \times R^{n_2}$ as in (8.3.4). Now test invariance becomes more stringent,

$$\delta(ax_1 + b, \dots, ax_{n_1} + b, ay_1 + b, \dots, ay_{n_2} + b) = \delta(\mathbf{x}, \mathbf{y}) \quad (8.3.9)$$

for all $a \neq 0, b$ □

Example 8.3.2. *The shift model (continued).* Suppose we observe X_1, \dots, X_n i.i.d. P from the shift model with fixed f_0 . Parametrize the shift model by $\theta \in R$, $\mathcal{P} = \{f_\theta \equiv f_0(\cdot - \theta) : \theta \in R\}$. Suppose we want to estimate θ with loss $l(\theta, a) \equiv \lambda(|\theta - a|)$, $\lambda : R^+ \rightarrow R^+$. Then,

$$l(\overline{g}_c\theta, a) = \lambda(|\theta + c - a|) = \lambda(|\theta - (a - c)|).$$

So, for equivariances, we must define $g_c^*(a) = a + c$ and \mathcal{G}^* is the shift group on R . Thus a *translation equivariant* estimate δ satisfies

$$\delta(x_1 + c, \dots, x_n + c) = \delta(x_1, \dots, x_n) + c, \quad \text{all } x_1, \dots, x_n, c. \quad (8.3.10)$$

□

Example 8.3.5. *The Gaussian one-sample problem.* Suppose we observe X_1, \dots, X_n i.i.d. $\mathcal{N}(\mu, \sigma^2)$. As in Example 8.3.4 this model is a group model for $\mathcal{G} = \{g_{ab} : x_i \rightarrow ax_i + b, 1 \leq i \leq n, a \neq 0, b \in R\}$. If $\theta = (\mu, \sigma^2)$, then

$$\overline{g}_{(a,b)}(\mu, \sigma^2) = (a\mu + b, a^2b^2),$$

a group operating on $R \times R^+$. Suppose we want to estimate μ with the equivariant Kullback-Liebler loss (Vol. I, page 169), which in this case is equivalent to scaled quadratic loss, $l(\theta, d) = (\mu - d)^2/\sigma^2$. Then, as in the previous example,

$$l(\overline{g}_{(a,b)}\theta, d') = \frac{(a\mu + b - d')^2}{a^2\sigma^2} = \frac{(\mu - \frac{d'-b}{a})^2}{\sigma^2}.$$

Thus we must define $g_{(a,b)}^*(d) = ad + b$ and \mathcal{G}^* is the affine group on R . Correspondingly, an estimate δ of μ is said to be *location-scale equivariant* iff

$$\delta(ax_1 + b, \dots, ax_n + b) = a\delta(x_1, \dots, x_n) + b, \text{ all } a \neq 0, b \in R. \quad (8.3.11)$$

8.3.3 Characterizing Invariant Tests

Invariant functions

Rather than restrict ourselves to invariant test functions we more generally define an *invariant function* $\psi : \mathcal{X} \rightarrow \mathcal{T}$, where \mathcal{T} is arbitrary, as one such that

$$\psi(gx) = \psi(x), \text{ all } g \in \mathcal{G}, x \in \mathcal{X}. \quad (8.3.12)$$

Invariant functions can equivalently be characterized as functions which are constant on orbits of \mathcal{G} . This leads naturally to the definition of a *maximal invariant* (function) $M : \mathcal{X} \rightarrow \mathcal{L}$, where \mathcal{L} is general, as a function which is invariant, but further, $M(x_1) \neq M(x_2)$ if x_1, x_2 are not in the same orbit. Thus M labels orbits uniquely. Clearly every invariant ψ is a function of M ; there exists $\lambda : \mathcal{L} \rightarrow \mathcal{T}$ such that

$$\psi(x) = \lambda(M(x)). \quad (8.3.13)$$

From its definition we see that M is far from unique. In fact, for those familiar with measure theory the appropriate object to consider is the invariant σ field

$$\mathcal{I} = \{A \text{ measurable } \subset \mathcal{X} : gA = A \text{ for all } g \in \mathcal{G}\}.$$

M simply induces \mathcal{I} and (8.3.13) can be interpreted as: "Every invariant function is measurable with respect to \mathcal{I} ."

Suppose that testing $H : \theta \in \Theta_0$ vs $K : \theta \in \Theta_1$ is invariant under \mathcal{G} as discussed in the previous section. We define $\varphi^*(x)$ (possibly randomized) as the (*uniformly*) *most powerful* (UMP) invariant level α , test of H vs K_1 if φ^* is invariant, has level α and is at least as powerful as any other level α invariant test (defined by φ invariant). Our discussion makes it plain that a UMP invariant (UMPI) test is simply the UMP test based on observing $M(X)$ only.

Here are three examples of maximal invariants.

Example 8.3.6. *The shift group.* We consider the group applied to $[R^d]^n$,

$$g_{\mathbf{c}} : (\mathbf{x}_1, \dots, \mathbf{x}_n) \rightarrow (\mathbf{x}_1 + \mathbf{c}, \dots, \mathbf{x}_n + \mathbf{c}).$$

If, for all \mathbf{c} , $\psi(\mathbf{x}_1 + \mathbf{c}, \dots, \mathbf{x}_n + \mathbf{c}) = \psi(\mathbf{x}_1, \dots, \mathbf{x}_n)$, let $\mathbf{c}_1 = -\mathbf{x}_1$. Then,

$$\psi(\mathbf{x}_1, \dots, \mathbf{x}_n) = \psi(\mathbf{0}, \mathbf{x}_2 - \mathbf{x}_1, \dots, \mathbf{x}_n - \mathbf{x}_1). \quad (8.3.14)$$

But the right hand side of (8.3.14) is clearly invariant with respect to shift. Evidently $M(\mathbf{x}_1, \dots, \mathbf{x}_n) \equiv (\mathbf{x}_2 - \mathbf{x}_1, \dots, \mathbf{x}_n - \mathbf{x}_1)$ is a maximal invariant since $(\mathbf{x}_2^0 - \mathbf{x}_1^0, \dots, \mathbf{x}_n^0 - \mathbf{x}_1^0)$

$\mathbf{x}_1^0) = (\mathbf{x}_2^1 - \mathbf{x}_1^1, \dots, \mathbf{x}_n^1 - \mathbf{x}_1^1)$ implies that $(\mathbf{x}_1^0, \dots, \mathbf{x}_n^0) = (\mathbf{x}_1^1 + \mathbf{c}, \dots, \mathbf{x}_n^1 + \mathbf{c})$ where $\mathbf{c} = \mathbf{x}_1^0 - \mathbf{x}_1^1$, which is just the condition required by maximality. Note that we could just as well have taken $M(\mathbf{x}_1, \dots, \mathbf{x}_n) = (\mathbf{x}_1 - \bar{\mathbf{x}}, \dots, \mathbf{x}_n - \bar{\mathbf{x}})$ where $\bar{\mathbf{x}}$ is the mean of the \mathbf{x}_i . Note also that $(\mathbf{x}_1, M(\mathbf{x}_1, \dots, \mathbf{x}_n))$ is an alternative representation of $(\mathbf{x}_1, \dots, \mathbf{x}_n)$. Thus, the value of $M(\mathbf{x}_1, \dots, \mathbf{x}_n)$ tells us which orbit we're on and \mathbf{x}_1 parametrizes points in the orbit, or equivalently \mathcal{G} . More generally, if \mathcal{G} is parametrized smoothly by $\theta \rightarrow g_\theta$ where $\theta \in \Theta$, an l dimensional manifold, and the dimension of \mathcal{X} is p , we expect $M(\mathbf{x})$ to range over a $p - l$ dimensional manifold. If $l \geq p$, then \mathcal{G} is transitive. \square

Example 8.3.7. *The orthogonal group.* Let $\mathcal{X} = R^d$ and \mathcal{G} be the group of orthogonal $d \times d$ matrices; the set of all $A_{d \times d}$ such that $AA^T = A^TA = J$, the identity. As is well known orthogonal transformations preserve Euclidean distances between points, $|A\mathbf{x} - A\mathbf{y}|^2 = |A(\mathbf{x} - \mathbf{y})|^2 = (\mathbf{x} - \mathbf{y})^T A^T A (\mathbf{x} - \mathbf{y}) = |\mathbf{x} - \mathbf{y}|^2$. This shows that $M(\mathbf{x}) \equiv |\mathbf{x}|^2$ is invariant. On the other hand $|\mathbf{x}|^2 = |\mathbf{y}|^2$ implies that there exists A orthogonal such that $A\mathbf{x} = \mathbf{y}$ (Problem 8.3.6). Thus, $|\mathbf{x}|^2$ is a maximal invariant. \square

Example 8.3.3 (continued). *The group of monotone transformations.* Suppose g is a continuous strictly monotone increasing function from R onto R and $x_{(1)} < \dots < x_{(n)}$. Evidently, $g(x_{(1)}) < \dots < g(x_{(n)})$. Thus if $\mathcal{X} = \{(\mathbf{x}_1, \dots, \mathbf{x}_n) : x_1 \neq x_2 \dots \neq x_n\}$ the ranks $R_j(x_1, \dots, x_n) = \sum_{i=1}^n 1(x_i \leq x_j)$, $1 \leq j \leq n$, are invariant. On the other hand, suppose $(R_1(\mathbf{x}), \dots, R_n(\mathbf{x})) = (R_1(\mathbf{y}), \dots, R_n(\mathbf{y}))$, $\mathbf{x}, \mathbf{y} \in \mathcal{X}$. Then, there exists $g \in \mathcal{G}$ such that $g(x_{(i)}) = y_{(i)}$, $1 \leq i \leq n$, where $x_{(1)}, \dots, x_{(n)}, y_{(1)}, \dots, y_{(n)}$ are the ordered \mathbf{x} 's and \mathbf{y} 's. Hence (R_1, \dots, R_n) is a maximal invariant (Problem 8.3.7). Note in this case a natural representation of (x_1, \dots, x_n) is $(R_1(\mathbf{x}), \dots, R_n(\mathbf{x}), x_{(1)}, \dots, x_{(n)})$. \square

We now construct UMPI tests.

Example 8.3.8. *Testing one shift family against another.* Let $\mathbf{X} \in R^d$ be distributed according to $P_{j,c}$, $j = 0, 1, \mathbf{c} \in R^d$ where $P_{j,c}$ has continuous case density $f_j(\cdot - \mathbf{c})$, $j = 0, 1$, and f_0, f_1 are of different types, that is, f_1 does not belong to the location family generated by f_0 . We want to test $H : P \in P_{0,c}$ for some \mathbf{c} vs $K : P \in P_{1,c}$ for some \mathbf{c} . The problem is clearly invariant under the shift group \mathcal{G} . By Example 8.3.6, we need only consider tests based on $M(\mathbf{X}) = (X_2 - X_1, \dots, X_d - X_1)$. Note that the distribution of $M(\mathbf{X})$ does not depend on \mathbf{c} and $M(\mathbf{X})$ has density on R^{d-1} given by

$$g_j(\mathbf{m}) = \int_{-\infty}^{\infty} f_j(u, m_1 + u, \dots, m_{d-1} + u) du. \quad (8.3.15)$$

Thus a UMPI level α test exists and is of the form

$$\text{Reject } H \text{ if } g_1(M(\mathbf{X})) > cg_0(M(\mathbf{X})), \text{ Accept } H \text{ if } g_1(M(\mathbf{X})) < cg_0(M(\mathbf{X}))$$

with c determined by g_0 and α . \square

This example suggests the result:

Proposition 8.3.1. The distribution of any maximal invariant $M(X)$ of a group \mathcal{G} leaving a model $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$ invariant depends on θ only through $\overline{M}(\theta)$, i.i.d. as (X, Y)

the maximal invariant of the group $\overline{\mathcal{G}}$ induced on Θ by \mathcal{G} .

Proof: See Problem 8.3.8 or Lehmann and Romano (2005).

Example 8.3.9. *The Gaussian linear model with known variance.* Suppose that as in Section 6.1, $\mathbf{Y}_{n \times 1} = \boldsymbol{\mu}_{n \times 1} + \boldsymbol{\varepsilon}$ where $\boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \sigma_0^2 J)$ and $\boldsymbol{\mu} \in V$, a linear subspace of R^n of dimension d . We want to test $\boldsymbol{\mu} = \mathbf{0}$ vs $\boldsymbol{\mu} \neq \mathbf{0}$. This testing problem is left invariant by the group \mathcal{G} of orthogonal matrices operating on R^n , since $A\boldsymbol{\mu} = \mathbf{0}$ iff $\boldsymbol{\mu} = \mathbf{0}$ and $A\boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \sigma_0^2 A J A^T) = \mathcal{N}(\mathbf{0}, J)$ if A is orthogonal. By Example 8.3.6 $M \equiv |\mathbf{X}|^2 = \sum_{j=1}^d X_i^2$ is maximal invariant. Under H , M has a $\sigma_0^2 \chi_d^2$ distribution. More generally, $M \sim \sigma_0^2 \chi_d^2(|\boldsymbol{\mu}|^2 / \sigma_0^2)$ where $\chi_d^2(\theta^2)$ is a noncentral χ^2 distribution with noncentrality parameter θ^2 . It may be shown (Problem 8.3.9) that the $\chi_d^2(\theta^2)$ family is Monotone Likelihood Ratio (MLR) in θ^2 and M . Therefore the test,

$$\text{Reject } H \text{ iff } \frac{M}{\sigma_0^2} \geq \chi_d(1 - \alpha)$$

where $\chi_d(1 - \alpha)$ is the $(1 - \alpha)$ quantile of χ_d^2 , is UMP invariant by Theorem 4.3.1. Extension of this result to the case σ^2 unknown and linear hypotheses are given in Lehmann and Romano (2005). \square

Example 8.3.10. *The nonparametric two-sample problem.* Let X_1, \dots, X_{n_1} be i.i.d. F , X_{n_1+1}, \dots, X_n be i.i.d G where F, G are continuous but otherwise unknown. All the X 's are independent. The testing problem, $H : F = G$ vs $K : F \neq G$, is invariant under the transformation group of Example 8.3.3, applied to $\mathbf{X} = (X_1, \dots, X_n)^T$. Thus $\mathbf{R} \equiv (R_1, \dots, R_n)^T$ is a maximal invariant. Under H , we have already seen that $P[(R_1, \dots, R_n) = (i_1, \dots, i_n)] = 1/n!$ for all permutations $\{i_1, \dots, i_n\}$ of $\{1, \dots, n\}$. Under the alternative, however, although there is a formula for the distribution of \mathbf{R} , there is no UMP invariant test (Problem 8.3.10). However, as was first noted by Lehmann (1953) there are semiparametric group submodels $\mathcal{P} \equiv \{P_{\theta, F} : G = q(F, \theta), \theta \in R\}$ for which (R_1, \dots, R_n) has a distribution depending on θ only (Problem 8.3.11). For instance, Savage (1956) considers

$$G = 1 - (1 - F)^\theta, \tag{8.3.16}$$

$\theta \geq 1$, a model which includes $F(t) = 1 - e^{-\lambda t}$, $G(t) = 1 - e^{-\lambda \theta t}$, the exponential lifetime two-sample model. The Lehmann-Savage⁽²⁾ model is a special case of the Cox proportional hazard model, Example 9.1.4, with λ being the hazard rate of F and $\lambda\theta$ that of G . See Problems 1.1.12 and 1.1.13. By arguing as in Example 8.3.8, for the model $G = 1 - (1 - F)^\theta$, there is a UMP invariant (wrt F) test for $H : \theta = 1$ vs $K : \theta > 1$, which in this case can be computed (Problem 8.3.10(c)). Unfortunately, it depends on θ_1 and we need a new concept: A test ψ of $H : \theta = \theta_0$ vs $K : \theta > \theta_0$ is *locally MP* if there exists $\varepsilon > 0$ such that ψ is MP for all $\theta \in (\theta_0, \theta_0 + \varepsilon)$. A formula for computing such tests is given in Problems 8.3.10-8.3.12. For model (8.3.16) there is a locally most powerful (see Problems 8.3.12 and 8.3.13) UMP invariant test called the *Savage exponential scores test*.

Let $a_E(j) = E(X_{(j)})$ where $X_{(1)} < \dots < X_{(n)}$ are the exponential, $\mathcal{E}(1)$, order statistics (see Problem 8.3.13). Then the Savage exponential scores test is given by

$$\text{Reject } H \text{ iff } \sum_{i=n_1+1}^n a_E(R_i) < c$$

with c determined under $F = G$ by the level α and the uniform distribution on permutations. Hoeffding (1951) considered the optimal rank test for the Gaussian two-sample problem. Note that all rank tests are permutation tests and similar and they have the attractive feature that when $H : F = G$ holds their distributions do not depend on $F = G$. They are distribution-free.

Example 8.3.11. *Invariant regression tests. Copula models.* As in Section 6.1 consider the framework where the i th measurement (response) Y_i among n independent observations has a continuous distribution F_i and we want to investigate whether F_i depends on a vector $\mathbf{z}_i = (z_{i1}, \dots, z_{id})^T$ of available constants (predictor values), $1 \leq i \leq n$, $d < n$, where $\mathbf{z}_{i1}, \dots, \mathbf{z}_{in}$ are not collinear. We consider the testing problem $H : F_i = F$, $1 \leq i \leq n$, vs $K : F_i \neq F_k$, some i, k , which is invariant under the group of continuous increasing transformations of Example 8.3.3 applied to $\mathbf{Y} = (Y_1, \dots, Y_n)^T$. Thus the rank vector $\mathbf{R} = (R_1, \dots, R_n)^T$ of \mathbf{Y} is maximal invariant. Consider semiparametric group submodels of the form

$$F_i(\cdot) = C_i(F(\cdot)), \quad 1 \leq i \leq n,$$

where C_i , which is called a *copula*, is a continuous df on $(0, 1)$ which is known except for a parameter θ_i that depends on \mathbf{z}_i ; and F is an unknown continuous baseline df.

The term “*copula*,” introduced by Sklar (1959), is a measure of the dependence between variables X_1, \dots, X_p . It is defined as

$$C(u_1, \dots, u_p) = P(G_1(X_1) \leq u_1, \dots, G_p(X_p) \leq u_p), \quad p \geq 2,$$

where the marginal df’s G_1, \dots, G_p are assumed to be continuous. Here we extend this definition to the $p = 1$ case: the *regression copula* for $Y_i \sim F_i$ w.r.t. the hypothesis $H : F_i = F$, $1 \leq i \leq n$, for continuous F is

$$C_i(u) = P(F(Y_i) \leq u), = F_i F^{-1}(u), \quad 1 \leq i \leq n.$$

The copula $C_i(u)$ is a measure of the dependence of the distribution F_i on \mathbf{z}_i . The term “*baseline distribution*” df refers to the hypothesis distribution F that does not depend on \mathbf{z}_i .

Set $V_i = F(Y_i)$, then $R_i = \text{Rank}(V_i)$ because the ranks are invariant under increasing transformations. The df of V_i is

$$C_i(F(F^{-1}(v))) = C_i(v), \quad 0 \leq v \leq 1,$$

and the distribution of \mathbf{R} only involves $C_i(\cdot)$. Useful choices are the *Gaussian copula* $C_i(v) = \Phi(\Phi^{-1}(v) - \theta_i)$ and $\theta_i = \beta^T \mathbf{z}_i$, in which case $\Phi^{-1}(V_i) = \Phi^{-1}(F(Y_i)) \sim$

$\mathcal{N}(\theta_i, 1)$ and we may assume $F_i = \mathcal{N}(\theta_i, 1)$ when we deal with rank tests. For this copula, when $d = 1$ and $\theta_i = \alpha + \beta z_i$, the locally UMP invariant (w.r.t. α, F) test of $H_0 : \beta = 0$ vs $H_1 : \beta > 0$ is based on the *normal scores* statistic

$$T_\Phi = n^{-\frac{1}{2}} \sum_{i=1}^n a_\Phi(R_i)(z_i - \bar{z}_i),$$

where $a_\Phi(k) = E(z^{(k)})$ with $Z^{(1)} < \dots < Z^{(n)}$ the $\mathcal{N}(0, 1)$ order statistics (Problem 8.3.14(d)). The normal scores $a_\Phi(k)$, which are also called the Fisher-Yates scores, can be closely approximated by $\Phi^{-1}\{(k - 3/8)/(n + 1/4)\}$.

The locally UMP invariant test for the *logistic copula* $C_i(u) = L(L^{-1}(u) - \theta_i)$ with $L(t) = \{1 + \exp(-t)\}^{-1}$ and $\theta_i = \alpha + \beta z_i$ is based on the *uniform scores* statistic

$$T_U \equiv n^{-\frac{1}{2}} \sum_{i=1}^n \frac{R_i}{n+1}(z_i - \bar{z}).$$

Critical values for the tests based on T_Φ and T_U can be obtained from the distribution of the ranks or their asymptotic distributions: T_Φ/s and $\sqrt{12}T_U/s$ are asymptotically $\mathcal{N}(0, 1)$ under H with

$$s^2 = \sum_{i=1}^n (z_i - \bar{z})^2 / (n-1)$$

provided (z_i) , $1 \leq i \leq n$, satisfy the Lindeberg condition

$$\frac{\max_i (z_i - \bar{z})^2}{\sum_{i=1}^n (z_i - \bar{z})^2} \rightarrow 0.$$

See Section 9.5 and Hajek and Sidak (1967), Section V.1.5. □

The parameter β in the model $\theta_i = \beta^T \mathbf{z}_i$ in the previous example is only identifiable if $d \leq n$ and $\{(z_{1j}, \dots, z_{nj})^T, 1 \leq j \leq d\}$ are not collinear; see Problem 1.1.9. However, modern technology has made it possible to generate data with d very large, so called “*big data*.” In particular, $d > n$. We turn to this case:

Example 8.3.11 (continued). *Invariant high dimensional data analysis.* One approach to the $d > n$ case is to replace $\beta^T \mathbf{z}_i$ in the preceding copula model with $\mathbf{a}^T \mathbf{z}_i = \sum_{j=1}^d a_j z_{ij}$, where the a_j are selected to maximize the spread of $\{\mathbf{a}^T \mathbf{z}_i : 1 \leq i \leq n\}$ subject to $\mathbf{a}^T \mathbf{a} = 1$. This makes sense if we regard $\mathbf{a}^T \mathbf{z}_i$ as predictors because regression analysis is most effective when the predictors are spread out. Thus we choose \mathbf{a} to maximize the sample variance of $\{\mathbf{a}^T \mathbf{z}_i : 1 \leq i \leq n\}$. That is, we use (see Section B.10.12 and Problem 8.3.24) the first sample *eigenvector*

$$\hat{\mathbf{a}} = \arg \max \{\mathbf{a}^T \hat{\Sigma} \mathbf{a} : \mathbf{a}^T \mathbf{a} = 1\}$$

where $\hat{\Sigma}$ is the sample covariance matrix of $(z_{ij})_{n \times d}$. Let

$$t_i = \hat{\mathbf{a}}^T \mathbf{z}_i, \quad 1 \leq i \leq n.$$

Here t_1, \dots, t_n , which are called the *first principal component* sample values, can be computed using available *principal component analysis (PCA)* software, e.g. “eigensoft.” Now we use the semiparametric copula regression submodel $F_i(\cdot) = C_i(F(\cdot))$ with C_i having parameter $\eta_i = \alpha + \beta t_i$. This is an example of a *sparse* model. Several principal components (Problem 8.3.24) can also be used. Typically, ten or fewer are used. The locally UMP invariant PCA tests of $H_0 : \beta = 0$ vs $H_1 : \beta > 0$ for the Gaussian and logistic regression copulas are the normal scores and uniform scores tests, respectively, with z_i replaced by t_i .

Remark 8.3.2. $\widehat{\Sigma}$ and the eigenvector $\widehat{\mathbf{a}}$ are known to be inconsistent estimates of their population counterparts when $d > n$, unless the model for the data is restricted to a submodel. One approach is to assume that the predictor is a random vector $\mathbf{Z} \in R^d$ whose covariance matrix Σ is restricted; see Bickel and Levina (2008 a, b). A similar approach is to restrict the ordered population eigenvalues $\lambda_1 \geq \lambda_2 \geq \dots$ (as defined in Section B.10.12 and Problem 8.3.24) of Σ . Such approaches are referred to as *sparse PCA*, e.g. see Patterson et al. (2006), Paul (2007), El Karoui (2008), Amini and Wainwright (2009), Jung and Marron (2009), Johnstone and Lu (2009), Witten, Tibshirani and Hastie (2009), Shen, Shen and Marron (2011), Birnbaum, Johnstone, Nadler and Paul (2013), and Hastie, Tibshirani, and Wainwright (2015). These references focus on the estimation of the covariance matrix of \mathbf{Z} and its eigenvalues and eigenvectors. For papers that focus on the association between individual predictors Z_k and a response Y , see Price et al. (2006), Lin and Zheng (2011), and Yang, Doksum and Tsui (2014). \square

Example 8.3.11 (continued). *Invariant tests in transformation models.* Consider the transformation regression model with

$$h(Y_i) = \theta_i + \varepsilon_i, \quad 1 \leq i \leq n,$$

where $\theta_i = \boldsymbol{\beta}^T \mathbf{z}_i$, ε_i has a known continuous distribution G , and h is an increasing continuous unknown function. In this case $h^{-1}(Y_i) \sim \theta_i + \varepsilon_i$. Thus if $G = \Phi$ and we use rank tests, we are in the realm of the Gaussian copula model and can use the normal scores rank tests described earlier in this example. The *proportional hazard* and *proportional odds* models are transformation regression models. See Problem 8.3.14.

Example 8.3.12. *Invariant tests of independence. Bivariate copulas.* Consider the problem of testing the independence of two random variables X and Y with continuous df $F(x, y)$. The testing problem is invariant under increasing continuous transformations $h(X)$, $g(Y)$, and for a sample $(X_1, Y_1), \dots, (X_n, Y_n)$ i.i.d. as (X, Y) the maximal invariant is $((R_1, S_1), \dots, (R_n, S_n))$ where $R_i = \sum_{k=1}^n 1(X_k \leq X_i)$ and $S_i = \sum_{k=1}^n 1(Y_k \leq Y_i)$ are the ranks. Sklar (1959) introduced models of the form

$$F_\theta(x, y) = C_\theta(F_1(x), F_2(y))$$

for some continuous df (copula) C_θ on $[0, 1] \times [0, 1]$ which is known except for the parameter θ , where F_1 and F_2 are the marginal df's of X and Y .

A useful choice is the bivariate Gaussian copula

$$\begin{aligned} C_\rho(u, v) &= \Phi_\rho(\Phi_1^{-1}(u), \Phi_2^{-1}(v)), \quad \Phi_1 = \mathcal{N}(\mu_1, \sigma_1^2), \\ \Phi_2 &= \mathcal{N}(\mu_2, \sigma_2^2), \quad \Phi_\rho = \mathcal{N}(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho). \end{aligned}$$

The locally UMP invariant test of $H_0 : \rho = 0$ vs $H_1 : \rho > 0$ is based on the bivariate normal scores statistic $\sum_{i=1}^n a_\Phi(R_i)a_\Phi(S_i)$ where $a_\Phi(k)$, $1 \leq k \leq n$, are the normal scores (see Problem 8.3.16). The null distribution is obtained by rearranging the pairs $(R_1, S_1), \dots, (R_n, S_n)$ so that $R_1 < R_2 < \dots < R_n$ and noting that the resulting Y -ranks \mathbf{S}^* has $P(\mathbf{S}^* = \mathbf{s}) = 1/n!$ for each permutation \mathbf{s} of $(1, \dots, n)$. Klaassen and Wellner (1997) show that in the Gaussian copula model the *normal scores correlation coefficient*

$$\widehat{\rho}_\Phi = \frac{\sum_{i=1}^n a_\Phi(R_i)a_\Phi(S_i)}{\sum_{i=1}^n a_\Phi^2(i)}$$

is an asymptotically semiparametrically efficient estimate of

$$\rho_\Phi = \text{Corr}(\Phi^{-1}(F_1(X)), \Phi^{-1}(F_2(Y)))$$

for each $\rho \in (-1, 1)$ uniformly in F_1, F_2 over the class of all regular estimates. It can be shown that for the bivariate normal copula model, $|\rho_\Phi|$ is the maximum monotone correlation coefficient (Problem 8.3.16) in the sense that

$$|\rho_\Phi| = \sup_{a,b} \text{corr}(a(X), b(Y))$$

for a and b monotone. Approximate critical value for $H : \rho = 0$ can be obtained from $n^{-\frac{1}{2}}\widehat{\rho}_\Phi \sim \mathcal{N}(0, 1)$.

For more on rank tests see the problems, Hajek and Sidak (1967), Lehmann (2006), and Lehmann and Romano (2005).

8.3.4 Characterizing Equivariant Estimates

Characterizing equivariant estimates is relatively simple in the case of the shift group. Suppose as in Example 8.3.2 we observe $\mathbf{X}_1, \dots, \mathbf{X}_n \in \mathbb{R}^d$ i.i.d. $f(\cdot - \boldsymbol{\theta})$ for f and $\boldsymbol{\theta}$ unknown. We want to estimate $\boldsymbol{\theta}$ and our loss function is of the form $\lambda(|\boldsymbol{\theta} - \mathbf{a}|)$, with λ increasing. As we argued for $d = 1$ an appropriately equivariant estimate obeys

$$\delta(\mathbf{x}_1 + \mathbf{c}, \dots, \mathbf{x}_n + \mathbf{c}) = \mathbf{c} + \delta(\mathbf{x}_1, \dots, \mathbf{x}_n). \quad (8.3.17)$$

Examples of equivariant estimates are $\delta^* = \overline{\mathbf{X}}$, or $(\text{med}_i X_{1i}, \dots, \text{med}_i X_{di})^T$, or even \mathbf{X}_1 . Note that if δ, δ^* are equivariant then $\delta - \delta^*$ is invariant by (8.3.17). Thus, if $M(\mathbf{x})$ is a maximal invariant under \mathcal{G} , any equivariant estimate can be written for some ψ as

$$\delta(\mathbf{x}_1, \dots, \mathbf{x}_n) = \delta^*(\mathbf{x}_1, \dots, \mathbf{x}_n) + \psi(M(\mathbf{x}_1, \dots, \mathbf{x}_n)). \quad (8.3.18)$$

Taking $\delta^*(\mathbf{x}_1, \dots, \mathbf{x}_n) = \mathbf{x}_1$ gives us a particularly simple expression in (8.3.18). From this formula it is easy to deduce that there is a uniformly best equivariant estimate and derive its form, at least for $\lambda(t) = t^2$ and similar loss functions.

Theorem 8.3.1. *If there exists an equivariant estimate for the shift group with finite risk, then there is a unique (Uniform Minimum Risk Equivariant) UMRE estimate, given by*

$$\delta^*(\mathbf{x}_1, \dots, \mathbf{x}_n) = \mathbf{x}_1 - \psi^*(M(\mathbf{x})) \quad (8.3.19)$$

where M is a maximal invariant and $\psi^*(M) = \arg \min_c E_0(\lambda(\mathbf{X}_1 + \mathbf{c})|M)$.

Proof. Write, keeping $f = f_0$ fixed,

$$R(\boldsymbol{\theta}, \boldsymbol{\delta}) = R(\mathbf{0}, \boldsymbol{\delta}) = E_0 \lambda(\mathbf{X}_1 + \psi(M(\mathbf{X}_1, \dots, \mathbf{X}_n))) . \quad (8.3.20)$$

The reason for the first identity is that $\mathbf{X}_1 - \boldsymbol{\theta}$ and $\psi(M(\mathbf{X}_1, \dots, \mathbf{X}_n))$ have a distribution not depending on $\boldsymbol{\theta}$. But

$$E_0 \lambda(\mathbf{X}_1 + \psi(M)) = E_0 \{E_0(\lambda(\mathbf{X}_1 + \psi(M))|M)\} .$$

Since ψ is arbitrary the result follows. \square

If we specialize to $\lambda(t) = t^2$ we readily obtain, by Theorem 1.4.1 (Problem 8.3.17),

$$\delta^*(\mathbf{X}_1, \dots, \mathbf{X}_n) = \mathbf{X}_1 - E_0(\mathbf{X}_1|M) . \quad (8.3.21)$$

If \mathbf{X}_1 has density f_0 we can obtain a very suggestive form,

$$\delta^*(\mathbf{X}_1, \dots, \mathbf{X}_n) = \frac{\int_{R^d} \boldsymbol{\theta} \prod_{i=1}^n f_0(\mathbf{X}_i - \boldsymbol{\theta}) d\boldsymbol{\theta}}{\int_{R^d} \prod_{i=1}^n f_0(\mathbf{X}_i - \boldsymbol{\theta}) d\boldsymbol{\theta}} . \quad (8.3.22)$$

We derive (8.3.22). Let $M(\mathbf{X}_1, \dots, \mathbf{X}_n) = (\mathbf{X}_2 - \mathbf{X}_1, \dots, \mathbf{X}_n - \mathbf{X}_1)$. Then, the conditional density of $\mathbf{X}_1 | M$ under $\boldsymbol{\theta} = \mathbf{0}$ is

$$f(\mathbf{u}|M) = \frac{f_0(\mathbf{u}) \prod_{i=1}^{n-1} f_0(\mathbf{m}_i + \mathbf{u})}{\int_{R^d} f_0(\mathbf{v}) \prod_{i=1}^{n-1} f_0(\mathbf{m}_i + \mathbf{v}) d\mathbf{v}} .$$

So,

$$E_0(\mathbf{X}_1|M) = \frac{\int_{R^d} \mathbf{u} f_0(\mathbf{u}) \prod_{i=1}^{n-1} f_0(\mathbf{m}_i + \mathbf{u}) d\mathbf{u}}{\int_{R^d} f_0(\mathbf{v}) \prod_{i=1}^{n-1} f_0(\mathbf{m}_i + \mathbf{v}) d\mathbf{v}}$$

and

$$\mathbf{X}_1 - E_0(\mathbf{X}_1|M) = \frac{\int_{R^d} (\mathbf{X}_1 - \mathbf{u}) f_0(\mathbf{u}) \prod_{i=1}^{n-1} f_0(\mathbf{m}_i + \mathbf{u}) d\mathbf{u}}{\int_{R^d} f_0(\mathbf{v}) \prod_{i=1}^{n-1} f_0(\mathbf{m}_i + \mathbf{v}) d\mathbf{v}} .$$

We obtain (8.3.22) by changing variables, from \mathbf{v} to $\boldsymbol{\theta} = \mathbf{X}_1 - \mathbf{v}$.

The UMRE estimate (8.3.22) known as *Pitman's estimate* is the (improper) Bayes estimate of $\boldsymbol{\theta}$ when $\boldsymbol{\theta}$ has prior "density," $\pi(\boldsymbol{\theta}) \equiv \text{constant}$, the "uniform distribution" on R^d . In particular,

Example 8.3.13. *Estimating the mean of a multivariate Gaussian distribution.* Let $\mathbf{X} \sim \mathcal{N}(\boldsymbol{\mu}, \sigma_0^2 J_P)$ where J_P is the identity matrix. By sufficiency, because $\bar{\mathbf{X}}$ is also Gaussian, we can consider $n = 1$ and suppose $\mathbf{X} = \boldsymbol{\theta} + \mathbf{e}$ where $\mathbf{e} \sim \mathcal{N}_p(\mathbf{0}, J_p)$. Then, we obtain readily that \mathbf{X} is the UMRE estimate of $\boldsymbol{\theta}$, just as we have seen it is UMVU. We will consider this and other estimates of $\boldsymbol{\mu}$ in Section 8.3.6. \square

8.3.5 Minimality for Tests: Application to Group Models

Recall Theorems 3.3.2 and 3.3.3 where we saw that minimax procedures δ^* were Bayes or approximately Bayes with respect to prior distributions which concentrated on the set of parameters that produce the maxima of the risk of δ^* . We have already noted that invariant and equivariant procedures have large sets of constant risk, the orbits of $\overline{\mathcal{G}}$ in Θ . Thus, best equivariant procedures are natural candidates for minimality.

Testing:

We begin with a general study of minimality for testing. Here, the asymmetry of the two types of error leads to an asymmetric formulation of minimality not covered in Chapter 3.

Definition 8.3.1. φ^* is *minimax level α* for testing $H : \theta \in \Theta_0$ vs $K : \theta \in \Theta_1$ iff φ^* is level α and

$$\inf\{E_\theta \varphi^*(x) : \theta \in \Theta_1\} \geq \inf\{E_\theta \varphi(x) : \theta \in \Theta_1\}$$

for all other level α tests φ .

Thus the minimax test maximizes the minimum power, where for 0-1 loss, power = 1 - risk. Such ψ^* is sometimes called a *maxmin* test.

Suppose we are given a prior distribution π for θ on $\Theta_0 \cup \Theta_1 \equiv \Theta$ such that

- i) $\Pi(\Theta_0) = \lambda = 1 - \Pi(\Theta_1)$, $0 < \lambda < 1$
- ii) θ given $\theta \in \Theta_j$ has distribution Π_j , $j = 0, 1$

and a loss function l_u given for $0 < u < 1$ by

$$l_u(\theta, 1) = u1(\theta \in \Theta_0), \quad l_u(\theta, 0) = 1(\theta \in \Theta_1).$$

From a simple extension of Example 3.2.2 (Problem 8.3.18) we see that Bayes tests are of the form, for $0 < \lambda < 1$,

$$\begin{aligned} \varphi_c(x) &= 1 \quad \text{if} \quad \frac{\int_{\Theta_1} p(x, \theta) \Pi_1(d\theta)}{\int_{\Theta_0} p(x, \theta) \Pi_0(d\theta)} > c \\ &= 0 \quad \text{if the inequality is reversed} \end{aligned} \tag{8.3.23}$$

where $c = \lambda u / (1 - \lambda)$. The argument in Example 3.2.2 makes it clear that (8.3.23) doesn't depend of the finiteness of Θ . We can apply Theorem 3.3.2 to obtain

Theorem 8.3.2. Suppose φ_c given by (8.3.23) is size α , and that

$$\begin{aligned} \Pi_0\{\theta : E_\theta \varphi_c(X) = \sup_{\Theta_0} E_{\theta'} \varphi_c(X)\} &= 1 \\ \Pi_1\{\theta : E_\theta(1 - \varphi_c(X)) = \sup_{\Theta_1} E_{\theta'}(1 - \varphi_c(X))\} &= 1. \end{aligned} \tag{8.3.24}$$

Then, φ_c is minimax level α .

Proof. Let $\beta = \sup_{\Theta_1} E_{\theta'}(1 - \varphi_c(X))$ and

$$u = \frac{\alpha}{\beta}, \quad \frac{\lambda}{1 - \lambda} = \frac{c}{u}. \quad (8.3.25)$$

Note that φ_c is Bayes for Π , l_u and that for $R(\theta, \varphi_c) = E_\theta[l_u(\theta, \varphi_c(X))]$,

$$\sup_{\Theta_0} R(\theta, \varphi_c) = \sup_{\Theta_1} R(\theta, \varphi_c) = \sup_{\Theta} R(\theta, \varphi_c). \quad (8.3.26)$$

Now, φ_c Bayes, (8.3.26), and Theorem 3.3.2 yield that φ_c is minimax for l_u and that the Bayes risk equals the maximum risk. Then, given a competing φ of level α , by hypothesis,

$$\begin{aligned} \lambda\alpha + (1 - \lambda)u\beta &= \lambda \int E_\theta \varphi_c(X) \Pi_0(d\theta) + (1 - \lambda)u \int E_\theta(1 - \varphi_c(X)) \Pi_1(d\theta) \\ &\leq \lambda \int E_\theta \varphi(X) \Pi_0(d\theta) + (1 - \lambda)u \int E_\theta(1 - \varphi(X)) \Pi_1(d\theta) \\ &\leq \lambda\alpha + (1 - \lambda)u \sup_{\Theta_1} E_\theta(1 - \varphi(X)) \end{aligned}$$

and the result $\beta \leq \sup_{\Theta_1} E_\theta(1 - \varphi(X))$ follows.

Corollary 8.3.1. If $\Theta_1 = \{\theta_1\}$, φ_c is given by (8.3.23) and the first condition of (8.3.24) holds; then φ_c is UMP level α for $H : \theta \in \Theta_0$ vs $K : \theta = \theta_1$.

It may be shown that the conditions of Theorem 8.3.2 and Corollary 8.3.1 are also necessary for minimaxity and UMP properties of test rules under regularity conditions on the model \mathcal{P} and compactness of Θ_0, Θ_1 (Blackwell and Girshick (1954, 1979)).

Example 8.3.14. Minimality in testing in the Gaussian linear model with σ^2 known. $H : \mu = \mathbf{0}$ vs $K : |\mu| = r_0$. Consider the test of Example 8.3.9, which we have shown to be UMP invariant under the orthogonal group. Note that the power function of this test depends on $|\mu|/\sigma_0$ only and hence (8.3.24) holds automatically. We need to exhibit Π_0 and Π_1 . Without loss of generality take $\sigma_0 = 1$. Evidently Π_0 is point mass at $\mathbf{0}$. For Π_1 it is natural to choose the uniform distribution on the sphere surface $\{\mu : |\mu| = r_0\}$. Then

$$L(\mathbf{X}) \equiv \int_{\Theta_1} \frac{p(\mathbf{X}, \mu)}{p(\mathbf{X}, \mathbf{0})} d\Pi_1(\mu) = e^{\frac{1}{2}|\mathbf{X}|^2} E(e^{-\frac{1}{2}|\mathbf{X} - r_0 \mathbf{U}|^2} | \mathbf{X}) \quad (8.3.27)$$

where \mathbf{X}, \mathbf{U} are independent, $\mathbf{X} \sim \mathcal{N}(\mathbf{0}, J)$, and \mathbf{U} is uniform on the surface of the unit sphere. Now (8.3.27) becomes

$$L(\mathbf{X}) = e^{-\frac{r_0^2}{2}} E(e^{r_0(\mathbf{X}, \mathbf{U})} | \mathbf{X}) = e^{-\frac{r_0^2}{2}} E(e^{r_0(|\mathbf{X}|(\mathbf{X}/|\mathbf{X}|, \mathbf{U}))} | \mathbf{X}).$$

Since the distribution of \mathbf{U} is invariant under rotations we can replace $\mathbf{X}/|\mathbf{X}|$ by, say, $(1, 0, \dots, 0)^T$ without changing the result. Thus $L(\mathbf{X})$ is a monotone increasing function

of $|\mathbf{X}|$. We conclude that the UMP invariant test is of the form (8.3.23) and hence we can apply Theorem 8.3.2 to conclude minimaxity. \square

Example 8.3.15. A UMP test for a composite multivariate Gaussian hypothesis. Let $\mathbf{X}_1, \dots, \mathbf{X}_n$ be i.i.d. $\mathcal{N}_p((\vartheta, \boldsymbol{\eta}^T)^T, \Sigma_0)$. The scalar ϑ and $\beta \equiv (\eta_2, \dots, \eta_p)^T$ are unknown; Σ_0 is nonsingular but known. We wish to test $H : \vartheta = \vartheta_0$ vs $K : \vartheta = \vartheta_1 > \vartheta_0$. Note that both hypothesis and alternative are composite and multivariate so that the UMP test theory of Chapter 4 doesn't apply. First reduce by sufficiency to the case $n = 1$ by considering $\bar{\mathbf{X}}$. Consider the case $\Sigma_0 = J$. Define Π_0, Π_1 as corresponding to point masses at $(\vartheta_0, \boldsymbol{\eta}_0)^T$ and $(\vartheta_1, \boldsymbol{\eta}_0)^T$. Then, if $\mathbf{X} \equiv (X_1, \dots, X_p)^T$,

$$\begin{aligned} L(\mathbf{X}) &\equiv \frac{\int_{\Theta_1} p(\mathbf{X}, \boldsymbol{\theta}) d\Pi_1(\boldsymbol{\theta}) d\Pi_1(\boldsymbol{\theta})}{\int_{\Theta_0} p(\mathbf{X}, \boldsymbol{\theta}) d\Pi_0(\boldsymbol{\theta})} \\ &= \exp -\frac{1}{2} \{ (\mathbf{X} - \boldsymbol{\theta}_1)^T (\mathbf{X} - \boldsymbol{\theta}_1) - (\mathbf{X} - \boldsymbol{\theta}_0)^T (\mathbf{X} - \boldsymbol{\theta}_0) \} \end{aligned}$$

where $\boldsymbol{\theta}_j = (\vartheta_j, \eta_2, \dots, \eta_p)^T, j = 0, 1$. Then,

$$\log L(\mathbf{X}) = (\vartheta_1 - \vartheta_0) X_1 - \frac{1}{2} (|\boldsymbol{\theta}_1|^2 - |\boldsymbol{\theta}_0|^2) \quad (8.3.28)$$

and thus the family of Bayes tests φ_c for the loss function $l_u(\theta, \varphi)$ is equivalent in this case to the family

$$\{\text{Reject } H \text{ iff } X_1 > d\}.$$

But under H , $X_1 \sim \mathcal{N}(\vartheta_0, 1)$. The conditions of Corollary 8.3.1 are clearly satisfied since the tests have power depending on ϑ only. It follows (Problem 8.3.16(a)) that the size α test for this hypothesis is, in fact, UMP for $H : \vartheta \leq \vartheta_0$ vs $K : \vartheta > \vartheta_0$. Now consider the case of general Σ_0 . Make the well-specified linear transformation,

$$\mathbf{X} \rightarrow \mathbf{U} = (X_1, \mathbf{X}_{[2,p]}^T - \text{Cov}(X_1, \mathbf{X}_{[2,p]})(\text{Var}X_1)^{-1} X_1)^T, \quad (8.3.29)$$

where $\mathbf{X}_{[2,p]} \equiv (X_2, \dots, X_p)^T$. Note that $\Sigma_R \equiv \text{Cov}(X_1, \mathbf{X}_{[2,p]}) = (\sigma_{12}, \dots, \sigma_{1p})$ and $\text{Var}X_1 = \sigma_{11}$ so that the second term has dimension $p - 1$. Write $\mathbf{U} \equiv (U_1, \dots, U_p)^T \equiv A\mathbf{X}$ and

$$\Sigma_0 = \begin{pmatrix} \sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}.$$

It follows that $\mathbf{U} \sim \mathcal{N}_p(\mathbf{U}, A\Sigma_0 A^T)$ where $\mathbf{U} \equiv A\vartheta$. Since $\mu_1 = \vartheta_1$ we still test $H : \vartheta = \vartheta_0$ vs $K : \vartheta = \vartheta_1$. Most significantly, $A\Sigma_0 A^T$ is of the form

$$\begin{pmatrix} \sigma_{11} & \mathbf{0}^T \\ \mathbf{0} & \Sigma_{22}^* \end{pmatrix}$$

where $\Sigma_{22}^* = \text{Var}(U_2, \dots, U_p)^T$. It is easy to see (Problem 8.3.16(b)) that in this case, just as with the special case $\Sigma_0 = J = \text{diag}(1, \dots, 1)$, the Bayes test with the same prior as before is of the form $\{\text{Reject } H \text{ iff } X_1 > c\}$. Therefore this test is UMP for $H : \vartheta \leq \vartheta_0$

vs $K : \vartheta > \vartheta_0$ no matter what Σ_0 is. This is the likelihood ratio test and is both UMP invariant and unbiased (Problem 8.3.19(c),(d)).

Remark 8.3.3.

- (a) It turns out, see Lehmann and Romano (2005), that some of the classical likelihood ratio tests for $H : \sigma \leq \sigma_0$, etc. when we observe X_1, \dots, X_n i.i.d. $\mathcal{N}(\mu, \sigma^2)$ are UMP and not just UMP unbiased, as we saw in Section 8.2, but most are not.
- (b) The theory of minimax testing and, as we shall see, estimation in group models is very closely tied to the existence of (invariant) Haar measures on groups—see Nachbin (1965) for a treatment. If \mathcal{G} can be smoothly indexed by a Euclidean parameter $\theta \in K$, a compact set in R^d , these can be taken as probability distributions, P_L, P_R on \mathcal{G} . They have the property that if A (measurable) $\subset \mathcal{G}$ then

$$\begin{aligned} P_L(gA) &= P_L(A) \\ P_R(Ag) &= P_R(A) \end{aligned}$$

for all A and $g \in \mathcal{G}$. Here $gA = \{h = gb, b \in A\}$ and similarly for Ag . The pair P_L, P_R are unique, although $P_L \neq P_R$ is possible if \mathcal{G} is not commutative. If $\mathcal{G} = \{g_\theta : \theta \in \Theta\}$ and $\overline{\mathcal{G}}$ is the group induced on Θ we can produce a prior distribution on any orbit of \mathcal{G} by considering $\overline{g}\theta_0$ for θ_0 fixed and $\overline{g} \sim P_L$ or P_R . It turns out that P_R is the Haar measure important in our context. In our Example 8.3.9, where $\overline{\mathcal{G}}$ is the orthogonal group, indeed $P_L = P_R$ is the uniform distribution on \mathcal{G} and not surprisingly this generates the uniform prior on a sphere (an orbit of \mathcal{G}). If \mathcal{G} is not compact, improper Haar measures arise. We discuss this further in the next subsection.

A very complete treatment of the theory of UMP and minimax tests is given by Lehmann and Romano (2005).

8.3.6 Minimax Estimation, Admissibility, and Steinian Shrinkage

We have discussed minimax estimation briefly in Chapter 3. In this section we want to go further and foreshadow some phenomena discussed in Chapters 10 and 11. We begin by recalling Section I.6 in simplified form.

Suppose X_1, \dots, X_p are independent Gaussian with common known variance σ_0^2 and $EX_i \equiv \mu_i$ unknown and arbitrary. Equivalently, $\mathbf{X} \sim \mathcal{N}_p(\boldsymbol{\mu}, \sigma_0^2 J)$ where $J = \text{diag}(1, \dots, 1)$. We want to estimate $\boldsymbol{\mu}$ with quadratic loss. That is, if $\mathbf{d} = (d_1, \dots, d_p)^T$, $l(\boldsymbol{\mu}, \mathbf{d}) = |\boldsymbol{\mu} - \mathbf{d}|^2 = \sum_{i=1}^p (\mu_i - d_i)^2$. This is a special case of Example 8.3.2 with $f_0 \leftrightarrow \mathcal{N}_p(\mathbf{0}, \sigma_0^2 J)$. It is easy to see that Pitman's estimator here is \mathbf{X} itself since $M(\mathbf{X})$ is degenerate ($n = 1$) and so $E_0(\mathbf{X}|M(\mathbf{X})) = E_0\mathbf{X} = \mathbf{0}$. We know that \mathbf{X} is UMVU and that, by Problem 3.3.20, \mathbf{X} is minimax. If we have a multivariate normal sample, $\bar{\mathbf{X}}$ is Gaussian, so the remarks about \mathbf{X} also apply to $\bar{\mathbf{X}}$.

Remark 8.3.4. As with testing, minimaxity and equivariance here can be tied to Haar's measure on the shift group. X_1, \dots, X_p are the formal Bayes estimates with respect to

the “uniform distribution” on R^p . Since $\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} dt_1 \dots dt_p = \infty$ the uniform is an improper prior. It corresponds to a Haar measure which is not a probability distribution. If $\lambda(A) \equiv \int_A dt_1 \dots dt_p$ then $\lambda(A + c) = \lambda(c + A) = \lambda(A)$. Corresponding Haar measures on scale and other groups generate priors, usually improper, which play an important role in Bayesian inference, being viewed as “uninformative.” Various arguments can be raised for their use including frequentist efficiency. There is now little controversy in the literature that these methods are as legitimate as maximum likelihood. There is more controversy about the nonsubjective interpretation of posterior probabilities calculated under such assumptions. Arguments in favor of the subjective point of view may be found in Bernardo and Smith (1994), Berger(1985), and Gelman, Carlin, Stern and Rubin (1995). \square

We introduce a decision theoretic notion now only rarely considered which played an important role in the discovery of the practical weaknesses of the minimax principle. A decision rule δ^* in a given decision problem is said to be *admissible* iff there exists no other δ such that $R(P, \delta) \leq R(P, \delta^*)$ for all $P \in \mathcal{P}$ with $<$ for some P . Otherwise δ^* is said to be inadmissible. By refining the methods of Section 3.3 it is possible to obtain usable criteria for proving admissibility (Blyth (1951)). This is also possible using the information inequality (Hodges and Lehmann (1951)). This work and relations to complete class theorems such as those mentioned in Section 1.3 are discussed in some detail in Lehmann and Casella (1998)—see also Schervish (1995).

We do not proceed further into the depths of this topic but make the trivial point that if a procedure has constant risk and is minimax but inadmissible then any improving procedure is necessarily also minmax. Stein’s remarkable (1956(b)) discovery was that although for the risk function $E(|\mathbf{X} - \boldsymbol{\mu}|^2)$, \mathbf{X} is minimax for all p , it is admissible only for $p = 1, 2$. In a sense this can be viewed as an indication that Robbins’ (1964) empirical Bayes view that, for p large, \mathbf{X} is improvable in fact already holds for $p > 2$.

The basic tool in this analysis which will prove its value more broadly is an identity due to Stein (1981) giving what has become known as Stein’s unbiased risk estimate.

Theorem 8.3.3. Let $\mathbf{X} \sim \mathcal{N}(\boldsymbol{\mu}, J)$ and $\boldsymbol{\delta} : R^p \rightarrow R^p$ be an estimate of $\boldsymbol{\mu}$. Define

$$\Delta(\mathbf{x}) = \boldsymbol{\delta}(\mathbf{x}) - \mathbf{x}. \quad (8.3.30)$$

Suppose that, for $1 \leq j \leq p$, the j th component of Δ is piecewise continuously differentiable with respect to x_j and

- (i) $E_{\boldsymbol{\mu}}|\Delta(\mathbf{x})|^2 = \sum_{j=1}^p E_{\boldsymbol{\mu}}\Delta_j^2(\mathbf{X}) < \infty$
and
- (ii) $\sum_{j=1}^p E \left[\left| \frac{d\Delta_j}{dx_i}(\mathbf{X}) \right| \right] |X_j - \mu_j| < \infty$

Then, the quadratic loss risk of $\boldsymbol{\delta}$ is given by

$$E_{\boldsymbol{\mu}}|\boldsymbol{\delta}(\mathbf{X}) - \boldsymbol{\mu}|^2 = p + E_{\boldsymbol{\mu}}|\Delta(\mathbf{X})|^2 + 2E_{\boldsymbol{\mu}} \sum_{j=1}^p \frac{\partial \Delta_j}{\partial x_j}(\mathbf{X}). \quad (8.3.31)$$

Proof. The proof relies on a lemma important not only for this result but, more generally, as the foundation of a large body of results in probability referred to as the Stein-Chen method (Stein (1981)).

Lemma 8.3.1. Let $X \sim \mathcal{N}(\mu, \sigma^2)$. Let $g : R \rightarrow R$ be a continuous, piecewise differentiable function such that

- (i) $E|g'(X)| < \infty$,
- (ii) $E|X - \mu|g(X) < \infty$.

Then,

$$Eg'(X) = E \frac{(X - \mu)}{\sigma^2} g(X). \quad (8.3.32)$$

Proof. Use integration by parts to get

$$\begin{aligned} \frac{1}{\sigma} \int_{-A}^A g'(x) \varphi\left(\frac{x - \mu}{\sigma}\right) dx &= g(x) \frac{1}{\sigma} \varphi\left(\frac{x - \mu}{\sigma}\right) \Big|_{-A}^A - \frac{1}{\sigma} \int_{-A}^A g(x) \varphi'\left(\frac{x - \mu}{\sigma}\right) dx \\ &= \frac{1}{\sigma} g(x) \varphi\left(\frac{x - \mu}{\sigma}\right) \Big|_{-A}^A + \int_{-A}^A g(x) \frac{(x - \mu)}{\sigma^2} \varphi'\left(\frac{x - \mu}{\sigma}\right) dx \end{aligned}$$

and the result follows since $g(x)\varphi\left(\frac{x-\mu}{\sigma}\right) \Big|_{-A}^A \rightarrow 0$ as $A \rightarrow \infty$. \square

The reason for the utility of this elementary lemma is that it holds for arbitrary g . A simple generalization provides a unique characterization of $\mathcal{N}(\mu, r^2)$. Condition (ii) may be dispensed with (Stein (1981)).

Proof of Theorem 8.3.3. By assumption,

$$\begin{aligned} &E_{\boldsymbol{\mu}} |\delta(\mathbf{X}) - \boldsymbol{\mu}|^2 \\ &= E_{\boldsymbol{\mu}} |\mathbf{X} - \boldsymbol{\mu}|^2 + 2E_{\boldsymbol{\mu}} \boldsymbol{\Delta}^T(\mathbf{X})(\mathbf{X} - \boldsymbol{\mu}) + E_{\boldsymbol{\mu}} |\boldsymbol{\Delta}(\mathbf{X})|^2 \\ &= p + 2 \sum_{j=1}^p \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \Delta_j(\mathbf{x})(x_j - \mu_j) \prod_{i=1}^p \varphi(x_i - \mu_i) dx_i + E_{\boldsymbol{\mu}} |\boldsymbol{\Delta}(\mathbf{X})|^2. \end{aligned}$$

Write

$$\begin{aligned} &\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \Delta_j(\mathbf{x})(x_j - \mu_j) \prod_{i=1}^p \varphi(x_i - \mu_i) dx_i \\ &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \prod_{i \neq j} \varphi(x_i - \mu_i) dx_i \int_{-\infty}^{\infty} \Delta_j(\mathbf{x})(x_j - \mu_j) \varphi(x_j - \mu_j) dx_j. \end{aligned}$$

Apply Stein's identity (8.3.32) to $\Delta_j(\mathbf{x})$ viewed as a function of x_j with $\{x_i : i \neq j\}$ fixed. The theorem follows. \square

This identity does indeed provide an unbiased estimate of the risk (mean squared error) of any estimate δ satisfying the conditions of Theorem 8.3.4. It can be reinterpreted as saying that

$$\hat{R} \equiv p + |\Delta(\mathbf{X})|^2 + \sum_{j=1}^p \frac{\partial \Delta_j}{\partial x_j}(\mathbf{X}) \quad (8.3.33)$$

is an unbiased estimate of $R(\mu, \delta) \equiv E_\mu |\delta(\mathbf{X}) - \mu|^2$. The application of (8.3.33) to studying inadmissibility of \mathbf{X} with respect to a competitor $\delta(\mathbf{X})$ is given by

Corollary 8.3.2. *Suppose δ satisfies the conditions of Theorem 8.3.4, and Δ defined by (8.3.30) is such that*

$$2 \sum_{j=1}^p \frac{\partial \Delta_j}{\partial x_j}(\mathbf{x}) + |\Delta|^2(\mathbf{x}) \leq 0, \quad (8.3.34)$$

with strict inequality on a set of positive probability. Then for all μ ,

$$R(\mu, \delta(\mathbf{X})) < R(\mu, \mathbf{X}) = p. \quad (8.3.35)$$

Note that still,

$$\sup_{\mu} R(\mu, \delta(\mathbf{X})) = p \quad (8.3.36)$$

Stein's original (1956(b)) proposal of an estimate satisfying (8.3.34) for $p \geq 3$ was

$$\delta_s(\mathbf{X}) = \left(1 - \frac{p-2}{|\mathbf{X}|^2}\right) \mathbf{X} \quad (8.3.37)$$

where (8.3.35) was established.

The nature of δ_s is interesting. It always lies along the vector \mathbf{X} . For large values of $|\mathbf{X}|$ it essentially coincides with \mathbf{X} but for $|\mathbf{X}| > p-2$ it is shrunk more and more towards $\mathbf{0}$ and, for $0 < |\mathbf{X}| < p-2$, it points in the direction opposite to that of \mathbf{X} . The last property is patently unreasonable and, indeed, Stein (1981) showed that the Stein positive part estimate,

$$\delta_s^+(\mathbf{X}) = \left(1 - \frac{p-2}{|\mathbf{X}|^2}\right)_+ \mathbf{X}, \quad (8.3.38)$$

strictly improves $\delta_s(\mathbf{X})$ and hence \mathbf{X} itself. Here $x_+ \equiv \max(x, 0)$. $\delta_s^+(\mathbf{X})$ has a very natural interpretation. The standard test of $\mu = 0$, Reject iff $|\mathbf{X}|^2 > p-2$ is carried out. If the test accepts, estimate μ by $\mathbf{0}$, else use $[1 - (p-2)/|\mathbf{X}|^2]\mathbf{X}$.

Next we establish Stein's result for δ_s using his identity. Here,

$$\begin{aligned}\Delta_s(\mathbf{X}) &= -\frac{(p-2)}{|\mathbf{X}|^2} \mathbf{X} \\ \frac{\partial \Delta_{js}}{\partial x_j} &= -(p-2) \left(\frac{1}{|\mathbf{X}|^2} - 2 \frac{x_j^2}{|\mathbf{X}|^4} \right)\end{aligned}$$

Thus, if $\Delta_s \equiv (\Delta_{1s}, \dots, \Delta_{ps})^T$

$$\begin{aligned}|\Delta_s|^2 + 2 \sum_{j=1}^p \frac{\partial \Delta_{js}}{\partial x_j}(\mathbf{x}) &= \frac{(p-2)^2}{|\mathbf{X}|^2} - \frac{2p(p-2)}{|\mathbf{X}|^2} + \frac{4(p-2)}{|\mathbf{X}|^2} \\ &= \frac{(p-2)(2-p)}{|\mathbf{X}|^2} < 0 \text{ if } p \geq 3\end{aligned}\tag{8.3.39}$$

and Corollary 8.3.2 shows that δ_s renders \mathbf{X} inadmissible. The same argument establishes that δ_s^+ renders \mathbf{X} inadmissible. Unfortunately showing that δ_s^+ renders δ_s inadmissible requires showing that the difference of their risks satisfies

$$E_{\mu} \left(|\Delta_s|^2 - |\Delta_s^+|^2 + 2 \sum_{j=1}^p \frac{\partial(\Delta_{sj} - \Delta_{sj}^+)}{\partial x_j} \right) \geq 0.\tag{8.3.40}$$

Here $\Delta_s^+, \Delta_{sj}^+$ are defined analogously to Δ_s, Δ_{sj} with δ_s replaced by δ_s^+ .

An enormous literature has grown up from Stein's results. One of the major conclusions of the theory is that "shrinking" towards linear submodels, not just $\mathbf{0}$, is an improvement for relatively small values of p . An important method in applied regression analysis, "Ridge regression," invented independently, successfully utilizes shrinkage towards regression models of dimension $k < p$ if there are p regressors. A very extensive discussion of this literature and many more results may be found in Lehmann and Casella (1998). We shall return to this topic in Chapters 10 and 11. We close this section, following Efron and Morris (1973), by showing how δ_s may be motivated from a parametric empirical Bayes point of view.

Empirical Bayes and the James-Stein estimates δ_s

Suppose $\mathbf{X} \sim \mathcal{N}_p(\boldsymbol{\mu}, J)$ as we have assumed throughout this subsection. Now, following the empirical Bayes point of views of Section I.6, suppose μ_1, \dots, μ_p are i.i.d. $\mathcal{N}(0, \sigma^2)$ with σ^2 unknown. If σ^2 were known the Bayes estimate of $\boldsymbol{\mu}$ is seen from Example 3.2.1 to be

$$\boldsymbol{\delta}_B(\mathbf{X}) = (\delta_B(X_1)), \dots, \delta_B(X_p))^T$$

where $\delta_B(x) = \sigma^2 x / (1 + \sigma^2)$. Can we estimate σ^2 from our data? Note that, unconditionally, the X_i are i.i.d. $\mathcal{N}(0, 1 + \sigma^2)$. Thus $\hat{\sigma}^2 \equiv \frac{|\mathbf{X}|^2}{p} - 1$ is the UMVU, MLE of σ^2 and we

are led to

$$\widehat{\delta}_B(\mathbf{X}) \equiv \left(1 - \frac{p}{|\mathbf{X}|^2}\right) \mathbf{X}. \quad (8.3.41)$$

This is close to but not quite the James-Stein (1961) estimate. It may in fact be shown, see Theorem 5.1, Lehmann and Casella (1998), for instance, that $\delta_c(\mathbf{X}) \equiv \left(1 - \frac{c(p-2)}{|\mathbf{X}|^2}\right) \mathbf{X}$ uniformly improves \mathbf{X} provided that $p \geq 3$ and $0 < c < 2$. Thus, the estimate (8.3.41) uniformly improves \mathbf{X} for $p > 4$.

Remark 8.3.5. We have seen that the Stein-type estimators of a vector parameter with $p \geq 3$ have a smaller average mean squared error than that of the usual unbiased estimator. However, the mean square error of any one of the components in the Stein vector estimate may be larger than that of the usual unbiased estimate. See Rao and Shinozaki (1978).

Summary: This section explores connections between equivariance, invariance, minimax decision rules, and group models. We consider models generated by location groups, location-scale groups, affine groups, and monotone transformation groups. It is found that when we restrict decision rules to those that have the same equivariance properties as the groups generating the models, we can in some cases derive uniformly minimum risk equivariant estimates and equivariant minimax tests. Testing problems that are invariant with respect to the group of increasing transformations lead to rank tests as the invariant tests. We consider rank tests that are locally UMP invariant for semi-parametric model subgroups called copulas. We also discuss weaknesses of the “minimax equivariant” principle by presenting results of Stein that show the inadmissibility of such minimax procedures.

8.4 PROBLEMS AND COMPLEMENTS

Problems for Section 8.2

1. Let $X \sim \mathcal{B}(n, \theta)$. Show that $X(n - X)/n(n - 1)$ is the UMVU estimate of $\theta(1 - \theta)$.
2. *Variable Selection.* Methods for selecting the covariates to be included in a regression model are often based on unbiased estimates of mean squared estimation error (or equivalently (Problem I.7.5), mean squared prediction error).
 - (a) In Problem I.7.3, show that $\widehat{R}(\boldsymbol{\mu}, \widehat{\boldsymbol{\mu}}_1) - \widehat{R}(\boldsymbol{\mu}, \widehat{\boldsymbol{\mu}}_0)$ is an UMVU estimate of $R(\boldsymbol{\mu}, \widehat{\boldsymbol{\mu}}_1) - R(\boldsymbol{\mu}, \widehat{\boldsymbol{\mu}}_0)$ when model (I.8.1) holds.
 - (b) In Problem I.7.4, show that $\widehat{R}_p - \widehat{R}_q$ is an UMVU estimate of $R_p - R_q$ when model (I.8.2) holds.
3. Let X_1, \dots, X_n be a sample from $\mathcal{U}(\theta_1, \theta_2)$ where θ_1, θ_2 are unknown.
 - (a) Show that $T(\mathbf{X}) = (\min(X_1, \dots, X_n), \max(X_1, \dots, X_n))$ is sufficient.

(b) Assuming that $T(\mathbf{X})$ is complete, find a UMVU estimate of $(\theta_1 + \theta_2)/2$.

(c) Show that $T(\mathbf{X})$ is complete.

4. Let $\mathbf{N} = (N_1, \dots, N_k)$ have a $\mathcal{M}(n, \theta_1, \dots, \theta_k)$ distribution, $k \geq 2, \theta_1, \dots, \theta_k$ unknown. Find a UMVU estimate of $\theta_2 - \theta_1$.

5. Let X_1, \dots, X_n be a sample from a $\mathcal{N}(\mu, \sigma^2)$ population.

(a) Show that if σ^2 is known and μ is not, then \bar{X} is an UMVU estimate of μ .

(b) Show that the UMVU estimate of σ^2 if $\mu = \mu_0$ is known, and σ^2 is not, is

$$\frac{1}{n} \sum_{i=1}^n (X_i - \mu_0)^2 .$$

6. Let X_1, \dots, X_n be as in Problem 4.2.5 and $\sigma = 1$. Find the UMVU estimate of $P_\mu[X_1 \geq 0] = \Phi(\mu)$.

Hint. (X_1, \bar{X}) has a bivariate normal distribution. Apply Theorem B.4.2.

7. Let X_1, \dots, X_n be a sample from a $\mathcal{P}(\theta)$ distribution. Find the UMVU estimate of $P_\theta[X_1 = 0] = e^{-\theta}$.

8. Let X_1, \dots, X_n be a sample from a $\Gamma(p, \lambda)$ population where both p and λ are unknown. Find the UMVU estimate of p/λ .

9. Let $\mathbf{X} = (X_1, \dots, X_n)$ be a $\mathcal{N}(\mu, \sigma^2)$ sample. Show that if $n \geq 2$, \mathbf{X} though sufficient is not complete.

Hint. Consider $S(\mathbf{X}) = X_2 - X_1$.

More instructive counterexamples may be found in Stigler (1972). See Problem 18.

10. Let $\mathbf{X} = (X_1, \dots, X_n)$ be a $\mathcal{U}(0, \theta]$ sample, $\theta > 0$. Show that $M_n = X_{(n)}$ is the MLE of θ while $T^* = [(n+1)/n]M_n$ is the UMVU estimate of θ . Show that, if R denotes mean squared error, there exists an estimate T such that $R(\theta, T) < \min\{R(\theta, M_n), R(\theta, T^*)\}$. Thus M_n and T^* are inadmissible.

Hint. Consider $T = \frac{n+2}{n+1}M_n$. Note that $R(\theta, T) = \frac{1}{(1+n)^2}\theta^2$.

11. Suppose that T_1 and T_2 are two UMVU estimates of $q(\theta)$ with finite variances. Show that $T_1 = T_2$.

Hint. If T_1 and T_2 are unbiased so is $\frac{T_1+T_2}{2}$. Use the correlation inequality.

Problems 12, 13, and 14 give UMVU estimates of genetic trait parameters

12. Consider the genetic example of Problem 2.4.6 where X has the zero *truncated binomial* distribution $\mathcal{B}_1(n, \theta)$ with

$$p(x, \theta) = \frac{\binom{n}{x} \theta^x (1-\theta)^{n-x}}{1 - (1-\theta)^n}, \quad x = 1, \dots, n .$$

- (a) Show that X is complete and sufficient for θ .
- (b) Show that the expected number of family members with the trait given that at least one has the disease is $E_\theta(X) = n\theta/[1 - (1 - \theta)^n]$ and thus conclude that (X/n) is the UMVU estimate of $q(\theta) = \theta/[1 - (1 - \theta)^n]$.

13. For reasons similar to those of Problem 2.4.6 the *zero truncated Poisson distribution*

$$p(x, \theta) = \frac{\theta^x e^{-\theta}/x!}{1 - e^{-\theta}}, \quad x = 1, 2, \dots$$

is sometimes used as a model. Note that $p(x, \theta) = P(Y = x|Y \geq 1)$ where $Y \sim \mathcal{P}(\theta)$. Show that the UMVU estimate of $q(\theta) = P_\theta(Y \geq 1) = 1 - e^{-\theta}$ is given by

$$\begin{aligned} T^* &= 0 \text{ if } x \text{ is odd} \\ &= 2 \text{ if } x \text{ is even.} \end{aligned}$$

Hint. Show that if $E_\theta(T(X)) = q(\theta)$ then T must equal T^* .

14. Consider the model of Problems 2.4.6 and 8.2.12. Suppose that among the offspring of k sisters, exactly one of the children has the disease. Let X denote the number of children with the trait in the family in which one child has the disease; let Y denote the number of children with the trait in the i th of the other families; and let θ denote the true proportion of children with the trait. A reasonable model is one in which X, Y_1, \dots, Y_r are independent, X has the $\mathcal{B}_r(n, \theta)$ distribution of Problem 8.2.2, and $Y_i \sim \mathcal{B}(n_i, \theta)$, where $r = k - 1$ and n, n_1, \dots, n_r denotes the number of offspring in the families. Show that the UMVU estimate of θ is

$$\frac{\binom{n+N-1}{T-1} - \binom{N-1}{T-1}}{\binom{n+N}{T} - \binom{N}{T}}$$

where

$$T = X + \sum_{i=1}^r n_i \quad \text{and} \quad N = \sum_{i=1}^r n_i.$$

Hint. Use Theorem 8.2.2 to show that T is complete and sufficient. Write $\sum_{i=1}^r Y_i$ as $\sum_{i=1}^N Z_i$ where Z_1, \dots, Z_N are the indicators of independent $\mathcal{B}(1, \theta)$ events. Now compute $E(Z_1|T)$.

15. Estimating the Probability of Early Failure. It is sometimes reasonable to think of the lifetime (time to first repair) of a piece of equipment as a random variable following an exponential distribution with parameter λ which is unknown. Suppose n “identical” pieces of equipment are run and the failure times X_1, \dots, X_n are observed. We want to estimate the probability of early failure, that is, $P_\lambda[X_1 \leq x] = 1 - e^{-\lambda x}$ for some fixed x . Show that the UMVE is

$$\begin{aligned} T^* &= 1 - \left(1 - \frac{x}{\sum_{i=1}^n X_i}\right)^{n-1} \quad \text{if} \quad \sum_{i=1}^n X_i \geq x \\ &= 1 \quad \text{otherwise.} \end{aligned}$$

Hint. $T = \sum X_i$ is complete, sufficient. Set $S(X_1) = 1[X_1 \leq x]$, then

$$E(S(X_1)|T = t) = P[X_1 \leq x|T = t] = P\left[\frac{X_1}{T} \leq \frac{x}{T}|T = t\right].$$

By the substitution theorem for conditional expectations, (B.1.16), the right hand side equals $P[(X_1/T \leq (x/t))|T = t]$. Now, X_1/T is independent of T and has a beta, $\beta(1, n-1)$, distribution because X_1 and $(X_2 + \dots + X_n)$ are independent and the second of these variables has, by Corollary B.2.2, a $\Gamma(n-1, \lambda)$ distribution; next apply Theorem B.2.3. Therefore, if $b_{1,n-1}$ denotes the $\beta(1, n-1)$ density,

$$E(S(X_1)|T = t) = \int_0^{(x/t)} b_{1,n-1}(u) du.$$

Since $b_{1,n-1}(u) = (n-1)(1-u)^{n-2}$ for $0 < u < 1$, the result follows.

16. In Problem 15, consider the UMVU estimate T^* and the MLE $\tilde{T} = 1 - \exp[-x/\bar{X}]$. Show that as $n \rightarrow \infty$, $\sqrt{n}(T^* - \tilde{T}) \xrightarrow{P} 0$.

17. Let X_1, \dots, X_n be a sample from a $\mathcal{N}(\mu, \sigma^2)$ population. We want to estimate the proportion $q(\theta)$ of the population below a specified limit c , where $\theta = (\mu, \sigma^2)$.

(a) Show that the MLE of $q(\theta)$ is $\Phi[(x - \bar{X})/\hat{\sigma}]$ where \bar{X} and $\hat{\sigma}^2$ are the sample mean and variance.

(b) Show that the UMVU estimate of $q(\theta)$ is given by

$$\begin{aligned} T^*(X) &= 0 \quad \text{if } kV \leq -1 \\ &= G\left(\frac{1}{2} + \frac{1}{2}kV\right) \quad \text{if } -1 < kV < 1 \\ &= 1 \quad \text{if } kV \geq 1 \end{aligned}$$

where $k = \sqrt{n}/(n-1)$, $V = (c - \bar{X})/s$, and G is the $\beta\left(\frac{1}{2}, \frac{1}{2}(n-2)\right)$ d.f.

Hint. An unbiased estimate is $S(X) = 1[X_1 \leq c]$. Now compute

$$E(S(X)|\bar{X}, \hat{\sigma}^2) = P(X_1 \leq c|\bar{X}, \hat{\sigma}^2).$$

18. (Stigler (1972)). Consider the family \mathcal{P} of distributions P_θ , $\theta = 1, 2, \dots$, defined by

$$P_\theta[X = i] = \frac{1}{\theta} \text{ if } i = 1, \dots, \theta, = 0 \text{ otherwise.} \tag{8.4.1}$$

(a) Show that \mathcal{P} is complete. *Hint.* Use induction.

(b) Show that $2X - 1$ is the UMVU estimate of θ .

- (c) Show that the family $P_0 = \mathcal{P} - \{P_k\}$ is not complete when k is a positive integer.
Hint. Consider $E(v(X))$ where

$$\begin{aligned} v(i) &= 0 \quad \text{for } i \neq k, k+1 \\ &= 1 \quad \text{for } i = k \\ &= -1 \quad \text{for } i = k+1. \end{aligned}$$

- (d) Show that $2X - 1$ is not UMVU if $P_\theta \in \mathcal{P}_0$.

Hint. Consider

$$\begin{aligned} T(X) &= 2X - 1, \quad X \neq k, k+1 \\ &= 2k, \quad X = k, k+1. \end{aligned}$$

Remark. This model is not regular in the sense of Sections 1.1.3 and I.0.

19. Consider model (8.4.1). For what $\alpha \in (0, 1)$ can you find a similar test of $H : \theta = 10$ vs $K : \theta > 10$?

20. Establish (8.2.12).

21. Establish the converse of Theorem 8.2.3.

22. Lehmann and Scheffé (1950, 1955) show that if a complete sufficient statistic exists, it must be minimal sufficient. Assume this result. Let X_1, \dots, X_n be a sample from $\text{Unif}(\theta - 0.5, \theta + 0.5)$.

- (a) Show that $(X_{(1)}, X_{(n)})$ is minimal sufficient.

Hint. Use the factorization theorem. See Example 1.5.1 (continued).

- (b) Show that $(X_{(1)}, X_{(n)})$ is not complete.

Hint. Take $v(s, t) = t - s - [(n-1)/(n+1)]$ and use Problem B.2.9.

- (c) Show that no complete sufficient statistic exists.

23. Establish (8.2.17).

24. Suppose $(X_1, Y_1), \dots, (X_n, Y_n)$ are i.i.d. as $\mathcal{N}(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho)$. Consider the two statistics:

$$\begin{aligned} T &= (\sum X_i, \sum Y_i, \sum X_i^2, \sum Y_i^2)^T \\ \hat{\rho} &= \frac{\sum(Y_i - \bar{Y})(X_i - \bar{X})}{[\sum(X_i - \bar{X})^2 \sum(Y_i - \bar{Y})^2]^{\frac{1}{2}}}. \end{aligned}$$

If $\rho = 0$, are T and $\hat{\rho}$ independent? Why? Why not?

25. Let $(X_i, Y_i)^T, i = 1, 2, \dots, n$, be independent vectors with joint density

$$f_{X_i, Y_i}(x_i, y_i | \tau, \sigma) = \frac{y_i^2 e^{-y_i/\tau - x_i y_i/\sigma \tau}}{\tau^3 \sigma}$$

for $x_i, y_i > 0$ and $\tau, \sigma > 0$.

- (a) Give a complete sufficient statistic (T_1, T_2) for $\theta = (\tau, \sigma)$.
- (b) Give the moment generating function for (T_1, T_2) and give the corresponding joint distribution.
- (c) Derive the maximum likelihood estimators $\hat{\tau}, \hat{\sigma}$.
- (d) Calculate $E(\hat{\tau}), E(\hat{\sigma})$ and construct UMVU estimators of τ and σ .
- 26.** Let Y_1, \dots, Y_n be i.i.d. from the uniform distribution $U(0, \theta)$ with an unknown $\theta \in (1, \infty)$. Suppose that we only observe
- $$X_i = \begin{cases} Y_i & \text{if } Y_i \geq 1, \\ 1 & \text{if } Y_i < 1, \end{cases} \quad i = 1, \dots, n.$$
- (a) Derive a UMVUE of θ .
- (b) Find the MLE of θ and $\eta \equiv P(Y_1 > 1)$.
- (c) Derive a UMP test of size α for testing $H : \theta \leq \theta_0$ versus $K : \theta > \theta_0$, where θ_0 is known and $\theta_0 > (1 - \alpha)^{-1/n}$.

Problems for Section 8.3

- 1.** Show that for the affine group on R^d , (8.3.2) defines \bar{g} uniquely; that $\bar{\mathcal{G}}$ is a group, and that $\bar{\mathcal{G}}$ is isomorphic to \mathcal{G} .
- 2.** Suppose $\mathbf{X} \sim P_0 = \mathcal{N}_d(\mathbf{0}, J)$, $J = \text{diag}(1, \dots, 1)_{d \times d}$. Show that the transformation model induced by P_0 and the affine group is $\{\mathcal{N}_d(\boldsymbol{\mu}, \Sigma); \boldsymbol{\mu} \in R^d, \Sigma \in S\}$, where $S = \text{class of } d \times d \text{ nonsingular matrices}$.
- 3.** Establish the claim of Example 8.3.2.
- 4.** Establish the claim of Remark 8.3.1(a).
- 5.** Establish the claim of Remark 8.3.1(c).
- 6.** If $\mathbf{x}, \mathbf{y} \in R^d$, $|\mathbf{x}|^2 = |\mathbf{y}|^2$ there exists A orthogonal such that $A\mathbf{x} = \mathbf{y}$.
Hint. Construct an orthonormal basis where the first member is $\frac{\mathbf{x}}{|\mathbf{x}|}$ mapping $\frac{\mathbf{x}}{|\mathbf{x}|}$ onto $(1, 0, \dots, 0)^T$. See Vol. I, page 496.
- 7. (a)** Show that the ranks are indeed maximal invariant in Example 8.3.2.
(b) Suppose $\mathcal{X} = R^n$ and \mathcal{G} is as before. Exhibit a maximal invariant in this case.
- 8.** Establish Proposition 8.3.1.
- 9.** Show that the $\mathcal{X}_d^2(\theta^2)$ family is a monotone likelihood ratio family in the parameter θ^2 .
Hint. See Problem B.3.12.
- 10.** Suppose (X_1, \dots, X_n) have joint continuous case density f_1 and that $f_0(x_1, \dots, x_n) = \prod_{i=1}^n h(x_i)$ for some univariate continuous case density h that we can choose. Let $X_{(1)} <$

$\dots < X_{(n)}$ be the order statistics of the X_i and (R_1, \dots, R_n) the ranks. Assume that $f_0 > 0$ whenever $f_1 > 0$.

(a) Show that the following Hoeffding's formula holds.

$$P_1[R_1 = r_1, \dots, R_n = r_n] = \frac{1}{n!} E_{f_0} \left\{ \frac{f_1(X_{(r_1)}, \dots, X_{(r_n)})}{h(X_{(r_1)}) \dots h(X_{(r_n)})} \right\}.$$

Hint. Use formula (2.4.8) and note that $(X_{(1)}, \dots, X_{(n)})$ and (R_1, \dots, R_n) are independent under f_0 .

Apply Hoeffding's formula to deduce that when

$$f_1(x_1, \dots, x_n) = \prod_{i=1}^{n_1} h(x_i) \prod_{i=n_1+1}^n g(x_i).$$

(b) $P_1[R_1 = r_1, \dots, R_n = r_n] = \frac{1}{n!} E_0 \left\{ \prod_{i=n_1+1}^n \frac{g}{h}(X_{(r_i)}) \right\}$.

(c) Suppose $h = H'$, $g = G'$ and $G = 1 - (1 - H)^\theta$, $\theta > 0$. Show that the UMP (uniformly in F_0) rank test of $H : \theta = 1$ vs. $K : \theta = \theta_1 > 1$ where θ_1 is known, is given by rejecting for large values of

$$T_n(\theta_1) = E_0 \left\{ \prod_{i=n_1+1}^n (1 - U_{(r_i)})^{\theta_1 - 1} \right\},$$

where $U_{(1)} < \dots < U_n$ are the order statistics of a sample from $\mathcal{U}(0, 1)$.

(d) Set $h_0(u) = 1$, $g_1(u) = 2u$, $g_2(u) = 2(1-u)$, $0 < u < 1$, $n_1 = n_2 = 2$, and $\alpha = 1/6$. Show that the MP level α rank test for $H : h = g = h_0$ vs $K : h = h_0, g = g_1$ rejects H iff $R_3, R_4 \in \{3, 4\}$ while the MP level α rank test for H vs (h_0, g_2) rejects H iff $R_3, R_4 \in \{1, 2\}$.

Hint. $P(R_3, R_4 \in \{3, 4\}) = \binom{4}{2} P(X_1 < X_2 < X_3 < X_4)$, etc.

(e) Describe the MP level $\alpha = k/(6)$ rank test for testing H vs (h_0, g_1) in model (d) when $n_1 = n_2 = 3$ and (i) $k = 1$, (ii) $k = 2$.

(f) Same as (e) except for testing H vs (h_0, g_2) .

11 (a) Show that the distribution of the rank vector (R_1, \dots, R_n) in the two-sample problem of Example 8.3.10 depends on (F, G) only through GF^{-1} . Thus if h is a given continuous and strictly increasing function on $[0, 1]$, then $\{P_{\theta, h} : G = h^\theta(F), \theta > 0\}$ is a group submodel where the distribution of the ranks depends only on θ and h .

Hint. Then rank vector is invariant under the transformation $X_i \rightarrow F(X_i) \equiv X'_i$. Here X'_i has the $\mathcal{U}(0, 1)$ distribution for $i = 1, \dots, n_1$, and the distribution GF^{-1} for $i = n_1 + 1, \dots, n$.

(b) In Example 8.3.10, show that GF^{-1} is maximal invariant.

(c) Set $\bar{F} = 1 - F$ and $\bar{G} = 1 - G$. Show that the Lehmann alternative $\{P_{\theta, h} : \bar{G} = \bar{F}^\theta\}$ includes the exponential two-sample model.

Hint. If $U \sim \mathcal{U}(0, 1)$, then $-\lambda^{-1} \log(1 - U) \sim \mathcal{E}(\lambda)$.

(d) Suppose $X \sim F$ and $Y \sim G$. Y and G are said to be *stochastically larger* than

X and F iff $P(Y \geq t) \geq P(X \geq t)$ for all $t \in R$, that is, iff $G(t) \leq F(t)$, $t \in R$. In the two sample framework of Example 8.2.7, a test $\varphi(\mathbf{x}, \mathbf{y})$ of $H : F = G$ vs $K : "G \text{ is stochastically larger than } F \text{ is } \text{monotone if } y'_j \geq y_j, 1 \leq j \leq n, \text{ implies } \varphi(\mathbf{x}, \mathbf{y}') \geq \varphi(\mathbf{x}, \mathbf{y})"$. Show that:

- (i) All monotone tests are rank tests. That is, functions of the ranks R_1, \dots, R_n of Y_1, \dots, Y_n . Also show that all tests of the form $\varphi(\mathbf{X}, \mathbf{Y}) = 1(\sum_{j=1}^n a(R_j) \geq c)$ are monotone provided $a(k') \geq a(k)$ for $k' \geq k$.
- (ii) All monotone tests have monotone power. That is, $E_{F,G}\psi(\mathbf{X}, \mathbf{Y}) \geq E_{F,H}\psi(\mathbf{X}, \mathbf{Y})$ if G is stochastically larger than H . Thus monotone tests are unbiased.

Remark. Wang (1996) developed a test of equality of distributions against nonparametric stochastically ordered alternatives.

12. Hoeffding (1951) and Hajek and Sidak (1967) show that for rank tests of $H_0 : \theta = \theta_0$ vs $K : \theta > \theta_0$, the locally UMP rank test is, under regularity conditions, based on $(\delta/\delta\theta)P(\mathbf{R} = r)|_{\theta=\theta_0}$.

(a) Show that for allowable levels α the locally UMP rank test for the model in Problem 8.3.10(c) is the Savage exponential scores test.

(b) Show that UMP level α permutation test of $H : F = G$ vs. $K : F(t) = 1 - e^{-\lambda t}$, $G(t) = 1 - e^{-\lambda \theta t}$, $\theta > 1$, is given by rejecting for small values of $\sum_{i=n+1}^n x_i$ with critical value $c_n(x_{(1)}, \dots, x_{(n)})$.

(c) Show that the test in (b) is uniformly more powerful in this parametric family of alternatives than the MP rank test given in (a).

13 (a) Show that if $X_{(1)} < \dots < X_{(n)}$ are the exponential, $\mathcal{E}(1)$, order statistics, then $E(X_{(j)}) = \sum_{i=n+1-j}^n i^{-1}$.

Hint. See Problem B.2.14.

(b) In (a), show that if $j = [nu + 1]$, $0 < u < 1$, then $E(X_{(j)}) = -\log(1 - u) + o(1)$.
Hint. $X_{(j)} = -\log(1 - U_{(j)})$, where $U_{(1)} < \dots < U_{(n)}$ are uniform, $\mathcal{U}(0, 1)$, order statistics. Taylor expand around $U_{(j)} \cong u$. See Problem B.2.9.

14. Copulas. Consider Example 8.3.11.

(a) Let $C_i(F(\cdot))$ be the copula model with $d = 1$, $\theta_i = \exp\{\beta z_i\}$ and the *Lehmann-Savage copula*

$$C_i(u) = 1 - [1 - u]^{\theta_i}, \quad 0 < u < 1.$$

Use Hoeffding's formula to show that the locally UMP invariant test of $H_0 : \beta = 0$ vs $H_1 : \beta > 0$ is based on the *exponential scores* statistic

$$T_E = n^{-\frac{1}{2}} \sum_{i=1}^n a_E(R_i)(z_i - \bar{z})$$

where $a_E(k) = E(Z^{(k)})$ with $Z^{(1)} < \dots < Z^{(n)}$ exponential order statistics.

Hint. Set $K(t) = 1 - \exp(-t)$. The df of $W = K^{-1}F(Y_i)$ is $K(w/\theta_i)$. (Approximate critical values for T_E can be obtained from $T_E/s \sim \mathcal{N}(0, 1)$, $s^2 = \Sigma(z_i - \bar{z})^2/(n - 1)$).

(b) In the transformation model $h(Y_i) = \theta_i + \varepsilon_i$, set $h(y) = \log\{-\log[1 - F(y)]\}$ for some unknown continuous df F and suppose $\varepsilon_i \sim 1 - \exp\{-\exp(t)\}$. Let $Y_i \sim F_i$ and assume that F_i and F have continuous case densities f_i and f . Show that F_i satisfies the proportional hazard rate model where $\lambda(y|\mathbf{z}_i) = \theta_i \lambda(y)$ with $\lambda(y|\mathbf{z}_i) = f_i(y)/[1 - F_i(y)]$ and $\lambda(y)/[1 - F(y)]$ being regression and baseline hazard rates.

(c) Show that the models of (a) and (b) are the same.

(d) Show that for $d = 1$ and $\theta_i = \alpha + \beta z_i$ in the Gaussian copula model, the normal scores statistic is locally UMP invariant. Use Hoeffding's formula.

(e) Same as (d) except logistic copula and uniform scores statistic.

(f) *Proportional odds model.*

(i) Let $C_i(F(\cdot))$ be the copula model with $d = 1$, $\theta_i = \alpha + \beta z_i$. Bell and Doksum (1966, Table 8.1), considered

$$C_i(u) = \frac{u}{u + (1 - u)\exp(\theta)} .$$

Use Hoeffding's formula to show that the uniform scores test is locally UMP invariant for testing $H_0 : \beta = 0$ vs $H_1 : \beta > 0$.

(ii) The *odds ratio* is defined as $r_i = F_i/(1 - F_i)$. Let $r = F/(1 - F)$ where F is a continuous baseline df. Show that for the model in (i),

$$r_i(y) = \exp(\theta_i)r(y) .$$

This is called the *proportional odds model*.

(iii) In the monotone transformation model $h(Y_i) = \theta_i + \varepsilon_i$ with $d = 1$ and $\theta_1 = \alpha + \beta z_i$, set $h(y) = \log\{F(y)/[1 - F(y)]\}$ for some unknown continuous df F and assume that ε_i has a logistic distribution. Show that this model is equivalent to the model in (i) and (ii).

(g) *Random effect proportional hazard.* Bell and Doksum (1966) considered

$$\begin{aligned} C_i(u) &= \frac{e^{\theta_i u} - 1}{e^{\theta_i} - 1}, & \theta_i \neq 0 \\ &= u, & \theta_i = 0 . \end{aligned}$$

Sibuya (1968) and Nabeya and Miura (1972, unpublished) showed that $C_i(F(\cdot))$ (called the SINAMI model) is a random effects proportional hazard model $\lambda(y|\mathbf{z}_i) = \Delta_i \lambda(y)$ with Δ_i having a zero truncated Poisson (θ_i) distribution. Show that when $d = 1$ and $\theta_i = \alpha + \beta z_i$, the uniform scores test is locally UMP invariant for testing $H_0 : \beta = 0$ vs $H_1 : \beta > 0$.

(h) Same as (f) except

- (i) $C_i(u) = (1 - \theta_i)u + \theta_i u^2$. Show that the uniform scores test is locally UMP invariant.
- (ii) $C_i(u) = (1 - \theta_i)u + \theta_i u^m$, $m \geq 2$ an integer. Find the locally UMP invariant test when m is known.

15. Monotone regression tests. In the regression framework where we observe $(z_1, Y_1), \dots, (z_n, Y_n)$ with $z_1 < \dots < z_n$ nonrandom and Y_1, \dots, Y_n independent. Let “ $\tilde{<}$ ” denote stochastic inequality. A test $\varphi(\mathbf{y})$ of $H : F_i = F$, $1 \leq i \leq n$, vs $K : F_1 \tilde{<} \dots \tilde{<} F_n$ is *regression monotone* if $\varphi(\mathbf{y}') \leq \varphi(\mathbf{y})$ for all \mathbf{y}' and \mathbf{y} for which $i < j$ and $y'_i \leq y'_j$ imply $y_i \leq y_j$. Show that

- (a) All monotone tests are rank tests.
- (b) All monotone tests have monotone power, that is, $\inf F_i^{-1}G_i(y) \leq F_j^{-1}G_j(y)$ for $i \leq j$, all y , and if $\varphi(\cdot)$ is monotone, then

$$E_F(\varphi(\mathbf{Y})) \geq E_G(\varphi(\mathbf{Y})).$$

Note that it follows that monotone tests are unbiased for H vs $K : F_1 \tilde{<} \dots \tilde{<} F_n$.

Hint. Let $Y'_i \sim G_i$, set $Y_i = F_i^{-1}(G_i(Y'_i))$, then $i < j$ and $Y'_i \leq Y'_j \implies Y_i \leq Y_j$.

- (c) Let R_1, \dots, R_n be the ranks of Y_1, \dots, Y_n . If $a(i)$ is nondecreasing in $i = 1, \dots, n$, then the test $\varphi(y) = 1[\sum_{i=j}^n z_i a(R_i) \geq c]$ is monotone.

16. Consider Example 8.3.12.

- (a) Show that if (X, Y) has the continuous case density $f(x, y)$, then (Hoeffding's formula for bivariate data)

$$P((\mathbf{R}, \mathbf{S}) = (\mathbf{r}, \mathbf{s})) = \frac{1}{n!} E \left\{ \sum_{i=1}^n \frac{f(V^{(r_i)}, W^{(s_i)})}{f_1(V^{(r_i)}) f_2(W^{(s_i)})} \right\}$$

where f_1 and f_2 are the marginal densities of f and $\{V^{(k)}\}$ and $\{W^{(k)}\}$ are the order statistics in independent samples from f_1 and f_2 .

- (b) Show that the bivariate normal scores statistic is locally UMP invariant for testing $H_0 : \rho = 0$ vs $H_1 : \rho > 0$ in the bivariate Gaussian copula. You may differentiate inside the expected value in Hoeffding's formula.

Hint. Because of the invariance of the ranks, assume $(X, Y) \sim \mathcal{N}(0, 0, 1, 1, \rho) \equiv f_\rho$. Use

$$\frac{\partial}{\partial \rho} \prod_{i=1}^n f_\rho = \left(\frac{\partial}{\partial \rho} \sum_{i=1}^n \log f_\rho \right) \prod_{i=1}^n f_\rho.$$

- (c) Consider the model with $g(X) = V + \theta Z$, $h(Y) = W + \theta Z$, where V and W have $\mathcal{N}(\mu_1, \sigma_1^2)$ and $\mathcal{N}(\mu_2, \sigma_2^2)$ densities, the df M of Z is arbitrary, V , W , and Z are independent, and g and h are increasing. Show that the bivariate normal scores statistic is locally UMP invariant for testing $H : \theta = 0$ vs $K : \theta > 0$.

Hint. For invariant (rank) statistics, we can assume

$$f(x, y; \theta) = \int_{-\infty}^{\infty} \varphi(x - \theta z) \varphi(y - \theta z) dM(z).$$

- (d) Show that for the bivariate Gaussian copula, $|\rho| = \sup_{a,b} \text{corr}(a(X), b(Y))$ where the sup is over monotone functions a and b .

17. Establish (8.3.21).

18. Establish (8.3.23).

19. Show in Example 8.3.13 that the size α test that rejects H iff $X_1 \geq c$ is

(a) UMP for testing $H : \theta \leq \theta_0$ vs $K : \theta \geq \theta_0$ when Σ_0 is known.

(b) The Bayes test for the loss function $l_c(u, \phi)$ and the priors π_0 and π_1 corresponding to point masses at $(\theta_0, \eta_0^T)^T$.

(c) UMP unbiased when Σ_0 is unknown.

(d) UMP invariant when Σ_0 is unknown.

20. Show that \mathbf{X} is minimax for quadratic loss when $\mathbf{X} \sim \mathcal{N}_p(\boldsymbol{\mu}, \sigma_0^2 J)$ where $J = \text{diag}(1, \dots, 1)_{p \times p}$.

21. Let N have the binomial distribution $\mathcal{B}(p, 10)$; given $N = n$, let X_1, \dots, X_n be i.i.d. $N(\mu, 1)$. The observations consist of (N, X_1, \dots, X_N) .

(i) If p has a known value p_0 , show that there exists neither a UMP test nor a UMP unbiased test of $H : \mu = 0$ against $\mu > 0$.

(ii) If p is unknown, show that there does exist a UMP unbiased test of $H : \mu = 0$ against $\mu > 0$.

22. Let X_1, \dots, X_n be i.i.d. $N(\mu, \sigma^2)$. State the MP test of $H : \sigma = \sigma_0$ against a simple alternative (μ_1, σ_1) for the two cases (i) $\sigma_1 > \sigma_0$; (ii) $\sigma_1 < \sigma_0$, and in each case describe the least favorable distribution for μ over the parameter set $\{(\mu, \sigma_0) : -\infty < \mu < \infty\}$.

23. Let X_1, X_2 be positive random variable with density

$$\frac{1}{\sigma^2} f\left(\frac{x_1}{\sigma}, \frac{x_2}{\sigma}\right), \sigma > 0.$$

Assuming it exists, let $\delta(\mathbf{X})$ be the minimum risk scale equivariant estimate of σ , under the loss function $(d - \sigma)^2 / \sigma^2$.

(a) Show that

$$\delta(\mathbf{X}) = \frac{\int_0^\infty u^2 f(uX_1, uX_2) du}{\int_0^\infty u^3 f(uX_1, uX_2) du}.$$

(b) Suppose there exists a complete sufficient statistic $t(\mathbf{X})$ for the model. Show that $\delta(\mathbf{X})$ is itself a sufficient statistic.

24. Let \mathbf{C} be a $d \times d$ matrix. Suppose $\gamma_1, \dots, \gamma_d$ and $\mathbf{w}_1, \dots, \mathbf{w}_d$ satisfy $\mathbf{C}\mathbf{w}_j = \gamma_j \mathbf{w}_j$, $1 \leq j \leq d$. Then $\gamma_1, \dots, \gamma_d$ are called *eigenvalues* of \mathbf{C} and $\mathbf{w}_1, \dots, \mathbf{w}_d$ are called *eigenvectors*. See Section B.10. Consider the inner product defined by $(\mathbf{x}, \mathbf{y}) = \sum_i x_i y_i$ and the norm $|\mathbf{x}| = (\mathbf{x}, \mathbf{x})^{1/2}$. Let

$$\begin{aligned} \mathbf{v}_1 &= \arg \max \{ \mathbf{a}^T \mathbf{C} \mathbf{a} : |\mathbf{a}| = 1 \} \\ \mathbf{v}_2 &= \arg \max \{ \mathbf{a}^T \mathbf{C} \mathbf{a} : |\mathbf{a}| = 1, \mathbf{a} \perp \mathbf{v}_1 \} \\ &\vdots \\ \mathbf{v}_d &= \arg \max \{ \mathbf{a}^T \mathbf{C} \mathbf{a} : |\mathbf{a}| = 1, \mathbf{a} \perp [\mathbf{v}_1, \dots, \mathbf{v}_{d-1}] \} \end{aligned}$$

and $\lambda_j = \mathbf{v}_j^T \sum \mathbf{v}_j$, $1 \leq j \leq d$.

- (a) Show that $\mathbf{v}_1, \dots, \mathbf{v}_d, \lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_d$ are all well defined and that $\mathbf{C}\mathbf{v}_j = \lambda_j \mathbf{v}_j$, $1 \leq j \leq d$. Thus $\lambda_1, \dots, \lambda_d$ are eigenvalues and $\mathbf{v}_1, \dots, \mathbf{v}_d$ are eigenvectors. This is the Courant-Fischer theorem.
- (b) When \mathbf{C} is the covariance matrix \sum of a random vector $\mathbf{X} \in R^d$, the linear combination $L_j = \mathbf{v}_j \mathbf{X}$ is called the j th principal component. Show that $\text{Var}(L_j) = \lambda_j$. The term “principal components” is also used for the empirical $\hat{\mathbf{v}}_j, \hat{\lambda}_j$ which correspond to $\widehat{\sum}$, the empirical variance covariance matrix of a sample $\mathbf{X}_1, \dots, \mathbf{X}_n$. Sometimes the correlation matrix is used in place of \sum and $\widehat{\sum}$.
- (c) Establish the (a) part for the inner product $(\mathbf{x}, \mathbf{y}) = \sum_{i=1}^d \Pi_i x_i y_i$, where $\Pi_i > 0$ and $\sum_{i=1}^d \Pi_i = 1$.

25. Let X_1, \dots, X_n be i.i.d. discrete random variables with

$$P(X_1 = x) = \gamma(x)\theta^x/c(\theta), \quad x = 0, 1, 2, \dots,$$

where $\gamma(x) \geq 0$, $\theta > 0$ is unknown, and $c(\theta) = \sum_{x=0}^{\infty} \gamma(x)\theta^x$. Let Y_1, \dots, Y_n be i.i.d. random variables having the beta distribution with density

$$\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} y^{a-1} (1-y)^{b-1}, \quad 0 < y < 1,$$

where $a > 0$ and $b > 0$ are unknown. Assume that X_i 's and Y_i 's are independent. Suppose that we observed $Z_i = X_i + Y_i$, $i = 1, \dots, n$.

- (a) Show that $E X_1 = \theta c'(\theta)/c(\theta)$, where $c'(\theta)$ is the first order derivative of $c(\theta)$.
- (b) Obtain a complete and sufficient statistic for the unknown parameter (θ, a, b) .
- (c) Derive a uniformly minimum variance unbiased estimator of the parameter θ .
- (d) For a given $\alpha \in (0, \frac{1}{2})$, derive a uniformly most powerful test of size α for

$$H_0 : \theta \geq \theta_0 \text{ vs } H_1 : \theta < \theta_0,$$

where $\theta_0 > 0$ is a known value.

8.5 Notes

Note for Section 8.2.

- (1) The “similar” terminology arose from the notion that if ϕ was non randomized and corresponded to critical region C , then C was, under H , similar to the sample space \mathcal{X} property $P(\mathcal{X}) = 1$ in having probability $P(C) = \alpha$, $P \in \mathcal{P}_0$, not depending on P .

Notes for Section 8.3.

- (1) For general definitions of abstract groups and their properties we refer to, for instance, Birkhoff and MacLane (1998).
- (2) Lehmann (1953) proposed $G = F^\theta$. Savage (1956,1980) noted that if F is replaced by $S = 1 - F$ one obtains (8.3.16). He called this *Lehmann alternatives*.
- (3) A formulation of this type is in Wijsman (1990).

Chapter 9

INFERENCE IN SEMIPARAMETRIC MODELS

As we indicated in Section I.1, our concern in this chapter will be with methods for inference about parameters in non- and semiparametric models whose qualitative large sample behaviour is like that of the corresponding procedures for regular parametric models. In the i.i.d. case, estimates converge at rate \sqrt{n} to Gaussian distributions, and the behaviour of tests is governed by that of Gaussian processes. Section 9.1 introduces theory for estimates based on maximizing modified and empirical likelihoods in important semiparametric models such as linear models with stochastic covariates, biased sampling models, Cox proportional hazard models with censoring, and independent component analysis models. Section 9.2 and 9.3 deal, respectively, with asymptotic normality and efficiency for estimates. Section 9.4 similarly deals with asymptotic inference for tests in semiparametric models. We need to use the machinery of Chapter 7 to extend the techniques of Chapters 5 and 6. In Section 9.5 we show how the powerful notions of contiguity and local asymptotic normality can clarify power and information bound calculations.

9.1 Estimation in Semiparametric Models

In Sections 9.1, 9.2, and 9.3, we want to primarily address the question of how to optimally estimate regular parameters in semiparametric models. We have defined semiparametric models by example and will continue to do so. Loosely, they are models which are naturally parametrized by both Euclidean and infinite dimensional parameters; or where distributions in the model are subject to restrictions. Nonparametric models are loosely described as ones where every probability distribution is at least approximable by distributions in the model. We will consider parametric and nonparametric models to be special cases of semiparametric models.

9.1.1 Selected Examples

We have already considered a number of semiparametric models:

Section 3.5: The symmetric location model (3.5.1)

$$X_i = \mu + \varepsilon_i, \quad \varepsilon_i \sim \text{i.i.d. } F, \quad \mu \in R,$$

where the natural parameter is (μ, F) with F the error distribution function.

The gross error model

$$f(x) = F'(x) = (1 - \lambda) \frac{1}{\sigma} \varphi(x/\sigma) + \lambda h(x)$$

of Section 3.5.3 parametrized by $(\mu, \lambda, \sigma, h)$ where h is a density and μ is unidentifiable without restrictions on $h(\cdot)$ such as $h(x) = h(-x)$.

Section I.1.1 and 8.3: The two-sample model of Example 8.3.10 where we observed two independent samples with distributions F and G , is naturally parametrized by (F, G) , which is fully nonparametric, and the hypothesis model is $\{(F, G) : F = G\}$, which is semiparametric.

We pursue an important model which was introduced in Examples 6.2.1 and 6.2.2.

Example 9.1.1. *Semiparametric linear models with stochastic covariates.* Here, we observe $(\mathbf{Z}_1, Y_1), \dots, (\mathbf{Z}_n, Y_n)$ i.i.d. as (\mathbf{Z}, Y) , $\mathbf{Z} \in R^p$, $Y \in R$, with

$$Y = \alpha + \mathbf{Z}^T \boldsymbol{\beta} + \sigma \varepsilon.$$

Various semiparametric versions of this model are in use. Here are two:

- (a) \mathbf{Z} and ε are independent, $\varepsilon \sim F$, $\mathbf{Z} \sim H$, $F \in \mathcal{F}$, $H \in \mathcal{H}$, and \mathcal{F} , \mathcal{H} are general, and $\Sigma \equiv \text{Var}_H(\mathbf{Z}) \equiv E_H(\mathbf{Z} - E_H(\mathbf{Z}))(\mathbf{Z} - E_H(\mathbf{Z}))^T$ is nonsingular.
- (b) The joint distribution Q of $(\mathbf{Z}, \varepsilon)$ is arbitrary save that $E\varepsilon^2 < \infty$, $E(\varepsilon|\mathbf{Z}) = 0$, and $\text{Var}(\mathbf{Z})$ is nonsingular.

Thus we can think of model (a) as parametrized by $(\alpha, \boldsymbol{\beta}, \sigma, F, H)$ and model (b) by $(\alpha, \boldsymbol{\beta}, \sigma, Q)$. It is easy to see that in neither of these models is σ identifiable and α is not in the first as well (Problem 9.1.1). However, we have already seen in Example 6.2.2 that, if in addition to assumption (a) above, we assume F with density f symmetric about 0, then $\boldsymbol{\beta}$ and α are both identifiable and we exhibited estimating equations for $(\alpha, \boldsymbol{\beta})$. In fact, symmetry is only required for identifiability of α , and $\boldsymbol{\beta}$ is \sqrt{n} consistently estimable in the sense that $\widehat{\boldsymbol{\beta}} = \boldsymbol{\beta} + O_P(n^{-\frac{1}{2}})$ for model (a) by using the estimating equations of Example 6.2.2 with, for instance, f_0 the logistic density (Problem 9.1.1(d)).

In model (b), the LSE of Example 6.2.1 is \sqrt{n} consistent. To show this, write

$$\widehat{\boldsymbol{\beta}} = \boldsymbol{\beta} + \widehat{\Sigma}^{-1} \frac{1}{n} \sum_{i=1}^n (\mathbf{Z}_i - \bar{\mathbf{Z}})(Y_i - \mathbf{Z}^T \boldsymbol{\beta}) \tag{9.1.1}$$

where $\widehat{\Sigma} \equiv \frac{1}{n} \sum_{i=1}^n (\mathbf{Z}_i - \bar{\mathbf{Z}})(\mathbf{Z}_i - \bar{\mathbf{Z}})^T$. To check \sqrt{n} consistency, note that by consistency of continuous functions of sample moments

$$\widehat{\Sigma}^{-1} \xrightarrow{P} \Sigma^{-1}.$$

Further,

$$\begin{aligned} & \frac{1}{n} \sum_{i=1}^n (\mathbf{Z}_i - \bar{\mathbf{Z}})(Y_i - \mathbf{Z}^T \boldsymbol{\beta}) \\ &= \frac{1}{n} \sum_{i=1}^n (\mathbf{Z}_i - E(\mathbf{Z}))\sigma \varepsilon_i - (\bar{\mathbf{Z}} - E(\mathbf{Z})) \frac{1}{n} \sum_{i=1}^n \varepsilon_i \\ &= O_P(n^{-\frac{1}{2}}) + O_P(n^{-1}) \end{aligned}$$

since $E(\mathbf{Z} - E(\mathbf{Z}))\varepsilon = E(\mathbf{Z} - E(\mathbf{Z}))E(\varepsilon|\mathbf{Z})$ by B.1.20 and $E(\varepsilon|\mathbf{Z}) = 0$, so that \sqrt{n} consistency of $\hat{\boldsymbol{\beta}}$ holds in model (b). \square

In these two semiparametric linear models, the construction of \sqrt{n} consistent estimates of the interesting Euclidean parameter $\boldsymbol{\beta}$ proved fairly simple. However, it is not clear what alternative and possibly better procedures there might be. That is, we need to develop a concept of efficiency for semiparametric models that can be used to determine what a “best” estimate is. This will be done in the context of this example in Section 9.3.

Example 9.1.2. Biased Sampling. Stratification. It is often the case that, while we want to obtain information about a population F , or at least features of F such as the mean and variance, we are only able to observe a sample from $P_{(F,G)}$, where G is a finite or infinite dimensional parameter. We begin by considering the essentially nonparametric model of biased sampling from a single population. Here, if the distribution F of Z has density (discrete or continuous case) f , observations are not on Z but on X from

$$p_F(x) = \frac{w(x)f(x)}{W(F)} \quad (9.1.2)$$

where $W(F) = \int w(x) dF(x)$, and $w(\cdot)$ is assumed known.

An important special case is “length biased” sampling where $X \in R^+$, $X \sim F$, and $w(x) = x$. This arises classically if we are interested in, say, the proportion $f(2)$ of two-child families in a city with a known total number of households N . If we could sample households with at least one child, and Z is the number of children in a sampled household we could estimate $f(2) = P[Z = 2]$ directly. Suppose instead we sample n children at random and consider the proportion $p_F(2)$ of children coming from two-child families in this group. If the city is large, $f(j)$ is the proportion of j child families in the city, $j = 0, 1, 2, \dots$, and X is the number of children in a sampled household, then,

$$p_F(2) = P[X = 2] = 2f(2) / \sum_{j=1}^{\infty} j f(j). \quad (9.1.3)$$

In this case (9.1.2) holds with $w(x) = x$.

Based on observations from $p_F(\cdot)$ in (9.1.2), is F identifiable? This is true iff $w(x) > 0$ whenever $f(x) > 0$. How do we identify F in this case? We claim that (Problem 9.1.2)

$$f(x) = \frac{p_F(x)/w(x)}{\int \frac{1}{w(x)} dP_F(x)}. \quad (9.1.4)$$

Next we consider stratified populations where biased sampling may occur within each strata. A stratified population S is made up of subpopulations S_1, \dots, S_k called strata. We are interested in the density $f(z)$ of a random draw Z from S , but sample by first randomly selecting an index I with corresponding strata S_I . This situation arises if S is the set of all patients in a county and S_1, \dots, S_k are the patients in the k hospitals and clinics in the county. Length biased sampling may occur within each strata. In this case we observe $X = (I, Y)$ where $I = 1, \dots, k$, $P[I = j] = \lambda_j$, $1 \leq j \leq k$, are assumed known, and given $I = j$, Y has density

$$p_F(y|j) = \frac{w_j(y)f(y)}{W_j(F)} \quad (9.1.5)$$

where $W_j(F) = \int w_j(x)dF(x)$. Again, w_1, \dots, w_k are assumed known. In stratified sampling λ_j is the probability that Y will be from the j th strata S_j and $w_j(x) = 1(x \in S_j)$. We claim that F is identifiable iff $\sum_{j=1}^k \lambda_j w_j(x)f(x) > 0$ (Problem 9.1.2). \square

The next two examples come from the area of *survival analysis*, the study of lifetimes of subjects in clinical trials, experimental animals, pieces of equipment, etc.. The models in these examples apply more generally to experiments that involve “time to event data.” A description of lifetime distributions more suited to the subject than densities or distribution functions is the hazard rate.

Definition 9.1.1. The *hazard rate* $\lambda(\cdot)$ of a nonnegative, continuous variable T with density f and distribution function F is defined for $t \geq 0$ by

$$\begin{aligned} \lambda(t) &= \lim_{\Delta \rightarrow 0} \{P[t \leq T \leq t + \Delta | T \geq t]\}/\Delta \\ &= \lim_{\Delta \rightarrow 0} \frac{F(t + \Delta) - F(t)}{\Delta(1 - F(t))} = \frac{f(t)}{1 - F(t)} = \frac{d}{dt}(-\log S(t)) \end{aligned}$$

where $S(t) = P[T > t] = 1 - F(t)$ is defined as the *survival function*.

It follows that,

$$f(t) = \lambda(t) S(t) = \lambda(t) e^{-\int_0^t \lambda(s)ds} \equiv \lambda(t) e^{-\Lambda(t)} \quad (9.1.6)$$

where $\Lambda(t)$ is the *cumulative hazard function*. Note that

$$\Lambda(t) = -\log S(t). \quad (9.1.7)$$

The *discrete* version of the *hazard rate* is defined for a discrete variable T as

$$\lambda(t) \equiv P[T = t | T \geq t] = \frac{P[T = t]}{P[T \geq t]}. \quad (9.1.8)$$

By definition, $\lambda(t)$ is 0 unless $P[T = t] > 0$. The analogue of (9.1.6) becomes

Lemma 9.1.1. If $\sum_{j=1}^{\infty} P[T = t_j] = 1$, $t_1 < t_2 < \dots$, then

$$P[T \geq t_k] = \prod_{j=1}^{k-1} (1 - \lambda(t_j)), \quad P(T = t_k) = \lambda(t_k) \prod_{j=1}^{k-1} (1 - \lambda(t_j)). \quad (9.1.9)$$

Proof. $P(T \geq t_k) = \Pi_{j=1}^{k-1} \left\{ \frac{P[T \geq t_{j+1}]}{P[T \geq t_j]} \right\}$. Next note that

$$P[T \geq t_{j+1}] = P[T \geq t_j] - P[T = t_j].$$

The second equation follows from (9.1.8). \square

We will call (9.1.6) and (9.1.7) the *continuous case approximations* to (9.1.9). In particular, $\exp\{-\Lambda(t)\}$ approximates $\Pi_{j=1}^{k-1}(1 - \lambda(t_j))$, and vice versa.

Our next example introduces two different types of biased sampling.

Example 9.1.3. Censoring and truncation. A very common type of data in many contexts is censored or truncated data. In (right) censored sampling we wish to observe “times” T_1, \dots, T_n pertaining to n units sampled from a population. Instead, events occurring at times C_1, \dots, C_n may prevent some of T_1, \dots, T_n from being observed. These censoring times are such that $(T_1, C_1), \dots, (T_n, C_n)$ are independent and we observe $X_i = (Y_i, \delta_i)$ where

$$Y_i = \min(T_i, C_i), \quad \delta_i = 1(T_i \leq C_i), \quad 1 \leq i \leq n.$$

That is, we observe either the time of interest T or the censoring time C and are told which one is observed. A classical situation where this occurs is when T is the time an individual with some disease dies after entering a study of duration L . If $T \leq L$, the survival time is observed; if $T > L$, the individual is said to be lost to follow up. In this case $C = L$. Subtler is the case where an individual enters the study at a random time T_0 assumed independent of the survival time from entry. We then observe survival time T if $T \leq L - T_0$ and $C = L - T_0$ otherwise, which is assumed independent of T , and indeed $1(T \leq C)$ is also observable. An extensive discussion is given for instance in Andersen, Borgan, Gill and Keiding (1988,1993) and Kalbfleisch and Prentice (2002). If we assume that T and C are independent with distributions F, G which are arbitrary and if F, G have densities f, g , then the distribution of (Y, δ) is given by

$$\begin{aligned} P[Y \leq y, \delta = 1] &= \int_{-\infty}^y f(s)\bar{G}(s)ds \\ P[Y \leq y, \delta = 0] &= \int_{-\infty}^y g(s)\bar{F}(s)ds \end{aligned} \tag{9.1.10}$$

where $\bar{F} = 1 - F$, $\bar{G} = 1 - G$ are the appropriate survival functions. Thus, the model is parametrized by (F, G) . Can (F, G) be fully identified? The answer is yes, under mild conditions, as we shall see later in the continuation (Example 9.1.9) of this example.

Another important type of biased sampling is *truncation*. Here, the distribution F of T is, as usual, what is wanted, but what we do observe is Y where Y has the conditional distribution of T given $T \geq M$ where M , assumed independent of T , is an observed threshold. Thus, if Y has density f and df F and M has density g and df G , then the density of $X = (M, Y)$ is

$$p(m, y) = \frac{g(m)f(y)1(y \geq m)}{\bar{F}(m)}. \tag{9.1.11}$$

An important example is truncation in astronomical data where the luminosities L of stars are available only if their (visual) brightness B is above a minimal detection level since brightness depends both upon the luminosity and the distance to the star. B which is a function of the distance is genuinely random, and, under a hypothesis of homogeneity of star types in any given direction, can be assumed independent of L — see Babu and Feigelson (1996) for instance. The issue of identifiability of F and G arises and again the answer is affirmative under natural conditions — see Example 9.1.9. \square

Example 9.1.4. *The Cox proportional hazard model.* Analogues of the Gaussian linear model that were meant to model event times T were initially built around the exponential, $\mathcal{E}(\lambda)$, family of distributions. Specifically, given a vector of covariates \mathbf{Z} which might include factors such as age, weight, and treatment indicator, the response lifetime variable T was modelled so that the natural parameter, the conditional hazard rate $\lambda(t|\mathbf{Z} = \mathbf{z})$, was a function of \mathbf{Z} and a parameter $\boldsymbol{\beta}$. Since λ is positive, a specification such as

$$\lambda(t|\mathbf{z}) = \exp\{\mathbf{z}^T \boldsymbol{\beta}\}$$

is natural. If, for instance, we have two groups, $\mathbf{z} = (z_1, z_2)^T$, $z_1 \equiv 1$, and $z_2 = 0$ or 1 as the observation comes from group 1 or 2, then this is just a model for two samples from arbitrary exponential distributions. This is a generalized linear model (Section 6.5) with link function

$$g(\mu) = -\log \mu ,$$

since $\mu = 1/\lambda$ is the mean of $\mathcal{E}(\lambda)$. But the $\mathcal{E}(\lambda)$ family tends to be a poor model for lifetimes that experience wear, in view of its memoryless property: the conditional distribution of $T - t$ given $T \geq t$ is the same as the marginal distribution of T .

The $\mathcal{E}(\lambda)$ family is characterized by $\lambda(t) \equiv \text{constant}$. Cox (1972) proposed an important semiparametric generalization of this model. He introduced an unknown continuous, positive *baseline hazard rate* $\lambda(t)$ on $[0, \infty)$ and then postulated the conditional hazard rate of the distribution of the nonnegative continuous variable T given $\mathbf{Z} = \mathbf{z}$ to be of the form

$$\lambda(t|\mathbf{z}) = \lambda(t) \exp\{\mathbf{z}^T \boldsymbol{\beta}\} . \quad (9.1.12)$$

This *Cox* $(\boldsymbol{\beta}, \lambda)$ model is in widespread use throughout biostatistics in part because of the interpretation of the coefficients β_j : By writing the model as $\log \lambda(t|\mathbf{z}) = \sum_{j=1}^d \beta_j z_j + \log \lambda(t)$, we see that β_j is the change in the hazards on the log scale as z_j is perturbed one unit.

More generally, we consider

$$\lambda(t|\mathbf{z}) = \lambda(t)r(\mathbf{z}, \boldsymbol{\beta}) , \quad (9.1.13)$$

where $r > 0$ is a known function. The model as specified by its conditional hazard rates and the marginal distribution of \mathbf{Z} is evidently semiparametric. Although $\lambda(\cdot)$ is arbitrary, just as with the linear model (a) of Example 9.1.1, it far from covers all $\lambda(t|\mathbf{z})$. It, in fact, contains a powerful assumption corresponding to the additive treatment effect assumption of the linear model: the relative effect of \mathbf{Z} on the distribution of T is $\lambda(t|\mathbf{z})/\lambda(t|\mathbf{0}) =$

$r(\mathbf{z}, \boldsymbol{\beta})/r(\mathbf{0}, \boldsymbol{\beta})$, which doesn't depend on λ or t . That is, (9.1.13) is a *proportional hazard rate* model.

The first question we ask is whether $\boldsymbol{\beta}$ is identifiable. We will give a full answer to this question later. For the time being, we note identifiability in an important special case. Suppose \mathbf{Z} is discrete, $\mathbf{Z} \in \{\mathbf{z}_0, \dots, \mathbf{z}_k\}$, $P[\mathbf{Z} = \mathbf{z}_j] > 0$, $0 \leq j \leq k$, and that (9.1.12) holds. Write $r_j(\boldsymbol{\beta}) \equiv r(\mathbf{z}_j, \boldsymbol{\beta})$ and let $S(t) \equiv e^{-\Lambda(t)}$ be the survival function corresponding to λ and assume $r(\mathbf{z}, \mathbf{0}) \equiv 1$. Then,

$$P[T > t | \mathbf{Z} = \mathbf{z}_j] = S^{r_j(\boldsymbol{\beta})}(t), \quad (9.1.14)$$

a model essentially proposed for the two-sample case by Lehmann. See Example 8.3.10. If \mathbf{Z} is discrete as above and we observe $(\mathbf{Z}_1, T_1), \dots, (\mathbf{Z}_n, T_n)$, let $N_j = \sum_{i=1}^n 1(\mathbf{Z}_i = \mathbf{z}_j)$ and let

$$\widehat{S}_j(t) \equiv \frac{1}{N_j} \sum_{i=1}^n 1(T_i > t, \mathbf{Z}_i = \mathbf{z}_j), \quad j = 0, \dots, k$$

be the empirical conditional survival functions. By the Glivenko-Cantelli theorem, as $n \rightarrow \infty$,

$$\sup_t |\widehat{S}_j(t) - S^{r_j(\boldsymbol{\beta})}(t)| \xrightarrow{P} 0$$

and $\boldsymbol{\beta}$ is identifiable iff $\boldsymbol{\beta} \rightarrow (r_0(\boldsymbol{\beta}), \dots, r_k(\boldsymbol{\beta}))$ is 1–1. In fact, $r_0(\boldsymbol{\beta}), \dots, r_k(\boldsymbol{\beta})$ and thus $\boldsymbol{\beta}$ can be estimated \sqrt{n} consistently in many ways — for instance, by picking a value t such that $0 < \widehat{S}_0(t) < 1$ and estimating $r_j(\boldsymbol{\beta})$ by

$$\widehat{r}_{j,t} = \frac{\log \widehat{S}_j(t)}{\log \widehat{S}_0(t)}. \quad (9.1.15)$$

Which is the best choice of t ? It turns out there is a “best” estimate of $\boldsymbol{\beta}$ in this model but it is not from (9.1.15). We shall develop it in Section 9.3. \square

Example 9.1.5. *The Independent component analysis (ICA) model. Principal components.*

As we have seen in Appendix B.6 of Volume I, the multivariate Gaussian $\mathcal{N}(\boldsymbol{\mu}, \Sigma)$ model can be written as

$$\mathbf{X}_{d \times 1} = A\mathbf{Z} + \boldsymbol{\mu}$$

where \mathbf{Z} is a vector of i.i.d. $\mathcal{N}(0, 1)$ variables and A and $\boldsymbol{\mu}$ are unknown. We know that this A may not be identifiable since $\text{Var}(\mathbf{X}) = AA^T \equiv \Sigma$ and $E(\mathbf{X}) = \boldsymbol{\mu}$ identify the distribution and A may have many more parameters than Σ and $\boldsymbol{\mu}$.

We will develop a particular choice of A which is identifiable by using the representation (see Appendix B.10)

$$\Sigma = \mathbf{O}\Lambda\mathbf{O}^T = \sum_{j=1}^d \lambda_j \mathbf{v}_j \mathbf{v}_j^T$$

where $\mathbf{O} = (\mathbf{v}_1, \dots, \mathbf{v}_d)^T$, which is orthogonal and $d \times d$, Λ is the diagonal matrix of eigenvalues $\lambda = (\lambda_1, \dots, \lambda_d)$ of Σ , with $\lambda_1 \geq \dots \geq \lambda_d > 0$. See Problem 8.3.24.

Here $\lambda_1, \dots, \lambda_d$ have corresponding orthonormal eigenvectors, $\mathbf{v}_1, \dots, \mathbf{v}_d$ with $|\mathbf{v}| = 1$, $\mathbf{v}_i \perp \mathbf{v}_j$ if $i \neq j$ and

$$\Sigma \mathbf{v}_j = \lambda_j \mathbf{v}_j, \quad 1 \leq j \leq d. \quad (9.1.16)$$

If the \mathbf{v}_j are chosen so that the first nonzero element of \mathbf{v}_j is positive, they are uniquely defined if $\lambda_1 > \dots > \lambda_d > 0$. If there are ties between some of the λ 's, let $\lambda_1^{(0)} > \dots > \lambda_k^{(0)}$ be the ordered distinct eigenvalues, then the set of $\{\mathbf{v}_j : \lambda_j = \lambda_i^{(0)}\}$, $1 \leq i \leq k$, is the identifiable parameter. Viewed in this way, \mathbf{O} and Λ are identifiable and so is $A = \mathbf{O}\Lambda^{\frac{1}{2}}$ and we have the ICA model

$$\mathbf{X} = \mathbf{O}\Lambda^{\frac{1}{2}}\mathbf{Z} + \boldsymbol{\mu}. \quad (9.1.17)$$

The linear combinations $\mathbf{v}_1^T \mathbf{X}, \dots, \mathbf{v}_d^T \mathbf{X}$ of X 's are the *principal components* of \mathbf{X} with $\mathbf{v}_j^T \mathbf{X}$ representing the contribution of $\mathbf{v}_j \mathbf{X}$ to the representation

$$\mathbf{X} = \boldsymbol{\mu} + \sum_{j=1}^d (\mathbf{v}_j^T \mathbf{X}) \mathbf{v}_j.$$

Here $\mathbf{v}_1^T \sum \mathbf{v}_1 = \sup_{\mathbf{a}} \{\text{Var}(\mathbf{a}^T \mathbf{X}) : |\mathbf{a}| = 1\}$. That is, the vector $\mathbf{a} \equiv \mathbf{v}_1$ of $\mathbf{v}_1^T \sum \mathbf{v}_1$ provides the linear combination $\sum a_i X_i$ with the largest possible variance subject to $|\mathbf{a}| = 1$. See Section B.10.1.2, Example 8.3.11 (continued), and Problem 8.3.24. Moreover, $\mathbf{v}_2^T \sum \mathbf{v}_2$ has the same property among the set of vectors \mathbf{a} orthogonal to \mathbf{v}_1 , and so on.

There are a number of natural semiparametric generalizations of sampling from this model: $\mathbf{X}_1, \dots, \mathbf{X}_n$ i.i.d. $\mathcal{N}_d(\boldsymbol{\mu}, \Sigma)$. The most important appears to be $\mathbf{X} = A\mathbf{Z}$ where A is an unknown parameter and $\mathbf{Z} = (Z_1, \dots, Z_d)^T$ with the Z_j independently distributed and with the distribution Q_j of Z_j arbitrary. Thus, if $\mathbf{Q} = (Q_1, \dots, Q_d)$,

$$\mathcal{P} = \{P_{(A, \mathbf{Q})} : P_{(A, \mathbf{Q})} \text{ is the distribution of } \mathbf{X} = A\mathbf{Z} \text{ with } \mathbf{Q} \text{ "arbitrary"}\}. \quad (9.1.18)$$

It is easy to see that the Q_j are identified at most up to location and scale. What is interesting, but certainly not obvious, is that if all the Q_j are *not* Gaussian, then, as we shall see later, A is identified up to a permutation and scale change of its columns. This is equivalent to saying that any parameter $q(A)$ such that $q(AD) = q(A)$, $q(A\pi) = q(A)$ for all diagonal matrices D and permutation matrices Π is identifiable. In the engineering literature algorithms estimating such parameters have proven very effective in a number of important problems — see Hyvarinen, Karhunen and Oja (2001). This methodology is referred to as *Independent component analysis* (ICA). The simplest problem leading to ICA corresponds to the situation where Z_1, \dots, Z_d represent the output of independent sources which are superimposed when received by d different observers, X_1, \dots, X_d . The task is to separately identify the signal coming from the sources. In practice, the Z_1, \dots, Z_d are often time series, as are the X_1, \dots, X_d , and the \mathbf{X}_j are observations taken at different times t_j , $j = 1, \dots, n$. However, the methods developed in the i.i.d. case work in the time series situation as well. The only adjustment needed is to the estimates of the variance of our estimates of A . \square

We now turn to methods of estimation in semiparametric models.

9.1.2 Regularization. Modified Maximum Likelihood

There is a fundamental difficulty when trying to apply the maximum likelihood method or the plug-in approach to some non- and semiparametric models. The difficulty occurs for models \mathcal{P} that do not contain the empirical probability \widehat{P} and where the parameter of interest, say $\nu(P)$, is not defined at \widehat{P} . In the case of the MLE, this means that the maximum of the likelihood is not assumed. Here is an example.

Example 9.1.6. We illustrate with X_1, \dots, X_n i.i.d. $P \in \mathcal{P} = \{\text{All probabilities } P \text{ with continuous densities on } R\}$. If we parametrize \mathcal{P} by p , the density of P , then the likelihood for an observed vector $\mathbf{x} = (x_1, \dots, x_n)^T$ is

$$L_{\mathbf{x}}(p) = \prod_{i=1}^n p(x_i).$$

Nothing prevents us from making $p(x_i)$ arbitrarily large for all $i = 1, \dots, n$. For instance, consider the distributions $\prod_{i=1}^n p_k(x_i)$ in \mathcal{P} with, for $k = 1, 2, \dots$,

$$p_k(x) = \frac{1}{n} \sum_{i=1}^n \frac{1}{\sigma_k} \phi\left(\frac{x - x_i}{\sigma_k}\right)$$

where ϕ is the standard normal density and $\sigma_k \rightarrow 0$ as $k \rightarrow 0$. Thus,

$$\sup_{\mathcal{P}} L_{\mathbf{x}}(p) = \infty$$

is not assumed for any p and this approach does not yield an estimate of the infinite dimensional parameter $\nu(P) = p$, nor simple one dimensional parameters such as $\nu(P) = \int p(x)dx$. \square

In this example, it is intuitively clear that sequences $\{p_k\}$ which make $L_{\mathbf{x}}(p) \rightarrow \infty$ converge weakly to $\widehat{P} = n^{-1} \sum_{i=1}^n \delta_{x_i}$, the empirical probability distribution, which does not have a density. From a measure theoretic point of view, the members of \mathcal{P} are dominated by Lebesgue measure, but the “maximum likelihood” \widehat{P} is with probability 1 undominated. Kiefer and Wolfowitz (1956) proposed an approach which they called *nonparametric maximum likelihood* which leads to choosing the empirical distribution as the estimate of P , when we enlarge \mathcal{P} to include all discrete distributions. This approach only makes sense in measure theoretic terms and we do not present it, but we shall discuss modified maximum likelihood methods related to *empirical likelihood* (Owen (1988, 2001)) in Section 9.1.3.

Let \mathcal{P} be an arbitrary specified semiparametric model and let \mathcal{M}_0 be the union of the closure of \mathcal{P} with the set of all discrete distributions. Let $\bar{\mathcal{P}}_0$ be the closure of \mathcal{M}_0 under (say) weak convergence. Then, given observations x_1, \dots, x_n of i.i.d. X_1, \dots, X_n , consider

$$\mathcal{P}_{\mathbf{x}} = \{P \in \bar{\mathcal{P}}_0 : P\{x_1, \dots, x_n\} = 1, P \text{ corresponds to } P \in \mathcal{P}\},$$

the set of all members of $\bar{\mathcal{P}}_0$ with support $\{x_1, \dots, x_n\}$ that satisfy the model restrictions of \mathcal{P} . Now $\mathcal{P}_{\mathbf{x}}$ can be parametrized by

$$\mathbf{p} = (p_1, \dots, p_n)^T \in \left\{ \mathbf{p} : p_i = P[X = x_i], 1 \leq i \leq n, \text{ some } P \in \mathcal{P}_{\mathbf{x}} \right\}, \quad (9.1.19)$$

a subset of the simplex in R^n . We identify \mathcal{P}_x with this set. In most applications, \mathcal{P}_x is a smooth parametric model indexed by $p(\cdot; \theta)$, $\theta \in R^d$, $d \leq n$, and maximum likelihood over \mathcal{P}_x is of the usual type. That is,

$$\hat{\theta} = \arg \max \left\{ \frac{1}{n} \sum_{i=1}^n \log p(x_i; \theta) : \sum_{i=1}^n p(x_i; \theta) = 1 \right\}.$$

Here $p(\cdot; \theta)$ is determined by the restrictions that define the distributions in \mathcal{P} . Sometimes, we can only require $0 < P\{x_1, \dots, x_n\} < 1$ with $1 - P\{x_1, \dots, x_n\}$ being assigned to ∞ . This happens, as we shall see, in censored data problems.

We define the *empirical* or *modified likelihood estimate* of P for the model \mathcal{P} by

$$\hat{P}_e = \arg \max \{ \Pi_{i=1}^n p_i : P \in \mathcal{P}_x \}. \quad (9.1.20)$$

We will refer to \hat{P}_e as the *nonparametric MLE* (NPMLE). The NPMLE of a parameter $\nu(P)$ is defined by $\hat{\nu} = \nu(\hat{P}_e)$ when $\hat{\nu}(\hat{P}_e)$ exists.

Example 9.1.7. We illustrate (9.1.20) with $\mathcal{P} = \{\text{All } P \text{ with continuous densities on } R\}$. Then $\bar{\mathcal{P}}_0 = \mathcal{M} \equiv \{\text{All distributions on } R\}$. When there are no ties among $\{x_1, \dots, x_n\}$, \mathcal{P}_x is naturally parametrized by

$$\Theta \equiv \{(p_1, \dots, p_{n-1}) : p_j \geq 0, \sum_{j=1}^{n-1} p_j \leq 1, 1 \leq j \leq n-1\}.$$

We can write

$$L_x(p) = \Pi_{i=1}^n p_i$$

with $p_n = 1 - \sum_{j=1}^{n-1} p_j$. Because this is a multinomial likelihood with one observation in each category, it is uniquely maximized over Θ by $p_1 = \dots = p_n = n^{-1}$. See Example 2.2.8. The NPMLE of P is indeed \hat{P} , the empirical distribution. \square

Remark 9.1.1. Ties. Categorical data. Ties occur when X is discrete, when there is roundoff, and when data are collected in categories (e.g. age groups or geographic strata). In the case of ties, we let $\{x_1^{(0)}, \dots, x_m^{(0)}\}$ be the distinct x_i 's or indicators of the categories $\{1, \dots, m\}$ and let

$$\mathcal{P}_x = \{P \in \bar{\mathcal{P}}_0 : P\{x_1^{(0)}, \dots, x_m^{(0)}\} = \sum_{j=1}^m p_i = 1\}.$$

This \mathcal{P}_x is parametrized by

$$p = (p_1, \dots, p_m)^T \in \{\mathbf{p} : p_j = P(X = x_j^{(0)}), 1 \leq j \leq m, \text{ some } P \in \mathcal{P}_x\}.$$

The empirical likelihood of X_1, \dots, X_n in the case of ties or categorical data is

$$L_x(\mathbf{p}) = \prod_{j=1}^m p_j^{n_j},$$

where $n_j = \sum_{i=1}^n \mathbf{1}(x_i = x_j^{(0)})$. In the case of two categories (e.g. treatment and control), $m = 2$, $p_2 = 1 - p_1$, and $L_{\mathbf{x}}(\mathbf{p}) = p_1^{n_1}(1 - p_1)^{n-n_1}$ is the Bernoulli likelihood.

For tied or categorical data, the estimate of P for the model \mathcal{P} is

$$\widehat{P}_e = \arg \max \left\{ \prod_{j=1}^m p_j^{n_j} : P \in \mathcal{P}_{\mathbf{x}} \right\}.$$

When $\bar{\mathcal{P}}_0$ is $\{\text{all distributions on } R\}$ the solution (NPMLE) is $\widehat{p}_j = n_j/n$, $1 \leq j \leq m$. See Problem 2.2.30. \square

There are many cases where the NPMLE of a parameter $\boldsymbol{\theta}$ can be obtained as the solution of a generalized estimating equation of the form $Q(\boldsymbol{\theta}, \widehat{P}) = 0$, where $\boldsymbol{\theta}$ solves $Q(\boldsymbol{\theta}, P) = 0$, and $Q(\boldsymbol{\theta}, P)$ is not necessarily linear in P . That is, Q may not satisfy

$$Q(\boldsymbol{\theta}, P) = \int v(\boldsymbol{\theta}, \mathbf{x}) dP(\mathbf{x})$$

for any function v . Stratification is an example.

Example 9.1.8. Biased sampling. Stratification (Example 9.1.2). Here \mathcal{P} is the collection of all P with densities $\prod_1^n p_F(x_i)$ where p_F satisfies (9.1.2), that is,

$$\mathcal{P}_{\mathbf{x}} = \left\{ (p_1, \dots, p_n) : p_i = \frac{w(x_i)f_i}{W(F)} \right\}$$

where $W(F) = \sum_{i=1}^n w(x_i)f_i$ and the parameter of interest is $\boldsymbol{\theta} = (f_1, \dots, f_n)^T$. Here (p_1, \dots, p_n) range freely over the n simplex. As we saw in Example 9.1.7, the maximum is attained for $\widehat{p}_i = n^{-1}$, $1 \leq i \leq n$, and, by inspection, a solution to $\widehat{p}_i = w(x_i)f_i/W(F)$ is

$$\widehat{f}_i = w^{-1}(x_i) \left(\sum_{k=1}^n w^{-1}(x_k) \right)^{-1}.$$

We can write this empirical likelihood estimate $\widehat{f}_e(x)$ of $f(x)$ in terms of the nonparametric empirical probability \widehat{P} as

$$d\widehat{F}_e(x) = w^{-1}(x)d\widehat{P}(x)/\int w^{-1}(y)d\widehat{P}(y)$$

in agreement with (9.1.4).

We next consider the more general stratified model with $X = (I, Y)$. Here

$$p_{(I,Y)}(j, y) = \lambda_j \frac{w_j(y)f(y)}{W_j(F)}, \quad 1 \leq j \leq k, y \in R.$$

It turns out that $\mathcal{P}_{\mathbf{x}}$ is unsatisfactory since the observed x_1, \dots, x_n force a coupling of the values of I and Y which is incompatible with the model distribution. However, we can define a *modified likelihood* naturally as follows:

Suppose $w_j(y) > 0$ for all y . Then the possible support of a member of \mathcal{P} which puts positive mass on all the observed (I_i, Y_i) , $i=1,\dots,n$, is $\{1, \dots, k\} \times \{y_1, \dots, y_n\}$. The point (a, y_b) in the support of $P \in \mathcal{P}$ is assigned probability

$$p(a, y_b) = \lambda_a \frac{w_a(y_b)}{W_a(F)} f_b; \quad W_a(F) = \sum_{b=1}^n f_b w_a(y_b).$$

The condition $w_j(y) > 0$ for all j is sufficient to make F identifiable. The modified likelihood is

$$\prod_{a,b} \lambda_a \frac{w_a(Y_b)}{W_a^{N_{a+}}} f_b^{N_{a+}}$$

where $N_{ab} = 1(I_b = a)$, $N_{a+} = \sum_{b=1}^n N_{ab}$, $N_{++} = \sum_{a=1}^k N_{ab}$, $W_a \equiv W_a(F)$. The modified likelihood equations under the condition $\sum_b f_b = 1$ introduced through a Lagrange multiplier γ yield

$$\begin{aligned} \hat{\lambda}_a &= \frac{N_{a+}}{n} \quad \text{for } a = 1, \dots, k \\ \frac{N_{++}}{\hat{f}_b} - \sum_{a=1}^k \frac{N_{a+}}{\hat{W}_a} w_a(Y_b) &= \gamma. \end{aligned}$$

Multiplying by \hat{f}_b and summing we get

$$n = \gamma + \sum_{a=1}^k \frac{N_{a+}}{\hat{W}_a} \sum_{b=1}^n w_a(Y_b) \hat{f}_a = \gamma + n.$$

Hence a solution is $\gamma = 0$ and

$$\hat{f}_b = \frac{N_{++}}{\sum_{a=1}^k \frac{N_{a+}}{\hat{W}_a} w_a(Y_b)}. \quad (9.1.21)$$

Then, for $j = 1, \dots, k$, summing $w_j(Y_b) \hat{f}_b$ we obtain the following estimate of $W_j(F)$,

$$\hat{W}_j = \sum_{b=1}^n [w_j(Y_b) N_{++} (\sum_{a=1}^k \frac{N_{a+}}{\hat{W}_a} w_a(Y_b))^{-1}]. \quad (9.1.22)$$

If we let \hat{P}_1 be the empirical df of I and \hat{P}_2 that of Y we can rewrite (9.1.22) as

$$\hat{W}_j - \int [w_j(y) (\int \frac{w_a(y)}{\hat{W}_a} d\hat{P}_1(a))^{-1}] d\hat{P}_2(y) = 0.$$

This is an equation yielding a generalized estimating equation or *generalized M estimate* $\hat{\mathbf{W}} \equiv (\hat{W}_1, \dots, \hat{W}_k)^T$. That is $\hat{\mathbf{W}} = \mathbf{W}(\hat{P})$ where P is the distribution of X , \hat{P} is the empirical probability, and $\mathbf{W}(P)$ solves $\mathbf{Q}(\mathbf{W}, P) = \mathbf{0}$. Here $\mathbf{Q}_{k \times 1}$ is the function

$$\mathbf{Q}(\mathbf{W}, P) \equiv \mathbf{W} - \int \mathbf{w}(y) [\int \frac{w_a(y)}{W_a} dP_1(a)]^{-1} dP_2(y) \quad (9.1.23)$$

where P is the joint probability distribution of $(I, Y)^T$ and P_1, P_2 are the respective marginals of I and Y . It may be shown, if $0 < P[\delta = 0] < 1$ (Problem 9.1.7), that $\widehat{\mathbf{W}}$ exists and is unique with probability tending to 1. Once we have computed $\widehat{\mathbf{W}}$ we have our NPMLE \widehat{P}_e of P

$$\begin{aligned} d\widehat{P}_e(i, y) &= \widehat{\lambda}_i w_i(y) d\widehat{F}_e(y) \\ d\widehat{F}_e(y) &= d\widehat{P}_2(y) \left(\int \frac{w_a(y)}{\widehat{W}_a} d\widehat{P}_1(a) \right)^{-1} \end{aligned}$$

and we can, in principle, estimate any parameter $\nu(P)$, $P \in \mathcal{P}$, by $\nu(\widehat{P}_e)$. \square

Example 9.1.9. *Censoring and truncation (Example 9.1.3).* We develop the empirical likelihood theory for censoring. Introduce the hazard rates,

$$\lambda_F = \frac{f}{F}, \quad \lambda_G = \frac{g}{G}$$

and note that we can write the joint density of (Y, δ) (for measure theory aficionados, with respect to the appropriate measure!) as

$$p_{(F,G)}(y, \delta) = \lambda_F^\delta(y) \lambda_G^{1-\delta}(y) \exp\{-(\Lambda_F(y) + \Lambda_G(y))\} \quad (9.1.24)$$

To write the likelihood for $\mathcal{P}_{\mathbf{x}}$, where $X_i = (Y_i, \delta_i)$, $Y_i = \min(T_i, C_i)$, and $\delta_i = 1(T_i \leq C_i)$, we take as possible values for Y the ordered Y_i denoted by $y_{(1)}, \dots, y_{(n)}$. The corresponding values of δ are $(\delta_{(1)}, \dots, \delta_{(n)})$, indicating whether $Y_{(i)}$ is censored or not. We will use the discrete form of the hazard rates of T, C at $y_{(i)}$ calling them $\lambda_{Fi}, \lambda_{Gi}$.

Consider $y_{(i)}, \delta_{(i)}$, $i = 1, \dots, n$, coming from $X_i = x_i$, $1 \leq i \leq n$, as fixed values, and $\lambda_{Fi}, \lambda_{Gi}$ as unknown parameters. Then, for $X = (Y, \delta) \sim P \in \mathcal{P}_{\mathbf{x}}$, using (9.1.8) and (9.1.9), and assuming $P(T_i = C_i) = 0$, $1 \leq i \leq n$,

$$\begin{aligned} p_X(y_{(j)}, \delta_{(j)}) &= P^\delta[T = y_{(j)}] P^\delta[C \geq y_{(j)}] \cdot P^{1-\delta}[C = y_{(j)}] P^{1-\delta}[T \geq y_{(j)}] \\ &= [\lambda_{Fj} \prod_{k=1}^{j-1} (1 - \lambda_{Fk})]^\delta [\prod_{k=1}^{j-1} (1 - \lambda_{Gk})]^\delta \\ &\quad \cdot [\lambda_{Gj} \prod_{k=1}^{j-1} (1 - \lambda_{Gk})]^{1-\delta} [\prod_{k=1}^{j-1} (1 - \lambda_{Fk})]^{1-\delta} \end{aligned} \quad (9.1.25)$$

where we write δ for $\delta_{(j)}$. We make a further modification of the likelihood by noting that under our assumptions the supports $\{t_1, \dots, t_n\}$ and $\{c_1, \dots, c_n\}$ of the empirical versions of F and G are disjoint. Then, with $\varepsilon_{ij} \equiv 1(Y_i = y_{(j)})$ the log empirical likelihood for a sample X_1, \dots, X_n is

$$\begin{aligned} l_{\mathbf{x}}(\lambda_F, \lambda_G) &= \sum_{i=1}^n \sum_{j=1}^n \varepsilon_{ij} \{ (\delta_{(j)} \log \lambda_{Fj} + (1 - \delta_{(j)}) \log \lambda_{Gj}) \\ &\quad + \sum_{k=1}^{j-1} [\log(1 - \lambda_{Fk}) + \log(1 - \lambda_{Gk})] \} . \end{aligned} \quad (9.1.26)$$

Let $Y_{(L)}$ be the largest $y_{(i)}$ with $\delta_{(i)} = 1$. For $i > L$, λ_{Fi} does not appear in the sum. We have no information about the distribution of times larger than the largest uncensored time other than an estimate of the survival function at $T_{(L)}$. For $i \leq L$, we obtain after some algebra (Problem 9.1.14), if there are no ties, that is, $\sum_{i=1}^n \varepsilon_{ij} = 1$ for all j , that the maximizer of (9.1.26) is

$$\hat{\lambda}_{Fi} = \frac{\delta_{(i)}}{n - i + 1}. \quad (9.1.27)$$

If $\delta_{(i)} = 0$, then $\hat{P}[T = y_{(i)}] = 0$. Using (9.1.9), we arrive at the classical *Kaplan-Meier estimate* of the survival function $S(y) = P(T > y)$ for $y \leq Y_{(L)}$,

$$\hat{S}_{KM}(y) = \prod\left\{(1 - \frac{\delta_{(i)}}{n - i + 1}) : y_{(i)} \leq y\right\}. \quad (9.1.28)$$

Remark 9.1.2.

- 1) When there are no ties, if \hat{P} corresponds to \hat{S}_{KM} , then $\hat{P}[T > Y_{(L)}] = 0$ iff $L = n$.
- 2) The formula (9.1.25) can be extended to the case where P is discrete to begin with. That is, there are ties but we still suppose the supports of T and C are disjoint or, as Lawless (1982) views it, apparent ties between censoring times and deaths are broken by moving censoring times infinitesimally to the right.
- 3) If we replace (9.1.25) by its continuous approximation, (9.1.24), formally given by

$$p_X(y_{(j)}, \delta_{(j)}) = \lambda_{Fj}^\delta \lambda_{Gj}^{1-\delta} \exp\left\{-\sum_{k=1}^j (\lambda_{Fk} + \lambda_{Gk})\right\}, \quad (9.1.29)$$

we arrive at the same formula for $\hat{\lambda}_{Fj}$ (Problem 9.1.13), but a new expression,

$$\hat{S}_{NA}(y_{(j)}) = \exp\left\{-\sum_{k=1}^j \hat{\lambda}_{Fk}\right\} \quad (9.1.30)$$

as an estimate of S . This is the *Nelson-Aalen estimate* which we will discuss further later.

To study identifiability of λ_F , λ_G , F , and G , consider \mathcal{P} defined as all P of the form (9.1.24) and note that the unique maximizer over \mathcal{P} of, (Problem 9.1.13)

$$K(P_{(F,G)}, P) \equiv \int \log p_{(F,G)}(y, \delta) dP(y, \delta)$$

for G fixed and $0 < P[\delta = 0] < 1$ is given by

$$\lambda_F(P)(y) \equiv \frac{p_1(y)}{P[Y \geq y]} \quad (9.1.31)$$

where

$$p_1(y) = \frac{d}{dy} P[Y \leq y, \delta = 1].$$

An analogous formula holds for G . From (9.1.31) and its analogue for G , we get F, G such that

$$P_{(F,G)} = P.$$

Hence (F, G) given by (9.1.31) and its G analogue is the unique maximizer of K and we have identifiability if $0 < P[\delta = 0] < 1$. Here the condition $P[\delta = 0] < 1$ is necessary for identifiability of F while $p_1(y)$ is not defined if $P[\delta = 1] = 0$. If $P[\delta = 0] = 0$, then G cannot be identified. Note that these identifiability results are density analogues of Problem 1.1.10, identifiability for the discrete case.

Truncation is easily handled by the same hazard rate parametrization.

$$\begin{aligned} p_{(M,Y)}(m,y) &= g(m) \frac{f(y)}{\bar{F}(m)} 1(y \geq m) \\ &= g(m) \lambda_F(y) \exp\{\Lambda_F(m) - \Lambda_F(y)\} 1(y \geq m). \end{aligned} \quad (9.1.32)$$

The NPMLE makes

$$\hat{\lambda}_F(y) = 0, \quad y \notin \{y_1, \dots, y_n\}$$

and if $y_{(1)} < \dots < y_{(n)}$,

$$\hat{\lambda}_F(y_{(j)}) = (N_j - j)^{-1}$$

where $N_j = \sum_{k=1}^n 1(M_k \leq y_{(j)})$, provided $N_j > j$. That is, we can only estimate $\hat{\lambda}_F(y_{(j)})$ for $j \geq J$ with $J = \text{first } k \text{ such that } y_{(k)} \neq M_{l_k}$ where $y_{l_k} \equiv y_{(j)}$. Equivalently we can only estimate $F(y)$ for $y \geq y_{(J)}$. We deduce that $\hat{\lambda}_F$ is defined with probability tending to 1 if $\bar{F}(y) > 0$ for all y . Moreover, f can be identified by

$$\lambda_F(y) = p_Y(y)(P[M \leq y] - P[Y \leq y])^{-1} \quad (9.1.33)$$

and if $G(m) > 0$ for all m , g can be similarly identified. We can also obtain

$$\begin{aligned} \hat{\bar{F}}(y) &= \Pi\{(1 - \hat{\lambda}_F(y_{(j)})) : y_{(j)} \leq y\} \\ \hat{P}_F[\{y_{(j)}\}] &= \hat{\lambda}_F(y_{(j)}) \hat{\bar{F}}(y_{(j)}). \end{aligned} \quad (9.1.34)$$

We leave details of these results to the reader (Problem 9.1.8). \square

Although empirical likelihood gives simple answers for nonparametric censoring and truncation, survival analysis models with no covariates, it is much less satisfactory for more complicated models such as that of Cox because of the awkward formula (9.1.9). Simpler solutions are obtained by a natural approximation to the empirical likelihood, yet another modified likelihood, suggested by (9.1.6) — but whose real foundations lie in counting process theory — see Aalen (1978), Andersen et al. (1993), and Kalbfleisch and Prentice (2002)⁽³⁾. Replace, even in the discrete case, if T is a survival time, $p(t)$ by $\lambda(t) \exp\{-\Lambda(t)\}$, where Λ is the cumulative hazard rate. We illustrate the method in

Example 9.1.10. The Cox model (Example 9.1.4). We enlarge \mathcal{P} of the proportional hazard Cox model to include all distributions of (\mathbf{Z}, T) with T discrete and with hazard rates, given $\mathbf{Z} = \mathbf{z}$,

$$\lambda(t|\mathbf{z}) = r(\mathbf{z}, \boldsymbol{\beta})\lambda(t)$$

where $\lambda(t)$ and $\lambda(t|\mathbf{z})$ are defined as in (9.1.8) and (9.1.13). Here \mathbf{Z} has density (discrete or continuous) $h(\cdot)$. We apply the empirical likelihood approach to this model. Given $x_1 = (\mathbf{z}_1, t_1), \dots, x_n = (\mathbf{z}_n, t_n)$, we can parametrize $\mathcal{P}_{\mathbf{x}}$ by $\boldsymbol{\theta} \equiv (h_1, \dots, h_n, \lambda_1, \dots, \lambda_n, \boldsymbol{\beta})$ with

$$h_i = h(\mathbf{z}_i), \quad \lambda(t_i|\mathbf{z}_i) \equiv \lambda_i r(\mathbf{z}_i, \boldsymbol{\beta}), \quad 1 \leq i \leq n,$$

where $h_i \geq 0, \lambda_i \geq 0 : 1 \leq i \leq n$, and $\boldsymbol{\beta} \in R^d$. Using (9.1.9), we obtain the modified likelihood,

$$L_{\mathbf{x}}(\boldsymbol{\theta}) = \prod_{i=1}^n h_i \lambda_i r(\mathbf{z}_i, \boldsymbol{\beta}) \prod_j \{(1 - r(\mathbf{z}_i, \boldsymbol{\beta}) \lambda_j) : t_j < t_i\}. \quad (9.1.35)$$

We assume that $\mathbf{h} = (h_1, \dots, h_n)^T$ is not related to $(\boldsymbol{\beta}, \boldsymbol{\lambda})$, in which case we can treat $\prod_{i=1}^n h_i$ as a constant c . See Example 6.2.1. Or we can replace h_j with its NPMLE $\widehat{h}_j = n_j/n$ from Remark 9.1.2, in which case $\prod \widehat{h}_j$ is a constant c . Thus we replace $\prod_{i=1}^n h_i$ with c and $\boldsymbol{\theta}$ with $(\boldsymbol{\lambda}, \boldsymbol{\beta})$ in what follows.

Our strategy will be to first fix $\boldsymbol{\beta}$ and find

$$\widehat{\boldsymbol{\lambda}}(\boldsymbol{\beta}) = \arg \max_{\boldsymbol{\lambda}} L_{\mathbf{x}}(\boldsymbol{\beta}, \boldsymbol{\lambda}).$$

Next consider $l_{\mathbf{x}}(\boldsymbol{\beta}) = L_{\mathbf{x}}(\boldsymbol{\beta}, \widehat{\boldsymbol{\lambda}}(\boldsymbol{\beta}))$, which is called the *profile empirical likelihood*. The *profile* NPMLEs are now $\widehat{\boldsymbol{\beta}} = \arg \max l_{\mathbf{x}}(\boldsymbol{\beta})$, and $\widehat{\boldsymbol{\lambda}}(\widehat{\boldsymbol{\beta}})$. Maximizing $L_{\mathbf{x}}(\boldsymbol{\theta})$ with respect to $\boldsymbol{\lambda}$ for $\boldsymbol{\beta}$ fixed, we get, solving $\partial \log L_{\mathbf{x}}(\boldsymbol{\theta}) / \partial \lambda_k = 0$,

$$\frac{1}{\lambda_k} = \sum_{i=1}^n \frac{r(\mathbf{z}_i, \boldsymbol{\beta})}{1 - r(\mathbf{z}_i, \boldsymbol{\beta}) \lambda_k} \mathbf{1}(t_k < t_i).$$

This equation is awkward. It has no explicit solution for λ_k , and computer solutions need to be produced for a large grid of $\boldsymbol{\beta}$'s. However, if we replace

$$\Lambda(t_i|\mathbf{z}_i) = \prod_j \{(1 - r(\mathbf{z}_i, \boldsymbol{\beta}) \lambda_j) : t_j < t_i\}$$

in (9.1.35) by its continuous case approximation from (9.1.6), that is

$$\Lambda(t_i|\mathbf{z}_i) \cong r(\mathbf{z}_i, \boldsymbol{\beta}) \sum_{j=1}^n \lambda_j \mathbf{1}(t_j \leq t_i),$$

we get the approximate empirical likelihood,

$$\mathcal{L}_{\mathbf{x}}(\boldsymbol{\theta}) = c \prod_{i=1}^n r(\mathbf{z}_i, \boldsymbol{\beta}) \lambda_i \exp\{-r(\mathbf{z}_i, \boldsymbol{\beta}) \sum_{j=1}^n \lambda_j \mathbf{1}(t_j \leq t_i)\}. \quad (9.1.36)$$

From (9.1.36) and $\partial \log \mathcal{L}_{\mathbf{x}}(\boldsymbol{\theta}) / \partial \lambda_k = 0$, we find that the maximizers of $\mathcal{L}_{\mathbf{x}}(\boldsymbol{\theta})$ are

$$\widehat{\lambda}_k(\boldsymbol{\beta}) = \left(\sum_{i=1}^n r(\mathbf{z}_i, \boldsymbol{\beta}) \mathbf{1}(t_i \geq t_k) \right)^{-1}, \quad 1 \leq k \leq n. \quad (9.1.37)$$

By substituting $\widehat{\lambda}_k(\boldsymbol{\beta})$ for λ_k in (9.1.36) and noting that

$$\sum_{i=1}^n r(\mathbf{z}_i, \boldsymbol{\beta}) \sum_{j=1}^n \widehat{\lambda}_j(\boldsymbol{\beta}) 1(t_j \geq t_i) = n ,$$

we find that the approximate profile likelihood is proportional to

$$\mathcal{L}(\boldsymbol{\beta}) \equiv \mathcal{L}_{\mathbf{x}}(\boldsymbol{\beta}, \widehat{\lambda}(\boldsymbol{\beta})) = \prod_{i=1}^n \left[\frac{r(\mathbf{z}_i, \boldsymbol{\beta})}{n^{-1} \sum_{j=1}^n r(\mathbf{z}_j, \boldsymbol{\beta}) 1(t_j \geq t_i)} \right] . \quad (9.1.38)$$

The profile empirical likelihood estimate of $\boldsymbol{\beta}$ based on the continuous case approximation (9.1.6) is now defined by

$$\widehat{\boldsymbol{\beta}} = \arg \max_{\boldsymbol{\beta}} \mathcal{L}(\boldsymbol{\beta}) .$$

It coincides with the Cox (1972,1975) partial likelihood estimate. See Example 9.1.11. The Cox estimates $\widehat{\beta}_j$, $1 \leq j \leq d$, and their approximate standard errors are available in most statistical software packages.

We next introduce expressions that will be useful for identifiability and asymptotics. Rewriting (9.1.37) with t in place of t_k we get

$$d\widehat{\Lambda}(t, \boldsymbol{\beta}) = \left(\int r(\mathbf{z}, \boldsymbol{\beta}) 1(s \geq t) d\widehat{P}(\mathbf{z}, s) \right)^{-1} d\widehat{P}_2(t) \quad (9.1.39)$$

where \widehat{P}_2 is the marginal empirical distribution of T and $\widehat{P}(\cdot, \cdot)$ is the empirical of (\mathbf{Z}, T) . Given $(\mathbf{Z}, T) \sim P$, define P_1 and P_2 in general as the the marginal distributions of \mathbf{Z} and T . Let

$$S_0(t, \boldsymbol{\beta}, P) = \int r(\mathbf{z}, \boldsymbol{\beta}) 1(s \geq t) dP(\mathbf{z}, s) = E_P[r(\mathbf{Z}, \boldsymbol{\beta}) 1(T \geq t)] . \quad (9.1.40)$$

By taking log in (9.1.38) and using (9.1.39) we see that maximizing $\mathcal{L}(\boldsymbol{\beta})$ as a function of $\boldsymbol{\beta}$ is equivalent to maximizing $\Gamma(\boldsymbol{\beta}, \widehat{P})$ where

$$\Gamma(\boldsymbol{\beta}, P) = E_{P_1}[\log r(\mathbf{Z}, \boldsymbol{\beta})] - E_{P_2}[\log S_0(T, \boldsymbol{\beta}, P)] . \quad (9.1.41)$$

Define

$$\boldsymbol{\beta}(P) = \arg \max \Gamma(\boldsymbol{\beta}, P) ; \quad (9.1.42)$$

then the estimate $\widehat{\boldsymbol{\beta}}$ is

$$\widehat{\boldsymbol{\beta}} = \boldsymbol{\beta}(\widehat{P}) .$$

Note that we can regard $\widehat{\boldsymbol{\beta}}$ as an empirical plug-in estimate based on the generalized estimating equation $\dot{\Gamma}(\boldsymbol{\beta}, \widehat{P}) = 0$, where $\dot{\Gamma}$ is the gradient with respect to $\boldsymbol{\beta}$. This makes it easier to handle identifiability and asymptotics. See Section 9.2.

Identifiability of β means that, under the Cox model, if β_0 is true, $-\Gamma(\beta, P)$ has a unique minimum at $\beta = \beta_0$, that is, it is a contrast function as defined in Section 2.1.1. This holds iff the map $\beta \rightarrow \mathcal{L}_P(r(\mathbf{Z}, \beta))$ is 1–1 and $P[S_0(T, \mathbf{0}, P) = c] < 1$ for all c . A proof is sketched in Example 9.2.3 when $r(\mathbf{z}, \beta) = \exp\{\beta^T \mathbf{z}\}$. We shall show that, under the identifiability conditions, $-\Gamma(\beta, P)$ is strictly convex in β when $r(\mathbf{z}, \beta) = \exp\{\beta^T \mathbf{z}\}$ even when the Cox model does *not* hold. Thus, for this $r, \beta(P)$ defined by (9.1.42) is a well defined parameter for general P . Similarly, for this $r, -\Gamma(\beta, \hat{P})$ is strictly convex in β with probability tending to 1 so that $\hat{\beta}$ is well defined. Note that $\Gamma(\beta, P)$ is not linear in P so that M estimation theory as given in Section 6.2.1 can not be used directly for asymptotics. We return to this in Section 9.2.2. \square

We next turn to a widely used modified likelihood.

Example 9.1.11. *The Cox partial likelihood. Regression analysis for censored and tied data.* Consider Example 9.1.9 with censoring, but now assume we have available a covariate vector \mathbf{Z} . As before, let $\lambda(t|\mathbf{z})$ be the conditional hazard rate given $Z = \mathbf{z}$. We change notation and let $t_1 < \dots < t_m$ be the observed distinct failure times, leaving out censoring times.

The Cox partial likelihood $\prod_{i=1}^m q_i$ is a conditional empirical “likelihood,” where q_i is the probability that a failure occurs at time t_i computed conditionally for those patients whose failure or censoring times are at least t_i , that is, for patients with $y_j = \min\{t_j, c_j\} \geq t_i$. It follows from (9.1.8) that with $R_i = \{j : y_j \geq t_i\}$,

$$\Pi_{i=1}^m q_i = \Pi_{i=1}^m \frac{\lambda(t_i|\mathbf{z}_i)}{\sum_{j \in R_i} \lambda(t_i|\mathbf{z}_j)} = \Pi_{i=1}^m \left[\frac{\lambda(y_i|\mathbf{z}_i)}{\sum_{j:y_j \geq y_i} \lambda(y_i|\mathbf{z}_j)} \right]^{\delta_i} \quad (9.1.43)$$

provided that given \mathbf{Z}_i , $1 \leq i \leq n$, C_1, \dots, C_n are independent of T_1, \dots, T_n . As pointed out by Cox (1972, 1975), $\Pi_{i=1}^m q_i$ is not a “complete” likelihood, so the term “partial” is appropriate. The $\lambda(t_i)$ term cancels in $\Pi_{i=1}^m q_i$ under the proportional hazard assumption $\lambda(t_i|\mathbf{z}_i) = \lambda(t_i)r(\mathbf{z}_i, \beta)$, and we have the Cox partial likelihood for proportional hazard model:

$$\Pi_{i=1}^m q_i = \Pi_{i=1}^m \left[\frac{r(\mathbf{z}_i, \beta)}{\sum_{j:y_j \geq y_i} r(\mathbf{z}_j, \beta)} \right]^{\delta_i}. \quad (9.1.44)$$

When there is no censoring and there are no ties, this likelihood is equivalent to the modified profile likelihood (9.1.38). \square

In the next subsection we formalize the approximation behind the modified likelihoods (9.1.36) and (9.1.38).

9.1.3 Other Modified and Approximate Likelihoods

There are models where it is convenient to use discrete type modified likelihoods that do not require $\sum p_i = 1$. See e.g. Murphy (1994), Murphy, Rossini and van der Vaart (1997), van der Vaart (1998), Murphy and van der Vaart (2000), Zeng and Lin (2007), and Kosorok

(2008). We consider this approach but start with what we call the *delta method likelihood*, L_d . It leads to approximate likelihoods such as (9.1.36) and the Cox likelihood (9.1.44) in a natural way. This approach is based on a very close approximation to the ordinary likelihood and does not involve discrete type modified likelihoods.

Here is the justification of the delta method likelihood: Many semiparametric models can be written as models with parameters $\beta \in R^d$ and $\eta(\cdot)$; and a likelihood that depends on β , $\eta(\cdot)$, and $\eta'(\cdot)$, where $\eta'(y) \geq 0$. Then the likelihood of i.i.d. $(Y_1, \mathbf{Z}_1), \dots, (Y_n, \mathbf{Z}_n)$ can be expressed in terms of $\theta_{(i)}$, $\eta(y_{(i)})$, $\gamma_{(i)} \equiv \eta'(y_{(i)})$, and $\mathbf{z}_{(i)}$ where the $y_{(i)}$'s are the y_i 's ordered, $\mathbf{z}_{(i)}$ is the covariate vector corresponding to $y_{(i)}$, and $\theta_{(i)} = r(\mathbf{z}_{(i)}, \beta)$ with $r(\cdot, \cdot)$ known. To simplify the likelihood, rewrite $\eta(y_{(i)})$ as the telescoping sum

$$\eta(y_{(i)}) = \sum_{j=1}^n \eta_{(j)} 1(y_{(j)} \leq y_{(i)})$$

with

$$\eta_{(j)} = \eta(y_{(j)}) - \eta(y_{(j-1)}), \quad \eta(y_{(0)}) = 0.$$

Now the parameters are β , γ and η where $\gamma = (\gamma_{(1)}, \dots, \gamma_{(n)})^T$, $\eta = (\eta_{(1)}, \dots, \eta_{(n)})^T$. To eliminate γ consider the delta-approximation $\eta(y+d) \cong \eta(y) + d\eta'(y)$, that is,

$$\gamma_{(i)} = \eta'(y_{(i)}) \cong \frac{\eta(y_{(i)}) - \eta(y_{(i-1)})}{y_{(i)} - y_{(i-1)}} = \frac{\eta_{(i)}}{d_{(i)}} \quad (9.1.45)$$

where $d_{(i)} = y_{(i)} - y_{(i-1)}$, $y_{(0)} = 0$. We substitute $\gamma_{(i)} \cong \eta_{(i)}/d_{(i)}$ in the likelihood and get an approximate likelihood $L_d(\beta, \eta)$ that depends on β and η only so that we have reduced the dimension of the parameter space. In the proportional hazard model, with $\eta(y) = \Lambda(y)$, the resulting likelihood is equivalent to the approximate empirical likelihood (9.1.36) (Problem 9.1.15), and, with η replaced by $\hat{\lambda}$ of (9.1.38), is equivalent to the Cox partial likelihood (9.1.44).

The authors referred to in the first paragraph define a similar approximate likelihood $L_a(\beta, \eta)$ by replacing $\eta(y)$ with the step function $\eta[y] \equiv \sum_{i=1}^n \gamma_{(i)} 1(y_{(j)} \leq y)$ with the γ_i being the basic parameter. That is, L_a can be obtained from L_d by setting the spacings $d_{(i)}$ in L_d equal to one. The two approaches based on the δ likelihood and L_a are equivalent for estimation of β when all terms in involving $d_{(i)}$, $1 \leq i \leq n$, do not contain any parameters. See Problem 9.1.16 for a case where they are not equivalent. L_d has the advantage that it is a very close approximation (Problem 9.1.24) to the actual semiparametric likelihood and it does not use discrete models where continuous models are natural.

Murphy and van der Vaart (2000), and Zeng and Lin (2007) among others, have shown under suitable assumptions, that if we use the profile approach where we fix β and set

$$\hat{\eta}(\beta) = \arg \max_{\eta} \{L_a(\beta, \eta)\},$$

the estimate

$$\hat{\beta} = \arg \max \{L_a(\beta, \hat{\eta}(\beta))\}$$

is asymptotically semiparametrically efficient in a sense to be introduced in Section 9.3. We call $\widehat{\beta}$ and $\widehat{\eta} \equiv \widehat{\eta}(\widehat{\beta})$ the *maximum approximate profile likelihood estimates* (MAPLEs).

Example 9.1.12. *Semiparametric transformation models.* Consider the family of composite models where the df of Y given $\mathbf{Z} = \mathbf{z}$ is a composite of a parametric df $G(\cdot; \theta)$ and a nondecreasing function $\eta(\cdot)$. That is, consider the *transformation model*

$$F(y; \theta, \eta) = G(\eta(y); \theta), \quad \theta = r(\mathbf{z}, \beta) \in \Theta, \quad (9.1.46)$$

where $\beta \in R^d$, $y \in A \subset R$, and $\eta(y) \in [\underline{\eta}, \bar{\eta}]$ with $G(\underline{\eta}, \theta) = 0$, $G(\bar{\eta}, \theta) = 1$, all $\theta \in \Theta$. The proportional hazard model is of this form with $G(t; \theta) = 1 - \exp\{-\theta t\}$ and $\eta(y) = \Lambda(y) = -\log(1 - F(y))$ for an unknown baseline df F . More generally, interesting cases are $G(t; \theta) = G_1(\theta t)$, $\eta(y) = G_1^{-1}(F(y))$; and $G(t; \theta) = G_0(t - \theta)$, $\eta(y) = G_0^{-1}(F(y))$; for specified df's G_1 and G_0 (see Problem 9.1.17).

If $\eta(\cdot)$ has a derivative $\eta'(y) > 0$, and G has density $g(v; \theta)$, the exact likelihood is $\prod_{i=1}^n q_i$ with

$$q_i = \eta'(y_{(i)}) g\left(\sum_{j \leq i} \eta_{(j)}; \theta_i\right) \quad (9.1.47)$$

where $\eta_{(j)} = \eta(y_{(j)}) - \eta(y_{(j-1)})$, $\eta_{(0)} = 0$, and $\theta_i = r(\mathbf{z}_i, \beta)$. With the approximation $\eta'(y_{(i)}) \cong \eta_{(i)}/d_{(i)}$, $L_d(\beta, \eta)$ is equivalent to $\prod_{i=1}^n p_i$ with

$$p_i = \eta_{(i)} g\left(\sum_{j \leq i} \eta_{(j)}; \theta_i\right), \quad (9.1.48)$$

where we do not require $\sum_{i=1}^n p_i = 1$. In this case, $L_d(\beta, \eta)$ is constant in $d_{(i)} = y_{(i)} - y_{(i-1)}$, $1 \leq i \leq n$, and is equivalent to $L_a(\beta, \eta)$. Moreover, L_a only depends on the ranks of y_1, \dots, y_n , and thus the estimate of β also only depends on the ranks.

The MAPLEs are formally

$$\widehat{\beta} = \arg \max_{\beta} \prod_{i=1}^n \widehat{\eta}_i(\beta) g\left(\sum_{j \leq i} \widehat{\eta}_j(\beta); r(\mathbf{z}_i, \beta)\right), \quad \widehat{\eta} = \widehat{\eta}(\beta),$$

where

$$\widehat{\eta}(\beta) = \arg \max_{\eta} \left\{ \prod_{i=1}^n p_i : \underline{\eta} < \eta_j < \bar{\eta}, 1 \leq j \leq n \right\}.$$

As an example, suppose $G(y; \theta) = L(y - \theta)$ where $L(t)$ is the logistic distribution $[1 + \exp(-t)]^{-1}$ and $\eta(t) = \log\{F(t)/[1 - F(t)]\}$ for some baseline F ; then the model (9.1.47) is the proportional odds model (Problem 9.1.17) and $\widehat{\beta}$ is the profile NPMLE of Murphy, Rossini and van der Vaart (1998). Existence, consistency, and asymptotic normality of the MAPLEs for this model is given in that paper.

9.1.4 Sieves and Regularization

Empirical likelihood and its relatives are methods which give attractive answers in a number of situations, but their scope is, in fact, limited. To see this, suppose we try to apply it to the symmetric location model of Example 3.5.1. Here

$$\mathcal{P} = \{P : P \text{ has density } p(x) = f(x - \theta), \ f \text{ is arbitrary symmetric about } 0, \ \theta \in R\}.$$

Unfortunately, \mathcal{P}_x cannot be defined since, in general, there is no distribution concentrated on x which puts positive mass on all points, which is also symmetric about any point.

Sieves

We next consider a conceptually very powerful approach, the *method of sieves*, introduced into statistics by Grenander (1981). In the method of *sieves*, the approach is to approximate the model \mathcal{P} by a sequence of regular parametric models $\{\mathcal{P}_m\}$ called a sieve that in some sense converge to \mathcal{P} as $m \rightarrow \infty$. The distribution \hat{P}_m in the sieve that corresponds to the maximum likelihood estimate for P_m estimates P and provides a plug-in estimate $\nu(\hat{P}_m)$ of a parameter $\nu(P)$ in a semiparametric model. It is sometimes convenient to take for \hat{P}_m not maximum likelihood but minimum distance or some other estimate well behaved when $P \in \mathcal{P}_m$. These methods can be thought of as examples of *regularization*, an approach first proposed by Tikhonov (1963) in the context of providing stable solutions to “ill posed” differential equations. See Section 11.4.1.

The *method of sieves* is a method which, in principle, can be applied to any statistical problem where we have available an i.i.d. sample with empirical probability \bar{P} . Suppose, given a model \mathcal{P} for our data, we can determine $\mathcal{P}_1 \subset \mathcal{P}_2 \subset \dots$ such that

- (i) $\mathcal{P} \subset \overline{\cup_j^{\infty} \mathcal{P}_j}$, where \bar{A} denotes closure of A in at least the sense of weak convergence.
- (ii) \mathcal{P}_j is a regular parametric model

$$\mathcal{P}_j = \{P_{\boldsymbol{\theta}^{(j)}} : \boldsymbol{\theta}^{(j)} \in \Theta^{(j)} \subset R^{d_j}\}$$

where $\{d_j\}$ is a sequence strictly increasing to ∞ as $j \rightarrow \infty$ with $P_{\boldsymbol{\theta}^{(j)}}$ having density function $p(\cdot, \boldsymbol{\theta}^{(j)})$ and the map $\boldsymbol{\theta}^{(j)} \rightarrow p(\cdot, \boldsymbol{\theta}^{(j)})$ is smooth.

Let $\rho(P, Q)$ be a measure of the discrepancy between two probabilities P and Q . For a given fitting criterion of the form

$$\Pi_j(Q) = \arg \min \{\rho(P, Q) : P \in \mathcal{P}_j\},$$

the idea now is to find parameters $\eta_j(P) \in \Theta \subset \overline{\cup_{j=1}^{\infty} \Theta^{(j)}}$ such that if

$$\Pi_j(P) \equiv P_{\eta_j(P)} \in \mathcal{P}_j$$

then

$$\Pi_j(P) \rightarrow \Pi(P) = P \text{ as } j \rightarrow \infty$$

for $P \in \mathcal{P}$, and

$$\Pi_j(\hat{P}) = \Pi_j(P) + \Delta_{jn}$$

where $\Delta_{jn} = o_P(n^{-\frac{1}{2}})$ for all j . Here $\Pi_j(P) - P$ can be thought of as contributing to bias and $\Pi_j(\hat{P}) - \Pi_j(P)$ as contributing to variance. Typically Δ_{jn} cannot be made uniformly small in j for all n .

In the method of sieves, we choose a *model selection rule* which chooses \mathcal{P}_{J_n} as the “best” model according to some criteria (see Section I.7), and then act as if \mathcal{P}_{J_n} were true. That is, with $m = J_n$ and \hat{P} the empirical probability, we estimate P by

$$\hat{P}_m \equiv \Pi_m(\hat{P}).$$

The estimate of a parameter $\nu(P)$ is then $\hat{\nu} \equiv \nu(\hat{P}_m)$.

The three elements of the method of sieves for estimating a parameter $\nu(P)$, $P \in \mathcal{P}$ are:

- a) The choice of $\{\mathcal{P}_j\}$.
- b) The choice of $\eta_j(P)$. Frequently the first thing tried is the maximum likelihood (Kullback-Liebler) discrepancy of Section 2.2.2,

$$\eta_j(P) = \arg \max \left\{ \int \log p(x, \boldsymbol{\theta}^{(j)}) dP(x) : \boldsymbol{\theta}^{(j)} \in \Theta^{(j)} \right\}.$$

That is, for $P \in \mathcal{P}$, $\eta_j(P)$ is the parameter $\boldsymbol{\theta}$ in the j th model that makes $P_{\boldsymbol{\theta}}$ closest to P in the Kullback-Leibler sense.

- c) The choice J_n of j .

Unfortunately, the applications of the method of sieves usually lead to technical difficulties which have to be resolved by delicate empirical process theory arguments; see Bickel, Klaassen, Ritov and Wellner (1993,1998), van der Vaart and Wellner (1996), and van der Geer (2000) for theory and practice, examples, and references to the literature.

Regularization

Sieves are just a type of regularization. Consider the representation of maximum likelihood in terms of Kullbach-Leibler discrepancy given in Section 2.2.2. We start by writing

$$\frac{1}{n} \log L_x(p) = \int \log p(x) d\hat{P}(x).$$

We saw in Example 9.1.6 that

$$\arg \max \left\{ \int \log p(x) d\hat{P}(x) : P \text{ corresponds to } P \in \mathcal{P} \right\} \quad (9.1.49)$$

is not always well defined. Yet we know by Shannon's Lemma (2.2.1) that, quite generally, if $P_0 \in \mathcal{P}$,

$$P_0 = \arg \max \left\{ \int \log p(x) dP_0(x) : p \text{ corresponds to } P \in \mathcal{P} \right\}$$

is well defined. This suggests that we consider approximate solutions of the problem (9.1.49) which are “nice” and so will be apt to be closer to the “nice” true P_0 .

Sieves do this by restricting the model over which optimization is carried out. An alternative is to penalize solutions which are not “nice” by changing the quantity to be optimized. For instance, we assume that \mathcal{P} is all densities p with a derivative p' such that $\int [p'(x)]^2 dx < \infty$. It is natural to maximize

$$\int \log p(x) d\hat{P}(x) - \lambda_n \int [p'(x)]^2 dx. \quad (9.1.50)$$

For $\lambda_n > 0$ the maximizer of (9.1.50) may be shown to be unique and necessarily have a density with $\int [p'(x)]^2 dx < \infty$. On the other hand if $\lambda_n \rightarrow 0$ it is intuitively clear that the maximizer of the population version of (9.1.50) where \hat{P} is replaced by P should approximate the true p_0 . This method is studied further in Chapter 11. It is, in fact, closely linked to the idea of using the posterior mode of a Bayes prior on \mathcal{P} . It can also be viewed as a special case of the method of sieves where, however, the models $\{\mathcal{P}_n\}$ of the sieve are themselves infinite dimensional, $\mathcal{P}_m = \{p : \int [p'(x)]^2 dx \leq \varepsilon_m\}$. We discuss this further in Chapter 11 but note here the general principles. A review of regularization with discussion of various examples is given by Bickel and Li (2006).

Examples

We shall give an illustration of how sieves can be used in a, by now, classical example. In this example we are using a straightforward extension to two sequences $\{j_1\}$ and $\{j_2\}$.

Example 9.1.13. *Partial linear model.* Suppose $X = (U, Z, Y)$ and we postulate

$$Y = \beta Z + g(U) + \varepsilon, \quad (9.1.51)$$

where (U, Z) have unknown joint distribution H , $\varepsilon \sim \mathcal{N}(0, \sigma^2)$ independent of (U, Z) and g is arbitrary except for regularity conditions. We shall assume $(U, Z) \in R^2$, although the model and results are readily extendable to vector U and Z and $\varepsilon \sim F$ arbitrary. This model, introduced by Engle et al. (1986), arises if, for instance, Z takes on only a finite set of values such as the levels of a treatment, but U is a real valued covariate of less intrinsic interest, e.g. age. Thus $\nu(P) = \beta$ is the parameter of interest. Rewrite the model (with a new g) as

$$\begin{aligned} Y &= \beta(Z - E(Z|U)) + g(U) + \varepsilon \\ \mathcal{P} &= \{P_{(\beta, g, \sigma^2)} : \beta \in R, \sigma^2 > 0, g \text{ arbitrary}\}. \end{aligned}$$

Then, in analogy with ordinary linear regression, because

$$Y - E_P(Y|U) = \beta(Z - E_P(Z|U)) + \varepsilon$$

for $P \in \mathcal{P}$, we can formally define a parameter

$$\beta(P) = \frac{E_P(Z - E_P(Z|U))(Y - E_P(Y|U))}{E_P(Z - E_P(Z|U))^2} \quad (9.1.52)$$

such that

$$\beta(P_{(\beta, g, \sigma^2)}) = \beta.$$

Now (9.1.52) clearly gives a valid definition of $\beta(P)$ if $E_P(Z^2) < \infty$, $E_P g^2(U) < \infty$, and $E_P(Z - E_P(Z|U))^2 > 0$. Then P is identifiable provided $E_P(Y^2) < \infty$, $E_P(Z^2) < \infty$, and Z is not a function of U with probability 1. In fact, we shall argue that only the condition $0 < E_P(Z - E_P(Z|U))^2 < \infty$ is needed (Problem 9.1.12). Note that we cannot define $\beta(\hat{P})$ because $E_{\hat{P}}(Z|U)$ and $E_{\hat{P}}(Y|U)$ are not well defined. However, it is natural to plug-in regularized estimates of these quantities, call them $\hat{E}(Y|U = \cdot)$ and $\hat{E}(Z|U = \cdot)$, to obtain a procedure which works. For simplicity, we shall use the histogram density estimate applied to $p(y|u)$, $p(z|u)$. Assume that (U, Y) has a density on $[0, 1] \times [0, 1]$. Define $I_{jh} = (jh, (j+1)h]$ for each axis of the square as in Example I.5 so that $I_{j_1, h_1}^{(1)} \times I_{j_2, h_2}^{(2)}$ form a partition of the square for $0 \leq j_i \leq 1/h_i$, $i = 1, 2$. Let

$$\hat{p}_n(y|u) = \hat{P}[Y \in I_{j_2, h_2}^{(2)}(y) \mid U \in I_{j_1, h_1}^{(1)}(u)]$$

where $I_{j_1, h_1}^{(1)}(u)$ is the partition interval that contains u and $I_{j_2, h_2}^{(2)}(y)$ the one that contains y . This is an application of the method of sieves where \mathcal{P}_{j_1, j_2} corresponds to the model with all densities of (U, Y) and (Z, Y) constant on the rectangles described above and $\{\theta^{(j_1, j_2)}\}$ are the probabilities of these rectangles. Let $h_2 \rightarrow 0$. Strictly speaking we no longer have a continuous case conditional density when $h_2 = 0$, but rather the discrete density assigning equal mass to all Y_i such that $U_i \in I_{j_1, h_1}$. However, the procedure turns out to be satisfactory, yielding,

$$E_{\hat{P}_n}(Y|U = u) \equiv \hat{E}(Y|U = u) \equiv \frac{\sum_{i=1}^n Y_i 1(U_i \in I_{j_1, h_1}(u))}{\sum_{i=1}^n 1(U_i \in I_{j_1, h_1}(u))}. \quad (9.1.53)$$

We next let $h_1 \rightarrow 0$ and get a similar definition for $\hat{E}(Z|U = u)$. Now $\hat{\beta} \equiv \beta(\hat{P}_n)$ is obtained by plugging these \hat{E} 's into (9.1.52). These are the plug in estimates corresponding to the regularization given. We shall later study a simplified version of $\hat{\beta}$. In this example, choosing $\eta_j(P)$ amounts to choosing the *bandwidths* h_1 and h_2 . See Chapters 11 and 12. See also Engle et al (1986), Chen (1988), and Fan et al (1998).

□

Example 9.1.14. ICA (Example 9.1.5). The density of \mathbf{X} in the independent component analysis (ICA) model (9.1.18) is

$$p(\mathbf{x}; A, \mathbf{Q}) = |A|^{-1} \prod_{j=1}^d q_j(\mathbf{a}^{(j)} \mathbf{x}) \quad (9.1.54)$$

where $\mathbf{a}^{(j)}$ is the j th row vector of A^{-1} .

We begin with the problem of identifiability, following Kagan, Linnik and Rao (1973); see also Comon (1994). Suppose $(A_1, \mathbf{Q}_1), (A_2, \mathbf{Q}_2)$ lead to the same distribution for \mathbf{X} , where all components of \mathbf{Q}_l , $l = 1, 2$, have mean 0 and variance 1. Without loss of generality (Problem 9.1.10), we can take $A_2 = J$, the $d \times d$ identity. For $l = 1, 2$, let $\mathbf{Z}_l = (Z_{1l}, \dots, Z_{dl})^T$ be the vector of independent variables distributed as $(Q_{1l}, \dots, Q_{dl})^T$. Consider the characteristic function of $\mathbf{X} = A_1 \mathbf{Z}$,

$$\Psi(\mathbf{t}) = E[\exp\{i\mathbf{t}^T \mathbf{X}\}] = E[\exp\{i\mathbf{t}^T A_1 \mathbf{Z}\}] = \prod_{k=1}^d \psi_k^{(1)}(\mathbf{a}_k \mathbf{t}) \quad (9.1.55)$$

where $\psi_k^{(l)}$ is the characteristic function of Z_{kl} , $l = 1, 2$, and \mathbf{a}_k is the k th row vector of A_1^T . Similarly, since A_2 is the identity,

$$\Psi(\mathbf{t}) = \prod_{k=1}^d \psi_k^{(2)}(t_k) . \quad (9.1.56)$$

Taking logs in (9.1.55) and (9.1.56) appropriately, we obtain if $\log \psi_k^{(j)} \equiv \gamma_k^{(j)}$,

$$\sum_{k=1}^d [\gamma_k^{(1)}](\mathbf{a}_k \mathbf{t}^T) = \sum_{k=1}^d [\gamma_k^{(2)}](t_k)$$

for all \mathbf{t} in a neighborhood of $\mathbf{0}$. Differentiating twice with respect to t_b, t_c , we obtain

$$\sum_{k=1}^d [\gamma_k^{(1)}]''(\mathbf{a}_k \mathbf{t}^T) a_{bk} a_{ck} = \delta_{bc} \quad (9.1.57)$$

where $\mathbf{a}_k = (a_{1k}, \dots, a_{dk})$ and δ_{bc} is the Kronecker delta. By assumption, $[\gamma_k^{(1)}]''(0) = 1$ for all k . Hence, we see that $\mathbf{a}_1, \dots, \mathbf{a}_d$ are orthonormal. We can deduce from (9.1.57) that either $[\gamma_k^{(1)}]''(t)$ is constant, i.e. Z_k is Gaussian, or \mathbf{a}_k is plus or minus one of the coordinate vectors and conclude that, if at most one of the Z_k is Gaussian, identifiability of A and γ_k , $1 \leq k \leq d$ follows (Problem 9.1.10). This argument suggests that an estimate of A can be constructed as follows: Let A^* be the matrix obtained from A by scaling each column to have norm 1 and drop the requirement that the Z_k have variance 1. Let

$$\widehat{Q}(A^*) = \int |\widehat{\psi}(\mathbf{t}) - \prod_{k=1}^d \psi_k(\mathbf{a}_k^* \mathbf{t}^T, \mathbf{a}_k^{(k)})|^2 \exp\{-\frac{|\mathbf{t}|^2}{2}\} d\mathbf{t}$$

where $\widehat{\psi}$ is the empirical characteristic function $E_{\widehat{P}}[\exp\{i\mathbf{t}^T \mathbf{X}\}]$ of \mathbf{X} and $\psi_k(\cdot, \mathbf{a}_k^{(k)})$ is the empirical characteristic function of $\mathbf{a}_k^{(k)} \mathbf{X}$ where \mathbf{a}_k^* and $\mathbf{a}_k^{(k)}$ are the k th row vectors of A^* and $[A^*]^{-1}$. Chen (2003) shows that

$$\widehat{A}^* = \arg \min \widehat{Q}(A^*)$$

is a \sqrt{n} consistent estimate of A^* under mild conditions.

More significantly, we can also apply the method of sieves to obtain what turns out to be an efficient estimate of A^* . The idea is to construct suitable estimates of $q'_k/q_k(\cdot, A^*)$, the score functions of the densities of Z_1, \dots, Z_d assuming that A^* is the truth. A by now standard approach (see Stone, Hansen, Kooperberg and Truong (1997)) is to approximate $\log q_k(\cdot, A^*)$ by a linear combination of basis functions, e.g. splines. That is, we approximate an arbitrary density q by a member of the exponential family,

$$p(x, \boldsymbol{\theta}) = \exp\left\{\sum_{k=1}^d \theta_k w_k(x) - A_d(\boldsymbol{\theta})\right\}$$

where the linear span of the w_k is dense in the space of all functions for suitable metrics. See Section 11.4.2. Given the $\hat{q}_k(\cdot, A^*)$ we plug back into (9.1.54) and maximize the resulting likelihood in A^* . This is an application of the method of sieves. Appropriate A_n^* can be constructed and despite technical difficulties, this approach works — see Chen (2003).

We next consider a model that's a rival to the Cox model (Reid (1994)).

Example 9.1.15. *The accelerated failure time (AFT) model.* Consider the model where a failure time T is an accelerated version of a baseline failure time T_0 in the sense that

$$T = \alpha T_0 \quad ; \quad T_0 \geq 0 ,$$

where $\alpha = \exp\{-\boldsymbol{\beta}^T \mathbf{Z}\}$, T_0 corresponds to $\boldsymbol{\beta} = 0$, and T_0 is independent of the vector \mathbf{Z} of predictors. We can rewrite this model as

$$Te^{\boldsymbol{\beta}^T \mathbf{Z}} = T_0 .$$

For the case where $\mathbf{Z} = \mathbf{Z}(t)$ is a time dependent covariate vector, Cox and Oakes (1984, Section 5.2) and Zeng and Lin (2007) considered the model

$$\int_0^T e^{\boldsymbol{\beta}^T \mathbf{Z}(t)} dt = T_0 . \tag{9.1.58}$$

Let $\Lambda(t|\mathbf{z})$ and $\Lambda(t)$ denote the cumulative hazards of T and T_0 , then

$$\Lambda(t|\mathbf{z}) = \Lambda\left(\int_0^t e^{\boldsymbol{\beta}^T \mathbf{z}(s)} ds\right) . \tag{9.1.59}$$

As before, let C denote a censoring time set $\delta = 1(T \leq C)$, $Y = \min(T, C)$, assume that C is independent of T given \mathbf{Z} , and that the distribution of C given \mathbf{Z} does not involve $\boldsymbol{\beta}$ or Λ . Then the log likelihood for data $\{(Y_i, Z_i(Y_i), \delta_i; 1 \leq i \leq n\}$ is proportional to (Problem 9.1.20),

$$l(\boldsymbol{\beta}, \lambda) = n^{-1} \sum_{i=1}^n \left\{ \delta_i \boldsymbol{\beta}^T \mathbf{Z}_i(Y_i) + \delta_i \log \lambda(\theta_i) - \Lambda(\theta_i) \right\} \tag{9.1.60}$$

where $\lambda(\cdot) = \Lambda'(\cdot)$ and

$$\theta_i = \int_0^{Y_i} \exp\{\beta^T \mathbf{Z}_i(s)\} ds.$$

In this example, the profile approach where we fix β and maximize $l(\beta, \lambda)$ with respect to $\lambda(\cdot)$ does not produce useable estimates (see Problem 9.1.20). If we formally pass to the limit as $n \rightarrow \infty$ in (9.1.60) we obtain $\theta(Y) \equiv \theta(Y, \beta) = \int_0^Y \exp\{\beta^T Z(s)\} ds$. Note that $T_0 = \theta^{-1}(T, \beta)$. It is natural to consider $\theta(\cdot, \beta)$ as a nuisance parameter and regularize the profile likelihood by replacing $\frac{d}{dt} P[\delta = 1, Y \leq \theta^{-1}(t, \beta)]$ by a kernel smoothed version which is then estimated empirically. See Chapter 11 for such estimates. The resulting estimates of β are, under regularity conditions, semiparametrically efficient in the sense of Section 9.3 (Zeng and Lin (2007)). \square

Remark 9.1.3 (a). The proportional hazard and the AFT models coincide for the Weibull model *and* the periodic hazard rate model. See Problem 9.1.21.

(b). The proportional quantile hazard rate and AFT models are equivalent. See Problem 9.1.22. \square

Summary. We discuss in Section 9.1.1 several semiparametric models for a variety of common experimental frameworks including the symmetric location model, the linear model with stochastic covariates and unknown error distribution and biased sampling models. In the context of survival analysis we introduce the *survival function*, the *hazard function*, *censoring*, *truncation*, and the *Cox proportional hazard model*. We also consider the *Independent Component Analysis* (ICA) model where the basic variables are linear combinations of independent but not identically distributed variables whose distributions are “arbitrary.” We consider in Sections 9.1.2 and 9.1.3 maximum likelihood type methods for semiparametric models that are based on modifications of the maximum likelihood principle. They are usually based on approximating or replacing the arguments of the objective functions associated with the maximum likelihood (Kullback-Leibler divergence) with arguments that are “well behaved.” Success is achieved when \sqrt{n} consistent (generally efficient) estimates of Euclidean parameters are obtained when the empirical probability \hat{P} is plugged in for the population probability P . The first method considered is *modified likelihood* which leads to nonparametric MLEs and the second method considered in Section 9.1.4 is the *method of sieves*. We then illustrate these *regularized maximum likelihood* and sieve procedures in the semiparametric models introduced in Section 9.1.1. We obtain some classical procedures including the Kaplan-Meier, Nelson-Aalen, and Cox estimates in survival analysis.

9.2 Asymptotics. Consistency and Asymptotic Normality

In this section we establish a general consistency criteria and use it to prove consistency and asymptotic normality for some of the estimates we have suggested.

9.2.1 A General Consistency Criterion

We state a generalization of Theorem 5.2.3, which turns out to be broadly useful for proving consistency. Let $\|f\|_K \equiv \sup\{|f(\mathbf{t})| : \mathbf{t} \in K\}$, $K \subset R^d$, and let

$$\{Q_n(\mathbf{t}) : \mathbf{t} \in R^d\}$$

be a sequence of real valued stochastic processes such that

A1: $Q_n(\mathbf{t})$ is convex as a function of \mathbf{t} with probability 1.

A2: $\|Q_n - Q\|_K \xrightarrow{P} 0$ as $n \rightarrow \infty$
for all compact $K \subset R^d$, where the limit $Q : R^d \rightarrow R$ is a deterministic function.

A3: Q is strictly convex and continuous with $\lim\{Q(\mathbf{t}) : |\mathbf{t}| \rightarrow \infty\} = \infty$.

It may be shown (Rockafellar (1969), p.74) that A2 follows from

A2': $Q_n(\cdot)$ is separable and $Q_n(t) \xrightarrow{P} Q(t)$ for all t .

Let

$$\mathbf{t}_0 \equiv \arg \min Q(\mathbf{t}), \quad \hat{\mathbf{t}}_n = \arg \min Q_n(\mathbf{t}).$$

By convention, $\hat{\mathbf{t}}_n$ is chosen among the set of minimizers if $\arg \min$ is not unique and $\hat{\mathbf{t}}_n \equiv \mathbf{0}$ if the minimum is not assumed.

Proposition 9.2.1. If A1, A2, and A3 hold,

(i) \mathbf{t}_0 is uniquely defined.

(ii) $\hat{\mathbf{t}}_n \xrightarrow{P} \mathbf{t}_0$.

Proof. Condition A3 guarantees (i). For (ii), let $\epsilon > 0$, and note that “ $|\hat{\mathbf{t}}_n - \mathbf{t}_0| > \epsilon$ ” implies that the minimizer of $Q_n(\mathbf{t})$ is in $\{\mathbf{t} : |\mathbf{t} - \mathbf{t}_0| \geq \epsilon\}$. Thus

$$P(|\hat{\mathbf{t}}_n - \mathbf{t}_0| > \epsilon) \leq P(\inf\{Q_n(\mathbf{t}) : |\mathbf{t} - \mathbf{t}_0| \geq \epsilon\} \leq Q_n(\mathbf{t}_0)).$$

Now (ii) follows if we argue that as $n \rightarrow \infty$,

$$P(\inf\{Q_n(\mathbf{t}) : |\mathbf{t} - \mathbf{t}_0| \geq \epsilon\} > Q_n(\mathbf{t}_0)) \rightarrow 1. \quad (9.2.1)$$

Note that by A3,

$$\inf\{Q(\mathbf{t}) : |\mathbf{t} - \mathbf{t}_0| = \epsilon\} > Q(\mathbf{t}_0).$$

Therefore, by A2, as $n \rightarrow \infty$,

$$P(\inf\{Q_n(\mathbf{t}) : |\mathbf{t} - \mathbf{t}_0| = \epsilon\} > Q(\mathbf{t}_0)) \rightarrow 1. \quad (9.2.2)$$

Next suppose $\hat{\mathbf{t}}_n \in \{\mathbf{t} : |\mathbf{t} - \mathbf{t}_0| > \epsilon\}$. Then there exists $\lambda \in (0, 1)$ such that

$$\mathbf{t}_1 \equiv \lambda \hat{\mathbf{t}}_n + (1 - \lambda) \mathbf{t}_0 \in \{\mathbf{t} : |\mathbf{t} - \mathbf{t}_0| = \epsilon\}.$$

By convexity of Q_n ,

$$Q_n(\mathbf{t}_1) \leq \lambda Q_n(\hat{\mathbf{t}}_n) + (1 - \lambda)Q_n(\mathbf{t}_0) \leq \lambda Q_n(\mathbf{t}_0) + (1 - \lambda)Q_n(\mathbf{t}_0) = Q_n(\mathbf{t}_0).$$

It follows that $\hat{\mathbf{t}}_n \in \{\mathbf{t} : |\mathbf{t} - \mathbf{t}_0| > \epsilon\}$ implies that

$$\inf\{Q_n(t) : \mathbf{t} : |\mathbf{t} - \mathbf{t}_0| = \epsilon\} \leq Q_n(\mathbf{t}_0).$$

But by (9.2.2), the probability of this event tends to zero. We have established (9.2.1). \square

Proposition (9.2.1) is what we generally need for consistency arguments. For asymptotic normality particularly in relation to modified likelihood, we shall use this consistency argument in conjunction with the generalized delta method of Section 7.1.

9.2.2 Asymptotics for Selected Models

We begin with

Example 9.2.1 *Biased sampling asymptotics (Examples 9.1.2 and 9.1.8).* We consider the case of one population first. Let $\mathcal{X} = R$ and recall

$$\hat{F}_e(x) = \int_{-\infty}^x d\hat{F}_e(y) = \frac{\int_{-\infty}^x w^{-1}(y)d\hat{P}(y)}{\int_{-\infty}^{\infty} w^{-1}(y)d\hat{P}(y)}, \quad (9.2.3)$$

where \hat{P} is the NPMLE of P and P is assumed to have density $\mathbf{p}_E(x) = w(x)f(x)/W(F)$. If $E_P(w^{-1}(X))^2 < \infty$, then, using Theorem 7.1.5 (Problem 9.2.1),

$$\mathcal{E}_n(x) \equiv \sqrt{n}(\hat{F}_e(x) - F(x)) \implies Z(x)$$

where

$$Z(x) = \frac{\int_{-\infty}^x w^{-1}(y)dW_P^0(y)}{E_P w^{-1}(X)} - \frac{\int_{-\infty}^x w^{-1}(y)dP(y)}{(E_P w^{-1}(X))^2} \int_{-\infty}^{\infty} w^{-1}(y)dW_P^0(y)$$

and $W_P^0(\cdot)$ is the Brownian bridge on the support of f with respect to P .

More generally, for the stratified framework with several populations, it follows from Theorem 7.2.1 that if $E_P w^{-2}(Y, F) < \infty$ where

$$w(y, F) \equiv \sum_{j=1}^k \lambda_j \frac{w_j(y)}{W_j(F)},$$

then $\mathbf{W}(\hat{F})$ is well defined with probability tending to 1 and

$$\mathbf{W}(\hat{F}) = \mathbf{W}(F) + \int \mathbf{P}(x, F)d\hat{P}(x) + o_P(n^{-\frac{1}{2}}) \quad (9.2.4)$$

for Ψ given by (9.1.23). It follows from (9.2.4) that if $\widehat{F}_e(y) = \int_{-\infty}^y d\widehat{F}_e(u)$ where $d\widehat{F}_e$ is defined in Example 9.1.8 and

$$\mathcal{E}_n(y) = \sqrt{n}(\widehat{F}_e(y) - F(y))$$

then

$$\begin{aligned} \mathcal{E}_n(y) &= n^{\frac{1}{2}} \left(\int_{-\infty}^y w^{-1}(u, F) d(\widehat{F}_e - F)(u) \right) \\ &\quad - \int_{-\infty}^y dF(u) \left\{ c^{-2}(u, p) \left[\int_{-\infty}^{\infty} \frac{w_a(u)}{W_a} d(\widehat{P}_1 - P_1)(a) \right. \right. \\ &\quad \left. \left. - \int \int \frac{w_a(u)}{W_a^2} \psi_a(x, P) dP_1(a) d(\widehat{P} - P)(x) \right] \right\} + o_P(1) \end{aligned} \quad (9.2.5)$$

by repeated application of the chain rule where $x = (i, v)$, $\psi = (\psi_1, \dots, \psi_k)$ and the second term comes from (9.2.4). These results are sketched in the problems. See also Gill, Vardi and Wellner (1988). \square

Example 9.2.2. Asymptotics for the Kaplan-Meier and Nelson-Aalen estimates (Examples 9.1.3 and 9.1.9). The Kaplan-Meier estimate of the survival function can be approximated by the simpler Nelson-Aalen estimate which is easier to analyze, and is of the modified empirical likelihood method. For $t < Y_{(L)}$, using $\log(1 - a) = -a + \frac{1}{2}a^2 + o(a^2)$,

$$\begin{aligned} \log \widehat{S}_{KM}(t) &= \sum \left\{ \log \left(1 - \frac{\delta_{(i)}}{n - i + 1} \right) 1(Y_{(i)} \leq t) \right\} \\ &= - \sum \frac{\delta_{(i)}}{n - i + 1} 1(Y_{(i)} \leq t) + O_P \left\{ \sum_{i=1}^n \frac{\delta_{(i)} 1(Y_{(i)} \leq t)}{(n - i + 1)^2} \right\}. \end{aligned} \quad (9.2.6)$$

Recall from Theorem 7.2.3 that

$$\sup \left\{ |Y_{(i)} - H^{-1}(\frac{i}{n+1})| : \frac{i}{n} \leq 1 - \varepsilon \right\} = o_P(1)$$

where

$$H(x) = 1 - \bar{F}(x)\bar{G}(x)$$

is the distribution function of Y . Since $t < Y_{(L)}$, $\bar{H}(t) \equiv 1 - H(t) > 0$, and hence

$$\sum_{i=1}^n \frac{\delta_{(i)} 1(Y_{(i)} \leq t)}{(n - i + 1)^2} = O_P\left(\frac{1}{n}\right). \quad (9.2.7)$$

Define

$$\widehat{H}_1(t) = \frac{1}{n} \sum_{i=1}^n \delta_{(i)} 1(Y_{(i)} \leq t)$$

the (sub) empirical distribution function of the uncensored observations, let $\widehat{H}(t)$ be the empirical distribution function of the Y_i , $1 \leq i \leq n$, and let $\widehat{\bar{H}} = 1 - \widehat{H}$. Then,

$$\sum_{i=1}^n \frac{\delta_{(i)}}{n-i+1} 1(Y_{(i)} \leq t) = \int_{-\infty}^t \frac{d\widehat{H}_1(y)}{\widehat{H}(y-)} \equiv \widehat{\Lambda}_{NA}(t) \quad (9.2.8)$$

where $\widehat{\Lambda}_{NA}$ is the *Nelson-Aalen estimate* of the cumulative hazard function. This leads to the *Nelson-Aalen estimate* of the survival function of T ,

$$\widehat{S}_{NA}(t) = \exp\{-\widehat{\Lambda}_{NA}(t)\}. \quad (9.2.9)$$

From (9.2.6) and (9.2.7), it is clear that if $\bar{H}(\tau) > 0$, then

$$\sup\{|\widehat{S}_{KM}(t) - \widehat{S}_{NA}(t)| : t \leq \tau\} = o_P(n^{-\frac{1}{2}}). \quad (9.2.10)$$

Using (9.2.10) and empirical process arguments from Chapter 7, we will show

Theorem 9.2.1. *If $\bar{H}(\tau) > 0$, and F is continuous, then if a “*” subscript denotes either the KM or NA estimates*

$$(i) \sup\{|\widehat{S}_*(t) - S(t)| : t \leq \tau\} = O_P(n^{-\frac{1}{2}}).$$

(ii) *As a stochastic process on $(-\infty, \tau]$,*

$$\begin{aligned} \widehat{S}_*(t) &= S(t) + \int_{-\infty}^t \bar{H}^{-1}(y-) d(\widehat{H}_1 - H_1)(y) \\ &\quad - \int_{-\infty}^t (\bar{\widehat{H}}(y-) - \bar{H}(y-)) \bar{H}^{-2}(y-) dH_1(y) + o_P(n^{-\frac{1}{2}}). \end{aligned}$$

(iii) $n^{\frac{1}{2}}(\widehat{S}_*(\cdot) - S(\cdot)) \implies Z(\cdot)$ *as a stochastic process on $[0, \tau]$ to a mean 0 Gaussian process with*

$$\text{Cov}(Z(s), Z(t)) = \bar{K}(s)\bar{K}(t)(C(s \wedge t) - C(s)C(t))$$

where $K = S/[1 + C]$ and

$$C(t) \equiv \int_0^t \frac{1}{\bar{H}(y-)} d\Lambda_F(y) = \int_0^t \frac{1}{\bar{H}(y-)F(y-)} dF(y) \quad (9.2.11)$$

Proof. We prove (ii) which will imply (i) and (iii). By (9.2.7), it's enough to consider \widehat{S}_{NA} . But

$$\widehat{S}^{NA} - S = \exp\{-\widehat{\Lambda}_{NA}\} - \exp\{-\Lambda_F\} = -(\Lambda_{NA} - \Lambda_F) + O_P(\|\widehat{\Lambda}_{NA} - \Lambda_F\|_\infty^2) \quad (9.2.12)$$

where $\|g\|_\infty \equiv \sup\{|g(t)| : -\infty < t \leq \tau\}$. Further,

$$\begin{aligned} (\widehat{\Lambda}_{NA} - \Lambda_F)(t) &= - \int_{-\infty}^t \frac{d(\widehat{H}_1 - H_1)}{\bar{H}(y-)}(y) \\ &\quad + \int_{-\infty}^t (\bar{\widehat{H}}(y-) - \bar{H}(y-)) \bar{H}^{-2}(y-) dH_1(y) + \Delta_n(t) \end{aligned}$$

where

$$\begin{aligned}\Delta_n(t) &= - \int_{-\infty}^t \frac{(\bar{\hat{H}} - \bar{H})(y-)}{\hat{H}\bar{H}} d(\hat{H}_1 - H_1)(y) \\ &\quad + \int_{-\infty}^t \frac{(\bar{\hat{H}} - \bar{H})^2}{\hat{H}\bar{H}^2}(y) dH_1(y).\end{aligned}$$

Call the terms of $\Delta_n(\cdot)$, Rem_1 and Rem_2 , respectively. Now, given Donsker's theorem for $\hat{H}(\cdot)$, we see that

$$\text{Rem}_2 = O_P(\|\bar{\hat{H}} - \bar{H}\|_\infty^2 H_1(\tau) \sup\{[\bar{\hat{H}}(y)\bar{H}^2(y)]^{-1}, y \leq \tau\}) = O_P(n^{-1})$$

because $\inf\{\bar{\hat{H}}(y)\bar{H}^2(y) : y \leq \tau\} = \bar{\hat{H}}(\tau)\bar{H}^2(\tau)$ and $\bar{H}(\tau) > 0$ by assumption.

For the other remainder term, write, using the probability integral transforms, $\nu = \hat{H}_1(y)$ and $\nu = H_1(y)$,

$$\begin{aligned}\text{Rem}_1 &= \int_0^{\hat{H}_1(\cdot)} \frac{(\bar{\hat{H}} - \bar{H})(\hat{H}_1^{-1}(v)-)}{(\hat{H}\bar{H}(\hat{H}_1^{-1}(v)-)} dv - \int_0^{H_1(\cdot)} \frac{(\bar{\hat{H}} - \bar{H})(H_1^{-1}(v)-)}{\hat{H}\bar{H}(H_1^{-1}(v)-)} dv \\ &= \int_0^{H_1(\cdot)} \frac{\{(\bar{\hat{H}} - \bar{H})(H_1^{-1}(v)-) - (\bar{\hat{H}} - \bar{H})(\hat{H}_1^{-1}(v)-)\}}{\bar{H}^2(H_1^{-1}(v)-)} dv + o_P(n^{-\frac{1}{2}}).\end{aligned}$$

By Donsker's theorem, $\text{Rem}_1 = o_P(n^{-\frac{1}{2}})$ because $\|(\bar{\hat{H}}_1^{-1} - \bar{H}_1^{-1})\|_\infty \xrightarrow{P} 0$ (Problem 9.2.4).

This completes (ii). The weak convergence to a mean zero Gaussian process in (iii) is again a consequence of Donsker's theorem after an integration by parts. Finally (9.2.11) may be obtained after a fairly tedious computation — see Problem 9.1.16. In fact, to obtain (9.2.11) simply, we have to approach the problem from the point of view of counting processes. We refer to Aalen (1978), Andersen et al (1993) and Shorack and Wellner (1986, p. 308) for such a development. \square

Example 9.2.3. Cox estimate asymptotics (Examples 9.1.4 and 9.1.10). We would like to argue that under suitable conditions, the estimate $\hat{\beta} = \arg \max \Gamma(\beta, \hat{P})$ from (9.1.41) is uniquely defined, consistent, and asymptotically normal. Although this can be done, it is easier to consider a slight modification: In order to avoid technical difficulties in the right tail of the distribution of T , we limit ourselves to observations (Z_i, T_i) with $T_i \leq \tau$ for some finite fixed constant $\tau > 0$. Here τ can be interpreted as a preassigned time limit in the study.

Recall that

$$S_0(t, \beta, P) = E_P[r(\mathbf{Z}, \beta)1(T \geq t)].$$

and let

$$\Gamma_\tau(\beta, P) = \int_0^\tau -\log S_0(t, \beta, P) dP_2(t) + E[1(T \leq \tau) \log r(\mathbf{Z}, \beta)] \quad (9.2.13)$$

where P_1 and P_2 are the marginal probability distributions of \mathbf{Z} and T . Note that $\Gamma(\boldsymbol{\beta}, P)$ as defined earlier in (9.1.14) is just $\Gamma_\infty(\boldsymbol{\beta}, P)$. Suppose $r(\mathbf{z}, \boldsymbol{\beta}) = \exp\{\boldsymbol{\beta}^T \mathbf{z}\}$ and let

$$\boldsymbol{\beta}(P) = \arg \max\{\Gamma_\tau(\boldsymbol{\beta}, P)\}. \quad (9.2.14)$$

We will show that $\boldsymbol{\beta}(P)$ is identifiable if we use this r and make the assumptions

C : (i) $P[T = c(\mathbf{Z})] < 1$ for every function $c : R^d \rightarrow R$.

(ii) $E\mathbf{Z}\mathbf{Z}^T$ is nonsingular.

(iii) T is continuous with a positive density on $(0, \infty)$.

We next show that assumption C implies identifiability of $\boldsymbol{\beta}(P)$ when $r(\mathbf{z}, \boldsymbol{\beta}) = \exp[\boldsymbol{\beta}^T \mathbf{z}]$. We also show consistency and asymptotic normality of $\widehat{\boldsymbol{\beta}} \equiv \boldsymbol{\beta}(\widehat{P})$ under these assumptions. These results hold even when the Cox model for $\mathbf{X} = (\mathbf{Z}, T)$ is not satisfied. Here $\boldsymbol{\beta}(P)$ can be thought of as the parameter of the distribution in the Cox model “closest” to \mathcal{P} . See Section 2.2.2. Let $\mathbf{x} = (\mathbf{z}, t)$.

Theorem 9.2.2. Assume that (\mathbf{Z}, T) , $T \geq 0$, has a joint probability distribution P such that C is satisfied. Set $r(\mathbf{z}, \boldsymbol{\beta}) = \exp(\boldsymbol{\beta}^T \mathbf{z})$. Then $\boldsymbol{\beta}(P) = \arg \max \Gamma_\tau(\boldsymbol{\beta}, P)$ is identifiable. Moreover, let S_0 and $\ddot{\Gamma}_\tau$ denote gradient and Hessian of S_0 and Γ_τ with respect to $\boldsymbol{\beta}$, then

(i) $\widehat{\boldsymbol{\beta}}$ is uniquely defined with probability tending to 1 and $\widehat{\boldsymbol{\beta}} \xrightarrow{P} \boldsymbol{\beta}(P)$ as $n \rightarrow \infty$.

(ii) $\widehat{\boldsymbol{\beta}} = \boldsymbol{\beta}(P) + \int \psi(\mathbf{x}, P) d\widehat{P}(x) + o_P(n^{-\frac{1}{2}})$ where

$$\psi(\mathbf{x}, P) = -\ddot{\Gamma}_\tau^{-1}(\boldsymbol{\beta}, P) \gamma(\mathbf{x}, P)$$

with an expression for $\ddot{\Gamma}_\tau(\boldsymbol{\beta}, P)$ given in the proof and, writing $\boldsymbol{\beta} = \boldsymbol{\beta}(P)$,

$$\begin{aligned} \gamma(\mathbf{x}, P) = & \left\{ \left[\mathbf{z} - E[\mathbf{Z} \mathbf{1}(T \leq \tau)] \right] - \left[\frac{\dot{S}_0}{S_0}(t, \boldsymbol{\beta}, P) - \int_0^\tau \frac{\dot{S}_0}{S_0}(u, \boldsymbol{\beta}, P) dP_2(u) \right] \right. \\ & \left. - e^{\boldsymbol{\beta}^T \mathbf{z}} \int_0^t S_0^{-1}(u, \boldsymbol{\beta}, P) \left\{ \mathbf{z} - \left[\frac{\dot{S}_0}{S_0}(u, \boldsymbol{\beta}, P) \right] \right\} dP_2(u) \right\} \mathbf{1}(t \leq \tau). \end{aligned} \quad (9.2.15)$$

(iii) $n^{\frac{1}{2}}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \xrightarrow{\text{D}} \mathcal{N}(0, \Sigma(\boldsymbol{\beta}, P))$ where

$$\Sigma(\boldsymbol{\beta}, P) = \int \psi \psi^T(\mathbf{x}, P) dP(\mathbf{x}).$$

Proof. In view of Proposition 9.2.1, for identifiability and (i), it is enough to check that

(a) $\Gamma_\tau(\boldsymbol{\beta}, \widehat{P})$ is concave with probability 1.

- (b) $\Gamma_\tau(\beta, P)$ is continuous and strictly concave and $\Gamma_\tau(\beta, P) \rightarrow -\infty$ as $|\beta| \rightarrow \infty$.
- (c) $\|\Gamma_\tau(\beta, \hat{P}) - \Gamma_\tau(\beta, P)\| \xrightarrow{P} 0$.

We will establish (a) and (b) by calculating the gradient of $\Gamma_\tau(\beta, P)$,

$$\dot{\Gamma}_\tau(\beta, P) = - \int_0^\tau \frac{\dot{S}_0}{S_0}(t, \beta, P) dP_2(t) + E[\mathbf{Z}1(T \leq \tau)] ,$$

and the Hessian,

$$\ddot{\Gamma}_\tau(\beta, P) = \int_0^\tau S_0^{-2} (\dot{S}_0^2 - S_0 \ddot{S}_0)(t, \beta, P) dP_2(t) .$$

It follows that $\Gamma_\tau(\beta, P)$ is strictly concave if $\dot{S}_0^2 - S_0 \ddot{S}_0$ is negative definite (Section B.9). We show this for the case $Z \in R$. Recall that $S_0(t, \beta, P) = E_P[e^{\beta Z} 1(T \geq t)]$. Let $U = [e^{\beta Z} 1(T \geq t)]^{\frac{1}{2}}$, $V = [Ze^{\beta Z} 1(T \geq t)]^{\frac{1}{2}}$, and $W = Z[e^{\beta Z} 1(T \geq t)]^{\frac{1}{2}}$, then

$$\dot{S}_0^2 - S_0 \ddot{S}_0 = [E(V^2)]^2 - (EU^2)(EW^2) .$$

By Cauchy-Schwarz,

$$(EU^2)E(W^2) \geq [E(UW)]^2 = [E(V^2)]^2 ,$$

with equality iff $W^2 = a + bU^2$ for some a and $b \neq 0$. Equality leads to the equation

$$(Z^2 - b)e^{\beta Z} 1(T \geq t) - a = 0 .$$

Taking expectation with respect to $\mathcal{L}(T|z)$ gives

$$P(T \geq t|z) = ae^{-\beta z}/(z^2 - b) .$$

This is impossible unless $T = c(Z)$ for some function c , which is ruled out by $C(i)$. Thus $\Gamma_\tau(\beta, P)$ is strictly concave. The condition $\Gamma_\tau(\beta, P) \rightarrow -\infty$ as $|\beta| \rightarrow \infty$ follows from strict concavity. The preceding Cauchy-Schwarz argument shows that $\Gamma_\tau(\beta, \hat{P})$ is concave in β with probability one. So (a) and (b) hold. Note that these concavity results hold for all $\alpha \in [0, 1]$, where $\alpha = P(T \leq \tau)$.

Finally, showing condition (c) is an exercise in empirical process theory. First check that

$$\sup\{|S_0(t, \beta, \hat{P}) - S_0(t, \beta, P)| : t \leq \tau + \varepsilon\} \xrightarrow{P} 0$$

for $\varepsilon > 0$ sufficiently small and then apply the Glivenko-Cantelli Theorem.

To obtain (ii), first note that the maximizer $\hat{\beta}$ of $\Gamma_\tau(\beta, \hat{P})$ satisfies $\dot{\Gamma}_\tau(\hat{\beta}, \hat{P}) = 0$. Since $\hat{\beta}$ is consistent, we can expand $\dot{\Gamma}_\tau(\hat{\beta}, \hat{P})$ around $\hat{\beta} = \beta$, set the expansion equal to zero, and use the mean value theorem to obtain

$$\sqrt{n}(\hat{\beta} - \beta) = -\ddot{\Gamma}_\tau^{-1}(\beta^*, \hat{P})\sqrt{n}\dot{\Gamma}(\beta, \hat{P})$$

where $|\beta^* - \beta| \xrightarrow{P} 0$. In addition, we can show

$$\dot{\Gamma}_\tau(\beta, \widehat{P}) = \dot{\Gamma}_\tau(\beta, P) + \int \gamma(x, P) d\widehat{P}(x) + o_P(n^{-\frac{1}{2}}). \quad (9.2.16)$$

We establish (9.2.16) by using the infinite dimensional delta method of Section 7.2. That is, we compute the Gâteaux derivative of $\nu(P) \equiv \dot{\Gamma}_\tau(\beta, P)$ and show that it equals $\gamma(x, P)$. Then we show that $-\ddot{\Gamma}_\tau^{-1}(\beta, P)\gamma(x, P)$ satisfies condition (7.2.2) for being an influence function for $\widehat{\beta}$. Some of the details are in Appendix D4.

Remark 9.2.1 The formula for $\gamma(x, P)$ is valid when $\tau = \infty$ (Tsiatis (1981), Anderson and Gill (1982)). \square

In the special case of the Cox model we have

Proposition 9.2.2. Suppose (Z, T) has the Cox (β_0, λ) distribution and that C holds. Then $\beta(P) = \beta_0$.

Proof. We have shown that $\dot{\Gamma}_\tau(\beta, P) = 0$ has a unique solution thus it remains to show that $\dot{\Gamma}_\tau(\beta_0, P) = 0$. We give the proof when $d = 1$; the general case is only notationally more difficult. Let P_0 , P_{10} , and P_{20} denote the joint and marginal probability distributions of (Z, T) under the Cox (β_0, λ) model (9.1.12), respectively. We need to show that β_0 satisfies the equation

$$\int_0^\tau \frac{\dot{S}_0}{S_0}(t, \beta_0, P_0) dP_{20}(t) = E_0[Z 1(T \leq \tau)],$$

where E_0 is expectation under P_0 . By conditioning on Z and using the iterated expectation theorem (B.1.20), we find

$$\begin{aligned} dP_{20}(t) &= dE_0[P_0(T \leq t | Z)] = dE_0[1 - \exp\{-\Lambda(t)e^{\beta_0 Z}\}] \\ &= E_0[\lambda(t)e^{\beta_0 Z} \exp\{-\Lambda(t)e^{\beta_0 Z}\}] = S_0(t, \beta_0, P_0) d\Lambda(t). \end{aligned}$$

By also conditioning on Z , we find

$$\begin{aligned} \dot{S}_0(t, \beta_0, P_0) &= E_0[Z e^{\beta_0 Z} 1(T \geq t)] = E_0[Z e^{\beta_0 Z} P(T \geq t | Z)] \\ &= E_0[Z e^{\beta_0 Z} \exp\{-\Lambda(t)e^{\beta_0 Z}\}]. \end{aligned}$$

It follows that

$$\begin{aligned} \int_0^\tau \frac{\dot{S}_0}{S_0}(t, \beta_0, P_0) dP_{20}(t) &= \int_{-\infty}^\infty \int_0^\tau z e^{\beta_0 Z} \exp\{-\Lambda(t)e^{\beta_0 Z}\} d\Lambda(t) dP_{01}(z) \\ &= \int_{-\infty}^\infty z \left(\int_0^{v_\tau(z)} e^{-v} dv \right) dP_{01}(z), \end{aligned}$$

by making the integral the change of variable $v = \Lambda(t)e^{\beta_0 Z}$ and setting $v_\tau(z) = \Lambda(\tau)e^{\beta_0 Z}$. Because $V = \Lambda(T)e^{\beta_0 Z}$ given $Z = z$ has a standard exponential distribution (Problem 9.2.14), the right hand side is $\int_{-\infty}^\infty z P(T \leq \tau | z) dP_{01}(z) = E(Z 1(T \leq \tau))$. \square

Example 9.2.4. *Partial Linear model (Example 9.1.13).* Write

$$\begin{aligned}\widehat{\beta} - \beta(P) &= \sum_{i=1}^n \{ (Z_i - \widehat{E}(Z_i|U_i))[Y_i - \widehat{E}(Y_i|U_i)] \\ &\quad - \beta \sum_{i=1}^n (Z_i - \widehat{E}(Z_i|U_i)) \} \left\{ \sum_{i=1}^n (Z_i - \widehat{E}^2(Z_i|U_i))^2 \right\}^{-1}.\end{aligned}$$

Since

$$\varepsilon_i = Y_i - E(Y_i|U_i) - \beta(Z_i - E(Z_i|U_i))$$

we obtain

$$\begin{aligned}\widehat{\beta} - \beta(P) &= \frac{\sum_{i=1}^n [Z_i - \widehat{E}(Z_i|U_i)]\varepsilon_i}{\sum_{i=1}^n (Z_i - \widehat{E}(Z_i|U_i))^2} \\ &\quad + \{ \sum_{i=1}^n (Z_i - \widehat{E}(Z_i|U_i))(E(Y_i|U_i) - \widehat{E}(Y_i|U_i)) \\ &\quad + \beta \sum_{i=1}^n (Z_i - \widehat{E}(Z_i|U_i))\{\widehat{E}(Z_i|U_i) - E(Z_i|U_i)\}(\sum_{i=1}^n (Z_i - \widehat{E}(Z_i|U_i))^2)^{-1}.\end{aligned}\tag{9.2.17}$$

Moreover,

$$\begin{aligned}\sum_{i=1}^n (Z_i - \widehat{E}(Z_i|U_i))^2 &= \sum_{i=1}^n (Z_i - E(Z_i|U_i))^2 \\ &\quad + 2 \sum_{i=1}^n (Z_i - E(Z_i|U_i))(E(Z_i|U_i) - \widehat{E}(Z_i|U_i)) + \sum_{i=1}^n (\widehat{E}(Z_i|U_i) - E(Z_i|U_i))^2.\end{aligned}\tag{9.2.18}$$

It follows that

$$\widehat{\beta} = \beta(P) + \frac{1}{n} \sum_{i=1}^n \frac{[Z_i - E(Z_i|U_i)]\varepsilon_i}{E(\text{Var}(Z_i|U_i))} + o_P(n^{-\frac{1}{2}})\tag{9.2.19}$$

provided that we can show

$$\frac{1}{n} \sum_{i=1}^n (\widehat{E}(Z_i|U_i) - E(Z_i|U_i))^2 = o_P(n^{-\frac{1}{2}})\tag{9.2.20}$$

$$\frac{1}{n} \sum_{i=1}^n (Z_i - E(Z_i|U_i))(E(Z_i|U_i) - \widehat{E}(Z_i|U_i)) = o_P(n^{-\frac{1}{2}})\tag{9.2.21}$$

$$\frac{1}{n} \sum_{i=1}^n (Z_i - \widehat{E}(Z_i|U_i))(E(Y_i|U_i) - \widehat{E}(Y_i|U_i)) = o_P(n^{-\frac{1}{2}}).\tag{9.2.22}$$

It's relatively easy to show such results under an artificial modification of the definition of \widehat{E} which is nevertheless useful conceptually. Suppose we divide the sample into

$\{X_1, \dots, X_{n-m}\}$ and $\{X_{n-m+1}, \dots, X_n\}$ where $m \asymp n^\alpha$ with $\alpha < 1$ to be chosen later. Use $\{X_{n-m+1}, \dots, X_n\}$ only to form the histogram estimates of $E(Y|U=u)$ and $E(Z|U=u)$. Call these $\widehat{E}^{(2)}(\cdot|u)$. Then, define

$$\widehat{\beta}^* = \frac{\sum_{i=1}^{n-m} (Z_i - \widehat{E}^{(2)}(Z_i|U_i))(Y_i - \widehat{E}^{(2)}(Y_i|U_i))}{\sum_{i=1}^{n-m} (Z_i - \widehat{E}^{(2)}(Z_i|U_i))^2}. \quad (9.2.23)$$

The advantage is that, by the independence of $\widehat{E}^{(2)}(\cdot|u)$ and X_1, \dots, X_{n-m} , the analogues of (9.2.20)–(9.2.22) become fairly easy. Results, which we shall postpone to Chapter 11, enable us to conclude that then

$$E(\widehat{E}^{(2)}(Y_1|U_1) - E(Y_1|U_1))^2 = O(m^{-\frac{2}{3}}) = O(n^{-\frac{2}{3}\alpha}) \quad (9.2.24)$$

and a similar result for $\widehat{E}^{(2)}(Z_1|U_1)$ holds under moment assumptions on Z and Y , the smoothness of the densities of (U, Z) and (U, Y) , and tail assumptions on the density of U . It is then easy to verify (9.2.20)–(9.2.22) and hence that (9.2.19) holds for $\widehat{\beta}^*$. We leave the details to the problems and the validation of (9.2.19) without sample splitting to Chapter 11. The validity of (9.2.19) under different regularization estimates \widehat{E} is well known — see Chen (1988) and Fan et al (1998) for instance. \square

9.3 Efficiency in Semiparametric Models

As we noted in Sections 5.4.3 and 6.2.2, in regular parametric models, there is an asymptotic notion of being the “best” (efficient) estimate of a parameter $\nu(P) \in R$ (or R^d) among broad classes of estimates of ν which are asymptotically Gaussian in some uniform way. For instance, we showed in Theorem 6.2.1, that in a model $\mathcal{P} = \{P_\theta : \theta \in \Theta \subset R^d\}$ which is “regular,” among M estimates of θ which are “regular,” the “unique” asymptotically best procedure is one which is equivalent to maximum likelihood to order $o_P(n^{-\frac{1}{2}})$. “Regular” in these cases corresponds to assumptions A0–A6 of Section 6.2.2 holding for the candidates’ influence function Ψ and for the optimal influence function $\Psi_{\text{OPT}} = \widehat{l}(\cdot, \theta)$ which corresponds to the MLE. Note that the conditions on Ψ_{OPT} correspond to smoothness conditions on the model. We also demonstrated through Example 5.4.2 (Hodges’ example) that some conditions were always necessary.

The models considered in this subsection will include the important case in which the parameter ν is defined implicitly. That is, $\mathcal{P} = \{P_{(\nu, \eta)} : \nu \in R^d, \eta \in H\}$ where η ranges over a function space and ν is defined by $\nu(P_{(\nu, \eta)}) = \nu$. For example, in the symmetric location model (3.5.1), $X \sim P_{(\nu, \eta)}$ with density

$$p(x, \nu, \eta) = \eta(x - \nu), \quad x, \nu \in R,$$

where η is an arbitrary unknown symmetric density. Here $\nu(P_{(\nu, \eta)})$ is uniquely defined as, for instance, $\nu(P) = \int x dP(x)$, if P has a first moment. Similarly, in the Cox proportional hazard model (9.1.12), ν is identified with β and $\eta(\cdot)$ with $\lambda(\cdot)$. The partial linear model (9.1.51) has $\nu = \beta$ and $\eta(\cdot) = g(\cdot)$.

In this section we shall try to separate out assumptions on the model and on the candidate procedures and, using Hilbert space geometry, show how to extend results from Section 5.4.3 and 6.2.2 to semiparametric models.

Regular Models

Suppose \mathcal{P} is a general model for the distribution P of a random vector X . Consider a parameter $\nu : \mathcal{P} \rightarrow R$ defined on \mathcal{P} . Let \mathcal{Q} be a one dimensional regular parametric submodel of \mathcal{P} containing a point P_0 in its interior. That is,

- (i) $\mathcal{Q} \subset \mathcal{P}$.
- (ii) $P_0 \in \mathcal{Q}$.
- (iii) $\mathcal{Q} = \{P_t : |t| < 1\}$ where P_t have continuous or discrete densities $p(\cdot, t)$.

Let $l(x, t) = \log p(x, t)$. We define a model \mathcal{Q} to be *regular* if

- (a) The map $t \rightarrow p(x, t)$ is continuously differentiable on $(-1, 1)$ for (almost) all x .
- (b) $\dot{l}(x, t) \equiv \frac{1}{p(x, t)} \frac{\partial p}{\partial t}(x, t)$ is defined for (almost) all x , and for all $|t| < 1$

$$0 < I(t) \equiv E_{P_t}([\dot{l}(X, t)]^2) = \int \dot{l}^2(x, t)p(x, t)d\mu(x) < \infty$$

where $d\mu(x)$ indicates sum or integral as defined in Section I.1; or more generally, μ is a measure.

- (c) The map $t \rightarrow I(t)$ is continuous.

Remark 9.3.1. We only consider submodels of the form (iii) that satisfy the conditions (a), (b), and (c). These conditions are stronger than the Hellinger differentiability conditions of Le Cam given in Bickel, Klaassen, Ritov and Wellner (1993,1998) (abbreviated to BKRW), who show in Proposition 1, p. 13, that they imply the Le Cam conditions. They are easily seen to be implied by A0–A6 of Section 5.4.6 applied to $\psi(x, t) \equiv \dot{l}(x, t)$. \square

Note that $l(x, t)$ and its derivatives depend on \mathcal{Q} and its parametrization. That is, if $f(x, \tau) = p(x, t(\tau))$ where $\tau(\cdot)$ is one to one with inverse $t(\cdot)$, then $\nabla \log f$ at $\tau(t)$ is related to $\nabla \log p$ at t by the chain rule. However, we will take advantage of the equivariance of the Fisher information bound. See Problems 3.4.3 and 9.3.2.

Regular Estimates

We shall say $\hat{\nu}$ is a *regular* estimate of $\nu(P)$ on \mathcal{P} if

- (i) $\hat{\nu}$ is consistent on \mathcal{P} .
- (ii) $\hat{\nu}$ is uniformly locally *asymptotically linear* and Gaussian on all regular parametric submodels containing P_0 for all $P_0 \in \mathcal{P}$.

What we mean by (ii) is the following: Fix $P_0 \in \mathcal{P}$. Let \mathcal{Q} be any regular one dimensional submodel of \mathcal{P} containing P_0 and let

$$L_2^0(P_0) \equiv \{h : \int h^2 dP_0 < \infty \text{ and } \int hdP_0 = 0\}.$$

Then, there exists $\psi(\cdot, P_0) \in L_2^0(P_0)$ such that for X_1, \dots, X_n i.i.d. as $X \sim P_0$,

- (a) $\widehat{\nu} = \nu(P_0) + \frac{1}{n} \sum_{i=1}^n \psi(X_i, P_0) + o_{P_0}(n^{-\frac{1}{2}})$
- (b) For any sequence $t_n \rightarrow 0$ such that $t_n = O(n^{-\frac{1}{2}})$,

$$\mathcal{L}_{P_{t_n}}(\sqrt{n}(\widehat{\nu} - \nu(P_{t_n}))) \rightarrow \mathcal{N}(0, \sigma^2(\psi, P_0)) \quad (9.3.1)$$

where $\sigma^2(\psi, P_0) = \int \psi^2(x, P_0) dP_0(x)$.

We call $\psi(\cdot, P_0)$ the *influence function* of $\widehat{\nu}$ at P_0 in agreement with our previous convention in Section 7.2.1. Note that (b) is a weak uniformity condition: $\sqrt{n}(\widehat{\nu} - \nu(P_t))$ converges to the same limit in law under P_t as under P_0 as long as $t \rightarrow 0$ at rate $n^{-\frac{1}{2}}$.

Remark 9.3.2: Although conditions A0–A6 in Section 5.4.2 appear to purely pertain to P_0 and not be uniform, they in fact imply that (9.3.1) holds via Le Cam's theory of contiguity; see Section 9.5. However, the definition of regularity given here is stronger than that of BKRW. \square

Efficient Estimates

The following discussion sketches the derivation of a lower bound on the asymptotic variance of regular estimates given in Lemma 9.3.1 below.

Consider any regular estimate $\widehat{\nu}$ with corresponding influence function ψ and any regular one dimensional submodel \mathcal{Q} containing P_0 . Note that $q(t) = \nu(P_t)$ is a parameter on \mathcal{Q} . Hence, we know that the asymptotic variance of $\sqrt{n}\widehat{\nu}$ at P_0 is no smaller than the best we could do if we knew that the true P belonged to \mathcal{Q} rather than to \mathcal{P} . If $\widehat{\nu}$ were an M estimate of $q(t)$, $P_t \in \mathcal{Q}$, and if $I(t)$ denotes the Fisher information for P_t , then this would mean that

$$\sigma^2(\psi, P_0) \geq [q'(0)]^2 I^{-1}(0) \equiv I^{-1}(P_0 : \nu, \mathcal{Q}) \quad (9.3.2)$$

by an extension of Theorem 5.4.3 to estimation of $q(t)$ rather than just to estimation of t as in Proposition 3.4.2. Note that the definition of $I^{-1}(P_0 : \nu, \mathcal{Q})$ given at the end of (9.3.2) indicates that the bound does not depend on the parametrization of \mathcal{Q} (Problem 9.3.2).

Suppose

$$q(t) = \nu(P_t) = \nu(P_0) + \int \psi(x, P_0) dP_t(x) + o(t) \quad (9.3.3)$$

and

$$\frac{\partial}{\partial t} \int \psi(x, P_0) dP_t(x)|_{t=0} = \int \psi(x, P_0) \dot{l}(x, 0) dP_0(x). \quad (9.3.4)$$

Expression (9.3.3) says that the same influence function approximation is valid for the $\{P_t\}$ as for the empirical probability \hat{P} and (9.3.4) that integration and differentiation can be interchanged. Together, (9.3.3) and (9.3.4) imply that

$$q'(0) = \frac{\partial \nu}{\partial t}(P_t)|_{t=0} = \text{Cov}_{P_0}(\psi(X, P_0), \dot{l}(X, 0)). \quad (9.3.5)$$

It may be shown (see Section 9.5) that by Le Cam's contiguity theory, (9.3.5) is implied by regularity of \mathcal{Q} and $\hat{\nu}$ as we have defined these. It follows that (9.3.2) can be extended to all regular $\hat{\nu}$, \mathcal{Q} . That is,

Lemma 9.3.1. For all $P_0 \in \mathcal{P}$ and all ψ corresponding to regular $\hat{\nu}$ on \mathcal{P} ,

$$\begin{aligned} \sigma^2(\psi, P_0) &\geq \sup_{\mathcal{Q}} \{I^{-1}(P_0 : \nu, \mathcal{Q}) : \mathcal{Q} \subset \mathcal{P}, \mathcal{Q} \text{ regular 1 dimensional containing } P_0\} \\ &\equiv I^{-1}(P_0 : \nu, \mathcal{P}). \end{aligned} \quad (9.3.6)$$

Proof. By (9.3.5) and (9.3.2),

$$I^{-1}(P_0 : \nu, \mathcal{Q}) = [\text{Cov}_{P_0}(\psi(X, P_0)), \dot{l}(X, 0)/I(0)]^2 I(0),$$

where $I(0) = E_{P_0}([\dot{l}(X, 0)]^2)$ by (iii)(b). The result follows by Cauchy-Schwarz. \square

Note that $I^{-1}(P_0 : \nu, \mathcal{P})$ doesn't depend on any parametrization of \mathcal{P} .

Definition 9.3.1. We call an influence function $\psi^*(\cdot, P_0)$ corresponding to a regular estimate $\hat{\nu}^*$ of $\nu(P)$, $P \in \mathcal{P}$, achieving the bound in (9.3.6), the *efficient influence function* for ν in \mathcal{P} at P_0 and we call $\hat{\nu}^*$ an *efficient estimate* at P_0 . This ψ^* is often denoted by $\tilde{l}(\cdot, P_0 : \nu, \mathcal{P})$.

Adopting the terminology of Section 3.3, $\hat{\nu}^*$ is called *asymptotically minimax* over the class of regular one dimensional submodels \mathcal{Q} of \mathcal{P} . We will find such $\hat{\nu}^*$ for the class of regular estimates, but it is true generally. See BKRW.

Next we sketch the derivation of a method for finding such $\hat{\nu}^*$ given in Proposition 9.3.1 below. Evidently, for ψ^* as in Definition 9.3.1 and all ψ corresponding to regular $\hat{\nu}$ on \mathcal{P} , we need to show that

$$\sigma^2(\psi^*, P_0) \leq \sigma^2(\psi, P_0).$$

On any \mathcal{Q} as above, by Theorems 5.3.3 and 5.4.3, any $\hat{\nu}^*$ regular on \mathcal{Q} , achieving $I^{-1}(P_0 : \nu, \mathcal{Q})$, is uniquely characterized by its influence function,

$$\psi^*(x, P_0) = \frac{q'(0)}{I(0)} \dot{l}(x, 0). \quad (9.3.7)$$

Note that $\psi^*(\cdot, P_0)$ satisfying Definition 9.3.1 cannot depend on the parametrization of \mathcal{Q} . We can show (Problem 9.3.2) that if $f(x, \tau) = p(x, t(\tau))$ as above with $t(0) = 0$ and $t'(\tau) > 0$, then $\dot{l}_f = t'(0)\dot{l}_p$. Thus if we let

$$\dot{\mathcal{Q}}(P_0) \equiv [\dot{l}(\cdot, 0)] = \{c\dot{l}(\cdot, 0) : c \in R\}, \quad (9.3.8)$$

then, although ψ^* and $\dot{l}(\cdot, 0)$ depend on the parametrization of \mathcal{Q} , $\dot{\mathcal{Q}}(P_0)$ does not. For a parameter ν corresponding to a regular estimate $\hat{\nu}$, we may rephrase (9.3.7) as:

“ $\psi^*(\cdot, P_0)$ is the unique member of $\dot{\mathcal{Q}}(P_0)$ such that
 $E_{P_0}[\psi^*(X, P_0)]^2 = I^{-1}(P_0 : \nu, \mathcal{Q}).$ ”

(9.3.9)

Note that $\hat{\nu}^*$ corresponding to ψ^* in (9.3.9) may not be regular or even consistent on \mathcal{P} . Suppose, for instance, that $\mathcal{P} = \{\mathcal{N}(\mu, \sigma^2) : \mu \in R, \sigma^2 > 0\}$. Let $P_0 = \mathcal{N}(0, 1)$ and $\mathcal{Q} = \{\mathcal{N}(0, \sigma^2) : \sigma^2 > 0\}$. Then, $n^{-1} \sum_{i=1}^n X_i^2$ is an efficient estimate of σ^2 on \mathcal{Q} , but is inconsistent for all $P \in \mathcal{P}$ with mean $\neq 0$. However, if we can assume that $\hat{\nu}^*$ is regular on *all* of \mathcal{P} , we can conclude

Proposition 9.3.1. *If $\hat{\nu}^*$ is a regular estimate of $\nu(P)$, $P \in \mathcal{P}$, and its influence function $\psi^*(\cdot, P_0)$ is in $\dot{\mathcal{Q}}(P_0)$ for some regular 1 dimensional \mathcal{Q} , then*

$$\sigma^2(\psi^*, P_0) = I^{-1}(P_0 : \nu, \mathcal{P}),$$

so that $\hat{\nu}^*$ is efficient at P_0 .

Proof. ψ^* necessarily satisfies (9.3.9) for \mathcal{Q} and is efficient for \mathcal{Q} . Thus,

$$\sigma^2(\psi^*, P_0) = I^{-1}(P_0 : \nu, \mathcal{Q}).$$

By definition

$$I^{-1}(P_0 : \nu, \mathcal{P}) \geq I^{-1}(P_0 : \nu, \mathcal{Q}) = \sigma^2(\psi^*, P_0).$$

On the other hand, by Lemma 9.3.1, because $\hat{\nu}^*$ is regular on \mathcal{P} ,

$$\sigma^2(\psi^*, P_0) \geq I^{-1}(P_0 : \nu, \mathcal{P}),$$

and the proposition follows. \square

Adopting the terminology of Section 3.3, \mathcal{Q} is asymptotically the *least favorable* regular one dimensional submodel for estimating $\nu(P)$ using regular estimates.

We next give a simple example to help interpret the new concepts and to show connections to concepts in Volume I.

Example 9.3.1. Suppose $\mathbf{X} = (\mathbf{Z}, Y)$ with $\mathbf{Z} \in R^d$, $Y \in R$. Let $\mathbf{X}_1, \dots, \mathbf{X}_n$ be i.i.d. as $\mathbf{X} \sim P$ where

$$Y = \mathbf{Z}^T \boldsymbol{\beta} + \varepsilon.$$

Here $\varepsilon \sim \mathcal{N}(0, \sigma_0^2)$, σ_0^2 is known, \mathbf{Z} and ε are independent, and $\mathbf{Z}\mathbf{Z}^T$ is nonsingular (cf Examples 6.2.1 and 9.1.1). Let \mathcal{P} be the class of distributions of \mathbf{X} , $\boldsymbol{\beta} \in R^d$, and let $\mathcal{Q}_{\mathbf{a}}$ be the class of distributions $P_{t, \mathbf{a}}$ of \mathbf{X} with $\boldsymbol{\beta}_t = \boldsymbol{\beta}_0 + t\mathbf{a}$, where $\boldsymbol{\beta}_0 \in R^d$ and $\mathbf{a} \in R^d$ are fixed and $t \in R$ is the parameter. If $l_{\mathbf{a}}$ is the log likelihood for $\mathcal{Q}_{\mathbf{a}}$, then, as a function of

β_t ,

$$\begin{aligned} l_{\mathbf{a}}(\mathbf{x}, t) &= -\frac{1}{2} \sum_{i=1}^n [y_i - \mathbf{z}_i^T \boldsymbol{\beta}_t]^2 + \text{constant} \\ l_{\mathbf{a}}(\mathbf{x}, 0) &= \sum_{i=1}^n (\mathbf{z}_i^T \mathbf{a}) [y_i - \mathbf{z}_i^T \boldsymbol{\beta}_t] \equiv \sum_{i=1}^n (\mathbf{z}_i^T \mathbf{a}) e_i \\ I_{\mathbf{a}}(0) &= nE(\mathbf{a}^T \mathbf{Z}^T \mathbf{a})^2. \end{aligned}$$

If we let \mathbf{c}_j be the d -vector with 1 in the j th entry and 0 elsewhere, then the equations $l_{\mathbf{c}_j}(\mathbf{x}, 0) = 0$, $j = 1, \dots, d$, are the normal equations as in Example 2.1.1, and yield the MLE of $\boldsymbol{\beta}$ as in Examples 6.1.2 and 6.2.1.

As in Example 6.2.1, consider the parameter

$$\nu = \nu(P) = \beta_1.$$

Let $\widehat{\boldsymbol{\beta}} = (\widehat{\beta}_1, \dots, \widehat{\beta}_d)$ be the MLE and set $\widehat{\nu} = \widehat{\beta}_1$. Then $\widehat{\nu}$ is regular and by (6.2.22) the influence function is $\psi_1(\mathbf{x}, \boldsymbol{\beta})$ where ψ_1 is the first entry in $\boldsymbol{\psi} = (\psi_1, \dots, \psi_d)^T = E(\mathbf{Z}\mathbf{Z}^T)^{-1}\mathbf{Z}^T e$. For this $\mathcal{Q}_{\mathbf{a}}$ and this ν , $q(t) = \nu(P_{t,\mathbf{a}}) = \beta_{01} + a_1 t$; and

$$\psi^*(\mathbf{x}, P_0) = \frac{a_1 \sum_{i=1}^n (\mathbf{z}_i^T \mathbf{a}) e_i}{nE(\mathbf{a}^T \mathbf{Z}^T \mathbf{a})^2}, \quad I^{-1}(P_0 : \nu, \mathcal{Q}_{\mathbf{a}}) = \frac{a_1^2}{nE(\mathbf{a}^T \mathbf{Z}^T \mathbf{a})^2}.$$

Moreover, independently of ν ,

$$\dot{\mathcal{Q}}_{\mathbf{a}}(P_0) = \left\{ c \sum_{i=1}^n (\mathbf{z}_i^T \mathbf{a}) e_i : c \in R \right\}.$$

Finally, using calculus (Problem 9.3.3), we find

$$\sup \left\{ I^{-1}(P_0 : \nu, \mathcal{Q}_{\mathbf{a}}) : \mathbf{a} \in R^d \right\} = \left([nE(\mathbf{Z}\mathbf{Z}^T)]^{-1} \right)_{(1,1)}.$$

The right hand side coincides with the information bound in Example 6.2.1 and we can conclude that $\widehat{\beta}_1$ is efficient and

$$I^{-1}(P_0 : \nu, \mathcal{P}) = \left([nE(\mathbf{Z}\mathbf{Z}^T)]^{-1} \right)_{(1,1)}.$$

□

Remark 9.3.3. *The d dimensional case.* These concepts and results extend readily from $\nu(P) \in R$ to $\nu(P) \in R^d$ by using Section 6.2. In this case, (9.3.6) becomes a matrix inequality where $A \geq B$ for $d \times d$ matrices A and B means that $(A - B)$ is nonnegative definite. See Section B.10.2. Note that if we compute the influence function ψ_j^* of each $\widehat{\beta}_j$ in Example 9.3.1, then the influence function of $\widehat{\boldsymbol{\beta}}$ is the vector of influence functions $(\psi_1^*, \dots, \psi_d^*)^T$. See Theorem 6.2.2 for the parametric case. □

Remark 9.3.4. It may be shown that, in parametric models \mathcal{P} satisfying the conditions of Theorem 5.4.3 or 6.2.2, the MLE is efficient in the sense of Definition 9.3.2. For the 1 dimensional case this is clear, but for $\Theta \subset \mathbb{R}^d$, $d > 1$, it has to be shown that being best on all 1 dimensional submodels is equivalent to optimality in the full model (Problem 9.3.4). \square

Remark 9.3.5. This method of approaching efficiency in the semi- and nonparametric context is due to Stein (1956(a)). It is encompassed in a much more general context by Le Cam (1986) and Le Cam and Yang (1990). For more background, see BKRW (Bickel, Klaassen, Ritov and Wellner (1998)). \square

In Subsection 9.1.2, we have examined modified likelihood approaches to constructing estimates in semiparametric models. In this subsection, we check if such estimates are efficient, and determine potential efficient influence functions in advance of constructing estimates we hope to be efficient.

Definition 9.3.2. The *tangent set* $\dot{\mathcal{P}}^0(P_0)$ of \mathcal{P} at P_0 is the union of the set of all $\dot{\mathcal{Q}}(P_0)$ obtained by varying \mathcal{Q} over all regular 1 dimensional submodels of \mathcal{P} containing P_0 .

By definition, $\dot{\mathcal{P}}^0(P_0)$ is a subset of the set $L_2^0(P_0)$ of mean zero $L_2(P_0)$ functions.

Definition 9.3.3. The *tangent space* $\dot{\mathcal{P}}(P_0)$ of \mathcal{P} at P_0 is the linear closure of $\dot{\mathcal{P}}^0(P_0)$ in $L_2^0(P_0)$, that is, the smallest closed linear subspace of $L_2^0(P_0)$ which contains $\dot{\mathcal{P}}^0(P_0)$. Thus $\dot{\mathcal{P}}(P_0)$ is the closure of the set of functions of the form $\sum_{j=1}^m c_j \psi_j$, $\psi_j \in \dot{\mathcal{P}}^0(P_0)$.

We use this notation generically, i.e. \mathcal{P} can be any model and, in particular, any regular 1 dimensional \mathcal{Q} .

Note that $\dot{\mathcal{P}}(P_0)$ is not dependent on any ν . It is computed from $\dot{l}(\cdot, 0)$ over regular 1 dimensional parametric models. $\dot{\mathcal{P}}(P_0)$ will be used to find efficient estimates for semi-parametric models containing both d dimensional and function parameters.

Remark 9.3.6. The parameters set for P_t were defined in (iii) as $(-1, 1)$. By writing $t = (t/\epsilon)\epsilon$, we see that we may use $(-\epsilon, \epsilon)$ as the parameter set for t . See the proof of Proposition 9.3.2 for a case where this is useful.

Here is a result that will help interpret $\dot{\mathcal{P}}(P_0)$:

Proposition 9.3.2. Suppose $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$ is a d dimensional parametric model with $l_0(\cdot, \theta) = \log f(\cdot, \theta)$, which satisfies A0–A6 of Theorem 6.2.2, and suppose $\Theta \subset \mathbb{R}^d$ is open. Then, if $P_0 \equiv P_{\theta_0}$, $\theta_0 \in \Theta$,

$$\dot{\mathcal{P}}(P_0) = \left[\frac{\partial l_0}{\partial \theta_1}(x, \theta_0), \dots, \frac{\partial l_0}{\partial \theta_d}(x, \theta_0) \right],$$

where $[S]$ is the linear span of S , that is, the set of all linear combinations of elements in S .

Proof. The assumptions A0 – A6 imply that for all $\mathbf{a} \neq \mathbf{0}$, and ϵ sufficiently small, $t \rightarrow P_{\theta_0 + t\mathbf{a}}$, $|t| < \epsilon$, is a regular one dimensional parametric submodel \mathcal{Q} with $\dot{l}(\cdot, 0) =$

$\sum_{j=1}^d a_j \frac{\partial l_0}{\partial \theta_j}(\cdot, \boldsymbol{\theta}_0)$. Thus, $\left[\frac{\partial l_0}{\partial \theta_1}(\cdot, \boldsymbol{\theta}_0), \dots, \frac{\partial l_0}{\partial \theta_d}(\cdot, \boldsymbol{\theta}_0) \right] \subset \dot{\mathcal{P}}(P_0)$. On the other hand, if $p(x, t) = f(x, \boldsymbol{\theta}(t))$ is a one dimensional submodel, then

$$\dot{l}(x, 0) = \frac{\partial l_0}{\partial t}(x, \boldsymbol{\theta}(t))|_{t=0} = \sum_{j=1}^d \frac{\partial \theta_j}{\partial t}(0) \frac{\partial l_0}{\partial \theta_j}(x, \boldsymbol{\theta}_0)$$

belongs to $\left[\frac{\partial l_0}{\partial \theta_1}, \dots, \frac{\partial l_0}{\partial \theta_d} \right]$. \square

This proposition makes the rationale for calling $\dot{\mathcal{P}}$ the tangent space clear. If we think of \mathcal{P} as a smooth d dimensional manifold on the set of all probability distributions viewed as a much higher dimensional space, then the tangent space at P_0 is, as in the Euclidean case, the linear space spanned by all tangents of smooth curves (1 dimensional regular models) through P_0 .

Let $\Pi(h|\mathcal{L})$ denote the projection in $L_2(P_0)$ of any h in $L_2(P_0)$ onto the closed linear space $\mathcal{L} \subset L_2(P_0)$ with inner product $(h_1, h_2) = \text{Cov}_{P_0}(h_1(X), h_2(X))$ and norm $\|h\| = (h, h)^{\frac{1}{2}}$. That is

$$\pi(h|\mathcal{L}) = \arg \min \{E_{P_0}[h(x) - g(x)]^2 : g \in \mathcal{L}\}.$$

Such projections are exemplified in Section 1.4 and discussed in terms of Hilbert spaces in Appendix B.10.3. In particular, if \mathcal{L} is the space of functions of the form $a + bX$, Theorem 1.4.3 states that for X, Y with $E(X) = E(Y) = 0$,

$$\pi(Y|\mathcal{L}) = [\text{Cov}(X, Y)/EX^2]X = [(X, Y)/|X|^2]X.$$

The natural candidate for efficient influence function is given by

Proposition 9.3.3. *Suppose $\dot{\mathcal{P}}^0(P_0)$ is linear and $\psi(\cdot, P_0)$ is the influence function of any d dimensional regular estimate. Then, the efficient influence function is given by,*

$$\psi^*(\cdot, P_0) \equiv \tilde{l}(\cdot, P_0 : \nu, \mathcal{P}) = \Pi(\psi(\cdot, P_0)|\dot{\mathcal{P}}(P_0)). \quad (9.3.10)$$

Proof. We claim that the result holds if $\mathcal{P} = \mathcal{Q}$, a regular 1 dimensional submodel since by Theorem 1.4.3, (9.3.5), and (9.3.7),

$$\Pi(\psi(\cdot, P_0)|\dot{\mathcal{Q}}(P_0)) = \frac{(\psi(\cdot, P_0), \dot{l}(\cdot, 0))}{\|\dot{l}(\cdot, 0)\|^2} \dot{l}(\cdot, 0) = \frac{q'(0)}{I(0)} \dot{l}(\cdot, 0)$$

which we have shown is an efficient influence function for \mathcal{Q} . We now write $\dot{\mathcal{Q}}, \dot{\mathcal{P}}^0$ for $\dot{\mathcal{Q}}(P_0), \dot{\mathcal{P}}^0(P_0)$. Since \tilde{l} is itself an influence function,

$$\Pi(\tilde{l}|\dot{\mathcal{Q}}) = \Pi(\psi|\dot{\mathcal{Q}})$$

for all $\dot{\mathcal{Q}} \in \dot{\mathcal{P}}^0$. Since $\dot{\mathcal{P}}^0$ is linear, it follows that

$$\Pi(\tilde{l}|\dot{\mathcal{P}}) = \Pi(\psi|\dot{\mathcal{P}}).$$

But, by assumption and what we have demonstrated,

$$\begin{aligned}\|\tilde{l}\|^2 &= I^{-1}(P_0 : \nu, \mathcal{P}) \\ &= \sup_{\mathcal{Q}} \{ \|\Pi(\tilde{l}|\dot{\mathcal{Q}})\|^2 : \mathcal{Q} \subset \mathcal{P}, \mathcal{Q} \text{ regular 1 dimensional containing } P_0 \}.\end{aligned}$$

Thus, there exist $\dot{\mathcal{Q}}_m$, $m \geq 1$, such that

$$\|\tilde{l}\|^2 = \lim \|\Pi(\tilde{l}|\dot{\mathcal{Q}}_m)\|^2$$

which by Pythagoras' theorem B.10.3.1 gives

$$\|\tilde{l} - \Pi(\tilde{l}|\dot{\mathcal{Q}}_m)\|^2 \rightarrow 0.$$

Thus, $\tilde{l} \in \dot{\mathcal{P}}(P_0)$ and the result follows. \square

Remark 9.3.7. We assume that there exists a regular estimate with influence function $\psi(\cdot, P_0)$ for the parameter ν . Such estimates exist under mild conditions. See BKRW.

It is not hard to derive Theorem 3.4.3 from Proposition 9.3.3 and the standard formula for projection onto a finite dimensional linear space (B.10.20) (Problem 9.3.5). This proposition and some others we are about to state are used in a semi-rigorous fashion. As we shall see in the applications (called examples) to important models which follow, we calculate the structure of tangent spaces formally, usually ending up with linear spaces which we can only be sure are subspaces of the tangent space. However, if we can show that the influence function of a regular estimate $\hat{\nu}^*$ belongs to the subspace we have identified or that the projection of an arbitrary influence function on such a subspace is the influence function of a regular estimate $\hat{\nu}^*$, we can conclude without further ado that $\hat{\nu}^*$ is efficient.

The following results given in BKRW are to be taken in that formal spirit also. That is, equalities as stated are really true only under regularity conditions which are model dependent.

Proposition 9.3.4. *Additive property of the efficient score function. Suppose $\mathcal{P} = \{P_{(g,h)} : g \in \mathcal{G}, h \in \mathcal{H}\}$, where \mathcal{G}, \mathcal{H} may be Euclidean, function spaces, or abstract. Let $P_0 = P_{(g_0,h_0)}$ and let*

$$\mathcal{P}_1(P_0) \equiv \{P_{(g,h_0)} : g \in \mathcal{G}\}, \quad \mathcal{P}_2(P_0) \equiv \{P_{(g_0,h)} : h \in \mathcal{H}\}.$$

Then,

$$\dot{\mathcal{P}}(P_0) = \dot{\mathcal{P}}_1(P_0) + \dot{\mathcal{P}}_2(P_0).$$

We have encountered parameters ν that are defined implicitly by equations of the form $\nu(P_{\theta,h}) = \theta$. For instance, in linear regression or Cox regression the regression parameter vector β is defined this way. Similarly, so is the location parameter in the location problem. For such parameters there is a geometric way of obtaining efficient influence functions \tilde{l} via orthogonal projections of the score function for θ on tangent spaces corresponding to nuisance parameters..

Proposition 9.3.5. *The efficient influence function for implicitly defined parameters. Suppose $\mathcal{G} \equiv \Theta$ open $\subset R$, and let $l_\theta \equiv l(\cdot, P_{(\theta, h_0)})$, $\theta \in \Theta$, be the log likelihood for θ . Suppose $\nu(P_{(\theta, h)}) = \theta$ for all $h \in \mathcal{H}$. Let $P_0 \equiv P_{(\theta_0, h_0)}$ and $\dot{l} \equiv \frac{\partial l_\theta}{\partial \theta}(\cdot)|_{\theta=\theta_0}$. Then,*

$$\tilde{l}(\cdot, P_0 : \nu, \mathcal{P}) = \frac{\dot{l} - \Pi(\dot{l}|\dot{\mathcal{P}}_2(P_0))}{\|\dot{l} - \Pi(\dot{l}|\dot{\mathcal{P}}_2(P_0))\|^2}, \quad (9.3.11)$$

$$I^{-1}(P_0 : \nu, \mathcal{P}) = \|\dot{l} - \Pi(\dot{l}|\dot{\mathcal{P}}_2(P_0))\|^{-2}. \quad (9.3.12)$$

□

Note that \tilde{l} is a normalized version of the projection of \dot{l} onto the orthocomplement of the nuisance parameter tangent space. The normalization ensures that $(\tilde{l}, \tilde{l}) = 1$, which is a requirement of influence functions; see BKRW (1998) for details.

The tangent space of a nonparametric model

Proposition 9.3.6. *Let \mathcal{M}_0 be a model that contains the collection of all probabilities on R with continuous bounded densities. Then, $\mathcal{M}_0(P_0) = L_2^0(P_0)$.*

Before proving this proposition, we note its satisfying implication.

Corollary 9.3.1. *If $\nu : \mathcal{M}_0 \rightarrow R$ is a parameter for which a regular estimate $\hat{\nu}$ exists, then $\hat{\nu}$ is efficient.*

Proof: The result is immediate from Proposition 9.3.4, since every influence function $\psi(\cdot, P_0) \in L_2^0(P_0)$ and is its own projection. □

For instance, if $\nu = g(\mu)$, $\mu = E(X)$, then $\hat{\nu} = g(\bar{X})$ is nonparametrically efficient for the model \mathcal{M}_0 provided g is differentiable at μ . See Theorem 5.3.3.

Proof of Proposition 9.3.4. It is enough to show that $\mathcal{M}_0(P_0)$ contains all bounded h such that $\int h(x)dP_0(x) = 0$. Given such an h , let $p(x, t) = \exp\{th(x) - A(t, h)\}p(x, 0)$ where $A(t, h) = \log \int e^{th(x)}dP_0(x)$. Then, $p(\cdot, t)$ is a canonical exponential family and a regular submodel with

$$\dot{l}(x, 0) = \frac{\partial}{\partial t} \log p(x, t)|_{t=0} = h(x) - \frac{\partial A}{\partial t}(0, h).$$

But by Corollary 1.6.1,

$$\frac{\partial A}{\partial t}(0, h) = \int h(x) dP_0(x).$$

Thus $[\dot{l}(x, 0)]$ contains h with $\int h(x) dP_0(x) = 0$ and the result follows. □

Examples

We finish this section by obtaining candidate efficient influence functions for some of our examples, seeing whether we have obtained estimates efficient at particular P_0 or on all

of \mathcal{P} . As we have noted we usually do not check that we have computed the full tangent space but rather identify a subspace that gives the structure of what elements of the tangent space should look like. Then we check whether an estimate $\hat{\nu}$ obtained by some criteria such as likelihood or modified likelihood has an influence function that belongs to the subspace we have identified. The subspace is computed using the functional delta methods of Section 7.2.

Example 9.3.2. *The semiparametric linear model (Example 9.1.1). Tangent spaces and influence functions.* As we have noted we usually do not check that we have computed the full tangent space but rather identify the structure of what elements of the tangent space should look like. We consider two models that correspond to assumptions and notation given earlier in Example 9.1.1 as (a) and (b) (with $\sigma \equiv 1$, variability in σ is absorbed into the parameter f below).

Model (a). Here,

$$p(\mathbf{z}, y; \alpha, \beta, h, f) = h(\mathbf{z})f(y - \alpha - \beta^T \mathbf{z}).$$

Suppose P_0 corresponds to the parameter value $(\alpha_0, \beta_0, h_0, f_0)$ and consider formally 1 dimensional models, $\alpha_t = \alpha_0 + t\Delta_\alpha$, $\beta_t = \beta_0 + t\Delta_\beta$, $\log h_t = \log h_0(\mathbf{z}) + t\Delta_h(\mathbf{z})$, $\log f_t = \log f(y) + t\Delta_f(y)$. Formally, if P_t corresponds to the parameter value $(\alpha_t, \beta_t, h_t, f_t)$ and has density $p_t(\mathbf{x})$ with $\mathbf{x} = (\mathbf{z}^T, y)^T$, then

$$\frac{\partial}{\partial t} \log p_t(\mathbf{x})|_{t=0} = \Delta_h(\mathbf{z}) + \Delta_f(\varepsilon) + \Delta_\alpha \frac{f'_0}{f_0}(\varepsilon), \quad (9.3.13)$$

where $\int \Delta_h(\mathbf{z}) h_0(\mathbf{z}) d\mathbf{z} = \int \Delta_f(\varepsilon) f_0(\varepsilon) d\varepsilon = 0$ and $\varepsilon \equiv y - \alpha_0 - \beta_0^T \mathbf{z}$. Thus, we expect that

$$\dot{\mathcal{P}}(P_0) = \left\{ a(\mathbf{Z}) + b(\varepsilon) + \left[\mathbf{Z} \frac{f'_0}{f_0}(\varepsilon) \right] + \left[\frac{f'_0}{f_0}(\varepsilon) \right] \right\},$$

where a, b are arbitrary functions in $L_2(P_0)$. Now the LSE $\hat{\beta}$ has influence function

$$\psi(\mathbf{x}, P_0) = [\text{Var}(\mathbf{Z})]^{-1} (\mathbf{Z} - E(\mathbf{Z})) \varepsilon \quad (9.3.14)$$

which belongs to $\dot{\mathcal{P}}(P_0)$ if $f'_0/f_0(\varepsilon) = c\varepsilon$ for some $c \neq 0$, that is, if f_0 corresponds to a mean 0 Gaussian distribution.

It follows from our analysis in Example 6.2.2 that, if f_0 is true, the efficient influence function corresponds to the MLE obtained by *assuming* f_0 is true and known. Since we don't know f_0 , where does this leave us? It turns out that this is a situation where we can adapt in the sense of Example 6.2.1. That is, it is possible to construct estimates $\hat{\beta}$ such that

$$\hat{\beta} = \beta + \frac{1}{n} \sum_{i=1}^n [\text{Var}(\mathbf{Z})]^{-1} (\mathbf{Z}_i - E(\mathbf{Z})) \left(-\frac{f'_0}{f_0} \right) (\varepsilon_i) + o_P(n^{-\frac{1}{2}}) \quad (9.3.15)$$

by estimating f_0 properly. Since (9.3.15) is how the best estimate of β behaves if we knew f_0 (but not α), this is the best we can hope to do. We refer to Chapters 11 and 12 of this volume as well as BKRW and van der Vaart (1998) for such constructions.

Model (b). Here, for simplicity, take $z \in R$. Then,

$$p(z, y, \alpha, \beta, f) = f(z, y - \alpha - \beta z) \quad (9.3.16)$$

subject to

$$\int (y - \alpha - \beta z) f(z, y - \alpha - \beta z) dy = 0 \quad (9.3.17)$$

for all z but f is otherwise an arbitrary density. Fix α, β at their true values and consider the one dimensional submodel

$$\log f_t = \log f_0 + t\Delta(z, \varepsilon)$$

with $\varepsilon \equiv y - \alpha - \beta z$. Then,

$$\frac{\partial}{\partial t} \log f_t = \Delta(z, \varepsilon)$$

and (9.3.17) is, for all z ,

$$\int \varepsilon \Delta(z, \varepsilon) f(z, \varepsilon) d\varepsilon = 0$$

which is equivalent to

$$E(\varepsilon \Delta(Z, \varepsilon) | Z) = 0. \quad (9.3.18)$$

We see from this and (9.3.14) that the influence function of the LSE $\hat{\beta}$ does *not* belong to $\dot{\mathcal{P}}(P_0)$ unless $E(\varepsilon^2 | Z) = 0$. The efficient estimate which can, under some conditions, be constructed has influence function

$$\frac{(Z - EZ)}{I(P_0, \beta)} \left(\frac{f'_0}{f_0}(Z, \varepsilon) - \frac{E\left(\frac{f'_0}{f_0}(Z, \varepsilon)\varepsilon | Z\right)}{E(\varepsilon^2 | Z)} \varepsilon \right), \quad (9.3.19)$$

where

$$I(P_0, \beta) = E \left\{ (Z - EZ)^2 \left[\left(\frac{f'_0}{f_0} \right)^2(Z, \varepsilon) - \frac{E^2 \left(\frac{f'_0}{f_0}(Z, \varepsilon)\varepsilon | Z \right)}{E(\varepsilon^2 | Z)} \right] \right\}$$

and

$$\frac{f'_0}{f_0}(Z, \varepsilon) = \frac{\partial}{\partial v} \log f_0(Z, v) \Big|_{v=\varepsilon}.$$

For these claims see Problem 9.3.8.

Example 9.3.3. *Tangent spaces for biased sampling (Examples 9.1.2, 9.1.8, and 9.2.1).* If $k = 1$, $\dot{\mathcal{P}}(P_0) = L_2(P_0)$ by (9.1.4). For $k > 1$, proceed formally as usual. Write, since (w_1, \dots, w_k) , λ are known,

$$p_t(j, y) = \frac{\lambda_j w_j(y) f_t(y)}{W_j(F_t)}$$

and, given h such that $\int h(y) dF_0(y) = 0$,

$$f_t(y) = \exp\{th(y) - A(t, f_0)\} f_0(y).$$

$$W_j(F_t) = \int w_j(y) \exp\{th(y) - A(t, f_0)\} dF_0(y).$$

Then,

$$\dot{l}(f_t(y), 0) = \frac{\partial}{\partial t} \log p_t(j, y)|_{t=0} = h(y) - A'(0, f_0) - \frac{\int h(y) w_j(y) dF_0(y)}{W_j(F_0)}.$$

Thus, at least for any bounded h , if $X = (I, Y)$,

$$\dot{l}(X, 0) = h(Y) - E_0(h(Y)|I). \quad (9.3.20)$$

It is, in fact, true that the tangent space is

$$\dot{\mathcal{P}}(P_0) = \{a(I) + b(Y) \in L_2^0(P_0) : E_0(b(Y)|I) = -a(I)\}. \quad (9.3.21)$$

It is not hard to check that the influence function of the estimate, $\theta(\hat{F})$ of $\theta(F) = F(y)$, or more generally,

$$\theta(F) = \int v(y) dF(y)$$

is of the form specified in $\dot{\mathcal{P}}(P_0)$, and efficiency of $\theta(\hat{F})$ follows provided regularity can be shown. \square

Example 9.3.4. *Censoring and truncation tangent spaces (Examples 9.1.3, 9.1.9, and 9.2.2).* It turns out that, in these cases,

$$\dot{\mathcal{P}}(P_0) = L_2^0(P_0)$$

so that any regular estimate of a parameter is efficient for that parameter. This follows, since given any distribution P of (Y, δ) , there is a unique (F, G) such that $P = P_{(F, G)}$. This is a consequence of

$$f(y) = \lambda_F(y) e^{-\Lambda_F(y)}$$

and

$$\lambda_F(y) = \frac{h_1(y)}{H(y)}$$

where

$$h_1(y) = \frac{d}{dy} H_1(y), \quad H_1(y) = P[Y \leq y, \delta = 1], \quad H(y) = P[Y \leq y]$$

are determined by P and determine P (as functions) — see Problem 9.3.9. \square

Example 9.3.5. *The Cox model tangent space (Examples 9.1.4, 9.1.10, and 9.2.3).* Here the model is

$$p(\mathbf{z}, y; \beta, \lambda, h) = h(\mathbf{z})\lambda(y)r(\mathbf{z}, \beta)\exp\{-r(\mathbf{z}, \beta)\Lambda(y)\}.$$

Let P_0 correspond to the parameter values β_0, λ_0, h_0 . We will use the additive property of efficient score functions (Proposition 9.3.4) and compute tangent spaces for each parameter with the other parameters fixed then combine these tangent spaces. First, with $\lambda = \lambda_0$ and $h = h_0$ fixed, $\beta_t = \beta_0 + t\mathbf{a}$, $a \in R^d$, p_{1t} the density p corresponding to the parameters β_t, λ_0, h_0 , and $\Lambda_0(t) = \int_0^y \lambda_0(s)ds$, we find

$$\dot{\mathcal{P}}_1(P_0) : \frac{\partial}{\partial t} \log p_{1t} \Big|_{t=0} = \mathbf{a}^T \left\{ \frac{\dot{r}}{r}(\mathbf{z}, \beta_0) - \dot{r}(\mathbf{z}, \beta_0)\Lambda_0(y) \right\} \quad (9.3.22)$$

where \dot{r} is the gradient $\nabla r(\cdot, \beta)$ with respect to β . Similarly, let $\beta = \beta_0, h = h_0$ be fixed, let $b(y)$ be a function with $E[b(Y)] = 0$, and set $\lambda_t(y) = \lambda_0(y)\exp\{tb(y)\}$. If p_{2t} is the density for these parameters, we find

$$\dot{\mathcal{P}}_2(P_0) : \frac{\partial}{\partial t} \log p_{2t} \Big|_{t=0} = b(y) - r(\mathbf{z}, \beta_0) \int_0^y b(s)d\Lambda_0(s).$$

Next, with $\beta = \beta_0, \lambda = \lambda_0$ fixed and $\log h_t(\mathbf{z}) = \log h_0(\mathbf{z}) + c(\mathbf{z})t$ where $E[c(\mathbf{Z})] = 0$; and with p_{3t} the corresponding density, we find

$$\dot{\mathcal{P}}_3(P_0) : \frac{\partial}{\partial t} \log p_{3t} \Big|_{t=0} = c(\mathbf{Z}).$$

It follows formally that the tangent set $\dot{\mathcal{P}}(P_0) = \dot{\mathcal{P}}_1(P_0) + \dot{\mathcal{P}}_2(P_0) + \dot{\mathcal{P}}_3(P_0)$ is the class of functions of the form

$$\left[\mathbf{a}^T \left\{ \frac{\dot{r}}{r}(\mathbf{z}, \beta_0) - \dot{r}(\mathbf{z}, \beta_0)\Lambda_0(y) \right\} \right] + \left[b(y) - r(\mathbf{z}, \beta_0) \int_0^y b(s)d\Lambda_0(s) \right] + [c(\mathbf{z})] \quad (9.3.23)$$

for functions $b(\cdot)$ and $c(\cdot)$ with $E[b(Y)] = E[c(\mathbf{Z})] = 0$.

The influence function of the Cox estimate as given by (9.2.15) can be shown to be of this form for the Cox model when $r(\mathbf{z}, \beta) = \exp\{\beta^T \mathbf{z}\}$. In this case $\dot{r} = \mathbf{z} \exp(\beta_0^T \mathbf{z})$ and $(\dot{r}/r) = \mathbf{z}$, and with $Y = \text{failure time}$,

$$S_0(y, \beta, P) = E_P[e^{\beta^T \mathbf{z}} 1(Y \geq y)].$$

Let \dot{S}_0 stand for the gradient with respect to β . We proceed with $\beta \in R$, then $\dot{S}_0(y, \beta, P) = E_P[Z e^{\beta^T \mathbf{z}} 1(Y \geq y)]$. Now, in the case where $1(y \leq \tau) = 1$, differentiate (9.2.15) with respect to y and show that it may be written in the form (9.3.23). We can show that the derivatives are equal; see Problem 9.3.10. Efficiency follows. \square

Example 9.3.6: *Partial linear model tangent space (Examples 9.1.13 and 9.2.4).* Assuming that $\varepsilon \sim \mathcal{N}(0, \sigma^2)$, we obtain as members of $\dot{\mathcal{P}}(P_0)$, $\mathbf{Z}^T \varepsilon$, and $(\varepsilon^2 / \sigma_0^2) - 1$. The nonparametric g , appearing as g_t in a 1 dimensional submodel, yields

$$\frac{\partial}{\partial t} \frac{1}{2\sigma_0^2} (Y - \beta_0^T \mathbf{Z} - g_t(U))^2|_{t=0} = -\frac{\varepsilon}{\sigma_0^2} \frac{\partial g_t}{\partial t}(U)|_{t=0}.$$

This gives us

$$\dot{\mathcal{P}}(P_0) = \left[\frac{\varepsilon^2}{\sigma_0^2} - 1, h(U)\varepsilon, w(\mathbf{Z}), \mathbf{Z}\varepsilon \right]$$

for all h, w such that $E_0 h^2(U) < \infty$, $E_0 w^2(\mathbf{Z}) < \infty$. The influence function of $\hat{\beta}$ may be shown to be

$$[E \operatorname{Var}_0 (\mathbf{Z}|U)]^{-1} (\mathbf{Z} - E_0(\mathbf{Z}|U))\varepsilon$$

whose components do indeed belong to $\dot{\mathcal{P}}(P_0)$. \square

We leave further consideration of these examples and other models to the problems.

9.4 Tests and Empirical Process Theory

It is natural, when entertaining a parametric model, $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$, to consider its consistency with the data. If we have no strong *a priori* evidence of departure from the model in any particular direction, we naturally begin by considering $H : P \in \mathcal{P}$ versus $K : P \notin \mathcal{P}$. As usual, with a formulation as general as this one, nothing can be done. But begin with a blanket hypothesis such as our usual one, $\mathbf{X} = (X_1, \dots, X_n)$, X_j i.i.d F . Then we can identify \mathcal{P} with $\mathcal{F} = \{F_\theta : \theta \in \Theta\}$, $\Theta \subset R^d$ Euclidean, a regular parametric model and test $H : F \in \mathcal{F}$ vs $K : F \notin \mathcal{F}$, using measures of distance between F and \mathcal{F} as we discussed in Section 4.1 and Example I.2. Here failure to reject the hypothesized model only suggests that a model may be an adequate approximation. Recall from Section 1.1.1 that “Models, of course, are never true but fortunately it is only necessary that they be useful,” Box (1979). We pursue model testing in more detail here. Strictly speaking, we are dealing with parametric hypotheses in nonparametric models.

Example 9.4.1. *Testing goodness-of-fit to the Gaussian distribution (Example 4.1.6 and Section I.2 continued).* Consider the test statistic T_n introduced in Example 4.1.6 for testing $H : F(\cdot) = \Phi(\frac{\cdot - \mu}{\sigma})$ for some μ, σ vs $K : F$ not Gaussian,

$$T_n \equiv \sup_x |\widehat{F}(\bar{X} + \widehat{\sigma}x) - \Phi(x)| = \sup_x \left| \widehat{F}(x) - \Phi\left(\frac{x - \bar{X}}{\widehat{\sigma}}\right) \right|.$$

We claim that, under H ,

$$\mathcal{L}(T_n) \rightarrow \mathcal{L}(\sup_x |Z(x)|) \tag{9.4.1}$$

where $Z(\cdot)$ is a Gaussian process with mean 0 and, if $x \leq y$,

$$\operatorname{Cov}(Z(x), Z(y)) = \Phi(x)(1 - \Phi(y)) - \varphi(x)\varphi(y) - \frac{1}{2}xy \varphi(x)\varphi(y) \tag{9.4.2}$$

As we have noted, to obtain critical values for this statistic and some others we study below, this result is unnecessary since we can always employ the Monte Carlo method described in Examples 4.1.6 and 10.1.1. However, the proof gives us a sense of how the behaviour of this statistic differs from the Kolmogorov statistic of Example 4.1.5 and suggests a general approach for constructing and analyzing goodness-of-fit statistics for composite hypotheses.

To establish (9.4.1), assume without loss of generality that the true member of H under which we compute is $\mathcal{N}(0, 1)$. Then,

$$\begin{aligned} Z_n(x) &= \sqrt{n}(\widehat{F}(x) - \Phi(x)) + \sqrt{n} \left(\Phi\left(\frac{x - \bar{X}}{\widehat{\sigma}}\right) - \Phi(x) \right) \\ &\equiv \mathcal{E}_n(x) + \Delta_n(x), \end{aligned}$$

where $\mathcal{E}_n(\cdot)$ is the empirical process. Expand $\Delta_n(x)$ as a function of \bar{X} and $\widehat{\sigma}^2$ around 0 and 1 to get (Problem 9.4.1)

$$\Delta_n(x) = -\varphi(x)\sqrt{n}\bar{X} - \frac{x}{2}\varphi(x)\sqrt{n}(\widehat{\sigma}^2 - 1) + R_n(x) \quad (9.4.3)$$

where $\sup_x |R_n(x)| = O_P(n^{-1})$. Let $1_x(y) = 1(y \leq x)$ and

$$f_x(y) = 1_x(y) - \varphi(x)y - \frac{x\varphi(x)}{2}(y^2 - 1).$$

By definition

$$\mathcal{E}_n f_x(y) = \int f_x(y) d\mathcal{E}_n(y) = Z_n(x) - \varphi(x)\sqrt{n}\bar{X} \frac{x\varphi(x)}{2}\sqrt{n}(\widehat{\sigma}^2 - 1).$$

So if we define $\tilde{Z}_n(x) \equiv \mathcal{E}_n(f_x)$, (9.4.3) yields $\|\tilde{Z}_n - Z_n\|_\infty = o_P(1)$. It follows that $Z_n \Rightarrow Z_0$ where Z_0 is the limit of \tilde{Z}_n . It is easy to show that $\{f_x(\cdot) : x \in R\}$ satisfies the conditions of Theorem 7.1.5 and hence that $Z_0(x) = Z(x) = W_\Phi^0(f_x(\cdot)) \equiv W^0(\Phi(x))$ where, in the notation of Section I.1, $W^0(\cdot)$ is the Brownian bridge on $[0, 1]$. That $\text{Cov}(Z(x), Z(y))$ is as specified in (9.4.2) can be calculated directly (Problem 9.4.1).

Formula (9.4.2) shows that the limit distribution of T_n is different from that (see 7.1.24) of the Kolmogorov-Smirnov statistic corresponding to $H : F = \Phi$ but this gives little further insight. However, here is another derivation of (9.4.2). We introduce some notation; see Appendix B.10. If \mathcal{H} is a Hilbert space and $\mathcal{G} \subset \mathcal{H}$, let $[\mathcal{G}]$ be the smallest closed linear subspace of \mathcal{H} containing \mathcal{G} . If $\mathcal{G} = \{h_1, \dots, h_k\}$, then $[\mathcal{G}] = \{c_1h_1 + \dots + c_kh_k : \mathbf{c} = (c_1, \dots, c_k) \in R^k\}$. Take $\mathcal{H} = L_2(\Phi)$, $\mathcal{L} = [X, X^2 - 1]$ where $X \sim \mathcal{N}(0, 1)$. Then $[X] \perp [X^2 - 1]$ and for any $g(X) \in L_2(\Phi)$, if $\Pi(\cdot | \mathcal{L})$ denotes projection on the closed linear space \mathcal{L} ,

$$\Pi(g | \mathcal{L}) = \Pi(g | [X]) + \Pi(g | [X^2 - 1])$$

and hence

$$\Pi(g | \mathcal{L})(y) = E(g(X)X)y + \frac{Eg(X)(X^2 - 1)}{2}(y^2 - 1).$$

If $g(y) = 1(y \leq x)$ which we abbreviate as $1_x(y)$,

$$\begin{aligned} Eg(X)X &= \int_{-\infty}^x y\varphi(y)dy = -\varphi(x) \\ Eg(X)(X^2 - 1) &= \int_{-\infty}^x y^2\varphi(y)dy - \Phi(y) \\ &= - \int_{-\infty}^x y\varphi(y)dy - \Phi(y) = -x\varphi(y). \end{aligned}$$

Then $f_x(\cdot) = 1_x(\cdot) - \Pi(1_x|\mathcal{L})(\cdot)$ and, by the usual properties of projections,

$$\begin{aligned} \text{Cov}(W_\Phi^0(f_x), W_\Phi^0(f_y)) &= \text{Cov}_\Phi(f_x, f_y) = \text{Cov}_\Phi(1_x(X), 1_y(X)) \\ - \text{Cov}_\Phi(\Pi(1_x|\mathcal{L}), \Pi(1_y|\mathcal{L})) &= \text{Cov}_\Phi(\Pi^\perp(1_x(X)), \Pi^\perp(1_y(X))) \end{aligned} \quad (9.4.4)$$

where Π^\perp is projection on the orthocomplement of $\dot{\mathcal{P}}(P_0)$ in $L_2^0(P_0)$ and we write $\text{Cov}_\Phi(g, h)$ for $\text{Cov}(g(X), h(X))$ with $X \sim \mathcal{N}(0, 1)$. Thus, as we might expect in the limit, the process $Z_n(\cdot)$ is more concentrated than $\mathcal{E}_n(\cdot)$ in the sense that all linear functionals of $Z_n(\cdot)$ have smaller variances than the same functional of $\mathcal{E}_n(\cdot)$ in the limit. The implications of this calculation for power are intuitively correct. The tests based on $Z_n(\cdot)$ are less powerful against the same alternative than the corresponding tests based on $\mathcal{E}_n(\cdot)$. See Problem 9.4.5. We generalize this argument in the next example. \square

Other test statistics' limit distributions can be similarly derived. For instance, the *Cramér-von Mises* type statistic for this situation is given by (see Problem 4.1.1 and Example 7.1.6)

$$S_n \equiv n \int_{-\infty}^{\infty} (\hat{F}(x) - \Phi(\frac{x - \bar{X}}{\hat{\sigma}}))^2 d\Phi(x) \quad (9.4.5)$$

Clearly, $h \rightarrow \int_{-\infty}^{\infty} h^2 d\Phi$, is a continuous map from $L_\infty(R)$ to R . Hence,

$$\mathcal{L}_\Phi(S_n) \longrightarrow \mathcal{L}\left(\int_{-\infty}^{\infty} [W_\Phi^0(f_x(\cdot))]^2 d\Phi(x)\right).$$

This limit law has an explicit form — see Shorack and Wellner (1986), for instance.

Example 9.4.2. *Goodness of fit to a general smooth parametric model.* We assume that $\mathcal{P} \equiv \{P_\theta : \theta \in \Theta \subset R^d\}$ is a smooth parametric model satisfying the assumptions of Theorem 6.2.2. Let T be a set of functions on \mathcal{X} satisfying the assumptions of Theorem 7.1.5. Let, for $f \in T$,

$$Z_n(f) = \frac{1}{\sqrt{n}} \sum_{i=1}^n (f(X_i) - E_{\hat{\theta}}(f(X))) \quad (9.4.6)$$

where $\hat{\theta}$ is the MLE of θ . Here we compute $E_\theta(f(X))$ then substitute $\theta = \hat{\theta}$ to get $E_{\hat{\theta}}(f(X))$. It is reasonable that suitable goodness-of-fit statistics for $H : P \in \mathcal{P}$ should

be based on the magnitude of $|Z_n(\cdot)|$. For instance, if $\mathcal{X} = R$, $f(x) = 1_x$, the natural generalization of the Kolmogorov statistic is

$$T_n = \sup_x |\widehat{F}(x) - F_{\widehat{\boldsymbol{\theta}}}(x)| = \frac{1}{\sqrt{n}} \sup_x |Z_n(1_x)|.$$

We claim that, under H with the data \mathbf{X} generated by $P_0 \equiv P_{\boldsymbol{\theta}_0}$,

$$Z_n(\cdot) \implies Z_{\boldsymbol{\theta}_0}^0(\cdot). \quad (9.4.7)$$

where $Z_{\boldsymbol{\theta}_0}^0(f) = W_{\boldsymbol{\theta}_0}^0(f - \Pi_{\boldsymbol{\theta}_0}(f))$. Here $W_{\boldsymbol{\theta}_0}^0 = W_{P_0}^0$, $\Pi_{\boldsymbol{\theta}_0}(f) \equiv \Pi(f \mid [\nabla l(X, \boldsymbol{\theta}_0)])$ denotes projection in $L_2(P_0)$, and $\nabla l(X, \boldsymbol{\theta}_0) = \left(\frac{\partial l}{\partial \theta_1}(X, \boldsymbol{\theta}_0), \dots, \frac{\partial l}{\partial \theta_d}(X, \boldsymbol{\theta}_0) \right)^T$. Then,

$$\begin{aligned} \text{Cov}(Z_{\boldsymbol{\theta}_0}^0(f), Z_{\boldsymbol{\theta}_0}^0(g)) &= \text{Cov}_{\boldsymbol{\theta}_0}(f(X), g(X)) - \text{Cov}_{\boldsymbol{\theta}_0}(\Pi_{\boldsymbol{\theta}_0}(f)(X), \Pi_{\boldsymbol{\theta}_0}(g)(X)) \\ &= \text{Cov}_{\theta_0}(\Pi_{\boldsymbol{\theta}_0}^\perp(f), \Pi_{\boldsymbol{\theta}_0}^\perp(g)). \end{aligned} \quad (9.4.8)$$

We have discussed the consequences of this formula in terms of power above. The proof of (9.4.7) is a straightforward extension of (9.4.4). Write

$$Z_n(f) = \mathcal{E}_n(f) - \sqrt{n}(E_{\widehat{\boldsymbol{\theta}}} f(x) - E_{\boldsymbol{\theta}_0} f(x)).$$

Write the last term as

$$\sqrt{n} \int f(x)(p(x, \widehat{\boldsymbol{\theta}}) - p(x, \boldsymbol{\theta}_0)) d\mu(x).$$

We can justify Taylor expansion to get

$$\begin{aligned} &\sqrt{n}(E_{\widehat{\boldsymbol{\theta}}} f(x) - E_{\boldsymbol{\theta}_0} f(x)) \\ &= \left(\int f(x) \nabla l(X, \boldsymbol{\theta}_0) p(x, \boldsymbol{\theta}_0) d\mu(x) \right)^T \cdot \sqrt{n}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) + o_p(1). \end{aligned} \quad (9.4.9)$$

Now by (6.2.10), (9.4.9) is equal to

$$\begin{aligned} &[E_{\boldsymbol{\theta}_0} f(x) \nabla l(X, \boldsymbol{\theta}_0)]^T \frac{1}{\sqrt{n}} \sum_{i=1}^n I^{-1}(\boldsymbol{\theta}_0) \nabla l(X_i, \boldsymbol{\theta}_0) + o_p(1) \\ &= \mathcal{E}_n([E_{\boldsymbol{\theta}_0} f(X) \nabla l(X, \boldsymbol{\theta}_0)]^{-1} I^{-1}(\boldsymbol{\theta}_0) \nabla l(\cdot, \boldsymbol{\theta}_0)). \end{aligned} \quad (9.4.10)$$

But the argument of \mathcal{E}_n is just $\Pi(f \mid [\nabla l(X, \boldsymbol{\theta}_0)])$ by (B.10.20), and (9.2.7) follows. The validity of (9.4.8) comes from the general property (B.10.14) applied to $H = L_2(P_0)$ and $L = [\nabla l(X, \boldsymbol{\theta}_0)]$,

$$\text{Cov}_P(f - \Pi(f \mid \mathcal{L})(X), g(X)) = 0$$

for all $g \in \mathcal{L}$.

□

The result of this example is implicit in the work of Darling (1955). See also Durbin (1973). An excellent and very complete treatment of goodness-of-fit tests such as these is Chapter 5 of Shorack and Wellner (1986).

The parametric bootstrap

How do we use the result (9.4.7)? Suppose that all the constants implicit in Theorem 7.1.5 and the conditions of Theorem 6.2.2 do not depend on P_{θ} for θ in small open sets. This implies that all stochastic convergence and approximations that we claim are uniform in θ on these open sets. We want to set an approximate critical value for a test statistic, $T_n = q(Z_n(\cdot))$ where q is continuous on $l_\infty(T)$. It is natural to use the *parametric bootstrap test* which we define as follows.

- (i) Compute $\hat{\theta}$, T_n .
- (ii) Generate B samples of size n , $(X_{b1}^*, \dots, X_{bn}^*)$, $b = 1, \dots, B$ from $P_{\hat{\theta}}$.
- (iii) Compute T_{bn}^* , $b = 1, \dots, B$, the test statistic $T_n(X_{b1}^*, \dots, X_{bn}^*)$ for each of the B samples. For instance, if $T_n = \sup_x |\hat{F}(x) - F_{\hat{\theta}}(x)|$, then
 $T_{bn}^* = \sup_x |\hat{F}_b^*(x) - F_{\hat{\theta}_b}(x)|$ where \hat{F}_b^* is the empirical df of $(X_{b1}^*, \dots, X_{bn}^*)$ and $\hat{\theta}_b^*$ is the MLE computed on the basis of X_{bj}^* , $j = 1, \dots, n$.
- (iv) Reject H at level α iff $T_n > T_{(b(1-\alpha)]+1)}^*$ where $T_{(1)}^* \leq \dots \leq T_{(B)}^*$ are the ordered T_{nb}^* , $1 \leq b \leq B$.

The result (9.4.7) suggests that this procedure provides an approximate size α test if the true θ_0 belongs to the interior of Θ . Here is a sketch proof under the assumption that the limit law $\mathcal{L}_{\theta_0}(q(Z_{\theta_0}))$ of $T_n = q(Z_n)$ is continuous. Note first that

$$\frac{1}{B} \sum_{b=1}^B \mathbb{1}(T_{nb}^* > T_n) - P_{\hat{\theta}}[T_n^* > T_n] \xrightarrow{P} 0 \quad (9.4.11)$$

as $B \rightarrow \infty$ by Chebyshev's inequality. Both expressions in (9.4.11) are to be interpreted as conditional on X_1, \dots, X_n , that is, $T_n^* = T_n(X_{11}^*, \dots, X_{1n}^*)$ where the X_{1i}^* are i.i.d $P_{\hat{\theta}(X_1, \dots, X_n)}$, given X_1, \dots, X_n . But under our assumptions,

$$\sup\{|P_{\theta}[T_n > x] - P_{\theta}[q(Z_{\theta}) > x]| : x \in R, |\theta - \theta_0| \leq \varepsilon\} \longrightarrow 0. \quad (9.4.12)$$

Since $\hat{\theta} \rightarrow \theta_0$ in P_{θ_0} probability, (9.4.11) and (9.4.12) imply that if T_n and T'_n have the same distribution but are independent, $P_{\hat{\theta}}[T_n^* > T_n] - P_{\theta_0}[T'_n > T_n] \rightarrow 0$ in P_{θ_0} probability. This is what we needed to argue.

This technique is the first example of bootstrap techniques, Monte Carlo in the service of inference, that we will investigate further in Chapter 10.

To make matters concrete, we specialize to

Example 9.4.3. *Goodness of fit for the gamma family.* Suppose \mathcal{P} is the $\Gamma(p, \lambda)$ family, a possible model for lifetime distributions (see Lawless (1982)). We specialize to $T = \{1(0, u) : u \in R\}$ or equivalently to statistics based on

$$Z_n(u) = \sqrt{n}(\widehat{F}(u) - G_{\widehat{p}, \widehat{\lambda}}(u))$$

where $(\widehat{p}, \widehat{\lambda})$ is the MLE of (p, λ) and $G_{p, \lambda}(u)$ is the c.d.f. of $\Gamma(p, \lambda)$. In particular, consider $T_n = \sup_u |Z_n(u)|$, a Kolmogorov-Smirnov type statistic. Since

$$G_{p, \lambda}(u) = G_{p, 1}(\lambda u)$$

$$T_n = \sqrt{n} \sup_u |\widehat{F}\left(\frac{u}{\lambda}\right) - G_{\widehat{p}, 1}(u)|$$

so that we can simulate under $\lambda = 1$. However, no such simplification is possible for the parameter p . Thus, to get critical values for T_n we need to apply the parametric bootstrap and simulate from $G_{\widehat{p}, 1}(\cdot)$ as we discussed. \square

Remark 9.4.1.

(a). Goodness of fit tests are often used as diagnostics. Thus, for instance, in the Gaussian goodness-of-fit problem we might want to assess where the deviations from Gaussianity are significant by taking into account that the standard deviation of the empirical process at a point depends on that point. Under normality, $\widehat{F}(\mu + \sigma u)$ has variance $\Phi(1 - \Phi)(u)/n$. Thus it is natural to consider

$$\mathcal{S}_n = n^{\frac{1}{2}} \sup\{|\widehat{F}(\bar{X} + \widehat{\sigma}u) - \Phi(u)|[\Phi(1 - \Phi)]^{-\frac{1}{2}}(u) : |u| \leq M\}. \quad (9.4.13)$$

Then, not only is a large value of \mathcal{S}_n indicative of departure from Gaussianity, but where that departure occurs is weighted appropriately as in Example 4.4.3. We cannot take $M = \infty$ here (Problem 9.4.2). However, it is possible to consider statistics such as

$$n \int_0^1 \frac{\left(\widehat{F}(x) - \Phi\left(\frac{x - \bar{X}}{\widehat{\sigma}}\right)\right)^2}{\Phi(x)(1 - \Phi(x))} d\Phi(x)$$

which are asymptotically equivalent to the well known Shapiro-Wilk statistics — see Mecklin and Mundfrom (2004) for a review.

It is possible, as in Section 6.3, to make power calculations for such tests. Consider $H : F = F_0$. For a sequence of alternatives $\{F_n\}$ with $F_n(x) = F_0(x) + n^{-\frac{1}{2}}(\Delta(x) + o(1))$ uniformly in x , it is possible to obtain a limit distribution for $Z_n(\cdot)$ in terms of a Gaussian process. Unfortunately, the family of Δ we need to consider is too large and the resulting expressions generally analytically not tractable. For some calculation of this type see Hajek and Sidak (1967). These results may also be put in the general framework of contiguity, which we discuss further in Section 9.5. It is possible to extend these goodness-of-fit tests we have described to hypotheses such as that of independence of U and V where, in general, $X = (U, V)$ has an arbitrary distribution. Such a hypothesis is semiparametric. This extension is discussed in the problems. A general approach to constructing tests of goodness-of-fit to semiparametric hypotheses is given in Bickel, Ritov and Stoker (2003).

9.5 Asymptotic Properties of Likelihoods. Contiguity

As we have seen in Chapters 5 and 6 the asymptotic behavior of maximum likelihood estimates and tests and related confidence regions corresponds to the exact behavior of the same procedures specialized to the Gaussian linear model with known variance. In this section we will show how these results follow heuristically from properties of the likelihood which are by no means limited to i.i.d. observations. The theory originated in early work of Wald (1943) but was brought to full generality and understanding largely by Le Cam in a series of papers starting in 1956 and culminating in his (1986) treatise. Hájek (1972) also made important contributions. We shall derive results only under the strong conditions of Chapters 5 and 6 rather than under the elegant necessary and sufficient conditions of Le Cam. But we will state the general results in the appropriate Hilbert space context with references to their appearance and proofs in Le Cam and Yang (1990), Le Cam (1986), and Hájek and Sidák (1967). In this way we shall connect this theory with the semiparametric estimation theory of Sections 9.1–9.3.

We begin by motivating the abstract definitions which follow by a closer examination of the i.i.d. case. We need to take the point of view of Section 9.3 in which we introduced *local parametric models*. That is, we fix P_{θ_0} and then consider regular parametric submodels of the form $\mathcal{P}_\varepsilon \equiv \{P_\theta : |\theta - \theta_0| \leq \varepsilon\}$, where ε is arbitrarily small and θ belongs to R^d . We formalize this by reparameterizing the model using the sample size dependent scale of order $n^{-1/2}$ in the i.i.d. case. That is, we consider

$$\mathcal{P}_{M,n} \equiv \{P_{\theta_0 + \frac{\mathbf{t}}{\sqrt{n}}} : |\mathbf{t}| \leq M\} \quad (9.5.1)$$

for $M < \infty$ arbitrary. Note that P_{θ_0} could be any P residing in a semiparametric model. What matters is that we construct a sequence of shrinking regular parametric models centered at P_{θ_0} .

This enables us, following Le Cam, to think about local approximations to models and all related decision theoretic problems rather than just to distributions of estimates or tests. In particular we can formulate all the optimality properties of Chapters 5 (Sections 5.4.3 and 5.4.4), 6 (6.2.2), and Section 9.3 in terms of the local parameter \mathbf{t} . If θ_n^* is a regular estimate of θ , then, by the plug-in principle, $\mathbf{t}_n^* \equiv \sqrt{n}(\theta_n^* - \theta_0)$ is the corresponding estimate of \mathbf{t} and we can state properties in terms of \mathbf{t}_n^* . For instance, optimality of the MLE $\widehat{\theta}_n$ as defined in (9.3.1) and (9.3.2) now reads: Under regularity conditions, if \mathbf{t}_n^* is asymptotically Gaussian with mean $\mathbf{0}$ uniformly in \mathbf{t} on all $\mathcal{P}_{M,n}$, that is,

$$\mathcal{L}_{\theta_0 + \frac{\mathbf{t}_n^*}{\sqrt{n}}}(\mathbf{t}_n^*) \rightarrow \mathcal{N}_d(\mathbf{t}, \Sigma^*) \quad (9.5.2)$$

uniformly for $|\mathbf{t}| \leq M$, all $M < \infty$, then

$$\Sigma^* \geq I^{-1}(\theta_0)$$

and, if $\widehat{\mathbf{t}}_n$ is the MLE of \mathbf{t} ,

$$\mathcal{L}_{\theta_0 + \frac{\mathbf{t}_n}{\sqrt{n}}}(\widehat{\mathbf{t}}_n) \rightarrow \mathcal{N}_d(\mathbf{t}, I^{-1}(\theta_0))$$

uniformly for $|t| \leq M$, all $M < \infty$. We also note that for the model

$$\mathcal{N}_d \equiv \{\mathcal{N}_d(t, I^{-1}(\boldsymbol{\theta}_0)) : t \in R^d\}$$

the various asymptotic optimality statements become exact for fixed n . We can read this as saying that \mathcal{N}_d is an approximation in a strong sense to the set of local model $\{\mathcal{P}_{M,n} : M < \infty\}$ as $n \rightarrow \infty$.

We begin by introducing the fundamental notion of Local asymptotic normality (LAN) of general likelihoods. For $\Theta_n \subset R^d$, let $\mathcal{E}_n \equiv \{P_{\boldsymbol{\theta}}^{(n)} : \boldsymbol{\theta} \in \Theta_n\}$, ; $n \geq 1$, be a sequence of experiments (models). By this notation we mean that, for the n th experiment, we observe $\mathbf{X}^{(n)} \in \mathcal{X}^{(n)}$, $\mathbf{X}^{(n)} \sim P_{\boldsymbol{\theta}}^{(n)}$, $\boldsymbol{\theta} \in \Theta_n$. We can think of $\mathbf{X}^{(n)}$ as being a vector $(X_1, \dots, X_n)^T$ belonging to R^n . In the i.i.d. case, where the X_i are i.i.d. with common density $f(x, \boldsymbol{\theta})$, $P_{\boldsymbol{\theta}}^{(n)}$ has density

$$p_n(\mathbf{x}, \boldsymbol{\theta}) = \prod_{i=1}^n f(x_i, \boldsymbol{\theta}).$$

The parametrization of Θ_n we consider usually depends on n . It contains open balls around a fixed point $\boldsymbol{\theta}_0$ so that we can talk about $\boldsymbol{\theta}_n = (\boldsymbol{\theta}_0 + t\gamma_n) \in \Theta_n$, where $\gamma_n \downarrow 0$, and $|t| \leq M$ for some $M > 0$ and all t if n is large enough. Having $\gamma_n \downarrow 0$ is what makes our calculations local. In all “regular” i.i.d. cases, $\gamma_n = n^{-\frac{1}{2}}$. Note that we have already made local calculations like this in Section 5.4.4, in connection with testing.

The local likelihood ratio is defined as

$$L_n(t) = \frac{p_n(\mathbf{X}^{(n)}, \boldsymbol{\theta}_n)}{p_n(\mathbf{X}^{(n)}, \boldsymbol{\theta}_0)}, \quad \boldsymbol{\theta}_n \equiv \boldsymbol{\theta}_0 + t\gamma_n. \quad (9.5.3)$$

We suppress $\boldsymbol{\theta}_0$ in the notation and we use the conventions that if $p^{(n)}(\cdot, \boldsymbol{\theta}_0) = 0$ and $p^{(n)}(\cdot, \boldsymbol{\theta}_n) > 0$, then $L_n(t) = \infty$, and that $0/0 = 0$. Finally let

$$\Lambda_n(t) \equiv \log L_n(t).$$

Definition 9.5.1. \mathcal{E}_n is uniformly locally asymptotically normal (ULAN) at $\boldsymbol{\theta}_0$ iff there exist $d \times 1$ statistics $\{\mathbf{T}_n(\mathbf{X}^{(n)}, \boldsymbol{\theta}_0)\}$, $n \geq 1$, such that we can write

$$\Lambda_n(t) = \mathbf{t}^T \mathbf{T}_n(\mathbf{X}^{(n)}, \boldsymbol{\theta}_0) - \frac{1}{2} \mathbf{t}^T \Sigma_0 \mathbf{t} + R_n(t)$$

where

$$(i) \sup\{|R_n(t)| : |t| \leq M\} \xrightarrow{P_{\boldsymbol{\theta}_0}} 0 \text{ for all } M < \infty.$$

$$(ii) \mathcal{L}_{\boldsymbol{\theta}_0}(\mathbf{T}_n(\mathbf{X}^{(n)}, \boldsymbol{\theta}_0)) \longrightarrow \mathcal{N}_d(\mathbf{0}, \Sigma_0).$$

Remark 9.5.1. If $\mathbf{X}_1, \dots, \mathbf{X}_n$ are i.i.d. $\mathcal{N}_d(\boldsymbol{\theta}, \Sigma_0^{-1})$ and $\boldsymbol{\theta} = \mathbf{0}$, then

$$\Lambda_n(t) = \mathbf{t}^T \Sigma_0 (n^{-\frac{1}{2}} \sum_{i=1}^n \mathbf{X}_i) - \frac{1}{2} \mathbf{t}^T \Sigma_0 \mathbf{t}$$

and $n^{-\frac{1}{2}} \sum_{i=1}^n \Sigma_0 \mathbf{X}_i \sim \mathcal{N}_d(\mathbf{0}, \Sigma_0)$.

We have already encountered the ULAN property in Examples 7.1.1 and 7.1.2. We show there that if $\{P_\theta : \theta \in \Theta\}$ is a model satisfying the regularity condition of these examples, then $\mathcal{E}_n \equiv \{P_\theta^{(n)} : \theta \in \Theta\}$, where $P_\theta^{(n)}$ is the (product) joint distribution of $\mathbf{X}^{(n)} \equiv (X_1, \dots, X_n)$ i.i.d. P_θ , is ULAN at all θ_0 with

$$\mathbf{T}_n(\mathbf{X}^{(n)}, \theta_0) = n^{-1/2} \sum_{i=1}^n i(X_i, \theta_0),$$

and $\Sigma_0 = I(\theta_0)$, the information matrix.

We shall also have use for the following general property. Let $\mathcal{X}^{(n)}$ be a sequence of probability spaces. Typically $\mathbf{X}^{(n)} \in \mathcal{X}^{(n)}$ with $\mathcal{X}^{(n)}$ as a subset of R^n .

Definition 9.5.2. Given two sequences of probabilities $\{P^{(n)}\}, \{Q^{(n)}\}$, $n \geq 1$, on $\mathcal{X}^{(n)}$ we say that $\{Q^{(n)}\}$ is *contiguous* to $\{P^{(n)}\}$ iff for any sequence of events $\{A_n\}$, $n \geq 1$, such that $P^{(n)}(A_n) \rightarrow 0, Q^{(n)}(A_n) \rightarrow 0$ also.⁽²⁾

Contiguity has a very simple statistical interpretation. Any test of $H : P = P^{(n)}$ vs $K : Q = Q^{(n)}$ which asymptotically has probability of type I error 0 must asymptotically have probability of type II error 1.

Contiguity can be used to turn asymptotic results for a relatively simple $P^{(n)}$ into asymptotic results for a more complicated $Q^{(n)}$ because if an approximation S_n to a statistic T_n satisfies $T_n - S_n = o_P(1)$ under $P^{(n)}$, then $T_n - S_n = o_P(1)$ under $Q^{(n)}$ also. For instance:

Example 9.5.1. Consider the uniform score statistic

$$T_U = n^{-\frac{1}{2}} \sum \frac{R_i}{n+1} (z_i - \bar{z})$$

of Example 8.3.11. Under the hypothesis H , an approximation to T_U is

$$S_U = n^{-\frac{1}{2}} \sum U_i (z_i - \bar{z}), \quad U_i = F(Y_i).$$

Here $U_i \sim \mathcal{U}(0, 1)$ under H . By ordering the U_i we can write

$$S_U - T_U = n^{-\frac{1}{2}} \sum \left(U_{(i)} - \frac{i}{n+1} \right) (z_{(i)} - \bar{z})$$

where $z_{(i)}$ is the predictor value of the sample point with response $Y_{(i)}$. Because $E(U_{(i)}) = i/(n+1)$ we have, under H ,

$$E_H(T_U - S_U)^2 = \text{Var}(T_U - S_U).$$

Set $s^2 = n^{-1} \sum (z_i - \bar{z})^2$; then we can use Problem B.2.9 to show that

$$E[(T_U - S_U)/s]^2 \rightarrow 0$$

provided z_i , $1 \leq i \leq n$, satisfy Lindeberg's condition. See Problem 9.5.3. For contiguous alternatives Q_n , $(T_U - S_U)/s = o_P(1)$ also. By Slutsky's theorem we can find the asymptotic power of T_U for contiguous alternatives by computing $\lim_{n \rightarrow \infty} P(\sqrt{12}S_U/s \geq z_{1-\alpha})$. See Problem 9.5.4. \square

Define

$$L^{(n)}(\mathbf{X}^{(n)}) \equiv q^{(n)}(\mathbf{X}^{(n)})/p^{(n)}(\mathbf{X}^{(n)})$$

where $p^{(n)}$ and $q^{(n)}$ are densities of $P^{(n)}$ and $Q^{(n)}$ and $0/0 = 0$, $a/0 = \infty$, for $a > 0$.

A simple and powerful sufficient (and necessary) condition is embodied in

Proposition 9.5.1. (Le Cam's First Lemma). *Suppose*

- (i) $\mathcal{L}_{P^{(n)}} L^{(n)}(\mathbf{X}^{(n)}) \Rightarrow \mathcal{L}(L)$ with L finite valued, and
- (ii) $E(L) = 1$.

Then, $Q^{(n)}$ is contiguous to $P^{(n)}$.

We refer to Hájek and Sidák (1967), BKRW, and van der Vaart (2000) for a general proof of this lemma. We sketch a proof in Problem 9.5.1.

We next consider $L^{(n)}(\mathbf{X}^{(n)}) = L_n(\mathbf{t})$ as defined in (9.5.3) and, as a corollary, we obtain

Proposition 9.5.2. Suppose $\{P_{\theta}^{(n)} : \theta \in \Theta_n\}$, $n \geq 1$, is ULAN at θ_0 . Let $P^{(n)} \equiv P_{\theta_0}^{(n)}$ and $Q^{(n)} \equiv P_{\theta_n}^{(n)}$, where $\theta_n \equiv \theta_0 + \mathbf{t}\gamma_n$, $\gamma_n = n^{-\frac{1}{2}}$, and $|\mathbf{t}| \leq M < \infty$ all n . Then $Q^{(n)}$ is contiguous to $P^{(n)}$.

Proof. By Examples 7.1.1 and 7.1.2, $L_n \Rightarrow e^Z$ where $Z \sim \mathcal{N}(-\sigma^2/2, \sigma^2)$ and $\sigma^2 \equiv \mathbf{t}^T \Sigma_0 \mathbf{t}$. But, by (A.13.20), $Ee^{sZ} = \exp\{-(\sigma^2 s)/2 + (\sigma^2 s^2)/2\}$ which equals one if $s = 1$, and the result follows from Proposition 9.5.1 with $L = e^Z$. \square

The importance of this proposition is that the likelihood ratio is asymptotically determined by $T^{(n)}$ whatever θ is in the local model which suggests that $T^{(n)}$ is, in a suitable sense, asymptotically sufficient, so that any decision procedure has risk equivalent in the local asymptotic sense to the experiment in which one observes only $T^{(n)}(\mathbf{X}^{(n)}, \theta_0)$. To see that this last experiment is very simple we state and prove the fundamental

Proposition 9.5.3. (Le Cam's Third Lemma). *Suppose \mathcal{E}_n is a sequence of ULAN experiments at θ_0 , $\mathbf{T}_n \equiv \mathbf{T}_n(\mathbf{X}^{(n)}, \theta_0)$ is as in Definition 9.5.1, and $\mathbf{S}_n \equiv \{\mathbf{S}_n(\mathbf{X}^{(n)})\}$ is a sequence of p dimensional statistics such that*

$$\mathcal{L}_{\theta_0}((\mathbf{T}_n^T, \mathbf{S}_n^T)^T) \rightarrow \mathcal{N}_{d+p} \left(\begin{pmatrix} \mathbf{0}_{d \times 1} \\ \boldsymbol{\mu}_{p \times 1} \end{pmatrix}, \begin{pmatrix} \Sigma_{d \times d} C_{d \times p} \\ C_{d \times p}^T K_{p \times p} \end{pmatrix} \right).$$

Then, if $\theta_n = \theta_0 + \mathbf{t}\gamma_n$, with $\gamma_n = n^{-\frac{1}{2}}$,

$$\mathcal{L}_{\theta_n}(\mathbf{S}_n(\mathbf{X}^{(n)})) \rightarrow \mathcal{N}_p(\boldsymbol{\mu} + C^T \mathbf{t}, K). \quad (9.5.4)$$

An immediate consequence is

Proposition 9.5.4. Suppose \mathcal{E}_n is ULAN at $\boldsymbol{\theta}_0$ with $\mathbf{T}_n(\mathbf{X}^{(n)}, \boldsymbol{\theta}_0)$ and Σ_0 as in Definition 9.5.1. Then, if $\boldsymbol{\theta}_n = \boldsymbol{\theta}_0 + \mathbf{t}n^{-\frac{1}{2}}$

$$\mathcal{L}_{\boldsymbol{\theta}_n}(\Sigma_0^{-1}\mathbf{T}_n(\mathbf{X}^{(n)}, \boldsymbol{\theta}_0)) \rightarrow \mathcal{N}_d(\mathbf{t}, \Sigma_0^{-1}).$$

Remark 9.5.2. We discuss the implications of Proposition 9.5.4 before giving the proof of both Propositions 9.5.3 and 9.5.4. Since observing $\Sigma_0^{-1}\mathbf{T}_n$ and \mathbf{T}_n are equivalent this proposition says that the experiment, observing $\Sigma_0^{-1}\mathbf{T}_n$ for the model $\mathcal{P}_n = \{P_{\boldsymbol{\theta}_0+\mathbf{t}\gamma_n}^{(n)} : \mathbf{t} \in R^d\}$, is equivalent, in some approximate sense, to observing $\mathbf{W} = \mathbf{t} + \mathbf{Z}, \mathbf{t} \in R^d$ where $\mathbf{Z} \sim \mathcal{N}_k(\mathbf{0}, \Sigma_0^{-1})$; or equivalently, $\Sigma_0^{\frac{1}{2}}\mathbf{W} = \boldsymbol{\mu} + \mathbf{Z}, \boldsymbol{\mu} \in R^d$, for $\mathbf{Z} \sim \mathcal{N}_d(0, J)$, $J = d \times d$ identity. Coming back to our i.i.d. example, observing (X_1, \dots, X_n) for i.i.d. $P_{\boldsymbol{\theta}_0+\mathbf{t}n^{-1/2}}$ is asymptotically equivalent, in the rough sense we have discussed, to observing $I^{-1}(\boldsymbol{\theta}_0)n^{-1/2} \sum_{i=1}^n \dot{l}(X_i, \boldsymbol{\theta}_0)$ which, under regularity conditions, is asymptotically equivalent to observing $\hat{\mathbf{t}}_n \equiv n^{1/2}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0)$. Unfortunately, having $o_p(1)$ approximations for the likelihood ratio statistic in the local model is far from enough to justify the asymptotic sufficiency properties they suggest. We refer to Le Cam (1956, 1972, 1986) for the difficult arguments involved.

Proof of Proposition 9.5.3. The characteristic function of S_n is

$$\begin{aligned} & E_{\boldsymbol{\theta}_n} \exp\{i\mathbf{w}^T S_n\} \\ &= E_{\boldsymbol{\theta}_0} \exp\{i\mathbf{w}^T S_n + \Lambda_n(\mathbf{t})\} + O(P_{\boldsymbol{\theta}_0}[p^{(n)}(\mathbf{X}^{(n)}, \boldsymbol{\theta}_0) = 0]). \end{aligned} \quad (9.5.5)$$

The second term in (9.5.5) tends to 0 by contiguity. By Definition 9.5.1,

$$i\mathbf{w}^T S_n + \Lambda_n(\mathbf{t}) = i\mathbf{w}^T S_n - \mathbf{t}^T T_n - \frac{1}{2}\mathbf{t}^T \Sigma_0 \mathbf{t} + o_{P_{\boldsymbol{\theta}_0}}(1). \quad (9.5.6)$$

Let (\mathbf{U}, \mathbf{V}) denote the limit in law of (T_n, S_n) . The right hand side of (9.5.6) converges in law to $i\mathbf{w}^T \mathbf{V} + \mathbf{t}^T \mathbf{U} - (1/2)\mathbf{t}^T \Sigma_0 \mathbf{t}$. It can be shown (Problem 9.5.2) that

$$E \exp\{i\mathbf{w}^T \mathbf{V} + \mathbf{t}^T \mathbf{U}\} = \exp\{i\mathbf{w}^T \boldsymbol{\mu} - \frac{1}{2}(\mathbf{w}^T K \mathbf{V} + \mathbf{t}^T \Sigma_0 \mathbf{t} + i\mathbf{t}^T C \mathbf{w})\}. \quad (9.5.7)$$

By Hammersley's theorem (B.7.3) we can find $(\Lambda_n^*(\mathbf{t}), [S_n^*]^T)$ with the same distribution as $(\Lambda_n(\mathbf{t}), S_n^T)$ which converge in probability to $(\mathbf{t}^T \mathbf{U}^* - 1/2 \mathbf{t}^T \Sigma_0 \mathbf{t}, \mathbf{V}^*)$ with $(\mathbf{U}^*, \mathbf{V}^*)$ having the same Gaussian distribution as (\mathbf{U}, \mathbf{V}) . Then,

$$\exp\{i\mathbf{w}^T S_n^* + \Lambda_n^*(\mathbf{t})\} \xrightarrow{P} \exp\{i\mathbf{w}^T \mathbf{V}^* + \mathbf{t}^T \mathbf{U}^* - \frac{1}{2}\mathbf{t}^T \Sigma_0 \mathbf{t}\}.$$

The limit variable has, by (9.5.6), expectation $\exp\{i\mathbf{w}^T (\boldsymbol{\mu} + C^T \mathbf{t}) - \frac{1}{2}\mathbf{w}^T K \mathbf{w}\}$ which is the characteristic function of the postulated Gaussian limit distribution.

Finally, we can conclude that the limit of expectations is the expectation of the limit in (9.5.5) by Remark B.7.1. \square

Proof of Proposition 9.5.4. Let $S_n = \Sigma_0^{-1}T_n$. Then the conditions of Proposition 9.5.3 are satisfied with $\mu = 0$, $K = \Sigma_0^{-1}\Sigma_0\Sigma_0^{-1} = \Sigma_0^{-1}$, $C = J_{k \times k}$ where J is the identity. The result follows. \square

We give an application of Le Cam's Third Lemma which also serves to illustrate the asymptotic optimality properties of procedures that we associate with LAN.

Example 9.5.2. *Asymptotic efficiency of the Rao score test.* Recall the Rao test in a 1 dimensional regular model given by (5.4.55):

$$\text{“Reject } H : \theta = \theta_0 \text{ iff } S_n \equiv [nI(\theta_0)]^{-\frac{1}{2}} \sum_{i=1}^n \frac{\partial l}{\partial \theta}(X_i, \theta_0) \geq z_{1-\alpha}.” \quad (9.5.8)$$

This test was shown to be asymptotically optimal by a direct calculation in Problem 5.4.8. We reprove this result under weaker assumptions. By Example 7.1.1., for a regular one dimensional model, $T_n(\mathbf{X}^{(n)}, \theta_0) = n^{-1/2} \sum_{i=1}^n \frac{\partial l}{\partial \theta}(X_i, \theta_0)$. By the central limit theorem, under H ,

$$(T_n, S_n) \Rightarrow \mathcal{N}(0, 0, I(\theta_0), 1, 1).$$

By Le Cam's Third Lemma, if $\theta_n = \theta_0 + t/\sqrt{n}$, then

$$\mathcal{L}_{\theta_n}(T_n) \rightarrow \mathcal{N}(t\sqrt{I(\theta_0)}, 1).$$

Thus the asymptotic power function of the Rao test is $\beta(t) = 1 - \Phi(z_{1-\alpha} - t\sqrt{I(\theta_0)})$, which by Theorem 5.4.5 is optimal.

Note that the result is valid under the conditions of Example 7.1.1 which are weaker than those of Theorem 5.4.5. In particular nothing is assumed about the MLE. To see this we need only note that, by Example 7.1.1, the most powerful test of $H : \theta = \theta_0$ vs. $K : \theta = \theta_n$, “reject H iff $\Lambda_n(\theta_0, \theta_n) \geq c_n(\alpha, \theta_0)$,” is asymptotically equivalent to the Rao test both under H and under K by ULAN and contiguity (Problem 9.5.9).

We next turn to an extension of the Rao score test to models with nuisance parameters.

Example 9.5.3. Neyman's C_α test. Consider $H : \nu = \nu_0$ vs. $K : \nu > \nu_0$ for $p(\cdot, \boldsymbol{\theta}), \boldsymbol{\theta} \in R^d$ a regular model with $\boldsymbol{\theta}_0 = (\nu_0, \boldsymbol{\eta}^T)^T$ satisfying the conditions of the multivariate part of Example 7.1.1. Let

$$S_n(\nu_0, \boldsymbol{\eta}) = n^{-\frac{1}{2}} \sum_{i=1}^n l^*(X_i, \nu_0, \boldsymbol{\eta})$$

where l^* is the efficient score function defined by $l^* = \tilde{l}/\|\tilde{l}\|^2$ with $\tilde{l} = \dot{l} - \Pi(\dot{l}|\dot{\mathcal{P}}_2(P_0))$, $\dot{\mathcal{P}}_2(P_0) \equiv \{P_{(\nu_0, \boldsymbol{\eta})} : \boldsymbol{\eta} \in R^{d-1}\}$, and \tilde{l}, Π defined as in (9.3.10). Here $\Pi(\dot{l}|\dot{\mathcal{P}}_2(P_0))$ is the orthogonal projection of the ν score function \dot{l} on the tangent space generated by the nuisance parameter $\boldsymbol{\eta}$. As shown in Section 1.4, such projections can often be computed as conditional expectations. See Remark 1.4.6.

Let us assume that we can use the data to find out that we are in a neighborhood of the hypothesis, if it holds. That is, there exist $\hat{\boldsymbol{\eta}}$, such that for any $\boldsymbol{\theta}_0$,

$$\begin{aligned} n^{-\frac{1}{2}} \sum_{i=1}^n l^*(X_i, \nu_0, \boldsymbol{\eta}_0) &= n^{-\frac{1}{2}} \sum_{i=1}^n l^*(X_i, \nu_0, \hat{\boldsymbol{\eta}}) + o_{P_{\boldsymbol{\theta}_0}}(1), \\ (9.5.9) \end{aligned}$$

$$n^{-1} \sum_{i=1}^n [l^*]^2(X_i, \nu_0, \boldsymbol{\eta}_0) = n^{-1} \sum_{i=1}^n [l^*]^2(X_i, \nu_0, \hat{\boldsymbol{\eta}}) + o_{P_{\boldsymbol{\theta}_0}}(1).$$

Maximum likelihood estimates of $\boldsymbol{\eta}$ under H satisfy (9.5.9) under A0-A6 of Theorem 6.2.2. See Problem 9.5.10 for a suitable semiparametric example where (9.5.9) holds. If (9.5.9) holds, we can construct the following Neyman C_α test of H :

$$\text{Reject } H \text{ iff } S_n(\nu_0, \hat{\boldsymbol{\eta}}) \geq z_{1-\alpha} \sigma(\hat{\boldsymbol{\eta}})$$

where

$$\sigma^2(\boldsymbol{\eta}) \equiv n^{-1} \sum_{i=1}^n \left(l^*(X_i, \nu_0, \boldsymbol{\eta}) - \bar{l}^*(\nu_0, \boldsymbol{\eta}) \right)^2 \text{ and } \bar{l}^*(\nu, \boldsymbol{\eta}) = \frac{1}{n} \sum_{i=1}^n l^*(X_i, \nu, \boldsymbol{\eta}).$$

Using (9.5.9) and Slutsky's theorem we can show (Problem 9.5.12) that for all $\boldsymbol{\theta}_0$ satisfying H ,

$$S_n(\nu_0, \hat{\boldsymbol{\eta}}) \Rightarrow \mathcal{N}(0, [I^{11}(\boldsymbol{\theta}_0)]^{-1}) \quad (9.5.10)$$

where $I^{11}(\boldsymbol{\theta}_0)$ is the upper left entry of $I^{-1}(\boldsymbol{\theta}_0)$. Thus the test based on S_n is asymptotically level α for each $\boldsymbol{\theta}_0$. By contiguity it is asymptotically of level α uniformly for $\boldsymbol{\theta}$ that satisfy H and are in radius $n^{-\frac{1}{2}}$ balls of $\boldsymbol{\theta}_0$.

What about the power function of the C_α test? To apply the third lemma, set

$$\mathbf{T}_n = n^{-\frac{1}{2}} \Sigma \dot{\ell}(x_i, \boldsymbol{\theta}_0) \in R^d, \quad \mathbf{S}_n = [I''(\boldsymbol{\theta}_0)]^{\frac{1}{2}} \mathbf{S}(\nu_0, \boldsymbol{\eta}) \in R.$$

We know that

$$(\mathbf{T}_n^T, \mathbf{S}_n^T)^T \xrightarrow{\mathcal{L}} \mathcal{N} \left(\begin{pmatrix} \mathbf{0}_{d \times 1} \\ 0 \end{pmatrix}, \begin{pmatrix} I(\boldsymbol{\theta}_0) & C_{d \times 1} \\ C_{d \times 1}^T & 1 \end{pmatrix} \right),$$

where $C_{d \times 1} = ([I^{11}(\boldsymbol{\theta}_0)]^{-\frac{1}{2}}, \mathbf{0}^T)^T$ is obtained from the identities – see Problem 9.5.13.

$$\begin{aligned} E l^*(\mathbf{X}_1, \boldsymbol{\theta}_0) \frac{\partial l}{\partial \nu}(\mathbf{X}_1, \boldsymbol{\theta}_0) &= E[l^*]^2(\mathbf{X}_1, \boldsymbol{\theta}_0) = [I^{11}(\boldsymbol{\theta}_0)]^{-1} \\ E l^*(\mathbf{X}_1, \boldsymbol{\theta}_0) \frac{\partial l}{\partial \eta_j}(\mathbf{X}_1, \boldsymbol{\theta}_0) &= 0, \quad 1 \leq j \leq d-1. \end{aligned} \quad (9.5.11)$$

We conclude from Le Cam's Third Lemma that if $\boldsymbol{\theta}_n = \boldsymbol{\theta}_0 + \mathbf{t} n^{-1/2}$, $\boldsymbol{\theta}_0 \in H$, $\mathbf{t} = (t_1, \dots, t_d)^T$,

$$\mathcal{L}_{\boldsymbol{\theta}_n}([I^{11}(\boldsymbol{\theta}_0)]^{\frac{1}{2}} S_n(\boldsymbol{\theta}_0)) \rightarrow \mathcal{N}(t_1 [I^{11}(\boldsymbol{\theta}_0)]^{-\frac{1}{2}}, 1).$$

Thus the asymptotic power function of the C_α test is

$$\beta^*(\boldsymbol{\theta}) = \lim_n P_{\boldsymbol{\nu}_n} [S_n(\boldsymbol{\nu}_0, \hat{\boldsymbol{\eta}}) \geq z_{1-\alpha} \sigma(\hat{\boldsymbol{\eta}})] = 1 - \Phi(z_{1-\alpha} - t_1 [I^{11}(\boldsymbol{\theta}_0)]^{-\frac{1}{2}})$$

for $\boldsymbol{\theta}_n$ as given before. The convergence is uniform on $n^{-1/2}$ neighborhoods of points $\boldsymbol{\theta}_0$ in H . The details of this argument are given in the hint to Problem 9.5.14. The basic point is that, again, the generalization (9.5.10) of the Rao test (the *Neyman C_α test*) allowing for nuisance parameters, has, in the presence of ULAN, by Example 7.1.1 the same asymptotic optimality property as the standard test in the multivariate shift Gaussian model with known variance covariance matrix — see Example 8.3.10. See Problem 9.5.15 for an example. \square

We conclude with two final applications of Le Cam's Third Lemma pointing to its generality.

Example 9.5.4. *Testing for constant hazard rate.* Recall that the exponential distribution is characterized by constant hazard rate and plays the role of default distribution in reliability theory and survival analysis. We will, following Bickel and Doksum (1969), analyze the power of a class of tests which are expected to have power against the plausible alternative of increasing failure rate $r(x) = f(x)/[1 - f(x)]$. Proschan and Pyke (1966) proposed the use of a statistic which is a U statistic in the normalized spacings of the data, defined as follows. Let $X_{(1)} < \dots < X_{(n)}$ be the order statistics of a sample X_1, \dots, X_n i.i.d. P concentrating on R^+ with P continuous. Define the *normalized spacings* as

$$D_i = (n - i + 1)(X_{(i)} - X_{(i-1)})$$

with $X_{(0)} \equiv 0$. It may be shown that if $P = \mathcal{E}(\lambda)$ then D_1, \dots, D_n are themselves i.i.d. $\mathcal{E}(\lambda)$. (See Problem B.2.14.) On the other hand, intuitively if P has an increasing failure rate we expect the D_i to decrease in i stochastically — see Problems 1.1.14, 3.5.13, and 9.5.5. Proschan and Pyke proposed $T = \sum_{i < j} 1(D_{ni} \geq D_{nj})$ as a test statistic. Barlow and Proschan (1996), Bickel and Doksum (1969), and Bickel (1970) investigated statistics of the form

$$U = \sum_{i=1}^n c_i \left(\frac{D_i}{S} - 1 \right) \tag{9.5.12}$$

where $S = \sum_{i=1}^n D_i = \sum_{i=1}^n X_i$, and the c_i are decreasing in i . The motivation is that these statistics are scale-free measures of the covariance between the normalized spacings D_i/S and the c_i . We expect the covariance to be positive if the failure rate is increasing. Moreover, as we shall see below, Bickel (1970) shows that for suitable c_i the tests based on (9.5.12) are efficient against specific alternative families.

Because U is scale invariant its distribution under H doesn't depend on λ and we can set suitable critical values for U by Monte Carlo simulation under $\mathcal{E}(1)$. See Example 4.1.6. As we shall see, a simple Gaussian approximation is also valid.

Suppose that we have a regular parametric model $\mathcal{P} \equiv \{P_{(\lambda, \eta)} : \lambda \in R^+, \eta \in R\}$ with $P_{(\lambda, 0)} = \mathcal{E}(\lambda)$ for all λ — see Problem 9.5.6. We want to compute the power of the size α

test for $H : \eta = 0$ vs $K : \eta > 0$ based on rejecting for large values of U and compare it to the asymptotically MP test. Let

$$\dot{\mathbf{l}}(\cdot) = \left(\frac{\partial \mathbf{l}}{\partial \lambda} (\cdot, \lambda_0, 0), \frac{\partial \mathbf{l}}{\partial \eta} (\cdot, \lambda_0, 0) \right)^T.$$

To apply Le Cam's Third Lemma we need the asymptotic joint distribution of U and $T_n \equiv n^{-1/2} \sum_{i=1}^n \dot{\mathbf{l}}(X_i, \boldsymbol{\theta}_0)$ for $\boldsymbol{\theta}_0 = (\lambda_0, 0)^T$. We start by approximating T_n with a linear function of the D_i . Without loss of generality set $\lambda_0 = 1$, then

$$\begin{aligned} T &\equiv \sum_{i=1}^n \dot{\mathbf{l}}(X_i, \boldsymbol{\theta}_0) = \sum_{i=1}^n \dot{\mathbf{l}}(X_{(i)}, \boldsymbol{\theta}_0) \\ &= \sum_{i=1}^n G^{-1}\left(\frac{i}{n+1}\right) + \sum_{i=1}^n \mathbf{a}\left(\frac{i}{n+1}\right) \left(X_{(i)} - G^{-1}\left(\frac{i}{n+1}\right)\right) \\ &\quad + \frac{1}{2} \sum_{i=1}^n \mathbf{b}_n\left(\frac{i}{n+1}\right) \left(X_{(i)} - G^{-1}\left(\frac{i}{n+1}\right)\right)^2 \end{aligned} \quad (9.5.13)$$

where G is the df of X_i under $P_{\boldsymbol{\theta}_0}$, so that $G^{-1}(t) = -\log(1-t)$, and

$$\begin{aligned} \mathbf{a}\left(\frac{i}{n+1}\right) &= \frac{\partial \dot{\mathbf{l}}}{\partial x}\left(G^{-1}\left(\frac{i}{n+1}\right), \boldsymbol{\theta}_0\right) \\ \mathbf{b}_n\left(\frac{i}{n+1}\right) &= \frac{\partial^2 \dot{\mathbf{l}}}{\partial x^2}\left(\mu_n\left(\frac{i}{n+1}\right), \boldsymbol{\theta}_0\right) \end{aligned}$$

where μ_n lies between $X_{(i)}$ and $G^{-1}(i/n+1)$.

Write

$$\begin{aligned} X_{(i)} - G^{-1}\left(\frac{i}{n+1}\right) &= \sum_{j=1}^i \frac{(D_j - 1)}{n-j+1} + \left(\sum_{j=1}^i (n-j+1)^{-1} + \log(1 - \frac{i}{n+1}) \right) \\ &= \sum_{j=1}^i \frac{(D_j - 1)}{(n-j+1)} + O_P\left(\frac{1}{n}\right). \end{aligned} \quad (9.5.14)$$

The last statement follows from the well-known approximation,

$$\sum_{j=1}^n \frac{1}{j} = \log n + \gamma_0 + O\left(\frac{1}{n}\right),$$

where γ_0 is Euler's constant. Substituting (9.5.14) into (9.5.13) we get

$$\begin{aligned} n^{-\frac{1}{2}} \sum_{i=1}^n \dot{\mathbf{l}}(X_i, 1, 0) &= n^{-\frac{1}{2}} \sum_{i=1}^n \mathbf{d}\left(\frac{i}{n+1}\right) (D_i - 1) + n^{-\frac{1}{2}} \sum_{i=1}^n O\left(\frac{1}{n}\right) \\ &\quad + n^{-\frac{1}{2}} \sum_{i=1}^n \mathbf{b}_n\left(\frac{i}{n+1}\right) \left(X_{(i)} - G^{-1}\left(\frac{i}{n+1}\right)\right)^2 \end{aligned} \quad (9.5.15)$$

where

$$\mathbf{d}\left(\frac{i}{n+1}\right) = \frac{1}{n+1} \sum_{k=i+1}^n \mathbf{a}\left(\frac{k}{n+1}\right) \frac{1}{\left(1 - \frac{k}{n+1}\right)}.$$

We can show (Problem 9.5.7) that $\left(X_{(i)} - G^{-1}\left(\frac{i}{n+1}\right)\right)^2 = O_P(n^{-1})$ and thus, if \mathbf{b}_n is bounded, then

$$n^{-\frac{1}{2}} \sum_{i=1}^n \mathbf{l}(X_i, \boldsymbol{\theta}_0) = n^{-\frac{1}{2}} \sum_{i=1}^n \mathbf{d}\left(\frac{i}{n+1}\right) (D_i - 1) + O_P(n^{-\frac{1}{2}}). \quad (9.5.16)$$

Suppose $c_i \equiv c(i/(n+1))$ with the function $c(\cdot)$ continuous, $\int_0^1 c(u)du = 0$, and

$$\int_0^1 |\mathbf{d}(t)|^2 dt + \int_0^1 c^2(t)dt < \infty.$$

Then, since the D_i are i.i.d. $\mathcal{E}(1)$ under H we can apply the Lindeberg-Feller theorem (Appendix D.1) and the delta method to U . We obtain after some calculation (Problem 9.5.8) that

$$(\mathbf{T}, U) \Rightarrow \mathcal{N}_3 \left(\mathbf{0}, 0, I(\boldsymbol{\theta}_0), \int_0^1 c^2(t)dt, \int_0^1 \mathbf{d}^T(t)c(t)dt \right). \quad (9.5.17)$$

From this we may use Lemma 9.5.3 to deduce the power function of U and relate it to that of the asymptotically most powerful test of $H : P = \mathcal{E}(\lambda)$ vs. $K : P = P_{(\lambda, \eta)}$, $\eta \geq 0$ (Problem 9.5.8). Note that approximate critical values can be obtained because $U \Rightarrow \mathcal{N}(0, \int c^2(u)du)$. For more discussion and the relations of tests of type U to the Proschan-Pyke and other rank tests see Bickel and Doksum (1969) and Bickel (1970). \square

Our final example needs central limit theorems for dependent random variables which we do not pursue in this book. We include it to illustrate the generality of LAN.

Example 9.5.5. Autocorrelation. The most common model for temporal dependence of a sequence of observations is the autoregressive one — see Example 1.1.5, described by

$$X_i = \beta X_{i-1} + \varepsilon_i \quad i = 1, \dots, n$$

where the ε_i are i.i.d. with common density f . This only specifies the conditional distribution of X_2, X_3, \dots given X_1 . If f is $\mathcal{N}(0, \sigma^2)$ and $|\beta| < 1$ we can, by specifying

$$X_1 \sim \mathcal{N}(0, \sigma^2(1 - \beta^2)^{-1}),$$

make the sequence X_1, \dots, X_n a stationary, homogeneous, Gaussian Markov Chain. That is, (X_1, \dots, X_k) and $(X_{1+m}, \dots, X_{k+m})$ are identically distributed and the conditional distribution of X_i given X_1, \dots, X_{i-1} depends only on X_{i-1} and is the same for all i . In fact, if we add to the model specification on X_1 a non-zero mean as in Example 1.1.5 we

obtain the only examples of stationary, homogeneous Gaussian Markov Chains. We restrict ourselves to this simple case for most of this example.

Suppose we are interested in testing $H : \beta = 0$ vs. $K : \beta \neq 0$; that is, independence vs. autoregression. It is easy to see, if $p(x_1, \dots, x_n, \beta)$ is as given in Example 1.1.5, for $\mu = 0$ and σ^2 known, that

$$\frac{\partial}{\partial \beta} \log p(X_1, \dots, X_n, \beta) |_{\beta=0} = \frac{1}{\sigma^2} \sum_{i=2}^n X_{i-1} X_i = \sum_{i=1}^n \varepsilon_{i-1} \varepsilon_i. \quad (9.5.18)$$

Thus, (9.5.18) is the Rao score statistic. By letting $X_1 \sim \mathcal{N}(0, \sigma^2/(1 - \beta^2))$ we make the process stationary. We examine the asymptotic distribution of

$$\Lambda(tn^{-\frac{1}{2}}) \equiv \log \frac{p(X_1, \dots, X_n, tn^{-\frac{1}{2}})}{p(X_1, \dots, X_n, 0)}. \quad (9.5.19)$$

The usual Taylor expansion gives

$$\Lambda(tn^{-\frac{1}{2}}) = tn^{-\frac{1}{2}} \sum_{i=2}^n \varepsilon_{i-1} \varepsilon_i + \frac{t^2}{2n} \sum_{i=2}^n \varepsilon_{i-1}^2 + o_P(1). \quad (9.5.20)$$

We show the model is ULAN. To establish LAN we need to show that

$$n^{-\frac{1}{2}} \sum_{i=2}^n \varepsilon_{i-1} \varepsilon_i \xrightarrow{\mathcal{L}} \mathcal{N}(0, 1) \quad (9.5.21)$$

since we know, by the WLLN, that

$$\frac{1}{n} \sum_{i=2}^{n-1} \varepsilon_i^2 \xrightarrow{P} E\varepsilon_1^2 = 1.$$

It is clear that $n^{1/2} \sum_{i=2}^n E\varepsilon_{i-1} \varepsilon_i = 0$ and

$$\text{Var}(n^{-1/2} \sum_{i=2}^n \varepsilon_{i-1} \varepsilon_i) = \text{Var}(\varepsilon_1 \varepsilon_2) + 2 \sum_{i \neq j}^n \text{Cov}(\varepsilon_{i-1} \varepsilon_i, \varepsilon_{j-1} \varepsilon_j) = \text{Var}(\varepsilon_1 \varepsilon_2) = 1,$$

since $\text{Cov}(\varepsilon_{i-1} \varepsilon_i, \varepsilon_{j-1} \varepsilon_j) = E\varepsilon_{i-1} \varepsilon_i \varepsilon_{j-1} \varepsilon_j = 0$ by the independence of the ε_i .

That (9.5.20) holds under the hypothesis H of independence is a consequence of any one of the three following approaches:

- (i) Martingale central limit theorems (Hall and Heyde (1980))
- (ii) Central limit theorems for functions of Markov chains (Prakasa Rao (1983))
- (iii) Strongly mixing random variables (Ibragimov and Linnik (1971))

All of these essentially imply asymptotic normality of $\Lambda(tn^{-1/2})$ with natural parameters (mean and variance). Unfortunately, none of the tools we have developed in the independence case can be applied here directly.

To perform tests, we can use statistics which are more robust, for instance,

$$U^* \equiv n^{-\frac{1}{2}} \sum_{i=2}^n \psi(X_{i-1})\psi(X_i) \quad (9.5.22)$$

with a function ψ which downweights large X_i . Thus, under the assumption that ε has a double exponential distribution we arrive at (9.5.22) for the Rao statistic with $\psi(t) \equiv \text{sgn}(t)$ as the Rao score statistic. Le Cam's Third Lemma and a multivariate central limit theorem for sums of dependent random variables yield the asymptotic distribution of U^* as $\mathcal{N}(0, \pi^2/4)$. Using (i), (ii), or (iii) again yields

$$\begin{aligned} \mathcal{L}(T, U^*) &\xrightarrow{\mathcal{L}} \mathcal{N}(0, 0, 1, 1, E(\text{sgn}(\varepsilon_i\varepsilon_{i-1})\varepsilon_i\varepsilon_{i-1})/E\varepsilon_1^2) \\ &= \mathcal{N}(0, 0, 1, 1, (E|\varepsilon_1|)^2/E\varepsilon_1^2). \end{aligned}$$

Thus, as in the case of the ordinary one sample problem, if $\varepsilon \sim \mathcal{N}(0, 1)$, the asymptotic distribution of U^* is $\mathcal{N}(0, \pi^2/4)$.

The comparison in power of T and this U^* leads to the same conclusions as the asymptotic comparison of the sign test to the t test or of the median to the mean in the one sample problem. For more on inference under dependence we refer to Prakasa Rao (1983) and Ibragimov and Hasminskii (1981). \square

Summary. This section treats local approximations to models that can be used to find approximations to the properties of statistical decision procedures. For the regular model $\{P_{\theta} : \theta \in \Theta\}$, the local model is $\{P_{\theta_n} : \theta_n = \theta_0 + t/\sqrt{n}; |t| \leq M\}$ for some constant $M > 0$. More generally we consider sequences of models, called *experiments*, of the form $\mathcal{E}_n = \{P_{\theta}^{(n)} : \theta \in \Theta_n\}$ and define \mathcal{E}_n to be *uniformly locally asymptotically normal* (ULAN) at θ_0 if the log likelihood ratio for testing θ_0 vs $\theta_0 + t\gamma_n$, where $|\gamma_n| \downarrow$ and $|t| \leq M$, can be closely approximated by an asymptotically normal sequence of statistics. We relate the ULAN concept to the concept of contiguity: The sequence of probabilities $\{Q^{(n)}\}$ is *contiguous* to the sequence of probabilities $\{P^{(n)}\}$ iff for any sequence of events A_n such that $P^{(n)}(A_n) \rightarrow 0, Q^{(n)}(A_n) \rightarrow 0$ also. This concept is useful because if we can show that an approximation S_n to a vector T_n satisfies $T_n - S_n \xrightarrow{P} 0$ under a hypothesis H , then $T_n - S_n \xrightarrow{P} 0$ also for contiguous alternatives. In the case where $P^{(n)} = P_{\theta_0}^{(n)}$ is ULAN and $Q^{(n)} = P_{\theta_n}^{(n)}$ with $\theta_n = \theta_0 + t\gamma_n$, $|t| \leq M$, $Q^{(n)}$ can be shown to be contiguous to $P^{(n)}$. We establish Le Cam's Third Lemma which states that if T_n denotes the ULAN approximation to the local log likelihood ratio and if the joint distribution of T_n and some statistic S_n tends to the multivariate normal distribution for $P^{(n)}$, then the asymptotic distribution of S_n is asymptotically multivariate normal for $Q^{(n)}$. We illustrate the usefulness of these ideas by establishing the asymptotic efficiency of the Rao score and Neyman C_α tests and by developing asymptotic theory for a class of unbiased tests of exponentiality vs increasing failure rate based on normalized spacings.

9.6 PROBLEMS AND COMPLEMENTS

Problems for Section 9.1

1. In Example 9.1.1,

- (a) Show that β is identifiable in model (a).
- (b) Show that α is unidentifiable in model (a) and σ is unidentifiable in both (a) and (b).
- (c) Show that both α and β are identifiable in model (b).
- (d) Show that $\hat{\beta}$ as defined in Example 6.2.2 with f_0 the logistic density is \sqrt{n} consistent for β in model (a).
- (e) Give the influence functions of the estimates $\hat{\beta}$ and $\hat{\alpha} = \bar{Y} - \bar{Z}\hat{\beta}$ in model (b).

2. In Example 9.1.2,

- (a) Establish (9.1.4).
- (b) Prove that F is identifiable in both (9.1.2) and (9.1.5). What conditions are needed?

3. Parametric biased sampling. Simpler models result in more efficient inference when the sample size is small; see Section I.7 (although if seriously wrong the biases they bring can be overwhelming). Thus comparisons of procedures derived from parametric and semi-parametric models are of interest. In the length bias part of Example 9.1.2 let V be the number of strata with v elements and suppose that V has the Poisson (μ) density

$$h(v) = h_\mu(x) \equiv \frac{e^{-\mu} \mu^v}{v!}, \quad v = 0, 1, \dots.$$

An element is drawn at random from the union of all strata S_0, S_1 , where S_j has j elements and the size X of the strata is observed, where now $X \geq 1$. Let Z be a variable with $\mathcal{L}(Z) = \mathcal{L}(V|V \geq 1)$.

- (a) Show that Z has density

$$f(z) = f_\mu(z) \equiv \frac{e^{-\mu} \mu^z}{(1 - e^{-\mu})z!}, \quad z = 1, 2, \dots.$$

- (b) Show that $W(F) = E(Z) = \mu/[1 - e^{-\mu}]$ and thus

$$p_F(x) = \frac{e^{-\mu} \mu^{x-1}}{(x-1)!}, \quad x = 1, 2, \dots.$$

That is $\mathcal{L}(X) = \mathcal{L}(V + 1)$.

- (c) Recall that V and Z are not observed; instead X is. Show that the MLE of μ based on a sample X_1, \dots, X_n from $p_F(x)$ is $\hat{\mu} = \sum_{i=1}^n (X_i - 1)/n$ and thus, by the equivariance property of MLEs, the MLEs of $h(v)$ and $f(g)$ are $\hat{h}(v) = h_{\hat{\mu}}(v)$ and $\hat{f}(g) = f_{\hat{\mu}}(z)$.

Remark 1. The Poisson model includes households with no children and $\mu = E(V)$ is the expected number of children over all households.

- 4. Parametric stratified sampling.** In the stratified sampling part of Example 9.1.2, suppose $(I_1, Y_1), \dots, (I_n, Y_n)$ is a sample from $p_F(x)(j, y)$. Based on this sample, find the MLE of μ when Z is Poisson (μ), $k = 2$, $\mathcal{X}_1 = \{0, 1, 2, \dots\}$, $\mathcal{X}_2 = \{1, 2, 3, \dots\}$, and (a) $\lambda_1 = \lambda_2 = \frac{1}{2}$, (b) λ_1 and λ_2 general but known; where $\lambda_j = P(I = j)$, $j = 1, 2$.

- 5. Size biased class size.** Suppose we are interested in the mean size of undergraduate classes at a certain university. We sample n students and ask which of the following categories their current classes fall in: $[1, 20), [20, 40), [40, 60), \dots, [800, \infty)$. These categories are converted to the class sizes $k = \{10, 30, 50, \dots, 900\}$, and we define the mean from the perspective of the students as $\mu_S = \sum_{k=K} \lambda_k p_k$, where p_k is the proportion of students in classes of size k . Suppose we are interested in the mean class size μ_P from the perspective of the professors at the university. Explain why μ_P is smaller than μ_S and describe how to estimate μ_P .

- 6. (a)** In Example 9.1.10, show that $\beta \rightarrow \mathcal{L}_P(r(\mathbf{Z}, \beta))$ is 1–1 and $P[S_0(T, \mathbf{0}, P) = c] < 1$ for all c are necessary for identifiability of β in the Cox model.

- (b)** Show that under these conditions, if $\beta(P)$ is defined by (9.1.42), $\beta(P_{\beta_0}) = \beta_0$.

- 7.** In Example 9.1.8, show that \widehat{W} exists and is unique with probability tending to one provided $0 < P(\delta = 0) < 1$.

Hint. Use exponential family theory.

- 8.** In Example 9.1.9, prove (9.1.32) to (9.1.34) following the approach for censoring.

- 9.** For a semiparametric example where maximum likelihood fails, consider the model where $Y \in \{0, 1\}$, $\mathbf{Z} \in B \subset R^d$, B bounded, $U \in [0, 1]$, and

$$E(Y|Z, U) = P(Y = 1|Z, U) = L(\beta^T \mathbf{Z} + \eta(u)), \quad \beta \in R^d,$$

with $L(t) = 1/[1 + e^{-t}]$. This is a *partially linear logistic regression* model. Assume that $\eta(\cdot)$ is in the class of functions \mathcal{F}_k with $J^{(k)}(\eta) < \infty$, where $J^{(k)}(\eta) = \int [\eta^{(k)}(t)]^2 dt$, $k \geq 1$. Let $L(\beta, \eta)$ denote the likelihood based on (Y_i, Z_i, U_i) i.i.d. as (Y, Z, U) . Write $p(y, z, u) = p(y|z, u)p(u, z)$ and assume that $p(u, z)$ does not involve β . Show that

$$\sup\{L(\beta, \eta) : \eta \in \mathcal{F}_k\} = 1.$$

Thus any $\mathbf{b} \in B$ is a MLE of β . In this model efficient estimates can be obtained by maximizing the penalized likelihood $L(\beta, \eta) - \lambda^2 [J^{(k)}(\eta)]^2$ where $[J^{(k)}(\eta)]^2$ is a roughness penalty and λ^2 is a tuning parameter. See Mammen and van der Geer (1997).

- 10.** In Example 9.1.14,

- (a) Show that without loss of generality we can take J to be the $d \times d$ identity matrix.
 (b) Fill in the details in the proof that if

$$\text{Var } Z_k = 1, EZ_k = 0, 1 \leq k \leq d$$

and at most one of the Z_k is Gaussian, then A is identifiable. (Assume that if $EZ_k^2 < \infty$, $\gamma_k(\mathbf{t})$ can be defined and is twice differentiable in a neighbourhood of 0.)

Hint. If $\gamma_k'' \equiv c$, then it takes on an infinite number of distinct values.

11. In Example 9.1.4, the estimate $\hat{r}_{j,t}$ depends on selecting $t \equiv \hat{t}$ so that $0 < \hat{S}_0(\hat{t}) < 1$. Thus $t = \hat{t}$ is random. Find conditions under which $\hat{r}_{j,\hat{t}} \xrightarrow{P} r_j(\boldsymbol{\beta})$.

12. Show that in Example 9.1.13, if $E_P|Y| < \infty$, $0 < E_P(Z - E(Z|U))^2 < \infty$, then (9.1.52) is valid and β is identifiable for the given parametrization.

Hint. $Y - E_P(Y|U) = \beta(Z - E(Z|U)) + \varepsilon$.

13. Show in Example 9.1.9 that

- (a) If (9.1.25) is replaced by its continuous approximation (9.1.29), we arrive at the same estimating equation for λ_{F_j} but the expression (9.1.30) for S .

- (b) The unique maximizer of $K(P_{(F,G)}, P)$ is given by (9.1.31). Thus the parametrization sending (F, G) into P according to (9.1.24) is identifiable if $0 < P[\delta = 0] < 1$ by using (9.1.31).

Hint for (b). Use Shannon's inequality.

14. In the censored data framework of Example 9.1.9, show that the empirical likelihood estimate of the hazard rate λ_{F_i} at $y_{(i)}$ based on maximizing on (9.1.26) or (9.1.29) is $\hat{\lambda}_{F_i} = \delta_{(i)} / (n + 1 - i)$ when there are no ties.

15. Show that in the proportional hazard model, L_d of Example 9.1.12 is proportional to the approximate empirical likelihood (9.1.36).

16. In Example 9.1.12, show that the likelihoods L_d and L_a are different for the model with

$$\Lambda(y|\mathbf{z}) = a\theta\Lambda(y) + (1-a)\Lambda(\theta y),$$

where $\theta = \exp\{\beta^T \mathbf{z}\}$, $a \in (0, 1)$ is fixed, and $\Lambda(y) = -\log(1 - F(y))$ for some baseline df F on $[0, \infty)$.

17. The *odds ratio* is defined as $\Gamma(t|\mathbf{z}) = F(t|\mathbf{z})/[1 - F(t|\mathbf{z})]$ and the *proportional odds model* is defined by

$$\Gamma(t|\mathbf{z}) = \theta\Gamma(t), \quad \theta > 0, \quad t > 0,$$

where $\theta = r(\mathbf{z}, \boldsymbol{\beta})$ and $\Gamma(t) = F(t)/[1 - F(t)]$ for some baseline distribution F on $[0, \infty)$.

- (a) Show that if $G(u; \theta) = u/[u + \theta(1 - u)]$, $0 \leq u \leq 1$, then $G(F(y); \theta)$ follow a proportional odds model.

- (b) Show that if $G(y; \theta) = L(y - \theta)$ where $L(t) = [1 + e^{-t}]^{-1}$ is the logistic df and if $F(y|\mathbf{z}) = G(\eta(y); \theta)$, $\eta(y) = L^{-1}(F(y))$, then $F(y|\mathbf{z})$ follow a proportional odds model.
- (c) (i) Write an expression for the likelihood L_d of Example 9.1.12 for this model.
(ii) When $\theta = \exp\{\beta z\}$, $z \in R$, show that this likelihood is concave in η_1, \dots, η_n , where we write η_i for $\eta_{(i)}$.

Remark 2. Bell and Doksum (1966, Table 8.1, last row) considered optimal tests of $H : \theta = 1$ in the proportional odds model (a). This paper gives several more examples of models of the form $G(\eta(y); \theta)$. See Problems 8.3.14 and 9.1.25.

18. We noted that L_d and L_a give the same estimates of θ and β in model (9.1.46). However, they give different estimates of $\eta'(t)$. For instance, in the Cox proportional hazard model, the L_a estimate of $\lambda = \Lambda'$ is given in (9.1.37). It is increasing in $y \in \{y_{(1)}, \dots, y_{(n)}\}$ and converges to zero and thus is not consistent. On the other hand, the *Breslow estimate*

$$\hat{\Lambda}_B(y) = \sum_{k=1}^n \hat{\lambda}(y_{(k)}) \mathbf{1}(y_{(k)} \leq y),$$

which is the counting measure integral of $\hat{\lambda}(\cdot)$, is known to be consistent.

- (a) Show that the L_d estimate is

$$\hat{\lambda}_d(y_{(i)}) = \frac{1}{(y_{(i)} - (y_{(i-1)}) \sum_{j \geq i} \exp\{\hat{\beta}^T \mathbf{z}_{(j)}\}} , \quad y_{(0)} = 0, 1 \leq i \leq n.$$

If we define $\hat{\lambda}_d(y)$ to be $\hat{\lambda}_d(y_{(i)})$ on $[y_{(i)}, y_{(i-1)})$, $1 \leq i \leq n$, and zero elsewhere, then

$$\int_0^y \hat{\lambda}_d(s) ds = \hat{\Lambda}_B(y)$$

is again the Breslow estimate. Thus $\hat{\lambda}_d(\cdot)$ is the ordinary right derivative of the Breslow estimate.

- (b) Show, using an example, that $\hat{\lambda}_d(\cdot)$ is not a consistent estimate of $\lambda(\cdot)$. For consistent estimates of $\lambda(\cdot)$, see Chapter 11.

Hint. If $Y_{(i)}$ is an exponential, $\mathcal{E}(\lambda)$, order statistic, then $(n+1-i)(Y_{(i)} - Y_{(i-1)})$ has an $\mathcal{E}(\lambda)$ distribution (Problem B.2.14).

19. In the accelerated failure time model, show that the log likelihood is proportional to (9.1.60).

20. For the AFT model (9.1.58) the approximate log likelihood l_a is obtained by replacing $\Lambda(\cdot)$ by a step function with the jump size λ_i at each $v = \theta_i$, $1 \leq i \leq n$. Show that

(a) $l_a(\beta, \lambda) = n^{-1} \sum_{i=1}^n \{\delta_i \beta^T \mathbf{Z}_i(Y_i) + \delta_i \log \lambda_i - \sum_{j=1}^n \lambda_j \mathbf{1}[\theta_j \leq \theta_i]\}.$

(b) For fixed β , the maximizer is

$$\hat{\lambda}_k(\beta) = \frac{\delta_k}{\sum_{j=1}^n 1(\theta_j \geq \theta_k)}, \quad 1 \leq k \leq n.$$

(c) The approximate profile log likelihood is proportional to

$$l_a(\beta) = n^{-1} \sum_{i=1}^n \left\{ \delta_i \beta^T \mathbf{Z}_i(Y_i) - \delta_i \log \left[\sum_{j=1}^n 1(\theta_j \geq \theta_i) \right] \right\}.$$

Note that this objective function has an infinite maximum and does not achieve its maximum for finite β . Thus the approximate likelihood L_a without the $\sum p_i = 1$ condition does not work for this model.

21. *Models where the proportion hazard (PH) and accelerated failure time (AFT) models are equivalent.* Let $T_i \geq 0$ be a failure time whose continuous df F_i depends on a vector \mathbf{z}_i of fixed covariates. Let $\lambda_i(t) = f_i(t)/[1 - F_i(t)]$ where $f_i(t)$ is the continuous case density of T_i . Consider the following models for independent failure times T_1, \dots, T_n :

- (i) PH: $\lambda_i(t) = \Delta_i \lambda(t)$, some unknown baseline $\lambda(t)$; $\Delta_i = r(\mathbf{z}_i, \beta) > 0$, some known $r(\cdot, \cdot)$, unknown β .
- (ii) AFT: $F_i(t) = F(\tau_i t)$, some unknown baseline F ; $\tau_i = s(\mathbf{z}_i, \alpha) > 0$, some known $s(\cdot, \cdot)$, unknown α .
- (iii) Periodic hazard rate (PHR): $\lambda_i(t) = h_i(\log t)t^{\gamma_i-1}$, some known non-negative period function $h_i(\cdot)$ with period $c_i > 0$; $\gamma_i = q(\mathbf{z}_i, \theta) > -1$, some known $q(\cdot, \cdot)$, unknown θ .
- (iv) Weibull: $\lambda_i(t) = \gamma_i t^{\gamma_i-1}$, $\gamma_i = q(\beta_i, \theta) > -1$, some known $q(\cdot, \cdot)$, unknown θ . In this case, F_i is said to be the *Weibull distribution*.
- (v) Trigonometric PHR: $\lambda_i(t) = \exp\{\sin 2\pi \log t / \log \tau_i\}t^{\gamma_i-1}$, $\tau_i = s(\mathbf{z}_i, \alpha) > 0$ and $\gamma_i = q(\mathbf{z}_i, \theta_i) > -1$, where $s(\cdot, \cdot)$ and $q(\cdot, \cdot)$ are known, α and θ are unknown.

(a) Show that (i) and (ii) both hold iff

$$(vi) \quad \lambda(\tau_i t) = \Delta_i \lambda(t) / \tau_i, \text{ all } t > 0, \text{ some } \tau_i \neq 1, \tau_i > 0, \text{ some } \Delta_i > 0.$$

(b) Suppose $\lim_{t \downarrow 0} [\lambda(t)/t^a]$ exists and is positive for some $a > -1$. Show that (i) and (ii) hold iff F_i is a Weibull distribution.

(c) Show that if (iii) holds with $c_i = |\log \tau_i|$ and $\gamma_i = \log \Delta_i / \log \tau_i$, then (vi) holds so that the PHR model is both PH and AFT.

(d) Show that (iv) and (v) are examples of (iii).

Remark. Doksum and Nabeya (1984) show that a model is both PH and AFT iff it is PHR.

22. The quantile hazard rate (QHR) of a random variable $T \geq 0$ with hazard rate $\lambda(\cdot)$ is defined by $q(\alpha) = \lambda(t_\alpha)$, $0 < \alpha < 1$, where $t_\alpha = F^{-1}(\alpha)$ is the α th quantile of the distribution F of T . Let $q(\alpha|\mathbf{z})$ denote the quantile hazard rate for $(T|\mathbf{z})$. The proportional quantile hazard (PQH) model is defined by

$$q(\alpha|\mathbf{z}) = \theta q(\alpha),$$

where $q(\cdot)$ is a baseline QHR and $\theta = r(\mathbf{z}, \beta)$ for some known function $r(\cdot, \cdot)$. Show that $(T|\mathbf{z})$ follow the PQH model iff $(T|\mathbf{z})$ follow the AFT model where $(T|\mathbf{z}) \stackrel{\mathcal{L}}{=} \theta^{-1} T_0$ for some baseline variable T_0 .

23. The Hoeffding Rank Likelihood.

- (a) Consider Example 9.2.12 where $F_i(y) = G(\eta(y); \theta_i)$ for some parametric df $G(v; \theta)$ and increasing differentiable function $\eta(\cdot)$, where $\theta_i = r(\mathbf{z}_i, \beta)$ with \mathbf{z}_i a nonrandom covariate vector. Show that the Hoeffding rank likelihood of Problem 8.3.10(a) reduces to

$$L_R(\beta) = E_G \left(\prod_{i=1}^n [g(V^{(r_i)}; \theta_i)/g(V^{(r_i)})] \right) / n!,$$

where $g(v; \theta)$ is the continuous case density of $G(v; \theta)$ and $g(v) = g(v; \theta_0)$ with θ_0 corresponding to a null hypothesis (typically $\beta = 0$). Assume that $g(v) > 0$ whenever $\prod_{i=1}^n g(v; \theta_i) > 0$.

- (b) Show that if $g(y; \theta) = 1 - \exp\{-\theta y\}$ and $\eta(y) = -\log[1 - F_0(y)]$, $y > 0$, for some continuous df F_0 , then $L_R(\beta)$ is equivalent to the Cox likelihood (9.1.44). See Kalbfleish and Prentice (1973, 2002) who called $L_R(\beta)$ a marginal likelihood.

Hint for (b): Note that if $h(\cdot)$ is decreasing, $\text{Rank}(h(Y_i)) = n+1-R_i$. For the proportional hazard model, transform Y_i by $U_i = 1 - F_0(Y_i)$, then $f_{U_i}(u) = \theta_i u^{\theta_i-1}$, $0 < u < 1$. Hoeffding's formula with $f(u) = 1$ ($0 < u < 1$) shows

$$L_R(\beta) \propto \prod_{i=1}^n \theta_i \int_{0 < u_1 < u_2 < \dots < u_n < 1} \prod_{i=1}^n u_i^{\delta_i-1} du_1 \dots du_n,$$

where $\delta_i = \theta_{b_i}$ and b_i = index on the Y with rank $n + 1 - i$. Use this to show that

$$L_R(\beta) \propto \prod_{i=1}^n \frac{\theta_i}{\sum_{k: Y_k \geq Y_{(i)}} \theta_k}.$$

Remark 3. Asymptotic properties of maximum rank likelihood estimates were obtained by Bickel and Ritov (1997) who show that in a general class of transformation models, estimates based on L_R are asymptotically efficient in the sense of Section 9.3.

24. (a) Show that n^{-1} times the log likelihood for the Cox model in the uncensored case is

$$l(\boldsymbol{\xi}) = n^{-1} \sum_{i=1}^n y_i \log p_i, \quad \boldsymbol{\xi} = (\boldsymbol{\beta}, \lambda(\cdot))$$

where

$$p_i = \theta_i \lambda(t_i) \exp\{-\theta_i \Lambda(t_i)\}, \quad \theta_i = \exp(\boldsymbol{\beta}^T \mathbf{z}_i).$$

(b) Set $\lambda_i = \Lambda(t_i) - \Lambda(t_{i-1})$, $\Lambda(t_0) = 0$, then $\Lambda(t_i) = \sum_{j \leq i} \lambda_j$. Let l_d be l with $\lambda(t_i)$ replaced by λ_i/d_i , where $d_i = t_i - t_{i-1}$. Let F and $f = F'$ be the df and density of T . Assume that $\log \lambda(\cdot)$ is uniformly continuous and that $\lambda(\cdot)$ is nondecreasing. Show that

$$l(\boldsymbol{\xi}) - l_d(\boldsymbol{\xi}) = O_P(n^{-1} \log \lambda(T_n))$$

where T_n is the largest F order statistic, $T \sim F$.

Hint. By the mean value theorem, $\lambda_i = d_i \lambda(T_i^*)$ where $T_i \leq T_i^* \leq T_{i-1}$. Here T_{i-1} and T_i are order statistics. Next show that

$$l(\boldsymbol{\xi}) - l_d(\boldsymbol{\xi}) \leq n^{-1} \sum [\log \lambda(T_i) - \log \lambda(T_{i-1})] = n^{-1} \log \lambda(T_n).$$

Remark. It is known that $\lambda(T_n) \cong T_n f(T_n)$ in the sense that the ratio tends to one with probability one. See de Haan and Ferreira (2006), Theorem 5.4.1. Thus if $t f(t)$ is bounded, then $l(\boldsymbol{\xi}) - l_d(\boldsymbol{\xi}) = O(n^{-1})$ with probability one.

(c) Let l_a be l with $\lambda(t_i)$ replaced by λ_i , that is, let $\Lambda(t)$ be a step function with jumps of size λ_i . Show that in the exponential case where $\lambda(t) = 1/\mu$ is constant, $E[l(\boldsymbol{\xi}) - l_a(\boldsymbol{\xi})]$ does not tend to zero as $n \rightarrow \infty$.

Hint. See Problem B.2.14 and recall that if $T \sim E \times p(\mu)$ then $E(T_{(i)}) = \sum_{j=n+1-i}^n (1/j)$.

Remark. In this model l_d and l_a are equivalent functions of $\boldsymbol{\xi}$. However, l_d is more convenient for showing closeness to the semiparametric likelihood l .

25. Two sample plug-in semiparametrics. Suppose X_1, \dots, X_n are i.i.d. as $X \sim F$, Y_1, \dots, Y_n are i.i.d. as $Y \sim H$; the X 's and Y 's are independent. Consider the model where $H(y) = G(F(y); \theta)$ with F unknown for some parametric model $G(u; \theta)$, $0 \leq u \leq 1$. Assume temporarily that F is known; then if we set $U = F(Y)$, then V has distribution $G(u; \theta)$ and we can find the MLE $\hat{\theta}(F)$ of θ based on U_1, \dots, U_n , $U_i = F(Y_i)$. If F is unknown, set $\hat{\theta} = \hat{\theta}(\hat{F})$, where \hat{F} is the empirical of the X 's. Because $\sup_x |\hat{F}(x) - F(x)| = O_P(n^{-\frac{1}{2}})$, $\hat{\theta}$ is consistent when $\hat{\theta}(F)$ is a consistent estimate of $\theta(F)$ and $\theta(F)$ is a continuous function of F in the sup norm. For the following models, find

(a) The influence function of $\hat{\theta}(F)$ assuming F is known.

(b) The influence function of $\hat{\theta} = \hat{\theta}(\hat{F})$.

Models:

(i) Proportional hazards. $G(u; \theta) = 1 - [1 - u]^\theta$, $\theta > 0$.

- (ii) Lehmann. $G(u; \theta) = u^\theta, \theta > 0.$
- (iii) Proportional odds. $G(u; \theta) = u/[u + \theta(1 - u)], \theta > 0.$
- (iv) Normal transformation model. $G(u; \theta) = \Phi(\Phi^{-1}(u) - \theta), \theta \in R.$
- (v) SINAMI. $G(u; \theta) = (e^{\theta u} - 1)(e^\theta - 1)^{-1}, \theta \neq 0, = u \text{ when } \theta = 0.$
- (vi) Self-contamination. $G(u; \theta) = (1 - \theta)u + \theta u^c, 0 \leq \theta \leq 1, c > 1 \text{ is a constant.}$

Remark 4. These models were considered in Problem 8.3.14.

Problems for Section 9.2

1. In Example 9.2.1, show that $\mathcal{E}_n(\cdot) \Rightarrow Z(\cdot).$

Hint. Apply the delta method and show that the remainder is bounded uniformly.

2. Derive (9.2.5) using Theorem 6.2.1 and Theorem 7.1.4.

3. Derive (9.2.7).

4. Establish that Rem 1 = $o_P(n^{-\frac{1}{2}})$ in the proof of Theorem 9.2.1, using the equicontinuity of the empirical process and the consistency of the empirical quantile function.

5. Establish (9.2.11) by computing the covariance of the limiting process in part (ii) of Theorem 9.2.1.

Hint. Use integration by parts.

6. Verify in detail that the conditions of Theorem 9.1.3 imply statements (a) and (b) of the proof of Theorem 9.2.2 showing how Theorem 1.6.3 applies.

7. Verify the details of the proof that c) of Proposition 9.1.1 applies in the proof of (iii) of Theorem 9.2.2.

Hint. Use the fact that $n^{\frac{1}{2}}(S_0(\cdot, \beta, \widehat{P}) - S_0(\cdot, \beta, P))$ can be thought of as an empirical process and then apply the delta method to $\dot{\Lambda}_\alpha$.

8. Derive a result similar to Theorem 9.2.1 for the truncation estimates (9.1.34).

9. Give the details of the derivation of (9.2.17).

10. Show that (9.2.17) and (9.2.20)–(9.2.22) imply (9.2.19).

11. Verify (9.2.20)–(9.2.22) assuming the validity of (9.2.24).

12. Consider the Cox model with \mathbf{z} fixed and T having hazard rate $\lambda(t) \exp\{\beta^T \mathbf{z}\}$. Let

$$A = \{t : \log S_0(t, \beta, P) \text{ is strictly convex in } \beta\}.$$

Using the proof of Theorem 1.6.3, show that $\log S_0(t, \beta, P)$ is always strictly convex and analytic and $P[T \in A] > 0$ unless

$$P[T = c(\mathbf{z})] = 1,$$

which is ruled out by our assumptions.

13. Suppose $Z \sim \text{Uniform}(0, 1)$ and that $(T|z)$ has the exponential model constant failure rate $\lambda \exp\{\beta z\}$, $\lambda > 0, \beta \in R$.

- (a) Find the asymptotic distribution of the MLE $\hat{\beta}_{MLE}$ of β in this model.
- (b) Find the asymptotic distribution of the Cox estimate $\hat{\beta}_{COX}$ using Theorem 9.2.2.
- (c) Give the asymptotic relative efficiency (ARE) of $\hat{\beta}_{COX}$ with respect to $\hat{\beta}_{MLE}$ in the exponential model (see Problem 5.4.1(f) for the definition of ARE). Give the value of the ARE when $\tau = \infty$.

(You may use MATLAB or any other software in this problem.)

- 14.** Suppose that $(\mathbf{Z}, T) \sim Cox(\beta_0, \lambda)$ and that condition C holds. Show that, given $\mathbf{Z} = \mathbf{z}$, $V = \Lambda(T) \exp(\beta_0^T \mathbf{z})$ has a standard exponential distribution.

Hint. Set $\theta = \exp(\beta_0^T \mathbf{z})$. Note that

$$P(\theta \Lambda(T) \leq s | \mathbf{z}) = P(T \leq \Lambda^{-1}(s/\theta) | \mathbf{z}).$$

- 15.** Suppose $(Z_1, T_1), \dots, (Z_n, T_n)$ are i.i.d. as (Z, T) , where $T > 0$ is a survival time and $Z \in \mathcal{R}$, with $0 < E[Z^2] < \infty$ and $P(Z = 0) = 0$, is a predictor random variable. Let $H(z, t; \beta)$, $\beta \in \mathcal{R}$, denote the distribution of (Z, T) and assume that the marginal distribution of Z does not involve β . Consider the estimating equation

$$\frac{1}{n} \sum_{i=1}^n \psi(z_i, t_i; \beta) = 0,$$

where $\psi(z, t; \beta) = \frac{\dot{r}(z; \beta)}{r(z; \beta)} - \dot{r}(z; \beta)t$, and $r(\cdot; \beta) > 0$ is a 1-1 known function of β with derivative $\dot{r}(z; \beta) = \partial r(z; \beta)/\partial \beta$.

- (a) Suppose (Z, T) has distribution H_1 where H_1 is such that T given Z has the Weibull distribution with df

$$1 - \exp\{-r(z, \beta_0)t^\alpha\}, \quad \alpha > 0.$$

For what values of α does $E_{H_1}[\psi(Z, T; b)] = 0$ have the solution $b = \beta_0$?

- (b) Let $r(z; \beta) = \exp(-\beta z)$ and let α be the answer(s) to part (a). Is the solution $b = \beta_0$ in part (a) unique?

Hint. You may use the fact that if W has the Weibull distribution with df $1 - \exp\{-r(w^\alpha/\theta)\}$, then $E[W] = \theta^{1/\alpha} \Gamma(\alpha^{-1} + 1)$.

- (c) Let $D(b) = \frac{1}{n} \sum_{i=1}^n \psi(Z_i, T_i; b)$. Suppose $r(z, \beta) = \exp(-\beta z)$. Show that $D(b) = 0$ has a unique solution.

- (d) Let β_0 be the solution to $E_H[\psi(Z, T; b)] = 0$ when $r(z; \beta) = \exp(-\beta z)$ and let $\hat{\beta}$ be the solution to $D(b) = 0$ as detailed in (c). Derive the asymptotic distribution of $\sqrt{n}(\hat{\beta} - \beta_0)$.

Hint. Note that $P(\sqrt{n}(\hat{\beta} - \beta_0) \leq s) = P(D(b_n) \leq 0)$, where $b_n = \beta_0 + s/\sqrt{n}$.

Problems for Section 9.3

1. Show that (9.3.5) is implied by (9.3.3) and (9.3.4).

2.(a) Show that $I^{-1}(P_0 : \nu, \mathcal{Q})$ given in (9.3.2) is valid and independent of parametrization.
Hint. Let $t \rightarrow \tau(t)$ be a reparametrization of \mathcal{Q} . Let $f^*(\cdot, \tau) = p(\cdot, t(\tau))$ where $t(\tau)$ has inverse $\tau(t)$, $t(0) = 0$, and $t'(\tau) > 0$. Let $P_{f,\tau}$ denote the probability corresponding to $f(\cdot, \tau)$ and compute

$$\frac{\partial}{\partial \tau} \log f^*(\cdot, \tau) \Big|_{\tau=0}, \quad \frac{\partial}{\partial \tau} \nu(P_{f,\tau}) \Big|_{\tau=0}.$$

(b) Show that ψ^* and $\dot{\mathcal{Q}}$ as defined in Definition 9.3.1 and (9.3.8) do not depend on the parametrization of \mathcal{Q} .

Hint. Note that $\dot{l}_f = t(a)\dot{l}_p$.

3. In Example 9.3.1, show that $\sup\{I^{-1}(P_0 : \nu, \mathcal{Q}_a) : a \in R^d\} = ([nE(\mathbf{Z}\mathbf{Z}^T)]^{-1})_{(1,1)}$.

4. Show that in parametric models \mathcal{P} satisfying the conditions of Theorem 6.2.2, the MLE is efficient in the sense of Proposition 9.3.1.

Hint. Take any $a \in R^d, \theta_0$ and consider the one dimensional model $Q \equiv \{P_{\theta_0 + \lambda a} : |\lambda| \leq \varepsilon\}$.

5. Derive Theorem 3.4.3 from Proposition 9.3.2.

6.(a) Give the tangent spaces for the $\{\mathcal{G}(p, \lambda) : p > 0, \lambda > 0\}$, $\{\beta(r, s) : r > 0, s > 0\}$ models at a generic point of the parameter space.

(b) Suppose $p(x, \eta) = \exp\{\sum_{j=1}^d \eta_j T_j(x) - A(\eta)\}h(x)$, $\eta \in \mathcal{E}$. Compute the tangent space of this model at η_0 .

7. Suppose $X \sim P_{\theta, \eta}$, $\theta \in R$, $\eta \in R^d$, a regular parametric model satisfying A0 – A6 of Theorem 6.2.2 for $l(X, \theta, \eta)$. Let $\hat{\theta}$ be the MLE of θ . Show that $\hat{\theta}$ has influence function at (θ_0, η_0) given by

$$\tilde{l}(X, P_0 : \theta, \mathcal{P}) = \frac{\dot{l}_1 - \Pi(\dot{l}_1 | [\dot{l}_{21}, \dots, \dot{l}_{2d}])}{\|\dot{l}_1 - \Pi(\dot{l}_1 | [\dot{l}_{21}, \dots, \dot{l}_{2d}])\|^2}$$

where

$$\dot{l}_1(X) = \frac{\partial l}{\partial \theta}(X, \theta_0, \eta_0), \quad l_{2j}(X) = \frac{\partial l}{\partial \eta_j}(X, \theta_0, \eta_0), \quad 1 \leq j \leq d.$$

Hint. \tilde{l} is the first coordinate of $I^{-1}(\theta_0, \eta_0)(\dot{l}_1, \dot{l}_{21}, \dots, \dot{l}_{2d})^T$. Use B.10.20.

8. Suppose $X = (Z, Y)$, $Y = \beta Z + \varepsilon$, $E(\varepsilon | Z) = 0$ with the joint distribution of (Z, ε) otherwise arbitrary and $E\varepsilon^2 < \infty$, $0 < EZ^2 < \infty$.

(a) Show formally that the tangent space of this model at $(\alpha_0, \beta_0, \mathcal{L}_0)$ where \mathcal{L}_0 is the

joint distribution of (Z, ε) is given by

$$\left\{ h(Z, \varepsilon) + c \frac{f'_0}{f_0}(Z, \varepsilon) : \begin{array}{l} f_0 \text{ the joint density of } (Z, \varepsilon), Eh(Z, \varepsilon) = 0, \\ Eh^2(Z, \varepsilon) < \infty, E(\varepsilon h(Z, \varepsilon)|Z) = 0 \text{ and } c \in R \end{array} \right\},$$

and that the influence function of the LSE does not belong to the tangent space in general.

(b) Suppose that Z is two valued, $P[Z = 0] > 0$, $P[Z = 1] > 0$. Show that the estimate which is the MLE for β in the model

$$Y_i = \alpha + \beta Z_i + \sigma(Z_i) \varepsilon_i$$

where the ε_i are i.i.d. $\mathcal{N}(0, 1)$ and independent of the Z_i , the Z_i are treated as fixed, and $\sigma(\cdot)$ as known is efficient for β at P_0 with the ε_i as above.

(c) Suppose $\sigma^2(Z)$ is unknown but the ε_i are still Gaussian $\mathcal{N}(0, 1)$ independent of Z_i . How would you construct an efficient estimate?

Hint. (a) If $p_t(z, \varepsilon) = \exp\{th(z, \varepsilon) - \Delta(p_0, h)\}p_0(z, \varepsilon)$ and $\int \varepsilon p_t(z, \varepsilon) d\varepsilon = 0$ for all t , z , then $E(\varepsilon h(Z, \varepsilon)|Z) = 0$.

(c) The LSE is a \sqrt{n} consistent estimate of β .

9. In the censoring example (Example 9.3.4), let $\nu(P) = S(t_0)$ for some fixed t_0 .

(a) Show (formally) that the tangent space at (F, G) in this example is

$$\left\{ h(Y, \delta) \equiv \delta \Delta_1(Y) - \int_{-\infty}^Y \Delta_1(s) d\Lambda_F(s) + 1 - \delta \Delta_2(Y) - \int_{-\infty}^Y \Delta_2(s) d\Lambda_G(s), \right. \\ \left. Eh(Y, \delta) = 0, Eh^2(Y, \delta) < \infty \right\}.$$

(b) Show that the influence function of the empirical MLE of ν , $\exp\{-\int_0^Y \frac{d\widehat{H}_1(s)}{\widehat{H}(s)}\}$ is of this form.

10. Show that in Example 9.3.5 the influence function of the Cox estimate (9.2.15) is of the form (9.3.23).

Hint. Set up identity and differentiate to solve.

11. (a) Show that the influence function of $\tilde{F}(x)$ for \tilde{F} as given in Example 9.1.2 belongs to $\dot{\mathcal{P}}_2(P_0)$ as given in Example 9.3.2.

(b) Show that $\theta(\widehat{F}) \equiv \widehat{F}(x)$ is regular. You may assume $w_j \geq \varepsilon > 0$ for $1 \leq j \leq k$ if necessary.

12. In Example 9.2.4, show that $\widehat{\beta}^*$ given by (9.2.23) is efficient.

13. Consider the nonparametric regression model $Y = \mu(X) + e$, where $\mu(\cdot)$ is unknown, $E(e) = 0$, $\text{Var}(e) = \sigma^2$, X and e are independent with c_{in} with continuous case densities f and g . The nonparametric correlation is $\eta^2 \equiv \text{CORR}^2(\mu(x), Y)$. Assume g' and η^2 exist. Recall (Problem 7.2.25) that $\eta^2 = [E(\mu^2(x)) - \mu_Y^2]/\text{Var}(Y)$. Formally find the

tangent space $\dot{\mathcal{P}}(P_0)$ and show that it is of the form “(function of x) + (function of x times a function of e).”

Remark. By Problem 7.2.25 it follows that the influence function $\mu^2(x) + 2\mu(x)e$ of $\nu(P) = E(\mu^2(X))$ is in $\dot{\mathcal{P}}(P_0)$. Thus $\widehat{\nu} \equiv n^{-1} \sum_{i=1}^n \widehat{\mu}^2(X_i)$ will be an efficient estimate of $\nu(P)$ provided $\widehat{\mu}(\cdot)$ is an “accurate” estimate of $\mu(\cdot)$. See Doksum and Samarov (1995) and Aït-Sahalia, Bickel and Stoker (2001) for such estimates of $E(\mu^2(X))$ for multivariate X .

14. Suppose $(X, Y) \sim P \in \mathcal{P}$ where \mathcal{P} is the bivariate normal copula model with $(\Phi^{-1}(F(X)), \Phi^{-1}(G(Y))) \sim \mathcal{N}(0, 0, 1, 1, \rho)$.

(a) Show that if F and G are known, then the tangent space at $(0, 0, 1, 1, \rho_0)$ is

$$\dot{\mathcal{P}}_1(P_0) = [\rho_0 + (1 + \rho_0^2)zw - \rho_0(z^2 + w^2)] / (1 - \rho_0)^2$$

where $z = \Phi^{-1}(F(x))$ and $w = \Phi^{-1}(G(y))$.

Hint. $P(X \leq x, Y \leq y) = H(z, w)$ where $H = \mathcal{N}(0, 0, 1, 1, \rho)$.

(b) Suppose F and G are unknown, and that $f = F'$ and $g = G'$ exist. Show that the tangent space $\dot{\mathcal{P}}(P_0)$ at (P_0, F_0, G_0) is $\dot{\mathcal{P}}_1(P_0) + \dot{\mathcal{P}}_2(P_0) + \dot{\mathcal{P}}_3(P_0)$ where

$$\begin{aligned}\dot{\mathcal{P}}_2(P_0) &= \frac{b(x)}{f_0(x)} + \frac{b(x)f_0(x)}{\varphi(z)} \left(\frac{\rho_0 w - z}{1 - \rho_0^2} + z \right) \\ \dot{\mathcal{P}}_3(P_0) &= \frac{c(y)}{g_0(y)} + \frac{c(y)g_0(y)}{\varphi(w)} \left(\frac{\rho_0 z - w}{1 - \rho_0^2} + w \right)\end{aligned}$$

for some $b(x)$ and $c(y)$ in $L_2(P_0)$. Here φ , f_0 , and g_0 are the continuous case densities of Φ , F_0 , and G_0 .

Remark. Klaassen and Wellner (1997) show that the normal scores correlation coefficient of Example 8.3.12 is efficient for the bivariate normal copula model. That is, they show that the influence function of $\nu(\widehat{P})$ is in $\dot{\mathcal{P}}(P_0)$, where $\nu(P) = \text{Corr}(\Phi^{-1}(F(X)), \Phi^{-1}(G(Y)))$.

Problems for Section 9.4

1. Validate (9.4.3).

2. (a) Show that \mathcal{S}_n of (9.4.13) is well defined, i.e. finite even if $M = \infty$.

(b) Show that weak convergence theory cannot yield anything here since

$$\sup_x |Z(x)| [\Phi(1 - \Phi)]^{-\frac{1}{2}}(x) = \infty$$

where Z is given in Example 9.2.1.

Hint. Use the Law of the Iterated Logarithm,

$$\overline{\lim}_{t \rightarrow 0} |W(t)| t^{-\frac{1}{2}} |\log_2 t|^{-\frac{1}{2}} > 0$$

where $\log_2 \equiv \log \log$.

3. Suppose that $F_n(x) = F_0(x) + \Delta_n(x)/n^{\frac{1}{2}}$ where $|\Delta_n - \Delta|_\infty \rightarrow 0$ and let P_n be the distribution under which X_1, \dots, X_n are i.i.d. F_n .

(a) Show that, under P_n ,

$$\mathcal{E}_{n0}(\cdot) \equiv \sqrt{n}(\widehat{F}(\cdot) - F_0(\cdot)) \implies W^0(F_0(\cdot)) + \Delta(\cdot).$$

(b) Let F_0 be the $\mathcal{U}(0, 1)$ df and let ϕ_1, ϕ_2, \dots be the eigenfunctions of the kernel

$$h \rightarrow Kh, \quad h \in L_2(0, 1)$$

where $[Kh](s) = \int_0^1 h(u)(s \wedge u - su)du$, i.e.

$$\int_0^1 \phi_a(u)\phi_b(u)du = \delta_{ab}, \quad K\phi_a = \lambda_a\phi_a.$$

The ϕ and λ are well known to be

$$\phi_a(u) = \sqrt{2}\sin(a\pi u), \quad 0 \leq u \leq 1, \quad \lambda_a = (a\pi)^{-2}.$$

Show that if $\Delta \in L_2(0, 1)$, $\Delta = \sum_{k=1}^{\infty} (\Delta, \phi_k)\phi_k$ where $(\Delta, \phi_k) \equiv \int_0^1 \Delta(u)\phi_k(u)du$ and $\widehat{Z}_{an} \equiv \int \phi_a(u)\mathcal{E}_{n0}(u)$. Then, for all k

$$(\widehat{Z}_{1n}, \dots, \widehat{Z}_{kn}) \xrightarrow{\text{FIDI}} (Z_1, \dots, Z_k)$$

where Z_j are independent $\mathcal{N}(\mu_j, \lambda_j)$, $1 \leq j \leq k$, and $\mu_j \equiv (\Delta, \phi_j)$.

(c) Show that $\int_0^1 \mathcal{E}_{n0}^2(u)du \implies \sum_{k=1}^{\infty} Z_k^2 = \sum_{j=1}^{\infty} \lambda_j(Z'_j + \Delta_j\lambda_j^{-\frac{1}{2}})^2$ where Z'_j , $j \geq 1$, are i.i.d. $\mathcal{N}(0, 1)$.

Hint. (c) $E \int_0^1 (\mathcal{E}_n(u) - \sum_{k=1}^m Z_{kn}\phi_k(u))^2 du = \sum_{k=m+1}^{\infty} (\lambda_k + (\Delta, \phi_k)^2)$.

4. Extend the result of Problem 3 to the situation of Example 9.4.2. Suppose $T = \{1(-\infty, s] : s \in R\}$, let P_n correspond to $F_{\boldsymbol{\theta}_0}(x) + \Delta_n(x)/n^{\frac{1}{2}}$ and $|\Delta_n - \Delta|_{\infty} \rightarrow 0$. Show that if

$\widehat{Z}_n(\cdot) = \sqrt{n}(\widehat{F} - F_{\widehat{\boldsymbol{\theta}}_0})(\cdot)$, then, under P_n ,

(a) $\widehat{Z}_n(\cdot) \implies Z_{\boldsymbol{\theta}_0}^0(\cdot) + (\Delta - \Pi_{\boldsymbol{\theta}_0}(\Delta))(\cdot)$.

(b) In particular, show that if $\Delta \in [\frac{\partial l}{\partial \boldsymbol{\theta}_0}(X, \boldsymbol{\theta}_0) : 1 \leq j \leq d]$, then any test based on $\widehat{Z}_n(\cdot)$ has no power for the alternative P_n . Why is this qualitatively reasonable?

5. Show, extending the result of Problem 9.2.4, that for any fixed S , $T_{n1} \equiv \sqrt{n}(\widehat{F} - F_{\widehat{\boldsymbol{\theta}}})(s)$ has no greater, and usually less, power as a test statistic for $H : F = F_{\boldsymbol{\theta}_0}$ vs $K : F_n$ as in Problem 9.2.4 than does $T_{n2} \equiv \sqrt{n}(\widehat{F} - F_{\boldsymbol{\theta}_0})(s)$.

Hint. Compute critical values under H using Problem 9.2.4 and then show as in Chapter 5

that the power for T_{nj} is governed by a ratio of signal to noise, μ_j/σ_j . Show that μ_2^2/σ_2^2 can be written as $(a, b)^2/\|b\|^2$ for a suitable a, b while

$$\frac{\mu_1^2}{\sigma_1^2} = \frac{(a, b - \Pi_0 b)^2}{\|b - \Pi_0 b\|^2} = \frac{(a, \Pi_0^\perp(b))^2}{\|\Pi_0^\perp(b)\|^2}$$

where Π_0, Π_0^\perp are projection operators. Use the identity

$$\Pi(a|[b]) = \Pi(a|\Pi_0^\perp b]) + \Pi(a|\Pi_0 b])$$

and hence $\|\Pi(a|[b])\|^2 \geq \|\Pi(a|\Pi_0^\perp b])\|^2$.

6. Suppose $\hat{\nu}$ is an efficient estimate of $\nu(P)$ for $P \in \mathcal{P}$. Let $\tilde{l}(X, P_0 : \nu, \mathcal{P})$ be its influence function. Suppose that, for all $P_0 \in \mathcal{P}$, $\tilde{l}(\cdot, \hat{P} : \nu, \mathcal{P})$ is well defined and

$$E_0 \left(\frac{1}{n} \sum_{i=1}^n (\tilde{l}(X_i, \hat{P} : \nu, \mathcal{P}) - \tilde{l}(X_i, P_0 : \nu, \mathcal{P})) \right)^2 \rightarrow 0.$$

(a) Show how to construct asymptotic $1 - \alpha$ confidence bounds for $\nu(P)$, $P \in \mathcal{P}$ based on $\hat{\nu}$.

(b) Suppose that

$$n \text{Var}^* \hat{\nu}^* \xrightarrow{P_0} I^{-1}(P_0 : \nu, \mathcal{P})$$

where the * as usual indicate computation under the bootstrap distribution given \hat{P} as defined in Section 9.4. Show how to construct the asymptotic $1 - \alpha$ confidence bounds in this case.

Problems for Section 9.5

1. Establish Proposition 9.5.1.

Hint. Given A_n , by the Neyman–Pearson lemma, there exists $c_n < \infty$ such that $P^{(n)}[L_n > c_n] \leq P^{(n)}(A_n) \leq P^{(n)}[L_n \geq c_n]$ and $Q^{(n)}[L_n \geq c_n] \geq Q^{(n)}(A_n)$. If $P^{(n)}(A_n) \rightarrow 0$, suppose $c_n \uparrow c \leq \infty$. Since $P^{(n)}[L_n > c_n] \rightarrow 0$, $P[\Lambda \geq c] = 0$. For suitable $d < c$, $1 - Q^{(n)}[L_n \geq d] = Q^{(n)}[L_n < d] = E_{P^{(n)}}(\Lambda_n 1(L_n < d)) \rightarrow E\Lambda 1(\Lambda < d)$. Since $E\Lambda = 1$, $Q^{(n)}[L_n \geq d] \rightarrow E\Lambda 1(\Lambda \geq d)$. But $E(\Lambda 1(\Lambda \geq d)) \rightarrow 0$ as $d \uparrow c$ by the dominated convergence theorem.

2. Establish (9.5.7).

3. In Example 9.5.1, show that $E[(T_u - S_U)/s]^2 \rightarrow 0$ under Lindeberg's condition.

Hint. Put bounds on $\text{Var}(U_{(i)})$ and $\text{Cov}(U_{(i)}, U_{(j)})$ using Problem B.2.9. Note that the covariance can be obtained from

$$\text{Var}(Y - X) = \text{Var}X + \text{Var}Y - 2\text{Cov}(X, Y).$$

4. Let X_1, \dots, X_n be i.i.d. with continuous df F and density $f = F'$; and let Y_1, \dots, Y_n be i.i.d. with df $G(y - \theta)$. This is Example 9.5.1 with $z_i = 0, 1 \leq i \leq m, z_i = 1$,

$m + 1 \leq i \leq N$, where $N = m + n$. In this case T_U of Example 9.5.1 is the two-sample Wilcoxon statistic. Show that for $\theta = t/\sqrt{n}$, $t > 0$ and $\lambda = \lim_{N \rightarrow \infty} (n/N)$, if $\lambda \in (0, 1)$, the asymptotic power of T_U is

$$\lim_{N \rightarrow \infty} P(\sqrt{12}T_U \geq z_{1-\alpha}) = 1 - \Phi\left(z_{1-\alpha} - \left[\sqrt{\lambda(1-\lambda)} t/12\sigma \int f^2(x)dx\right]\right).$$

Hint. By contiguity it is enough to compute $\lim_{N \rightarrow \infty} P(\sqrt{12}S_U \geq z_{1-\alpha})$ where S_U is defined in Example 9.5.1.

5. Let $G(t) = 1 - \exp\{-t\}$, $t \geq 0$, and $r(t) = f(t)/[1 - F(t)]$ where $\{t : f(t) > 0\} = [0, \infty)$. Show that

- (a) If $r(t)$ is increasing on $[0, \infty)$ and $i < j$, then $P_F(D_i \leq D_j) \geq P_G(D_i \leq D_j) = \frac{1}{2}$.
- (b) Define F to have more increasing failure rate (IFR) than F_1 , written $F \underset{c}{>} F_1$, if $F_1^{-1}F$ is convex (cf. van Zwet (1964)). Show that if F_1 and F are increasing on $[0, \infty)$, 0 elsewhere, $F \underset{c}{>} F_1$ and $i < j$, then

$$P_F(D_i \leq D_j) \geq P_{F_1}(D_i \leq D_j).$$

- (c) A test $\phi(\cdot)$ is said to have a *monotone power* for IFR testing if

$$F \underset{c}{>} F_1 \implies E_F(\phi(\mathbf{X})) \geq E_{F_1}(\phi(\mathbf{X})).$$

In Example 9.5.4, let $D = (D_1, \dots, D_n)$ and $D' = (D'_1, \dots, D'_n)$ be two vectors of normalized spacings and let $\delta(\cdot)$ be a test function based on normalized spacings. Then $\delta(\cdot)$ is said to be *monotone wrt IFR testing* if $\delta(D') \leq \delta(D)$ for all D and D' with D'_i/D_i nondecreasing in i . Show that the tests $\delta_{U,v}(\cdot)$ that reject exponentiality in favor of K : “ F is IFR” when $U \geq v$, $v > 0$, have monotone power and are unbiased. (U is given by (9.5.12).)

Hint (a). Use (b) with $F_1 = G$ for the first inequality.

Hint (b). Since $F_1^{-1}F$ is increasing, $X'_{(i)} = F_1^{-1}F(X_{(i)})$ is the i th order statistic in a random sample from a population with distribution F_1 . Let $D'_i = (n - i + 1)(X'_{(i)} - X'_{(i-1)})$, $i = 1, \dots, n$. Since $F_1^{-1}F$ is convex, $i < j$ and $D'_i \geq D'_j$ implies $D_i \geq D_j$.

Hint (c). Let Z be a discrete random variable taking on the values $1, \dots, n$ and let P_0 and P_1 be probabilities such that $P_0(Z = k) = D_k / \sum D_i$ and $P_1(Z = k) = D'_k / \sum D'_i$. Let D and D' be such that D'_k/D_k is nondecreasing in k , then P_0 and P_1 have monotone likelihood ratio in k . Set $c_n(i) = c_i$. It follows from Theorem 4.3.1 (first established by Lehmann (1955)) that $E_{P_0}(-c_n(Z)) = -\sum c_n(i)D_i / \sum D_i \leq -\sum c_n(i)D'_i / \sum D'_i = E_{P_1}(-c_n(Z))$; thus the tests $\delta_{U,v}(\cdot)$ are monotone. Next use the hint to (b).

6. Let $G(t) = 1 - e^{-t}$; $t \geq 0$. Consider the models $P_{\eta,\lambda}^j$ with densities $\lambda f_j(\lambda x; \eta)$, $j = 1, 2, 3$, $x > 0$, $\lambda > 0$, $\eta \geq 0$, where, for $t \geq 0$.

$$\begin{aligned} f_1(t; \eta) &= (1 + \eta)t^\eta \exp\{-t^{(1+\eta)}\} && (\text{Weibull}) \\ f_2(t; \eta) &= (1 + \eta t) \exp\{-(t + \frac{1}{2}\eta t^2)\} && (\text{Linear FR}) \\ f_3(t; \eta) &= [1 + \eta G(t)] \exp\{-[t + \eta(t - G(t))]\} && (\text{Exponential FR}) \end{aligned}$$

- (a) Show that the FRs for $f_j(t; \eta)$, $j = 1, 2, 3$, are $(1 + \eta)t^\eta$, $(1 + \eta t)$, and $1 + \eta G(t)$.
- (b) Show that for the models $P_{\eta,\lambda}^j$, $j = 1, 2, 3$, the statistic $\sum_{i=1}^n d_1\left(\frac{i}{n+1}\right)(D_i - 1)$ as defined by (9.5.15) and (9.5.16) has $d_1(u)$ given by $-\log[-\log(1-u)]$, $\log(1-u)$, and $-u$, respectively (recall that $\lambda = 1$ in (9.5.16)).

7. Show that if $X_{(i)}$ is an $\mathcal{E}(1)$ order statistic, and if $G(t) = 1 - \exp(-t)$, $t \geq 0$, then

$$\left[X_{(i)} - G^{-1}\left(\frac{i}{n+1}\right)\right]^2 = O_P(n^{-1}).$$

Hint. Write $X_{(i)} = G^{-1}(U_{(i)})$ where $U_{(i)}$ is a uniform $(0, 1)$ order statistic. Expand $G^{-1}(U_{(i)})$ around $E(U_{(i)}) = i/(n+1)$ and use Problem B.2.9.

8. In Example 9.5.4, let $c_i = c\left(\frac{i}{n+1}\right)$ with $\int c(u) du = 0$ and $\int_0^1 c^2(u) du < \infty$.

(a) Establish (9.5.17).

(b) Set $U^* = \sum_{i=1}^n c_i(n^{-1}\lambda D_i - 1)$. Show that $n^{-\frac{1}{2}}(U - U^*) \xrightarrow{P} 0$ as $n \rightarrow \infty$ under the exponential hypothesis. It follows that U and U^* have the same asymptotic distributions for contiguous alternatives.

(c) Use LeCam's Third Lemma and (9.5.17) to give the asymptotic power function for contiguous alternatives of the test based on U .

(d) Use (c) above to find $c(\cdot)$ that maximizes the asymptotic power function of U .

9. In Example 9.5.2 assume the model satisfies the ULAN condition. Show that $\Lambda_n = S_n + o_P(1)$ under $K : \theta_n = \theta_0 + t/\sqrt{n}$.

10. Consider the partial linear model $Y = \beta Z + g(0) + \varepsilon$, $\varepsilon \sim N(0, \sigma^2)$, of Examples 9.1.13, 9.2.4, and 9.3.5.

(a) Establish formally analogues of (9.5.9). Identify ν_0 with β and $\boldsymbol{\eta}$ with $(\mu_1, \mu_2)^T$ where $\mu_1(u) = E(Z|U = u)$ and $\mu_2(u) = E(Y|U = u) = \beta\eta_1(u) + g(u)$. Recall that we observe X_1, \dots, X_n i.i.d. as $X = (Z, U, Y)$. Assume n even, set $k = n/2$, and use X_1, \dots, X_k to estimate $\mu_1(\cdot)$, $\mu_2(\cdot)$ and X_{k+1}, \dots, X_n to estimate the influence function. Use (11.6.1) to estimate $\mu_1(\cdot)$ and $\mu_2(\cdot)$.

(b) Determine sufficient conditions for your estimates of (a) to satisfy (9.5.9).

(c) Find the asymptotic power function of the Neyman c_α test (9.5.10) for $H : \beta = 0$ vs $K : \beta > 0$.

(d) Suppose (U, Z) has a $\mathcal{N}(0, 0, 1, 1, \rho)$ distribution. Discuss the behaviour of the power function in (c) as a function of ρ .

- 11.** Consider the logistic copula regression model of Example 8.3.11 with $\theta_i = \beta_0 + \beta_1 z_i$. Assume that the z_i , $1 \leq i \leq n$, satisfy Lindeberg's condition and that the baseline distribution F is known. We test $H : \beta = 0$ vs $\beta > 0$ using $S_U = n^{-\frac{1}{2}} \sum_{i=1}^n F(X_i)(z_i - \bar{z})$. Find the asymptotic power of this test for contiguous alternatives $\beta = t/\sqrt{n}$.

Remark. By Problem 9.5.3, the uniform scores rank statistic T_U has the same asymptotic power as S_U .

- 12.** Establish (9.5.10).

- 13.** Establish (9.5.11).

Hint. $l^* = \frac{\partial l}{\partial \nu_0} - \pi_0 \left(\frac{\partial l}{\partial \nu_0} | \dot{\mathcal{P}}_1(P_0) \right)$ is orthogonal to $\dot{\mathcal{P}}_1(P_0)$.

- 14.** Convergence of the power function $\beta_n(\theta)$ of the Neyman C_α test is uniform on $n^{-\frac{1}{2}}$ neighbourhoods of θ_0 .

Hint. Apply Le Cam's third lemma to $S_n(\nu_0, \hat{\eta})$ for sequences $(\nu_0 + \frac{s}{\sqrt{n}}, \eta_0 + \frac{t}{\sqrt{n}})$.

- 15. The Neyman-Scott model.** Suppose $X_{ij} = Y_{ij} + \sigma Z_{ij}$, $i = 1, \dots, n$, $j = 1, \dots, K$, $K \geq 2$ where $Y_{ij} \equiv Y_i$, independent of Z_{ij} and each other and distributed according to F unknown, and Z_{ij} are i.i.d. $\mathcal{N}(0, 1)$. Only X_{ij} are observed.

- (a) Construct a test for $H : \sigma^2 = 1$ vs $\sigma^2 > 1$, F unknown, which is size α and has asymptotic power $> \alpha$ for $\sigma^2 > 1$ independent of F .

Hint. Consider $U_i \equiv \frac{1}{K-1} \sum_{j=1}^k (X_{ij} - \bar{X}_i)^2$, i.i.d. $\sigma^2 \chi_{K-1}^2$.

- (b) Show that this test is asymptotically most powerful among all asymptotically level α tests of $F = N(\mu, \tau^2)$ for any μ, τ^2 .

Hint. Compute l^* .

- (c) **2. Neyman Scott Z.** Give a heuristic argument for or against the test being asymptotically most powerful at other F .

- (d) Suppose you treat the Y_{ij} as unknown constants μ_i . Show that you arrive to the same test.

- (e) However, show that under the assumption of (b) the MLE of σ^2 is not consistent.

- 16.** Suppose that $P_n(\mathbf{x}, \theta)$, $\theta \subset R$, satisfy the ULAN condition with $\mathbf{T}_n(\mathbf{X})$ is asymptotically $\mathcal{N}(0, I)$ under $P_n(\mathbf{x}, \theta)$, S_n is asymptotically ancillary if $\mathcal{L}_{\theta+tn^{-\frac{1}{2}}}(S_n) + \mathcal{N}(\mu, \sigma^2)$ independent of t . Show that T_n and S_n are then asymptotically independent.

9.7 Notes

Notes for Section 9.1.

(1) Owen (2001) makes his extension of empirical likelihood to (multiple) biased sampling by considering k independent samples of size n_j , with $n_j > 0$, $1 \leq j \leq k$, $\sum_{j=1}^k n_j = n$, and permitting an empirical likelihood model for each.

(2) The Cox model was introduced by Cox (1972,1975) in connection with yet another modified likelihood, partial likelihood. See Example 9.1.11. Other modifications, conditional and marginal likelihoods, are discussed in Kalbfleisch and Sprott (1970) and Kalbfleisch and Prentice (1973). See Problem 9.1.23.

Notes for Section 9.2.

(1) For an alternate proof of the concavity of $\Gamma_\tau(\beta, P)$, see Problem 9.2.12.

Note for Section 9.3.

(1) It may turn out that even though we have a candidate efficient influence function, no regular estimate exists — see BKRW.

Notes for Section 9.5.

(1) Le Cam's notion of LAN is uniform in a much weaker sense. He requires only $R_n(t_n) \xrightarrow{P} 0$ if $t_n \rightarrow t$. The notion we use is due to Hájek (1970). It simplifies proofs and applies to all common situations of interest in which Le Cam's weaker property applies.

(2) The notion of contiguity is an asymptotic analogue of the familiar measure theoretic, finite n property that $P^{(n)}$ dominate $Q^{(n)}$.

Chapter 10

MONTE CARLO METHODS

10.1 The Nature of Monte Carlo Methods

The importance of Monte Carlo methods as we have discussed before is that they enable us to construct estimates of features of probability distributions which are impossible or very difficult to compute analytically. Here “estimate” is not used in the sense we have before of a quantity depending on data (which is random through the data being used) to estimate a feature of an *unknown* probability distribution. Rather it is we who are generating random quantities which can be used to estimate features of a probability distribution specified in a *completely known* way.

We can couple the plug-in principle (Section 2.1.2) with Monte Carlo methods to obtain statistical estimates as well. That is, if we have data from an unknown probability distribution belonging to a model we can estimate that distribution in some way and then, if we have a feature (parameter) of the true distribution we want to estimate, use Monte Carlo to estimate the feature by random quantities generated using the estimated distribution. We shall discuss both Monte Carlo and statistical applications of Monte Carlo in this chapter.

Here are some typical examples of statistical situations where Monte Carlo is natural.

Example 10.1.1. *Goodness-of-fit tests.* Consider testing $H : F = F_0$ using a statistic $T_n(X_1, \dots, X_n)$ where X_1, \dots, X_n are i.i.d. with df F . We need an α critical value c_n such that $P_{F_0}[T_n \geq c_n] = \alpha$.

Monte Carlo Solution: Generate, on the computer⁽¹⁾, B i.i.d. sets of n i.i.d. observations $(X_{11}, \dots, X_{1n}), \dots, (X_{B1}, \dots, X_{Bn})$, from F_0 and form $T_{nb} \equiv T_n(X_{1b}, \dots, X_{nb})$, $1 \leq b \leq B$. Then, for large B , the empirical distribution of the T_{nb} , $1 \leq b \leq B$, is arbitrarily close to the distribution of $T_n(X_1, \dots, X_n)$ under F_0 by the law of large numbers. Thus if $T_{(1)} \leq \dots \leq T_{(B)}$ are the ordered T_{nb} and $[]$ is the greatest integer function, the natural estimate $T_{([B(1-\alpha)]+1)}$ of c_n tends to c_n as $B \rightarrow \infty$. (We assume c_n uniquely defined here.) See also Example 4.1.6. Recall again that B can be made as large as we please. Here are two applications:

(a) A simple application is to *permutation tests*. Consider the two sample problem where X_1, \dots, X_n are independent with X_1, \dots, X_m i.i.d. G_1 , X_{m+1}, \dots, X_n i.i.d. G_2 and $H : G_1 = G_2$. This hypothesis is not simple. However, as we have noted in Chapter 8, under H , if $X_{(1)} \leq \dots \leq X_{(n)}$ are the ordered X_1, \dots, X_n , the conditional distribution of

(X_1, \dots, X_n) given $X_{(1)} = x_{(1)}, \dots, X_{(n)} = x_{(n)}$, $x_{(1)} \leq \dots \leq x_{(n)}$, the permutation distribution is simply the uniform distribution on $\{(x_{(i_1)}, \dots, x_{(i_n)}) : (i_1, \dots, i_n)$, a permutation of $\{1, \dots, n\}\}$. This is our F_0 . Sampling from this distribution is easy. One simply selects i_1 uniformly from $\{1, \dots, n\}$, i_2 uniformly from $\{1, \dots, n\} - \{i_1\}$, and so on. A routine for doing this is implemented in *R* for instance as an option of the “sample” function. The two-sample t -statistic

$$T = \sqrt{\frac{m(n-m)}{n}} \left(\frac{\bar{Y} - \bar{X}}{s} \right)$$

of Example 8.2.7 is an example of a statistic T_n . The permutation distribution of T is appropriate when G_1 and G_2 are not Gaussian and $\min(m, n-m)$ is not large.

(b) A second application is to *testing for normality*. As in Example 4.1.6 and Section I.2, consider testing the goodness-of-fit hypothesis $H : F = \Phi(\frac{x-\mu}{\sigma})$ for some μ, σ , where Φ is the $\mathcal{N}(0, 1)$ distribution and say the statistic is an unorthodox one such as that of Shapiro–Wilk type, the residual sum of squares of a normal quantile plot for the data,

$$T_n(X_1, \dots, X_n) = \sum_{i=1}^n \left(Z_{(i)} - \Phi^{-1} \left(\frac{i}{n+1} \right) \right)^2, \quad Z_{(i)} = \frac{X_{(i)} - \bar{X}}{\hat{\sigma}}.$$

For this T_n , we can take $F_0 = \mathcal{N}(0, 1)$ because T_n is invariant under $X_i \rightarrow (X_i - \mu)/\sigma$. Approximations to the distribution of T_n under F_0 are known (Venter and de Wet (1972)) but the exact distribution is only sparsely tabulated. As in Example 4.1.6, to do Monte Carlo we need a set of B simple random samples of size n from $\mathcal{N}(0, 1)$. Most statistical packages including *R* can generate these. But for some F_0 , it is not clear a priori how this is done, if we view as most primitive that one can generate $\mathcal{U}(0, 1)$ observations easily.⁽¹⁾ We shall discuss some methods in Section 10.2.

Example 10.1.2. Estimating risks functions. Given a parametric model for $\mathbf{X} \in R^n$, $\{P_\theta : \theta \in \Theta\}$, a decision theoretic formulation with loss function $l(\theta, \delta)$, and a decision rule $\delta(\mathbf{X})$, we want to compute the risk $R(\theta, \delta) = E_\theta l(\theta, \delta(\mathbf{X}))$ for selected θ . As discussed in Section 5.1, this can typically not be done analytically since if \mathbf{X} , say, has density $p(\mathbf{x}, \theta)$,

$$R(\theta, \delta) \equiv E_\theta l(\theta, \delta(\mathbf{X})) = \int l(\theta, \delta(\mathbf{x})) p(\mathbf{x}, \theta) d\mathbf{x}$$

is an n -dimensional integral.

The Monte Carlo approach is just to generate $\mathbf{X}_1, \dots, \mathbf{X}_B$ i.i.d. P_θ and approximate $R(\theta, \delta)$ by

$$\frac{1}{B} \sum_{b=1}^B l(\theta, \delta(\mathbf{X}_b))$$

and again appeal to the law of large numbers. For instance take $\mathbf{X} = (X_1, \dots, X_n)$ i.i.d. with df $F_0(x - \theta)$ where F_0 is symmetric with center of symmetry 0. Let, for $0 \leq \alpha < \frac{1}{2}$,

$$\bar{X}_\alpha = (n - 2[n\alpha])^{-1} \sum_{j=[n\alpha]+1}^{n-[n\alpha]} X_{(j)}$$

be the alpha trimmed mean. Then, if $\hat{\theta} = \bar{X}_\alpha$ and $l(\theta, d) = (\theta - d)^2$,

$$E_\theta(\hat{\theta} - \theta)^2 = E_0(\hat{\theta}^2),$$

and we can approximate the constant risk function arbitrarily closely by taking B samples of size n , from F_0 , and computing the average of $l(\theta, \delta(\mathbf{X}_1)), \dots, l(\theta, \delta(\mathbf{X}_B))$ over these samples. We evaluate the performance of \bar{X}_α and other estimates by repeating this for several F_0 of various shapes such as the Gaussian, Laplace, and Cauchy df's. See Problem 3.5.9 and Andrews et al (1972).

In the case $\alpha = \frac{1}{2}$, when \bar{X}_α is interpretable as the median, it is possible to use numerical integration in one dimension to evaluate its MSE by formula (5.1.2). For $\bar{X}_0 \equiv \bar{X}$ one can compute the characteristic function numerically and then use Fourier inversion. Even for \bar{X}_α , by conditioning on $\bar{X}_{([n\alpha]+1)}$ and $\bar{X}_{(n-[n\alpha])}$ it is possible to use one- and two-dimensional numerical integrations (Problem 10.1.2). However, these methods become more and more elaborate and fail entirely, for instance, for general linear combinations of order statistics. Thus, Monte Carlo is appropriate. \square

Both of these examples should make it clear that, so far, the use of Monte Carlo is an arbitrarily sharp approximation to a high-dimensional integral or sum. If the dimension (sample size) were 1 or 2 we could use more accurate numerical integration procedures. But even for modest sample sizes numerical integration doesn't work. The great virtue of (simple random sampling) Monte Carlo methods is that their MSE

$$E \left(\frac{1}{B} \sum_{b=1}^B l(\theta, \delta(X_b)) - E_\theta l(\theta, \delta(X)) \right)^2 = \frac{\text{Var } l(\theta, \delta(X))}{B}$$

is not dependent on the dimension n and the numerator is bounded uniformly if l is.

We shall see that simple random sampling is not always applicable, but important variants can be used.

Example 10.1.3. Joint posterior distributions. Consider a Bayesian framework where θ has prior density π on R^k and X given $\theta = \theta$ has density $p(x|\theta)$. Here both π and $p(x|\theta)$ are given analytically. By Bayes' rule, the posterior density is

$$\pi(\theta | X = x) = p(x|\theta)\pi(\theta) / \int_{R^k} p(x|\mathbf{t})\pi(\mathbf{t})d\mathbf{t} . \quad (10.1.1)$$

The denominator of (10.1.1) presents a difficult high-dimensional integration problem if k is at all large (unless π is an analytically explicit conjugate prior as in Section 1.6.5). The problem becomes even more acute if we need, say, the marginal posterior density of θ_1 which involves a $k - 1$ -dimensional integration for each θ_1 .

It is often very useful to represent the posterior distribution of θ by the empirical distribution of i.i.d. observations $\theta_1, \dots, \theta_B$ from the posterior distribution. In this case, obtaining an approximation to the (marginal) posterior distribution of, say, $q(\theta)$ is easy. We just use its representation by the empirical distribution of $q(\theta_1), \dots, q(\theta_B)$. Unfortunately it turns out that, if k is at all large, in general, it is very difficult to obtain a genuine simple

random sample, $\theta_1, \dots, \theta_B$, as above. For instance, consider a generalization of Example 3.2.1. Let $\theta = (\mu, \sigma^{-2})$ and let X_1, \dots, X_n be i.i.d. $\mathcal{N}(\mu, \sigma^2)$ given θ . Put on θ the prior distribution making μ and σ independent with $\mu \sim \mathcal{N}(\nu_0, \tau_0^2)$ and $\sigma^{-2} \sim \Gamma(p_0, \lambda_0)$. It is, *a priori*, quite unclear how to generate independent observations from the resulting posterior distribution.

In such situations, methods of a type we shall discuss in Section 10.4 known as Markov chain Monte Carlo (MCMC) are used to generate $\theta_1^*, \dots, \theta_B^*$ which are approximately independent and individually have approximately the correct posterior distribution. \square

Summary. We illustrated the usefulness of Monte Carlo methods in three situations. The first is where for some completely specified distribution F_0 , the distribution $L_F(T_n)$ of a test statistic equals $L_{F_0}(T_n)$ for all distributions F in a null-hypothesis class \mathcal{F}_0 . Suppose $L_{F_0}(T_n)$ cannot be computed analytically. We can then generate B independent samples from F_0 , compute the values T_{n1}, \dots, T_{nB} for these B Monte Carlo samples, and then approximate the null-hypothesis distribution of T_n by the empirical distribution of T_{n1}, \dots, T_{nB} . This case illustrates the need for algorithms for generating samples from a given F_0 . The second situation is where we need to compute the risk $E_{\theta, \tau} l(\theta, \delta(\mathbf{X}))$ of a decision procedure δ for a model with principle parameter θ and nuisance parameter τ . For instance, if we use a location equivariant estimate (Problem 3.5.6) to estimate the center of symmetry θ of a distribution $F_\theta(x) = F_0(x - \theta)$, with squared error loss, the risk is not a function of θ , but it is a function of $\tau = F_0$ that can typically not be computed analytically. The risk can be estimated by generating B independent samples from F_0 , and computing the average of $l(\theta, \delta(\mathbf{X}_1)), \dots, l(\theta, \delta(\mathbf{X}_B))$ over these samples. Thus, to evaluate the performance of decision procedures δ , we need to generate Monte Carlo samples from various interesting distributional shapes F_0 . The final example is where in a Bayesian framework, we want the posterior distribution of a random parameter θ given the data, but this distribution cannot be computed analytically. One fruitful approach is to generate i.i.d. observations $\theta_1, \dots, \theta_B$ from the posterior distribution and to use the empirical distribution of $q(\theta_1), \dots, q(\theta_B)$ to approximate the distribution of a quantity $q(\theta)$ of interest. In Section 10.4 we will introduce *Markov chain Monte Carlo* (MCMC) methods which are the most commonly used procedures for approximately generating such Bayesian Monte Carlo samples.

10.2 Three Basic Monte Carlo Methods

Having seen some examples of the potential applicability of Monte Carlo methods we now turn to the basic problem of generating simple random samples from a distribution on R^k which is specified either by its density p_0 or some other feature. We also consider the problem of using samples from p_0 to estimate integrals involving $p \neq p_0$, or to obtain samples from $p \neq p_0$.

The fundamental methods discussed in this section and a number of other topics are treated much more extensively in the classical book of Hammersley and Handscomb (1965) and the more recent books by Ripley (1987) and Liu (2001).

10.2.1 Simple Monte Carlo

We begin by considering the problem of generating $X \sim F$, for $X \in \mathcal{X}$. If $\mathcal{X} = R$ then the first basic approach is to use the probability integral transform. If F is continuous, $F^{-1}(u) = \inf\{x : F(x) \geq u\}$, and $U \sim \mathcal{U}(0, 1)$, then $F^{-1}(U) \sim F$ because

$$P(F^{-1}(U) \leq x) = P(U \leq F(x)) = F(x).$$

If F is specified analytically, this approach is natural since F^{-1} can always be computed, say, by bisection. Of course, it's best if F^{-1} is itself analytically described.

Example 10.2.1. *Exponential and Cauchy variables.* $\mathcal{E}(\lambda)$ variables can be generated as $-\lambda^{-1} \log U$ since $F(x) = 1 - \exp(-\lambda x)$ and $1 - U \sim \mathcal{U}(0, 1)$. Cauchy variables (see Problem B.2.1) can be generated as $\tan(\pi(U - \frac{1}{2}))$. \square

This method becomes more powerful by using known distribution results.

Example 10.2.2. *Box–Muller–Knuth method.* We want to generate independent pairs of $\mathcal{N}(0, 1)$ variables (X_1, X_2) . This may be done as follows. Let $X_1 = R \cos A$, $X_2 = R \sin A$, where (R, A) are the polar coordinates of (X_1, X_2) . Then by Theorem B.3.1, R^2 has an $\mathcal{E}(\frac{1}{2})$ distribution and is independent of A which is $\mathcal{U}(0, 2\pi)$. Hence

$$\sqrt{-2 \log U_1} \cos(2\pi U_2), \sqrt{-2 \log U_1} \sin(2\pi U_2)$$

have the desired joint distribution if (U_1, U_2) are independent $\mathcal{U}(0, 1)$. \square

The Box–Muller–Knuth method is an example of a general principle. When a distribution F corresponds to a function of independent variables which can be generated already, e.g., $\mathcal{U}(0, 1)$ or $\mathcal{E}(\lambda)$, then observations from F can be generated.

Example 10.2.3. *Gamma and beta.* By Theorem B.3.1, the gamma, $\Gamma(p, \lambda)$, distribution, for $p = m/2$ where m is an integer, can be represented as $(2\lambda)^{-1} \sum_{i=1}^m Z_i^2$ where the Z_i are i.i.d. $\mathcal{N}(0, 1)$. Thus by generating a block of m i.i.d. $\mathcal{N}(0, 1)$ variables we can generate a $\Gamma(p, \lambda)$ variable. If m is even, $m = 2p$, we can generate the variable more simply from a block of p i.i.d. $\mathcal{E}(\lambda)$ variables using the explicit form of F^{-1} for exponential variables.

For positive integers m and n , set $r = m/2$ and $s = n/2$. Then we can go further since $V = X_1/(X_1 + X_2)$ will have a beta, $\beta(r, s)$, distribution if X_1, X_2 are independent $\Gamma(r, 1), \Gamma(s, 1)$. Thus using independent gamma variables we can generate $\beta(r, s)$ variables (Theorem B.2.3). To obtain general gamma and beta variables, see Example 10.2.6. \square

Example 10.2.4 *Sampling from a distribution with finite support.* Suppose \mathcal{X} is finite $\{x_1, \dots, x_k\}$ and we want to simulate observations of an \mathcal{X} valued random variable X such that $P[X = x_j] \equiv p_j$, $\sum_{j=1}^k p_j = 1$. Even though the x_j need not be real, we can do this as follows. Form the partition of $[0, 1]$ given by $I_1 \equiv [0, p_1], I_2 \equiv [p_1, p_1 + p_2], \dots, I_k \equiv [p_1 + \dots + p_{k-1}, 1]$. Suppose that U has a $\mathcal{U}(0, 1)$ distribution. Then, it is clear that if we define

$$g(u) = x_j \quad \text{iff} \quad u \in I_j, \quad j = 1, \dots, k$$

then $g(U) \sim X$. Since any distribution may be approximated by finite discrete distributions, this seems easy but, in fact, for selected x_1, \dots, x_k , the $P[X = x_j]$ are usually not given analytically. This method is basic in the more difficult problem of generating a sample $\mathbf{S} \equiv \{x_{i_1}, \dots, x_{i_n}\}$ of size n from a finite population x_1, \dots, x_N with probabilities of inclusion π_1, \dots, π_N , that is, as in Example 3.4.1,

$$P[x_i \in \mathbf{S}] = \pi_i, i = 1, \dots, N. \quad \square$$

10.2.2 Importance Sampling

Suppose we wish to approximate $I(h) = \int h(\mathbf{x})p(\mathbf{x})d\mathbf{x}$ for a specific h , when we are given, for $X \in R^k$, an analytically presented density $p(\mathbf{x})$. Thus if

$$h(\mathbf{x}) = 1(x_1 \leq t_1, \dots, x_k \leq t_k),$$

$I(h)$ would be the df $F(t_1, \dots, t_k)$, and if $h(\mathbf{x}) = x_1^{r_1} \dots x_k^{r_k}$, $I(h)$ would give us moments. There is a way of generating an unbiased Monte Carlo approximation to $\int h(\mathbf{x})p(\mathbf{x})d\mathbf{x}$ without in fact generating a sample from p by using the method of *importance sampling*.

Theorem 10.2.1. *Let p_0 be any density from which $X \in R^k$ can be easily generated. Suppose that P is absolutely continuous with respect to P_0 , that is, $p_0(x) = 0 \implies p(x) = 0$. Then, if X_1, \dots, X_B are i.i.d. according to p_0 ,*

$$\hat{I} \equiv \frac{1}{B} \sum_{b=1}^B \frac{p(X_b)}{p_0(X_b)} h(X_b) \quad (10.2.1)$$

is an unbiased estimate of $I \equiv \int p(\mathbf{x})h(\mathbf{x})d\mathbf{x}$. Moreover, \hat{I} has variance

$$V \equiv \frac{1}{B} \left(\int \frac{p^2(\mathbf{x})}{p_0(\mathbf{x})} h^2(\mathbf{x})d\mathbf{x} - I^2 \right). \quad (10.2.2)$$

Proof. Unbiasedness follows from

$$E \frac{p(X_b)}{p_0(X_b)} h(X_b) = \int_{[p_0 > 0]} \frac{p(\mathbf{x})}{p_0(\mathbf{x})} h(\mathbf{x}) p_0(\mathbf{x}) d\mathbf{x}.$$

The variance argument is similar. \square

The idea here is that one finds a density p_0 which qualitatively behaves like p but from which simulation is easy. The extent to which the estimate is good is precisely gauged by the size of p/p_0 , since this ratio, a natural measure of qualitative similarity, governs the variance.

The following example is given to illustrate simply the qualitative features of importance sampling. It is evidently not a situation where importance sampling would be used.

Example 10.2.5 Gaussian distributions. Suppose p_0 is $\mathcal{N}(0, 1)$ and p is $\mathcal{N}(0, \tau^2)$. Qualitatively both are symmetric densities but as τ^2 moves from 1, p becomes more and more peaked or diffuse. Take $h(x) = x$ for simplicity. Then we get, for $B = 1$,

$$V = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} x^2 \lambda^2 \exp\left\{-\frac{x^2}{2}(2\lambda^2 - 1)\right\} dx$$

for $\lambda = \tau^{-1}$. Thus $V = \infty$ if $\lambda^2 \leq \frac{1}{2}$, and by recognizing V as a multiple of the variance of a Gaussian distribution, $V = \lambda^2(2\lambda^2 - 1)^{-3/2}$ if $\lambda > \frac{1}{2}$. So if $\lambda = 1$, $V = 1$ as expected, while if $\lambda \rightarrow \infty$, $V \sim 2^{-3/2}\lambda^{-1} \rightarrow 0$. This last result is, in retrospect, not surprising, since $\text{Var}_\lambda(X) \rightarrow 0$ as $\lambda \rightarrow \infty$.

Note that it is not true that the best choice of p_0 for all h is $p_0 = p$ but rather p_0 proportional to $|h|p$ (Problem 10.2.2). This incidentally leads to the important observation that importance sampling is directed at integrating a specific function with respect to p rather than the general purpose of “representing” P by the empirical distribution of B observations identically and independently distributed (approximately) as P . Some methods related to importance sampling such as antithetic variables will be considered in the problems. \square

10.2.3 Rejective Sampling

The rejective sampling method due to von Neumann (1951) enables us, knowing the density p up to a constant, say $p(x) = aq(x)$, to generate X with density p given that one can generate observations from a density p_0 with the property that

$$\sup_x \frac{p}{p_0}(x) = c < \infty.$$

Necessarily $c > 1$ unless $p \equiv p_0$. Let

$$\pi(x) \equiv c^{-1} \frac{p}{p_0}(x) = \frac{q(x)/p_0(x)}{\sup[q(x)/p_0(x)]}, \quad 0 \leq \pi(x) \leq 1. \quad (10.2.3)$$

The method is to generate i.i.d. X_1, X_2, \dots consecutively from p_0 . Then consecutively generate independent Bernoulli variables I_1, I_2, \dots such that

$$\text{Prob}[I_j = 1 \mid X_j] = \pi(X_j).$$

Let τ be the first j such that $I_j = 1$. “Reject” all $X_j, j < \tau$; “accept” X_τ as an observation from p . Moreover, let Pr denote the probability corresponding to this sampling scheme. Then

Theorem 10.2.2. Under the given conditions,

- (a) τ has a geometric marginal distribution with parameter $1/c$, that is

$$Pr[\tau = j] = \frac{1}{c} \left(1 - \frac{1}{c}\right)^{j-1}, \quad j \geq 1, \quad (10.2.4)$$

and hence $E(\tau) = c$.

(b) $\Pr[\tau < \infty] = 1$.

(c) X_τ has density p .

Proof. (a) and (b) hold because the I_j are independent and

$$\Pr[I_j = 1] = \frac{1}{c} \int \frac{p}{p_0}(x) p_0(x) dx = \frac{1}{c}.$$

To establish (c), note that

$$\begin{aligned} \Pr(X_\tau \in A) &= \sum_{j=1}^{\infty} \Pr(\tau = j) \Pr(X_j \in A | \tau = j) \\ &= \sum_{j=1}^{\infty} \Pr(\tau = j) P(X_j \in A) \\ &= \sum_{j=1}^{\infty} \left(1 - \frac{1}{c}\right)^{j-1} \frac{1}{c} \int_A \frac{p(x)}{p_0(x)} p_0(x) dx \\ &= \int_A p(x) dx. \end{aligned}$$

□

Remark 10.2.1. For rejective sampling, the loss one incurs by using p_0 rather than p is naturally measured by how long it takes to generate X_τ , for instance, by the expected time to acceptance. Using (10.2.4) we can write

$$E(\tau) = c(p, p_0) \equiv \sup_x \frac{p(x)}{p_0(x)}, \quad (10.2.5)$$

We illustrate the properties of this method:

Example 10.2.6. Gaussian, gamma, and beta distributions by rejective sampling. By arguing as in Example 10.2.1, if $U \sim \mathcal{U}(0, 1)$ then $V \equiv \text{sgn}(U)(-\log|U|)$ has a Laplace (double exponential) density, $p_0(x) = \frac{1}{2}e^{-|x|}$. Suppose we want to generate $X \sim \mathcal{N}(0, 1)$ by rejective sampling from the Laplace distribution. Then,

$$c(p, p_0) = \sup_x \frac{p}{p_0}(x) = \sup_x \sqrt{\frac{2}{\pi}} \exp\left\{-\frac{x^2}{2} + |x|\right\} = \sqrt{\frac{2e}{\pi}} = 1.32.$$

Thus one obtains a Gaussian variable typically after 2 or 3 steps. This is clearly inefficient compared to the Box–Muller–Knuth method.

As a second example suppose we wish to generate $\text{beta}(r, s)$ variates. Suppose $r, s > 1$. If we take $p_0 \equiv 1$, on $(0, 1)$ we see that

$$c(p, p_0) = \frac{\Gamma(r+s)}{\Gamma(r)\Gamma(s)} \left(\frac{r}{r+s}\right)^{r-1} \left(\frac{s}{r+s}\right)^{s-1}.$$

It may be shown that (Problem 10.2.3), as $r, s \rightarrow \infty$,

$$\frac{r}{r+s} \rightarrow t, \quad 0 < t < 1, \quad c(p, p_0) \sim \sqrt{\frac{r+s}{2\pi t(1-t)}}. \quad (10.2.6)$$

Thus using the uniform distribution as p_0 is a poor method for r and s large and doesn't work for $r, s < 1$. This is not surprising since $p(0) = p(1) = 0$. On the other hand, using rejective sampling and p_0 given by $\text{beta}([2r]/2, [2s]/2)$ is reasonable (Problem 10.2.4).

An alternative is to first generate independent $\Gamma(r, 1)$ and $\Gamma(s, 1)$ variables and use the method of Example 10.2.3. For the gamma distribution $\Gamma(r, 1)$, $0 < r < 1$, we can take

$$p_0(x) = \frac{1}{2}(rx^{r-1}1(0 < x < 1) + e^{-x}), \quad x > 0 \quad (10.2.7)$$

from which it is easy to simulate (Problem 10.2.5) and then note that by Theorem B.2.3

$$\sup_x \frac{p}{p_0}(x) \leq \frac{2}{r\Gamma(r)} \leq 2.$$

□

This method is dependent on being able to find a p_0 from which it is easy to generate and for which $c(p, p_0)$ is close to one. This issue becomes acute when we need to generate a large number of i.i.d. variables having the desired density p , as we do, for instance, in Bayesian inference. The most general way of attacking this problem is through the Markov Chain Monte Carlo (MCMC) methods discussed in Section 10.4. In the next section we study a statistically very important example of the use of Monte Carlo methods where generation of the variables is simple.

Summary. We considered first the *simple Monte Carlo* method where desired samples from a given density p_0 can be obtained by using transformations of basic variables. Thus normal variables can be obtained as transformations of uniform variables, and gamma, beta, t , F , and multivariate normal variables can be obtained from independent normal variables using Sections B.3, B.4, and B.6. Next we considered *importance sampling* where a sample from P_0 can be used to estimate integrals of the form $E_{P_0}h(\mathbf{X})$ when P is absolutely continuous with respect to P_0 . Finally, we considered *rejective sampling* where variables with density p_0 can be used to generate a variable with density $p \neq p_0$ provided $c = \sup_x [p(x)/p_0(x)] < \infty$. Rejective sampling is useful when p is of the form $p(x) = aq(x)$ with q known and a unknown; and when c is close to one.

10.3 The Bootstrap

The “bootstrap,” formally introduced by Efron (1979) (see also Efron and Tibshirani (1993) and Shao and Tu (1995)), is the first broad application of Monte Carlo methods to inference. The three classical problems to which he applied the method remain excellent illustrations of the idea. The context throughout this section is simple random sampling, X_1, \dots, X_n i.i.d. as $X \sim P$. To avoid confusion, in this section, we shall write \hat{P}_n for the empirical probability \hat{P} to indicate its dependence on n .

10.3.1 Bootstrap Samples and Bias Corrections

We are given a parameter $\mu(P)$ defined for all $P \in \mathcal{P}$ and its plug-in estimate $\hat{\mu}_n \equiv \mu(\hat{P}_n)$, where we assume that \mathcal{P} is large enough so that all P with finite support such as \hat{P}_n are included. We wish to estimate the bias of $\hat{\mu}_n$,

$$\text{BIAS}_n(P) \equiv E_P \hat{\mu}_n - \mu(P).$$

If we assume for simplicity that $\mu(P)$ is bounded, $\text{BIAS}_n(P)$ is itself a parameter defined for all $P \in \mathcal{P}$. However, it has three special features. In general,

- a) It depends on n .
- b) It is computable from knowledge of P only through a multiple integral. If we write $\hat{\mu}_n$ as $\mu_n(X_1, \dots, X_n)$ for some known function μ_n , then

$$E_P(\hat{\mu}_n) = \int \cdots \int \mu_n(x_1, \dots, x_n) dP(x_1) \dots dP(x_n).$$

- c) We expect $\text{BIAS}_n(P)$ to tend to 0 as $n \rightarrow \infty$, but we expect $n\text{BIAS}_n(P)$ to tend to some limit $B(P) \neq 0$.

We have encountered this type of estimation problem for the special case

$$\mu(P) = h \left(\int g_1(x) dP(x), \dots, \int g_d(x) dP(x) \right) \quad (10.3.1)$$

in Theorem 5.3.2 and, in the general case, in Chapter 7. We applied the delta method for Euclidean and infinite dimensional parameters, respectively, to obtain approximations to $\text{BIAS}_n(P)$ of the form $B(P)/n$. We can then reasonably estimate $\text{BIAS}_n(P)$ by $B(\hat{P}_n)/n$. The difficulty with this program as we have discussed earlier is that even for $d = 1$ and certainly for $d > 1$, the approximation $B(P)/n$ is both tedious to compute analytically and may give little insight.

The alternative proposed by Efron is to estimate the parameter $\text{BIAS}_n(P)$ by the empirical plug-in method to obtain $\text{BIAS}_n(\hat{P}_n)$. Save for the dependence of the parameter on n , this is the plug-in idea discussed in Chapter 2. Unfortunately, computing the first term in

$$\begin{aligned} \text{BIAS}_n(\hat{P}_n) &= \int \mu_n(x_1, \dots, x_n) d\hat{P}_n(x_1) \dots d\hat{P}_n(x_n) - \mu(\hat{P}_n) \\ &= \frac{1}{n^n} \sum_{1 \leq i_1, \dots, i_n \leq n} \mu_n(X_{i_1}, \dots, X_{i_n}) - \mu(\hat{P}_n) \end{aligned}$$

is, in general, unfeasible practically. However, we can apply Monte Carlo “resampling” as follows.

Bootstrap Estimation of Bias

a) *Monte Carlo Step.* Generate B i.i.d. samples of size n from \hat{P}_n ; call them

$$\mathbf{X}_b^* = (X_{b1}^*, \dots, X_{bn}^*), \quad b = 1, \dots, B.$$

That is, given $X_1, \dots, X_n, X_{b1}^*, \dots, X_{bn}^*$ are i.i.d. as X^* with distribution $P^* = \hat{P}_n$,

$$P^*[X^* = X_j] = \frac{1}{n}, \quad j = 1, \dots, n.$$

Here, the superscript * is used to denote conditioning on the data $\mathbf{X} = (X_1, \dots, X_n)$. More informally, for fixed b , each X_{bj}^* is obtained by sampling with replacement from the population $\{X_1, \dots, X_n\}$. Let \hat{P}_{nb}^* denote the empirical distribution of the b th *bootstrap sample* $\mathbf{X}_b^* = (X_{b1}^*, \dots, X_{bn}^*)$.

b) *Estimation Step.* Estimate $\text{BIAS}_n(\hat{P}_n)$ by

$$\begin{aligned} \text{BIAS}_n^{(B)}(\hat{P}_n) &\equiv \frac{1}{B} \sum_{b=1}^B \mu(\hat{P}_{nb}^*) - \mu(\hat{P}_n) \\ &= \frac{1}{B} \sum_{b=1}^B \mu_n(X_{b1}^*, \dots, X_{bn}^*) - \mu(\hat{P}_n). \end{aligned}$$

Remark 10.3.1

a) Note that B is at our disposal and may be chosen as large as we wish. In particular, formally, $B = \infty$ yields $\text{BIAS}_n(\hat{P}_n)$ by the law of large numbers applied to the i.i.d. variables $\mu(\hat{P}_{nb}^*), \quad b = 1, \dots, B$.

b) This procedure is nothing else than Monte Carlo approximation of the integral of $\mu_n(x_1, \dots, x_n)$ with respect to the product measure $\hat{P}_n \times \dots \times \hat{P}_n$.

c) Even if $B = \infty$, we have no guarantee that the plug-in estimate $\text{BIAS}_n(\hat{P}_n)$ is good. \square

Given a biased estimate $\hat{\theta}_n = \theta(\hat{P}_n)$ of a parameter $\theta(P)$, we may wish to try to eliminate or reduce its bias (see Section 3.4.1). A natural approach is to use the *empirically bias corrected* estimate

$$\tilde{\theta}_n^* \equiv \hat{\theta}_n - \text{BIAS}_n(\hat{P}_n)$$

or the *bootstrap bias corrected* estimate

$$\hat{\theta}_{nB} = \hat{\theta}_n - \text{BIAS}_n^{(B)}(\hat{P}_n).$$

We illustrate these issues with an example, in which exact calculation is clearly superior in one special case but, in another special case of much more common type, the bootstrap is much simpler to apply.

Example 10.3.1. *Estimating the bias of sample moments.* To avoid technicalities, suppose $P \in \mathcal{M} = \{\text{All probabilities on } [0, 1] \text{ with 4 finite moments}\}$. Suppose $\mu(P) = \int x \, dP$, the population mean. Then, $\hat{\mu}_n = \mu(\hat{P}_n) = \bar{X}$. Now

$$E_P \hat{\mu}_n = E_P \bar{X} = E_P X = \mu(P).$$

Therefore, $\text{BIAS}_n(P) = 0 = \text{BIAS}_n(\hat{P}_n)$, but $\text{BIAS}_n^{(B)}(\hat{P}_n) = B^{-1} \sum_{b=1}^B \bar{X}_b^* - \bar{X}$. Then,

$$E^* \text{BIAS}_n^{(B)}(\hat{P}_n) = 0.$$

$$\text{Var}^* \text{BIAS}_n^{(B)}(\hat{P}_n) = (nB)^{-1} n^{-1} \sum_{i=1}^n (X_i - \bar{X})^2 = O_P((nB)^{-1})$$

where E^* and Var^* denote expected value and variance for the conditional probability $P^* = \mathcal{L}(X^* | \mathbf{X})$ defined in the Monte Carlo step. Note that P^* , E^* , and Var^* integrate out the Monte Carlo randomness that has been introduced.

Turn now to estimating the bias of $\sigma^2(\hat{P}_n)$, where $\sigma^2(P) \equiv \text{Var}_P(X)$. Here,

$$\sigma^2(\hat{P}_n) = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2.$$

From (3.4.2),

$$\begin{aligned} E_P \sigma^2(\hat{P}_n) &= \frac{n-1}{n} \sigma^2(P) \\ \text{BIAS}_n(P) &= -\frac{\sigma^2(P)}{n}. \end{aligned}$$

Now,

$$\text{BIAS}_n(\hat{P}_n) = -\frac{1}{n^2} \sum_{i=1}^n (X_i - \bar{X})^2.$$

Thus, the empirically bias corrected estimate is

$$\sigma^2(\hat{P}_n) - \text{BIAS}_n(\hat{P}_n) = \frac{n+1}{n^2} \sum_{i=1}^n (X_i - \bar{X})^2, \quad (10.3.2)$$

whose bias is $-\sigma^2(P)/n^2$. The bootstrap bias corrected estimate

$$\hat{\sigma}_{nB}^2(\hat{P}_n) = \sigma^2(\hat{P}_n) - \text{BIAS}_n^{(B)}(\hat{P}_n) = \frac{n-1}{n^2} \sum_{i=1}^n (X_i - \bar{X})^2$$

agrees to order n^{-1} with the usual unbiased estimate $(n-1)^{-1} \sum_{i=1}^n (X_i - \bar{X})^2$ (Problem 10.3.3).

Going further, let

$$K_3(P) = E_P(X - \mu(P))^3$$

be estimated by $K_3(\hat{P}_n) = n^{-1} \sum_{i=1}^n (X_i - \bar{X})^3$. Here

$$E_P K_3(\hat{P}_n) = K_3(P)(1 - \frac{1}{n^2}) .$$

Hence, $\text{BIAS}_n(\hat{P}_n) = -n^{-2}K_3(P)$ and again $K_3(\hat{P}_n) - \text{BIAS}_n(\hat{P}_n)$ is unbiased. However, consider the numerator of the kurtosis (see A.11.10),

$$K_4(P) = E_P X^4 - 3(E_P X^2)^2 \quad (10.3.3)$$

Now (see Problem 10.3.4)

$$E_P K_4(\hat{P}_n) = K_4(P) + \left(\frac{a_1}{n} + \frac{a_2}{n^2} + \frac{a_3}{n^3} \right) K_4(P) + \left(\frac{b_1}{n} + \frac{b_2}{n^2} + \frac{b_3}{n^3} \right) \sigma^4(P) \quad (10.3.4)$$

for appropriate constants $a_j, b_j; j = 1, 2, 3$. Here, $K_4(\hat{P}_n) - \text{BIAS}_n(\hat{P}_n)$ is biased. However, while

$$E K_4(\hat{P}_n) = K_4(P) + O\left(\frac{1}{n}\right) ,$$

$$E K_4(\hat{P}_n) - \text{BIAS}_n(\hat{P}_n) = K_4(P) + O_P\left(\frac{1}{n^2}\right) . \quad (10.3.5)$$

Note that the example illustrates that using $\text{BIAS}_n^{(B)}$ is easier than going through the exact calculation of $\text{BIAS}_n(\hat{P}_n)$. Moreover, computing $B(P)/n$ by the delta method for this last example is almost as tedious as computing $E_P K_4(\hat{P}_n)$, and using the approximation $\text{BIAS}_n^{(B)}(\hat{P}_n)$ is easier. \square

A natural question to ask is how large does B have to be so that (10.3.5) holds when $\text{BIAS}_n(\hat{P}_n)$ is replaced by $\text{BIAS}_n^{(B)}(\hat{P}_n)$ or, more generally, when is

$$E\theta(\hat{P}_n) - \text{BIAS}_n^{(B)}(\hat{P}_n) = \theta(P) + O_P\left(\frac{1}{n^2}\right)?$$

Calculations that answer these questions are in Hall (1986). When $\theta(P)$ is a smooth function of means as in Theorem 5.3.2,

$$E^* \text{BIAS}_n^{(B)}(\hat{P}_n) = \text{BIAS}_n(\hat{P}_n) \quad (10.3.6)$$

$$\text{Var}^* \text{BIAS}_n^{(B)}(\hat{P}_n) = O_P(n^{-1}B^{-1}) . \quad (10.3.7)$$

Here we again use Efron's convenient * notation for the conditional distribution $\mathcal{L}(X^*|\mathbf{X})$ of observations resampled from the data $\mathbf{X} = (X_1, \dots, X_n)$. Recall that P^*, E^* , etc. integrate out the Monte Carlo randomness we have introduced.

Since, under the conditions of Corollary 5.3.1,

$$\text{BIAS}_n(\hat{P}_n) = \text{BIAS}_n(P) + O_P(n^{-\frac{3}{2}})$$

Hall's and our conclusion is that for $\text{BIAS}_n^{(B)}(\hat{P}_n)$ to be asymptotically equivalent to $\text{BIAS}_n(P)$ to order $O_P(n^{-\frac{3}{2}})$, we need to take $B \asymp n^2$, where $a_n \asymp b_n$ as usual denotes $a_n = O(b_n)$ and $b_n = O(a_n)$. The proofs of these results are sketched in Problem 10.3.1.

In bias correction the bootstrap makes an asymptotically appropriate higher order correction. Of greater importance is the use of the bootstrap in estimation of variances of complex estimates and in setting confidence bands and regions, which we turn to in the next section.

10.3.2 Bootstrap Variance and Confidence Bounds

The standard formula for an upper asymptotic level $1 - \alpha$ upper confidence bound (UCB) for a parameter θ on the basis of the MLE $\hat{\theta}_n$ in a regular one dimensional parametric model with Fisher information $I(\theta)$ is given in Section 5.4.5 and is of the form

$$\hat{\theta}_n + \frac{z_{1-\alpha}}{\sqrt{nI(\hat{\theta}_n)}} .$$

More generally, if we have a regular, as defined in Section 9.3, asymptotically normal estimate $\hat{\mu}_n$ of a parameter $\mu(P)$ such that

$$\mathcal{L}_p(\sqrt{n}(\hat{\mu}_n - \mu(P))) \rightarrow \mathcal{N}(0, \sigma^2(P))$$

for all $P \in \mathcal{P}$ and $\hat{\sigma}_n$ is a (locally uniformly) consistent estimate of $\sigma(P)$ in \mathcal{P} , then $\hat{\mu}_n + z_{1-\alpha}\hat{\sigma}_n/\sqrt{n}$ is a natural asymptotic level $(1 - \alpha)$ UCB. Thus, for instance, if

$$\mathcal{P} = \{P_\theta : \theta \in R^d\}$$

and the MLE $\hat{\boldsymbol{\theta}}_n = (\hat{\theta}_{n1}, \dots, \hat{\theta}_{nd})^T$ behaves regularly, then, letting $\|I^{ij}\|$ denote the inverse of the information matrix I of Section 6.2.2,

$$\hat{\theta}_{n1} + z_{1-\alpha}\sqrt{\frac{I^{11}(\hat{\boldsymbol{\theta}}_n)}{n}} \tag{10.3.8}$$

is an asymptotic level $1 - \alpha$ UCB in \mathcal{P} . As we have noted in the past, construction of $\hat{\sigma}_n$ and computation of $I^{11}(\hat{\boldsymbol{\theta}}_n)$ is usually far from trivial. The bootstrap gives what appears to be an “automatic” solution.

Suppose that the standard deviation of $\hat{\mu}_n$, $SD_n(P)$, exists and normalizes $\hat{\mu}_n$ properly,

$$\mathcal{L}_p\left(\frac{\hat{\mu}_n - \mu(P)}{SD_n(P)}\right) \rightarrow \mathcal{N}(0, 1) \tag{10.3.9}$$

and

$$\sqrt{n}SD_n(P) \rightarrow \sigma(P) . \tag{10.3.10}$$

Then it is natural to try $SD_n(\hat{P}_n)$ as a consistent estimate of $SD_n(P)$, in the expectation that

$$SD_n(\hat{P}_n) - SD_n(P) = O_P(n^{-1}). \quad (10.3.11)$$

The resulting proposed confidence bound would, by the usual inversion argument, be

$$\hat{\mu}_n + z_{1-\alpha} SD_n(\hat{P}_n).$$

Calculation of SD_n is again a problem because

$$\begin{aligned} SD_n^2(P) &= \int \dots \int \hat{\mu}_n^2(x_1, \dots, x_n) dP(x_1) \dots, dP(x_n) \\ &- \left(\int \dots \int \hat{\mu}_n(x_1, \dots, x_n) dP(x_1) \dots, dP(x_n) \right)^2. \end{aligned}$$

However, getting a bootstrap approximation is simple.

Bootstrap Estimation of Variance.

- (i) Let $\{X_{bj}^* : 1 \leq b \leq B, 1 \leq j \leq n\}$ be bootstrap samples generated as in the Monte Carlo step of the estimation of $\text{BIAS}_n(P)$ and compute $\hat{\mu}_n(X_{b1}^*, \dots, X_{bn}^*) \equiv \hat{\mu}_{nb}^*$ and $\hat{\mu}_{n \cdot}^* = B^{-1} \sum_{b=1}^B \hat{\mu}_{nb}^*$.
- (ii) Estimate $SD_n(P)$ by

$$\widehat{SD}_n^{(B)} \equiv \left(\frac{1}{B} \sum_{b=1}^B (\hat{\mu}_{nb}^* - \hat{\mu}_{n \cdot}^*)^2 \right)^{\frac{1}{2}}.$$

Again, if $B = \infty$, $\widehat{SD}_n^{(B)} = SD_n(\hat{P}_n)$.

Example 10.3.2. *Estimation of the variance of the median and trimmed means.* Let $M_n \equiv \text{Median}(X_1, \dots, X_n)$. We have seen that (Problem 5.4.1), if X has density $f = F'$ and if $\theta(F) \equiv F^{-1}(\frac{1}{2})$, $f(\theta) > 0$, then

$$\mathcal{L}_F \left(\sqrt{n}(M_n - F^{-1}(\frac{1}{2})) \right) \rightarrow \mathcal{N} \left(0, \frac{1}{4f^2(\theta)} \right).$$

Consistent estimation of f , and thus of $\sigma(P) = 1/2f(\theta)$, is possible in this case, see Chapter 11. So also is estimation of $\sigma(P)$ by plugging \hat{P}_n into the formulae (5.1.2) and (5.1.3). Moreover, application of the bootstrap is easy. Similarly for the trimmed means \bar{X}_α given by (3.5.3), we have in Example 7.2.6 the result

$$\sqrt{n}(\bar{X}_\alpha - \mu_\alpha(P)) \rightarrow \mathcal{N}(0, \sigma_\alpha^2(P))$$

where $\bar{X}_\alpha = \mu_\alpha(\hat{P}_n)$,

$$\mu_\alpha(P) = \frac{1}{1-2\alpha} \int_{x_\alpha}^{x_{1-\alpha}} x dP(x), \quad \sigma_\alpha^2(P) = \text{Var}_P \psi_\alpha(X)$$

with ψ_α given by (7.2.42). Again $\sigma_\alpha^2(\hat{P}_n)$ is easily computed. However, while there is no particularly attractive formula to plug into for $SD_n^2(\hat{P}_n)$, the bootstrap is easy to apply in the estimation of variance step above. \square

There are a number of questions we need to ask:

- (i) For estimation of variances per se, when does plugging in \hat{P}_n yield a consistent estimate?
- (ii) When is $SD_n(P)$ an appropriate normalization to yield asymptotic normality for $(\hat{\mu}_n - \mu(P))$?
- (iii) How big does B need to be for $\widehat{SD}_n^{(B)}$ to approximate $SD_n(P)$ to the appropriate order $O_P(n^{-\frac{1}{2}})$?
- (iv) What is gained by using the bootstrap in cases where an alternative approximation exists?

It turns out that for the median and trimmed means, the answers to (i) and (ii) are affirmative under the natural conditions considered in Example 7.2.6. We summarize the affirmative answers to the first three questions for the simple case where $\hat{\mu}_n$ is a smooth function of a vector mean. Its proof is left to the Problem 10.3.1.

Theorem 10.3.1. *Under the conditions of Corollary 5.3.2 (a) and (b),*

$$\widehat{SD}_n = SD_n(\hat{P}_n) + O_P(n^{-\frac{1}{2}} B^{-\frac{1}{2}}).$$

Since, under the same conditions,

$$SD_n(\hat{P}_n) = SD_n(P) + O_P(n^{-1}).$$

$B \asymp n$ will give total errors of the same order as the error in $SD_n(\hat{P}_n)$.

The answer to question (iv), bias and variance estimation is unclear in terms of statistical performance. However, the computational simplification can be striking.

The Jackknife

The jackknife due to Quenouille (1949) and Tukey (1958) preceded the bootstrap as a “sampling” approach to estimating biases and variances of complex statistics. It is closely related to the sensitivity curve discussed in Section 3.5.3 (Volume I).

Given an i.i.d. sample $\mathbf{X}^{(n)} \equiv \{X_1, \dots, X_n\}$ with $X \sim F$, a parameter $\theta(F)$ and a sequence of estimates $\theta^{(m)}(X_1, \dots, X_m)$, $1 \leq m \leq n$, of $\theta(F)$, we want to estimate the bias

$$B_n(F) \equiv E_F \widehat{\theta}^{(n)} - \theta(F)$$

and the variance

$$\sigma_n^2(F) \equiv \text{Var}_F \widehat{\theta}^{(n)} .$$

The idea is to use the estimates $\widehat{\theta}^{(n-1)}(\mathbf{X}_{-i}) \equiv \widehat{\theta}_{(i)}$ obtained by evaluating $\widehat{\theta}^{(n-1)}$ at the subsamples

$$\mathbf{X}_{-i} \equiv \{X_1, \dots, X_{i-1}, X_{i+1}, \dots, X_n\}$$

of size $n - 1$ obtained by deleting one X_i at the time, $1 \leq i \leq n$. The estimates of bias and variance are

$$\begin{aligned}\widehat{B}_n &= (n-1)(\widehat{\theta}_{(.)} - \widehat{\theta}), \quad \widehat{\theta}_{(.)} = n^{-1} \sum_{i=1}^n \widehat{\theta}_{(i)}, \\ \widehat{\sigma}_n^2 &= \sum_{i=1}^n (\widehat{\theta}_{(i)} - \widehat{\theta}_{(.)})^2 .\end{aligned}$$

The relation of the jackknife idea to the sensitivity curve defined in Section 3.5 and Problem 7.2.2 can be seen from

$$SC(\mathbf{X}^{(n)}, \widehat{\theta}^{(n)}) = n(\widehat{\theta}^{(n)} - \widehat{\theta}^{(n-1)}) .$$

Since if $\theta(F)$ is smooth in F , the sensitivity curve is a crude approximation to n^{-1} times the influence function (Problem 7.2.2). Thus we can write

$$\widehat{\sigma}_n^2 \approx n^{-2} \sum_{i=1}^n (\psi(X_i, F) - \bar{\psi})^2 \approx n^{-1} \int \psi^2(x, F) dF(x) \quad (10.3.12)$$

where ψ is the influence function of $\widehat{\theta}^{(n)}$ as defined in Section 7.2 and

$$\bar{\psi} = n^{-1} \sum_{i=1}^n \psi(X_i, F) \approx 0 .$$

This approximation to $\widehat{\sigma}_n^2$ is exactly correct if $\theta(F)$ is linear in F so that $\widehat{\theta}^{(n)}$ is an average and, as we have seen in Problem 7.2.2, approximately correct generally. \widehat{B}_n can also be justified using higher order asymptotics. For more on the jackknife, see Efron and Tibshirani (1993), Chapter 11.

10.3.3 The General i.i.d. Nonparametric Bootstrap

Efron (1979)⁽¹⁾ proposed a general bootstrap principle along the following lines: Let \mathcal{P} denote a nonparametric class of probabilities on a set \mathcal{X} . We assume that \mathcal{P} contains all discrete distributions on \mathcal{X} . Let $\mathbf{X} = (X_1, \dots, X_n)$ denote i.i.d. observations from $P \in \mathcal{P}$. We are interested in a functional $T_n(\mathbf{X}, P)$ which is invariant under permutations $\pi(\mathbf{X})$ of the elements of \mathbf{X} , that is $T_n(\pi(\mathbf{X}), P) = T_n(\mathbf{X}, P)$ for each $\pi(\mathbf{X}) = (X_{i_1}, \dots, X_{i_n})$ with $i_j \neq i_k$ for $j \neq k$. By sufficiency of the empirical distribution, this restriction has

no force as long as we are only considering i.i.d. observations. In this case we also write $T_n(\widehat{P}_n, P)$ for $T_n(\mathbf{X}, P)$, where we have identified \widehat{P}_n with \mathbf{X} . The parameter $\lambda_n(P)$ we are interested in is calculable from the distribution of $T_n(\widehat{P}_n, P)$. That is, for some map θ from $\{\mathcal{L}_P(T_n(\widehat{P}_n, P)) : P \in \mathcal{P}\}$ to some set Θ , we can write

$$\lambda_n(P) = \theta(\mathcal{L}_P(T_n(\widehat{P}_n, P))). \quad (10.3.13)$$

Then, the plug-in principle leads to the estimate

$$\lambda_n(\widehat{P}_n) = \theta(\mathcal{L}^*(T_n(\widehat{P}_n^*, \widehat{P}_n)))$$

of $\lambda_n(P)$, where

$$T_n(\widehat{P}_n^*, \widehat{P}_n) = T_n(X_1^*, \dots, X_n^*, \widehat{P}_n)$$

and $*$ as usual denotes that we are dealing with an i.i.d. sample X_1^*, \dots, X_n^* from \widehat{P}_n . We will use bootstrap samples as follows to approximate \mathcal{L}^* and $\lambda_n(P)$:

a) *Approximating $\mathcal{L}^*(T_n(\widehat{P}_n^*, \widehat{P}_n))$.* Generate bootstrap samples $\mathbf{X}_b^* = (X_{b1}^*, \dots, X_{bn}^*)$, $b = 1, \dots, B$, as in the Monte Carlo Step of the bootstrap bias and variance approximation. Then compute $\{T_{nb}^*, 1 \leq b \leq B\}$ with $T_{nb}^* = T_n(X_{b1}^*, \dots, X_{bn}^*, \widehat{P}_n)$. Let δ_x be pointmass at x . Approximate $\mathcal{L}^*(T_n(\widehat{P}_n^*, \widehat{P}_n))$ by $\mathcal{L}_B^* \equiv B^{-1} \sum_{b=1}^B \delta_{T_{nb}^*}$, the empirical distribution of $\{T_{nb}^*, 1 \leq b \leq B\}$.

b) *Bootstrap Estimation of $\lambda_n(P)$.* The estimate $\lambda_n^{(B)}(\widehat{P}_n)$ of $\lambda_n(P)$ is given by $\theta(\mathcal{L}_B^*)$.

It is easy to see that estimation of the bias $\text{BIAS}_n(P)$ corresponds to taking

$$\begin{aligned} T_n(\widehat{P}_n, P) &= \widehat{\mu}_n - \mu(P), \\ \theta(F_T) &= \int z dF_T(z), \end{aligned}$$

where F_T is the df of $T_n(\widehat{P}_n, P)$. This gives

$$\theta(\mathcal{L}_P(T_n(\widehat{P}_n, P))) = E_P(\widehat{\mu}_n) - \mu(P).$$

With the same choice of $T_n(\widehat{P}_n, P)$, estimation of $SD_n(P)$ corresponds to

$$\theta(F_T) = \left(\int z^2 dF_T(z) - \left(\int z dF_T(z) \right)^2 \right)^{\frac{1}{2}}.$$

This formulation suggests an entirely different approach to setting confidence bounds and confidence regions.

Consider again setting a confidence bound on $\mu(P)$ using $\widehat{\mu}_n$. In Examples 4.4.1 and 4.4.2, to obtain confidence bounds for the mean and variance of a normal distribution, we used the existence of pivots, functions $T(\widehat{\mu}_n, \mu(P))$ whose distribution did not depend on P , with appropriate monotonicity properties. In fact, our general construction of asymptotic UCBs for $\mu(P)$ is based on an asymptotic pivot of the form $\sqrt{n}(\widehat{\mu}_n - \mu(P))/\widehat{\sigma}_n$ which tends to $\mathcal{N}(0, 1)$ for all $P \in \mathcal{P}$.

Consider $\hat{\mu}_n - \mu(P)$ as a potential pivot. Essentially only if $\mu(P)$ is a translation parameter of a translation parameter family $\{F(x - \mu) : \mu \in R\}$ is $\hat{\mu}_n - \mu(P)$ in fact a pivot. However, note that if $c_{n\alpha}(P)$ is the unknown α quantile of the distribution of $\hat{\mu}_n - \mu(P)$, then

$$P[\hat{\mu}_n - \mu(P) \geq c_{n\alpha}(P)] = P[\mu(P) \leq \hat{\mu}_n - c_{n\alpha}(P)] = 1 - \alpha$$

Thus estimating the parameter $c_{n\alpha}(P)$ to appropriate accuracy by \hat{c}_n should yield $\hat{\mu}_n - \hat{c}_n$ as an asymptotic $(1 - \alpha)$ UCB. The bootstrap yields an estimate $\hat{c}_n^{(B)}$ as follows: Let $T_n(\hat{P}_n, P) = \hat{\mu}_n - \mu(P)$, and let $\lambda_n(P)$ be the α quantile of $\mathcal{L}(\hat{\mu}_n - \mu(P))$. The concrete application of the bootstrap is, having generated $T_{nb}^* = \hat{\mu}_n(\mathbf{X}_b^*) - \mu(\hat{P}_n)$, $1 \leq b \leq B$, then the bootstrap approximation to $c_{n\alpha}(P)$ is the α th quantile of the empirical distribution of $\{T_{nb}^* ; 1 \leq b \leq B\}$, that is

$$\hat{c}_n^{(B)} = T_{n([B\alpha])}^*$$

where $T_{n(1)}^* \leq \dots \leq T_{n(B)}^*$ are the ordered $\{T_{nb}^*\}$ and $[]$ denotes the greatest integer function. If we specialize to $\hat{\mu}_n = \mu_n(\hat{P}_n)$, then

$$\hat{c}_n^{(B)} = \hat{\mu}_{n([B\alpha])}^* - \hat{\mu}_n$$

where $\hat{\mu}_{n(1)}^* \leq \dots \leq \hat{\mu}_{n(B)}^*$ are the ordered $\{\hat{\mu}_{nb}^*\}$. Thus, $\hat{\mu}_{n([B\alpha])}^*$ is the Monte Carlo estimate of the α quantile of the bootstrap distribution of $\hat{\mu}_n$. We thus obtain as a potential asymptotic $(1 - \alpha)$ UCB,

$$\hat{\mu}_n - (\hat{\mu}_{n([B\alpha])}^* - \hat{\mu}_n) = 2\hat{\mu}_n - \hat{\mu}_{n([B\alpha])}^*. \quad (10.3.14)$$

This is *not Efron's* so called *percentile method*.

Efron proposed the UCB

$$\hat{\mu}_{n([B(1-\alpha)+1])}^*.$$

It may be shown (Problem 10.3.6) that if $\sqrt{n}(\hat{\mu}_n - \mu(P))$ tends to a Gaussian distribution and bootstrap quantiles converge to population quantiles at the $n^{-\frac{1}{2}}$ rate, then Efron's method with $B = \infty$ yields an asymptotic $(1 - \alpha)$ UCB. Efron's motivation for his proposal here and for subsequent refinements is that $\hat{\mu}_{n([B(1-\alpha)+1])}^*$ is equivariant under monotone transformations. That is, if we use it as an asymptotic level $(1 - \alpha)$ UCB for $\mu(P)$ then $g(\hat{\mu}_{n([B(1-\alpha)+1])}^*)$ is the corresponding asymptotic level $(1 - \alpha)$ UCB for $g(\mu(P))$ if g is increasing. For more on these topics, we refer to Efron and Tibshirani (1993) and Hall (1997).

There is nothing special about the proposed pivot $\hat{\mu}_n - \mu(P)$. If we have an estimate of the asymptotic variance of $\hat{\mu}_n$, call it $\hat{\sigma}_n^2$, and

$$\mathcal{L}\left(\sqrt{n}\frac{(\hat{\mu}_n - \mu(P))}{\hat{\sigma}_n}\right) \rightarrow \mathcal{N}(0, 1),$$

then it is natural to apply the above process to

$$T_n(\hat{P}_n, P) = \sqrt{n}\frac{(\hat{\mu}_n - \mu(P))}{\hat{\sigma}_n}$$

to obtain a bootstrap estimate, say $\widehat{d}_{n\alpha}^{(B)}$, using $T_{nb}^* = (\widehat{\mu}_{nb}^* - \widehat{\mu}_n)/\widehat{\sigma}_{nb}^*$, where $\widehat{\sigma}_{nb}^*$ is $\widehat{\sigma}_n(\mathbf{X}_b^*)$, and invert to get $\widehat{\mu}_n - \widehat{\sigma}_n \widehat{d}_{n\alpha}^{(B)}$ as an asymptotic UCB. We illustrate this process concretely in a classical example.

Example 10.3.3. Bootstrap t. Suppose, $\mathcal{P} = \{P : 0 < E_P X^2 < \infty\}$ and we want to put confidence bounds on $\mu(P) = E_P X$. If we restrict P to Gaussian distributions, we are naturally (see Example 4.4.1 and Section 4.9.2) led to the t statistic defined, for s given in (10.3.2), by

$$T_n = \sqrt{n} \frac{(\bar{X} - \mu)}{s}$$

and the level $(1 - \alpha)$ UCB of Example 4.4.1, $\bar{X} + t_{n-1}(1 - \alpha)s/\sqrt{n}$, where $t_{n-1}(1 - \alpha)$ is the $1 - \alpha$ quantile of the \mathcal{T}_{n-1} distribution. If P is not Gaussian but is unknown, the distribution of T_n is also unknown and it is natural to use the bootstrap. Form the bootstrap samples $(X_{b1}^*, \dots, X_{bn}^*)$, the corresponding \bar{X}_b^* and s_b^* , $T_{nb}^* = \sqrt{n}(\bar{X}_b^* - \bar{X})/s_b^*$, $b = 1, \dots, B$, and the lower α quantile of the bootstrap distribution, $T_{n([B\alpha])}^* \equiv \widehat{d}_{n\alpha}^{(B)}$. The level $(1 - \alpha)$ bootstrap UCB is

$$\bar{X} - \widehat{d}_{n\alpha}^{(B)} \frac{s}{\sqrt{n}} .$$

Applying the same principle to the pivot $\sqrt{n}(\bar{X} - \mu)$, we arrive at the UCB,

$$2\bar{X} - \bar{X}_{([B\alpha])}^* ,$$

where $\bar{X}_{(1)}^* \leq \dots \leq \bar{X}_{(B)}^*$ are the ordered bootstrap sample means. Does either of these work in the sense of giving asymptotically correct coverage probability? Is one preferable to the other on theoretical grounds? We address questions such as these now. \square

10.3.4 Asymptotic Theory for the Bootstrap

Since asymptotic theory for statistics eventually rests on the law of large numbers and the central limit theorem, we start there as well. Suppose X_1, \dots, X_n are i.i.d. P with $\int x^2 dP(x) < \infty$. Let (X_1^*, \dots, X_n^*) be a nonparametric bootstrap sample. By B.1.3 the joint distribution of $(X_1, \dots, X_n, X_1^*, \dots, X_n^*)$ is completely determined by the marginal distribution of (X_1, \dots, X_n) and the conditional distribution of X_1^*, \dots, X_n^* given $X_i = x_i$, $1 \leq i \leq n$, specified as X_i^* i.i.d. with common distribution $\widehat{P}_n \equiv n^{-1} \sum_{i=1}^n \delta_{x_i}$. As we have noted, for any statistic $T_n(X_1, \dots, X_n)$, $\mathcal{L}^*(T_n(X_1^*, \dots, X_n^*))$, the conditional distribution of $T_n(X_1^*, \dots, X_n^*)$ given X_1, \dots, X_n , is a random quantity. Suppose $\mathcal{L}_P(T_n(X_1, \dots, X_n)) \Rightarrow \mathcal{L}_0$. It makes sense to ask if

$$P[\mathcal{L}^*(T_n(X_1^*, \dots, X_n^*)) \Rightarrow \mathcal{L}_0] = 1$$

which we define as *almost sure (a.s.) convergence in law of the bootstrap distribution of T_n* , or the weaker statement which is the one needed for statistics: For all $\varepsilon > 0$, as $n \rightarrow \infty$

$$P[\rho(\mathcal{L}^*(T_n(X_1^*, \dots, X_n^*)), \mathcal{L}_0) \geq \varepsilon] \rightarrow 0 \tag{10.3.15}$$

where ρ is a metric for weak convergence, i.e. $\mathcal{L}_n \Rightarrow \mathcal{L}_0$ iff $\rho(\mathcal{L}_n, \mathcal{L}_0) \rightarrow 0$. See Problem 10.3.8 for Mallows' metric which is an example of such a ρ . We call (10.3.15) *convergence in law in probability of the bootstrap distribution of T_n* . An elegant proof of (10.3.15) for certain T_n using Mallows' metric can be found in Bickel and Freedman (1981). See Problem 10.3.8.

The most basic result is

Theorem 10.3.2. Suppose $\mathbf{X}_1, \dots, \mathbf{X}_n$ are i.i.d. as $\mathbf{X} \in R^d$. Let P denote the joint distribution of $(\mathbf{X}_1, \dots, \mathbf{X}_n, \mathbf{X}_1^*, \dots, \mathbf{X}_n^*)$.

(a) If $E|\mathbf{X}| < \infty$ then, as $n \rightarrow \infty$,

$$P\left[\frac{1}{n} \sum_{i=1}^n \mathbf{X}_i^* \rightarrow E\mathbf{X}\right] = 1. \quad (10.3.16)$$

(b) If $E|\mathbf{X}|^2 < \infty$ and \mathbf{X} has positive definite covariance matrix Σ then

$$P[\mathcal{L}^*(n^{-\frac{1}{2}} \sum_{i=1}^n (\mathbf{X}_i^* - \bar{\mathbf{X}})) \Rightarrow \mathcal{N}(\mathbf{0}, \Sigma)] = 1. \quad (10.3.17)$$

Proof. (a) Note that $\mathbf{X}_1^*, \dots, \mathbf{X}_n^*$ is a double array because the empirical probability \hat{P}_n depends on n . Thus we need to use the uniform strong law of large numbers (SLLN) given in Appendix D.1. That is, we need to show that

$$E_P(|\mathbf{X}^*|1(|\mathbf{X}^*| \geq M)) = n^{-1} \sum_{i=1}^n |\mathbf{X}_i|1(|\mathbf{X}_i| \geq M)$$

converges a.s. to zero as $n \rightarrow \infty$. The right hand side converges a.s. to $E[|\mathbf{X}|1(|\mathbf{X}| \geq M)]$ by the SLLN, and this expression tends to zero as $M \rightarrow \infty$ by the dominated convergence theorem (Theorem B.7.5). Thus $P[\bar{\mathbf{X}}^* \rightarrow \bar{\mathbf{X}}] = 1$ by D.6 and D.7. Because $P[\bar{\mathbf{X}} \rightarrow E(\mathbf{X})] = 1$ by the SLLN, we can for each $\varepsilon > 0$ select N such that, a.s., for $n \geq N$, $|\bar{\mathbf{X}}^* - \bar{\mathbf{X}}| \leq \varepsilon/2$ and $|\bar{\mathbf{X}} - E(\mathbf{X})| \leq \varepsilon/2$; then, a.s., $|\bar{\mathbf{X}}^* - E(\mathbf{X})| \leq \varepsilon$ for $n \geq N$ and we have shown (a). To establish (b), we refer to the Lindeberg-Feller theorem for double arrays of independent variables as stated in Appendix D.1. According to this result (Theorem D.3), to establish (10.3.17) for $d = 1$, we need only check that $E^*(X_i^*) = \bar{X}$ and that for all $\varepsilon > 0$, as $n \rightarrow \infty$

$$\frac{1}{n} \sum_{i=1}^n E^*(X_i^*)^2 1(|X_i^*| \geq \varepsilon n^{\frac{1}{2}}) \xrightarrow{a.s.} 0 \quad (10.3.18)$$

$$E^*(X_i^* - \bar{X})^2 \xrightarrow{a.s.} \text{Var}(X). \quad (10.3.19)$$

The left hand side of (10.3.18) is

$$\Delta_n(\varepsilon) \equiv \frac{1}{n} \sum_{i=1}^n (X_i)^2 1(|X_i| \geq \varepsilon n^{\frac{1}{2}}). \quad (10.3.20)$$

Let $M > 0$ be arbitrary and select n so that $\varepsilon n^{\frac{1}{2}} > M$. Then $\Delta_n(\varepsilon) \leq U_n$, where

$$U_n \equiv \frac{1}{n} \sum_{i=1}^n (X_i)^2 \mathbf{1}(|X_i| > M). \quad (10.3.21)$$

By the SLLN, $U_n \xrightarrow{a.s.} E_P(U_n) = E_P[(X)^2 \mathbf{1}(|X| > M)]$. By the dominated convergence theorem, we can make $E_P(U_n)$ arbitrarily small by selecting M sufficiently large, thus (10.3.18) follows.

Since

$$E^*(X_i^* - \bar{X})^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2 = \frac{1}{n} \sum X_i^2 - (\bar{X})^2,$$

(10.3.19) follows from the SLLN and the result follows for $d = 1$. The general result is argued in the same way (Problem 10.3.9). \square

With this theorem, we can easily establish the conclusion we hoped for in Example 10.3.3 and the preceding discussion.

Example 10.3.3. Bootstrap t (Continued). By Theorem 10.3.2 (b), if $\sigma^2(P) = \text{Var}_P(X)$,

$$P[\sqrt{n}(\bar{X}^* - \bar{X}) \Rightarrow \mathcal{N}(0, \sigma^2(P))] = 1.$$

Thus the bootstrap approximation to the $(1 - \alpha)$ quantile of the distribution of $\sqrt{n}(\bar{X} - \mu)$ converges to $z_{1-\alpha}$ with probability one, which, by the argument leading to (10.3.14), establishes $2\bar{X} - \bar{X}_{[\infty, \alpha]}$ as an asymptotic $(1 - \alpha)$ UCB.

Next consider $(\bar{X}^* - \bar{X})/s^*$. By Theorem 10.3.2 (a), if $[s^*]^2 \equiv n^{-1} \sum_{i=1}^n (X_i^* - \bar{X}^*)^2$, then

$$s^* \xrightarrow{P} \sigma(P). \quad (10.3.22)$$

Here (10.3.22) follows from $n^{-1} \sum (X_i^* - \bar{X}^*)^2 = n^{-1} \sum (X_i^* - \bar{X})^2 + (\bar{X}^* - \bar{X})^2$,

$$\frac{1}{n} \sum_{i=1}^n (X_i^* - \bar{X})^2 \xrightarrow{P} \sigma^2(P),$$

and

$$(\bar{X}^* - \bar{X})^2 \xrightarrow{P} 0.$$

It follows from (10.3.22) and Slutsky's theorem that

$$\sqrt{n} \frac{(\bar{X}^* - \bar{X})}{s^*} \Rightarrow \mathcal{N}(0, 1)$$

in probability which yields that $\hat{d}_{n\alpha}^{(\infty)} \xrightarrow{P} z_\alpha$, and hence establishes the asymptotic validity of

$$\bar{X} - \hat{d}_{n\alpha}^{(\infty)} s / \sqrt{n} \quad (10.3.23)$$

as a $1 - \alpha$ UCB.

Next let $n \rightarrow \infty$ and let B temporarily be fixed. For each $\delta > 0$

$$\lim_{n \rightarrow \infty} P^*[\sup_t \left| \frac{1}{B^{\frac{1}{2}}} \sum_{b=1}^B (1(T_{nb}^* \leq t) - P^*(T_n^* \leq t)) \right| \geq \delta] \leq P[\|\mathcal{E}_B\|_\infty \geq \delta] \quad (10.3.24)$$

where $T_n^* = T_n(X_1^*, \dots, X_n^*)$ and \mathcal{E}_B is the empirical process for U_1, \dots, U_B i.i.d. $\mathcal{U}(0, 1)$. To establish (10.3.24) let $G_n(t)$ denote the distribution of T_n^* , and let U, U_1, \dots, U_B be i.i.d. $\text{Unif}(0, 1)$. Then the expression inside the sup in (10.3.24) has the same distribution as

$$B^{-\frac{1}{2}} \sum_{b=1}^B [1(U_b \leq G_n(t)) - P(U \leq G_n(t))] .$$

The sup over t of this expression is bounded above by $\|\xi_B\|_\infty$. To show equality, let H_n denote the distribution of T_n . Then

$$|G_n(t) - \Phi(t)| \leq |G_n(t) - H_n(t)| + |H_n(t) - \Phi(t)|$$

and the central limit theorem together with Polya's theorem (Theorem B.7.7) implies that $G_n(t)$ converges uniformly to $\Phi(t)$. The continuity of Φ shows that

$$\sup_t B^{-\frac{1}{2}} \sum_{b=1}^B [1(T_{nb}^* \leq t) - P^*(T_n^* \leq t)]$$

converges weakly to $\|\xi_B\|_\infty$. By Donsker's theorem,

$$\widehat{d}_{n\alpha}^{(B_n)} \xrightarrow{P} z_\alpha$$

where $B = B_n \rightarrow \infty$ as $n \rightarrow \infty$. We conclude that for such B the bootstrap t UCB

$$\bar{X} - \widehat{d}_{n\alpha}^{(B)} \frac{s}{\sqrt{n}}$$

is asymptotically valid in the sense that $P(\mu(P) \leq \bar{X} - \widehat{d}_{n\alpha}^{(B)} s / \sqrt{n}) \rightarrow (1 - \alpha)$ (Problem 10.3.11).

Which of these three is better and is either to be preferred to using

$$\bar{X} - z(\alpha) \frac{s}{\sqrt{n}} = \bar{X} + z(1 - \alpha) \frac{s}{\sqrt{n}}$$

as an approximate $(1 - \alpha)$ UCB? It turns out (see Hall (1997)), that

$$P[\bar{X} + z(1 - \alpha) \frac{s}{\sqrt{n}} \geq \mu(P)] = (1 - \alpha) + \Omega(n^{-\frac{1}{2}})$$

where $c_n = \Omega(a_n)$ means $c_n = O(a_n)$ and $a_n = O(c_n)$. Moreover

$$P[\mu(P) \leq \bar{X}_{[\infty(1-\alpha)+1]}^*] = (1 - \alpha) + \Omega(n^{-\frac{1}{2}}) ,$$

if $E_P|X_1|^3 < \infty$, while

$$P \left[\mu(P) \leq \bar{X} - \widehat{d}_{n\alpha}^{(\infty)} \frac{s}{\sqrt{n}} \right] = (1 - \alpha) + \Omega(n^{-1})$$

if $E_P X_1^4 < \infty$. This establishes that this bootstrap t UCB method (10.3.23) is better than the bootstrap percentile method if only probability of coverage is considered. These results are based on Edgeworth expansions and are beyond the scope of this book. We refer to Hall (1986) and (1997), where the effect of choice of B is also discussed. \square

The empirical process theory of Section 7.1.1 may also be generalized. The most general result, due to Giné and Zinn (1990), is the following:

Theorem 10.3.3. *Let $(X_1, \dots, X_n, X_1^*, \dots, X_n^*)$ be defined as before,*

$$\mathcal{E}_n(f) = n^{\frac{1}{2}} \left(\int f d(\widehat{P}_n - P) \right) ,$$

and

$$\begin{aligned} \mathcal{E}_n^*(f) &= n^{\frac{1}{2}} \left(\int f d(\widehat{P}_n^* - \widehat{P}_n) \right) \\ &= n^{-\frac{1}{2}} \left(\sum_{i=1}^n f(X_i^*) - \sum_{j=1}^n f(X_j) \right) . \end{aligned}$$

Then, if $\mathcal{E}_n \Rightarrow W_P^0$,

$$P[\mathcal{E}_n^* \Rightarrow W_P^0] = 1 ,$$

and conversely.

Note that since finite dimensional convergence has been established by Theorem 10.3.2, it is only tightness that is at issue. Some maximal inequalities also generalize to bootstrap samples. We refer to van der Vaart and Wellner (1996) for further discussion.

Here is an example of an application of the Giné-Zinn result which also follows from an earlier result of Bickel and Freedman (1981).

Example 10.3.4. *Confidence bands for a distribution function.* Suppose that $X \in R$ with distribution function F which is completely unknown. We want to set a simultaneous confidence band for F . By the equivalence of tests and confidence regions, we consider the Kolmogorov statistic (see Example 4.4.6)

$$T_n(F_0) \equiv \sup_x |\widehat{F}(x) - F_0(x)| .$$

If $c_n(F_0)$ is the $1 - \alpha$ quantile of the distribution of $T_n(F_0)$ under F_0 , we can have as a $1 - \alpha$ confidence region,

$$C_\alpha \equiv \{F : \sup_x \sqrt{n} |\widehat{F}(x) - F(x)| \leq c_n(F)\} . \quad (10.3.25)$$

If we assume F_0 is continuous, $c_n(F) \equiv c_n$ doesn't depend on F and thus

$$C_\alpha^0 = \widehat{F}(x) \pm c_n / \sqrt{n}$$

is a simultaneous $1 - \alpha$ confidence band for F . It turns out, Problem 10.3.12, that C_α^0 is, in fact, level $1 - \alpha$ for all F . However, if F is discrete, C_α^0 is too big, while C_α is difficult to compute and may not be a band. To remedy this, we can apply the bootstrap principle and estimate $c_n(F)$ by $c_n(\widehat{F})$ or rather the usual approximation, the $[B(1 - \alpha)] + 1$ th order statistic $c_{nB}(\widehat{F})$ of

$$T_{nb}^* \equiv \sqrt{n} \sup_x |\widehat{F}^*(x) - \widehat{F}(x)|, \quad 1 \leq b \leq B.$$

It follows from our discussion and the Giné-Zinn theorem that, indeed

$$\widehat{C}_\alpha \equiv \widehat{F} \pm \frac{c_{nB_n}(\widehat{F})}{\sqrt{n}}$$

is an asymptotic size $1 - \alpha$ confidence band for F arbitrary. For further discussion and some Monte Carlo simulations, see Bickel and Krieger (1989). \square

10.3.5 Examples Where Efron's Bootstrap Fails. The m out of n Bootstraps

Our results so far have not established whether, for instance, Efron's bootstrap which is drawn n times from $\{X_1, \dots, X_n\}$ will consistently estimate features of more irregular $T_n(\widehat{P}_n, P)$, such as $n\text{Var}_P\{\text{median}(X_1, \dots, X_n)\}$. In fact, the bootstrap is consistent in approximating the distribution of

$$\sqrt{n}(\text{med}(X_1, \dots, X_n) - F^{-1}\left(\frac{1}{2}\right))$$

and its variance. However, it is fairly easy to generate examples of inconsistency of the bootstrap, as we next illustrate.

Example 10.3.5. *The bootstrap distribution of the maximum.* Suppose X_1, \dots, X_n are i.i.d. real valued with df F such that $F(x) = 1$, $x \geq c$. Assume that X_1 has a density f with $f(c) > 0$. Then, (Problem 10.3.13),

$$n(c - \max(X_1, \dots, X_n)) \implies \mathcal{E}(f(c)), \tag{10.3.26}$$

in probability, where $\mathcal{E}(\lambda)$ denotes the exponential df with mean λ^{-1} . But,

$$n(\max(X_1, \dots, X_n) - \max(X_1^*, \dots, X_n^*))$$

does not converge in law in probability — see Problem 10.3.14. For this and many other examples of bootstrap failures, see Bickel, Götze and van Zwet (1997). \square

There is a fix for the bootstrap failures found independently by Götze (1993) and Politis and Romano (1994). For $m \leq n$, let $\mathbf{X}^{**} \equiv (X_1^{**}, \dots, X_m^{**})$ be a sample from $\mathbf{X} \equiv \{X_1, \dots, X_n\}$ taken *without replacement*. Suppose

$$\mathcal{L}_P(T_n(X_1, \dots, X_n)) \implies \mathcal{L}_0.$$

Then, if \mathcal{L}^{**} indicates the conditional distribution $\mathcal{L}(\mathbf{X}^{**} | \mathbf{X})$ of $(X_1^{**}, \dots, X_m^{**})$ given X_1, \dots, X_n ,

$$\mathcal{L}^{**}(T_m(X_1^{**}, \dots, X_m^{**})) \implies \mathcal{L}_0$$

in probability, provided only that $m \rightarrow \infty$, $m/n \rightarrow 0$. This m *out of n without replacement bootstrap*, also referred to as *subsampling*, is studied in Politis, Romano and Wolf (1999). The analogous with replacement m out of n bootstrap which includes that of Efron is studied in Bickel, Götze and van Zwet (1997).

The m out of n bootstraps can be seen as methods of regularization. Difficulties with the Efron bootstrap occur when a feature $\lambda_n(Q)$ of $\mathcal{L}_Q(T_n(\hat{P}_n, Q))$, which includes $Q = \hat{P}_n$ as a possible argument, is such that

$$\lambda_n(P) \rightarrow \lambda(P),$$

where $\lambda(P)$ is an irregular parameter. Then, while the “bias” $\lambda_n(P) - \lambda(P) \rightarrow 0$, the “variance” $\lambda_n(\hat{P}_n) - \lambda_n(P)$ can blow up. Taking $m(n)/n \rightarrow 0$ and $m(n) \rightarrow \infty$ still makes the “bias” $\lambda_{m(n)}(P) - \lambda(P) \rightarrow 0$ but reduces the “variance” arbitrarily depending on how slowly $m(n) \rightarrow \infty$. We illustrate this phenomena in Problem 10.13.15.

Summary. We introduced the *bootstrap*, which is a method for approximating distributions, estimates, critical values, and confidence procedures by drawing i.i.d. Monte Carlo samples from the empirical probability distribution \hat{P}_n . Let \mathcal{P} be a set of probabilities on a set \mathcal{X} , and suppose \mathcal{P} contains all discrete distributions. Let $\mathbf{X} = (X_1, \dots, X_n)$ be i.i.d. observations from $P \in \mathcal{P}$ and let $T_n(\mathbf{X}, P)$ be a functional which is invariant under permutations of \mathbf{X} . We also write $T_n \equiv T_n(\hat{P}_n, P)$, where we have identified \hat{P}_n with \mathbf{X} . We considered parameters that are features of $\mathcal{L}(T_n)$ such as the distribution $F_{T_n}(t) = P(T_n \leq t)$, quantiles $F_{T_n}^{-1}(\alpha)$, $\theta < \alpha < 1$, the expected value $E_P T_n$, and the variance $\text{Var}_P T_n$. More generally, we considered

$$\lambda_n(P) = \theta(\mathcal{L}_P(T_n(\hat{P}_n, P)))$$

where θ is a map from $\{\mathcal{L}_P(T_n(\hat{P}_n, P)) : P \in \mathcal{P}\}$ to a set Θ . The *bootstrap estimate* of $\lambda_n(P)$ is the empirical plug-in estimate $\lambda_n(\hat{P}_n)$. We can write $\lambda_n(\hat{P}_n)$ as $\theta(\mathcal{L}^*(T_n(\mathbf{X}^*, \hat{P}))$ where $\mathbf{X}^* = (X_1^*, \dots, X_n^*)$ is an i.i.d. sample from \hat{P} . This Monte Carlo sample, which we can think of as a sample drawn with replacement from $\{X_1, \dots, X_n\}$, is called a *bootstrap sample*. When \mathcal{L}^* and $\lambda_n(P_n)$ are difficult to compute, they can be approximated by the following procedure called the (Efron) *bootstrap*: Generate B independent bootstrap samples $\mathbf{X}_1^*, \dots, \mathbf{X}_B^*$ and compute $T_{nj}^* = T_n(\mathbf{X}_j^*, \hat{P}_n)$, $j = 1, \dots, B$. Now use the empirical distribution \mathcal{L}_B^* of $T_{n1}^*, \dots, T_{nB}^*$ to approximate $\mathcal{L}^*(T_n(\mathbf{X}^*, \hat{P}_n))$ and use

$\theta(\mathcal{L}_B^*)$ to approximate $\lambda_n(\hat{P}_n)$. We showed how the bootstrap can be used to estimate the bias and variance of estimates, how it can be used to find approximate critical values for tests, as well as approximate confidence bounds, intervals, and regions. We develop and discuss some asymptotic results that imply that in “regular” situations, bootstrap approximations are asymptotically valid. We also give an example of an “irregular” case where the bootstrap fails, and we introduce alternative bootstraps, *the m out of n with and without replacement bootstraps*. The latter can be used in all of the situations where the Efron bootstrap does not work.

10.4 Markov Chain Monte Carlo (MCMC)

10.4.1 The Basic MCMC Framework

We want to generate random variables X_1, X_2, \dots on a sample space \mathcal{X} distributed according to p , but direct generation is not feasible. Rejective sampling procedures generate from a sequence of i.i.d. variables, Y_1, Y_2, Y_3, \dots distributed according to p_0 , a variable X exactly distributed according to p . Evidently we obtain B variables, X_1, \dots, X_B i.i.d. according to p by accepting $X_1 = Y_{\tau_1}$ and then continuing to sample Y ’s, till we get $X_2 = Y_{\tau_2}$ and so on. Rejective sampling is not always practicable because there may not be a clear choice of p_0 such that $\sup_x p/p_0(x)$ is not too large. It turns out to be useful to simultaneously enlarge our field of action and our use of the term “sampling” in two directions:

- (i) We permit (Y_1, Y_2, \dots) to be generated according to a Markov sequence, a generalization of independent sampling. The required Markov chain theory is reviewed in Appendix D.5.
- (ii) We do *not* insist that the X_1, \dots, X_B we obtain be either exactly marginally distributed according to p or be exactly independent.

This generalization of rejective sampling introduced by Metropolis et al (1953) and developed by Hastings (1970) has become very important.

The basic idea of MCMC is to construct a homogeneous positive recurrent Markov chain with *transition kernel* $K(x_1, x_2)$, the conditional density $p(x_2|x_1)$ of X_2 evaluated at x_2 , given $X_1 = x_1$, such that p is the unique stationary density of K . Here the term “conditional density” refers to both the discrete and continuous case. Thus in the discrete case, $K(x_1, x_2) = P(X_2 = x_2|X_1 = x_1)$, and in the continuous case moving from x_1 to x_2 means drawing x_2 according to the distribution whose density is $p(x_2|x_1) = K(x_1, x_2)$.

In what follows we will use the continuous case notation and write integrals. In the discrete case, the integrals should be read as sums. For functions g on the sample space \mathcal{X} define the *total variation norm* by

$$\|g\|_1 \equiv \int |g(x)|dx$$

for continuous models and

$$\|g\|_1 = \sum_{i=1}^n |g(x_i)|$$

for discrete models. Specifically, we seek $K(x, y)$ from which, given $Y_1 = x$, we can generate consecutive Y_2, Y_3, \dots , simply such that the following holds.

- (a) For all y , $\int p(x)K(x, y)dx = p(y)$, that is, stationarity of p with respect to K .
- (b) Let Y_1 be distributed according to a selected p_0 . Let Y_2, \dots, Y_m, \dots be the Markov sequence obtained by using Y_1 as an initial value and K as the transition kernel. Then, if $p_m(y)$ is the density of Y_m , we require that there exist $c > 0$ and $0 < \rho < 1$, such that

$$\|p_m - p\|_1 \leq c\rho^m. \quad (10.4.1)$$

The idea is to begin with a relatively simple Y_1 , and for M large, run the chain to Y_M as in (b) above (*burn in*). Continue the chain and define $X_j = Y_{M+j}$, $j = 1, \dots, n, \dots$. Then X_1, \dots, X_n, \dots will be distributed approximately according to p by (a) and (b) (Problem 10.4.1). There are many variants, for instance, repeating the (b) step B times using independent Y_1 's and using the B resulting approximately i.i.d. as p "X₁'s".

Why should we use this extension of rejective sampling, rather than just stick to the former? The main reason is that, as we stated and we shall see in our examples, it is often the case that a p_0 with a reasonably small $c(p, p_0)$ is not readily available. The "uniform" p_0 almost invariably is very poor. MCMC in some sense allows us to adjust p_0 as we move so as to get closer to p on the first iterate we choose to retain.

10.4.2 Metropolis Sampling Algorithms

A general class of kernels producing a desired p was proposed by Metropolis et al (1953) and then generalized by Hastings (1970) and others. All have the following form:

- (1) Choose a *proposal kernel* $K_0(x_1, x_2)$ from which it is relatively easy to generate x_2 for a given x_1 . Given x_1 , select a candidate new value x_2 according to $K_0(x_1, x_2)$.
- (2) For a given function $r : R^2 \rightarrow [0, 1]$, such that $r(x, x) \equiv 1$, move to x_2 with probability $r(x_1, x_2)$, otherwise stay at x_1 . Call the resulting value y_1 .
- (3) Repeat with y_1 in place of x_1 to obtain y_2 , etc.

This leads to the new kernel

$$\begin{aligned} K(x_1, x_2) &= r(x_1, x_2)K_0(x_1, x_2), \quad \text{if } x_1 \neq x_2 \\ &= K_0(x_1, x_1) + \sum_{y \neq x_1} (1 - r(x_1, y))K_0(x_1, y) \\ &= 1 - \sum_{y \neq x_1} r(x_1, y)K_0(x_1, y), \quad \text{if } x_1 = x_2. \end{aligned} \quad (10.4.2)$$

The original Metropolis algorithm has $K_0(\cdot, \cdot)$ symmetric and

$$r(x_1, x_2) \equiv \min\left(1, \frac{p(x_2)}{p(x_1)}\right).$$

Intuitively, a symmetric K_0 has the uniform distribution as stationary distribution (Problem 10.4.2). Then this $r(x_1, x_2)$ nudges the generated x 's towards values more probable under p , so that, in the end, the relative frequency with which x is visited over a long time is approximately $p(x)$. Also note that, as for rejective sampling to compute r , we need only to know $p(\cdot)$ up to a constant.

It is evident that if $x_1 \neq x_2$, for the Metropolis original proposed K ,

$$\begin{aligned} p(x_1)K(x_1, x_2) &= \min(p(x_1), p(x_2))K_0(x_1, x_2) \\ &= \min(p(x_1), p(x_2))K_0(x_2, x_1) \\ &= p(x_2)K(x_2, x_1). \end{aligned} \tag{10.4.3}$$

The condition

$$p(x_1)K(x_1, x_2) = p(x_2)K(x_2, x_1), \text{ for all } x_1, x_2 \tag{10.4.4}$$

is called *detailed balance* and implies that p is the stationary distribution of $K(\cdot, \cdot)$; see Appendix D.5.

The following result gives a general prescription for constructing K .

Proposition 10.4.1 (Stein). *Suppose K_0 is a Markov kernel and*

$$K(x_1, x_2) = r(x_1, x_2)K_0(x_1, x_2), \quad x_1 \neq x_2. \tag{10.4.5}$$

Then K is itself a Markov kernel satisfying detailed balance if and only if $r(x_1, x_2)$ can be written as

$$\frac{\delta(x_1, x_2)}{p(x_1)K_0(x_1, x_2)} \tag{10.4.6}$$

where $\delta(x_1, x_2)$ is symmetric and, for all x_1 ,

$$\sum_{x_2 \neq x_1} \delta(x_1, x_2) \leq p(x_1). \tag{10.4.7}$$

Proof. Detailed balance holds by definition iff $p(x_1)K(x_1, x_2)$ is symmetric. To check symmetry note that, by (10.4.5) and (10.4.6),

$$\delta(x_1, x_2) = p(x_1)r(x_1, x_2)K_0(x_1, x_2) = p(x_1)K(x_1, x_2).$$

Thus detailed balance holds for K if and only if δ is symmetric. To establish (10.4.7), note that because we need

$$0 \leq K(x_1, x_1) = 1 - \sum_{x_2 \neq x_1} r(x_1, x_2)K_0(x_1, x_2),$$

we must have

$$\sum_{x_2 \neq x_1} r(x_1, x_2) K_0(x_1, x_2) = \sum_{x_2 \neq x_1} \frac{\delta(x_1, x_2)}{p(x_1)} \leq 1$$

and conversely. The proposition follows. \square

Remark 10.4.1. (a) The inequality (10.4.7) is implied by $r(x_1, x_2) \leq 1$ for all x_1, x_2 . This is necessary for our interpretation of r and of the implementation of step (2) of the algorithm but not for (10.4.4) and (10.4.5).

(b) The general Metropolis proposal is to use

$$r(x_1, x_2) \equiv \min \left(1, \frac{p(x_2) K_0(x_2, x_1)}{p(x_1) K_0(x_1, x_2)} \right) \quad (10.4.8)$$

which is evidently symmetric.

(c) If K_0 satisfies detailed balance with respect to p , then we may choose $K = K_0$.

The Metropolis condition and *a fortiori* (10.4.7) give little guidance as to how K_0 is to be chosen and, indeed, this is usually dictated by the structure of the particular situation we want to deal with and considerations which are both numerical and statistical.

To illustrate the Metropolis algorithms, we consider an interesting application. We shall pursue this somewhat complicated example further in Section 10.5.

Example 10.4.1. *A prediction problem.* Let,

$$Y_i = Z_i + \varepsilon_i, \quad i = 1, \dots, n$$

where we assume that (Z_1, \dots, Z_n) are observed and, for initial simplicity, that $\varepsilon_1, \dots, \varepsilon_n$ are i.i.d. with known $\mathcal{N}(0, 1)$ distribution. We do not, however, observe Y_1, \dots, Y_n but rather the order statistics $(Y_{(1)}, \dots, Y_{(n)})$ where $Y_{(1)} < \dots < Y_{(n)}$. Our goal is to predict $Y_i, \quad i = 1, \dots, n$ on the basis of this data. An example (Petty et al (2002)) where this question arises (with the distribution of the ε_i unknown) is observation of a set of n cars on a highway passing the first of two checkpoints at times $Z_1 < \dots < Z_n$ which label the cars, and again at a second checkpoint at times $Y_{(1)} < \dots < Y_{(n)}$. Observation is carried out by a device (loop detector) which can record the time of passing of a vehicle but not its identity. The model specifies that $\varepsilon_1, \dots, \varepsilon_n$, the travel times of the vehicles, are i.i.d. given (Z_1, \dots, Z_n) and we wish to predict $\varepsilon_i = Y_i - Z_i$ given this data. But, since Z_i are known, this is equivalent to predicting Y_i . Let (R_1, \dots, R_n) be the (unobserved) ranks of Y_1, \dots, Y_n defined by $Y_i = Y_{(R_i)}$. In what follows, all quantities depend on (Z_1, \dots, Z_n) , but we suppress this dependence. The best squared error predictor of Y_i is by Theorem 1.4.1,

$$E(Y_i | Y_{(1)}, \dots, Y_{(n)}) = \sum_{j=1}^n Y_{(j)} P[R_i = j | Y_{(1)}, \dots, Y_{(n)}] \quad (10.4.9)$$

and the $P[R_i = j | Y_{(1)}, \dots, Y_{(n)}]$ are known in principle. Unfortunately, the natural formula is computationally hopeless since it appears to involve $n!$ multiplications of n terms each and $n!$ additions,

$$\begin{aligned} P[R_i = j | Y_{(1)}, \dots, Y_{(n)}] &= \\ \frac{\sum\{A(Y_{(k_1)}, \dots, Y_{(k_n)}, \mathbf{Z}) : \text{All permutations } (k_1, \dots, k_n) \text{ of } \{1, \dots, n\} \text{ with } k_i = j\}}{\sum\{A(Y_{(k_1)}, \dots, Y_{(k_n)}, \mathbf{Z}) : \text{All permutations } (k_1, \dots, k_n)\}} \end{aligned} \quad (10.4.10)$$

where

$$A(Y_{(k_1)}, \dots, Y_{(k_n)}, \mathbf{Z}) = \prod_{j=1}^n \varphi(Y_{(k_j)} - Z_j).$$

However, we claim that the expression (10.4.10) can be approximated using MCMC. The idea is to generate B (approximately) independent ‘‘observations’’ (R_{1b}, \dots, R_{nb}) , $1 \leq b \leq B$, from the conditional distribution of (R_1, \dots, R_n) given $Y_{(1)}, \dots, Y_{(n)}$, using the Metropolis algorithm. We can then compute the natural estimate of the marginal distribution of R_i , $i = 1, \dots, n$, by

$$\hat{P}[R_i = j | Y_{(1)}, \dots, Y_{(n)}] = \frac{\#\{R_{ib} = j\}}{B}$$

and the natural predictor of Y_i ,

$$\hat{Y}_i \equiv \hat{E}(Y_i | Y_{(1)}, \dots, Y_{(n)}) = \sum_{j=1}^n Y_{(j)} \hat{P}[R_i = j | Y_{(1)}, \dots, Y_{(n)}].$$

To see that we can apply the Metropolis sampler, note that

$$P[R_1 = k_1, \dots, R_n = k_n | Y_{(1)}, \dots, Y_{(n)}] \propto \prod_{j=1}^n \varphi(Y_{(k_j)} - Z_j)$$

where the ‘‘unknown’’ proportionality constant is the marginal density of $(Y_{(1)}, \dots, Y_{(n)})$, $\sum\{A(Y_{(k_1)}, \dots, Y_{(k_n)}), \mathbf{Z}\} : \text{All permutations } (k_1, \dots, k_n)\}$. The state space for our Markov chain is thus $\{\text{All permutations } (k_1, \dots, k_n) \text{ of } \{1, \dots, n\}\}$.

What proposal distribution shall we use? One possibility is to take $K(x_1, x_2) = \frac{1}{n!}$ for all x_1, x_2 , that is, pick the proposal permutation uniformly at random. This seem unreasonable in the car example, since, given that cars have approximately the same speed and $Z_1 < \dots < Z_n$, the identity permutation $(1, \dots, n)$ should be favored. A possible type of proposal suggested by Ritov in Pasula et al (1999) and studied extensively by Ostland (1999) is to make only interchanges of a pair of cars possible on each step, that is,

$$K_0\{(k_1, \dots, k_n), (k'_1, \dots, k'_n)\} > 0$$

iff, for one and only one j , $k'_j = k_{j+1}$, $k'_{j+1} = k_j$, and $k'_i = k_i$ for $i \neq j, j + 1$. Call such moves *interchanges*. Further, it is assumed that all interchanges have equal probability $(n - 1)^{-1}$. Since K_0 is symmetric, the Metropolis algorithm can be described by

(1) Given $\mathbf{k} = (k_1, \dots, k_n)$, propose the interchange

$$(k_j, k_{j+1}) \rightarrow (k_{j+1}, k_j) = (k'_j, k'_{j+1}).$$

(2) Accept the interchange with probability

$$r(\mathbf{k}, \mathbf{k}') = \min\{\exp\{(Z_{j+1} - Z_j)(Y_{(k_j)} - Y_{(k_{j+1})})\}, 1\}.$$

Note that since, in the car example, $Z_{j+1} - Z_j > 0$, this procedure reasonably always accepts interchanges such that cars which follow each other at the beginning but do not follow each other at the end in the initial permutation do so after interchange. But, with probability depending (inversely) on the discrepancy between the arrival times at the beginning and at the end, we can also invert the natural order as given by the initial permutation. It may be shown that K is geometrically ergodic, that is, satisfies (10.4.1). Hints to proofs of the assertions of this example are given in the problems. A discussion and simulations of this and related models may be found in Ostland (1999). \square

10.4.3 The Gibbs Samplers

An important and well-analyzed class of kernels K dictated by problem structure is that of the *Gibbs samplers*. The simplest case of these can be characterized as special cases of Metropolis samplers with $K = K_0$ since this simple Gibbs sampler satisfies detailed balance.

Suppose that $X = (U, V)$ has distribution with density $p(\cdot, \cdot)$. U and V can be vectors, $U \in R^{d_1}$, $V \in R^{d_2}$, or abstract. We assume that we can easily generate observations from the conditional distributions, $\mathcal{L}(V|U = u)$ for all u , and $\mathcal{L}(U|V = v)$ for all v , and want to generate observations (approximately) from the marginal distribution of U , respectively V , or (see Remark 10.4.2) generate observations from the joint distribution of (U, V) . The Gibbs sampler, introduced by Geman and Geman (1984), prescribes a Markov chain on R^{d_1} whose stationary distribution is the marginal distribution of U as follows:

- (1) Draw $U_1 = u_1$ from some distribution on R^{d_1} .
- (2) Generate $V_1 = v_1$ according to the conditional distribution of V given $U = u_1$.
- (3) Draw $U_2 = u_2$ from the conditional distribution of U given $V = v_1$ and repeat (2) and (3) indefinitely using the output u generated in (3) as the input u generated in (2).

We claim that

Proposition 10.4.2. *Let U_1, U_2, \dots be generated as in (2) and (3) above. Then*

- (a) *The desired marginal density of U , call it $p(\cdot)$, is known up to a constant.*
- (b) *The chain has p as stationary density.*
- (c) *The transition kernel defined by (2) and (3) satisfies detailed balance.*

Proof. To establish (a), note that if $p(v|u)$ and $q(u|v)$ denote the conditional densities of V given U and U given V and q is the marginal density of V , then

$$p(v|u) = \frac{q(u|v)q(v)}{p(u)} \quad (10.4.11)$$

by Bayes' theorem. Hence

$$p(u) = \frac{q(u|v_0)}{p(v_0|u)} q(v_0)$$

for any fixed v_0 with $p(v_0|u) > 0$ and (a) holds with unknown constant $q(v_0)$.

To establish (b), note that the Markov kernel $K(u_1, u_2)$ we have described is

$$K(u_1, u_2) = \int p(v|u_1)q(u_2|v)dv \quad (10.4.12)$$

where we assume that all variables considered have continuous type densities, but this is immaterial. Essentially, the procedure is valid whenever the formalism makes sense. The general condition for stationarity is easily checked:

$$\begin{aligned} & \int p(u_1) \int p(v|u_1)q(u_2|v)dv du_1 \\ &= \int q(u_2|v) \left(\int p(u_1)p(v|u_1)du_1 \right) dv \\ &= \int q(u_2|v)q(v)dv = p(u_2). \end{aligned} \quad (10.4.13)$$

The proof of detailed balance is left to the problems (Problem 10.4.3). \square

Remark 10.4.2 Note that similarly, we can generate V_2, V_3, \dots satisfying Proposition 10.4.2. For the Gibbs sampler, $U \sim p(\cdot)$ and $V \sim q(\cdot)$ are, by (2) and (3), equivalent to $(U, V) \sim p(\cdot, \cdot)$ because $p(u, v) = p(u)q(v|u) = q(v)p(v|u)$. Thus $(U_M, V_M), (U_{M+1}, V_{M+1}), \dots$ are approximately from $\mathcal{L}(U, V)$.

Here is an important example of the use of the Gibbs sampler.

Example 10.4.2. Posterior distributions in the Gaussian model. We consider an extension of Example 1.6.12 where we observe X_1, \dots, X_n i.i.d. $\mathcal{N}(\mu, \sigma^2)$ and then prescribe a conjugate prior distribution for $\theta = (\mu, \sigma^2)$ by specifying that μ and σ^2 are independent with μ having a $\mathcal{N}(\mu_0, \tau_0^2)$ distribution and σ^{-2} having a $\Gamma(p_0, \lambda_0)$ distribution. This model falls under the framework of Example 10.1.3. As we have seen in Example 1.6.12, the posterior distribution of μ given X_1, \dots, X_n and σ^2 is

$$\mathcal{N} \left(\frac{\frac{n\bar{X}}{\sigma^2} + \frac{\mu_0}{\tau_0^2}}{\frac{n}{\sigma^2} + \frac{1}{\tau_0^2}}, \frac{1}{\frac{n}{\sigma^2} + \frac{1}{\tau_0^2}} \right).$$

This is $\mathcal{L}(U|V = v)$ with $U = \mu$ and $V = \sigma^{-2}$ in Gibbs sampler notation. The posterior distribution of σ^{-2} given μ and the data can be computed (Problem 10.4.5) and is the

gamma distribution $\Gamma(p_0 + \frac{1}{2}n, \lambda(\mathbf{X}, \mu))$, where

$$\lambda(\mathbf{X}, \mu) = \lambda_0 + \frac{1}{2} \sum (X_i - \mu)^2. \quad (10.4.14)$$

Thus $\mathcal{L}(V|U = u)$ is also available. So we can implement the Gibbs sampler once we know how to draw from the normal and gamma distributions as we have learned in Section 10.2. In this way we get, for instance, (approximate) observations from the posterior distribution of σ^2 given \mathbf{X} . As we noted we can use the Gibbs sampler to generate a bivariate Markov chain (U_m, V_m) , with U_m corresponding to μ and V_m to σ^{-2} yielding approximate observations from the posterior of (μ, σ^{-2}) given \mathbf{X} . We shall see how to use these methods in inference in Section 10.5. \square

Example 10.4.3. A bivariate normal example. Let

$$\mathcal{L}(Y|X = x) = \mathcal{N}(\rho x, (1 - \rho^2)) = \mathcal{L}(X|Y = x)$$

for $|\rho| < 1$. By Theorem B.4.2 in Volume I, these are the conditional distributions of the bivariate normal distribution $(X, Y) \sim \mathcal{N}_2(0, 0, 1, 1, \rho)$ and hence $X \sim Y \sim N(0, 1)$. We will observe the workings of the Gibbs sampler. Suppose we initiate

$$X^{(0)} \sim N(0, \sigma^2).$$

Then we can write

$$Y^{(0)} = \rho X^{(0)} + (1 - \rho^2)^{\frac{1}{2}} Z,$$

where $Z \sim N(0, 1)$ independent of X . Therefore,

$$Y^{(0)} \sim \mathcal{N}(0, \rho^2 \sigma^2 + (1 - \rho^2)).$$

At the next round,

$$\begin{aligned} X^{(1)} &= \rho Y^{(0)} - (1 - \rho^2)^{\frac{1}{2}} Z \\ X^{(1)} &\sim N(0, \rho^4 \sigma^2 + \rho^2(1 - \rho^2) + (1 - \rho^2)) = N(0, \rho^4 \sigma^2 + (1 - \rho^4)). \end{aligned}$$

Continuing, if at the k th stage

$$X^{(k)} \sim N(0, \sigma_k^2)$$

then (Problem 10.4.8)

$$X^{(k+1)} \sim N(0, \rho^4 \sigma_k^2 + (1 - \rho^4)),$$

and thus

$$\sigma_{k+1}^2 - 1 = \rho^4(\sigma_k^2 - 1).$$

Then, $\sigma_k^2 - 1$ is monotone decreasing or increasing according as $\sigma_k^2 \gtrless 1$ and converges to the limit

$$\sigma_\infty^2 - 1 = \rho^4(\sigma_\infty^2 - 1).$$

So $\sigma_\infty^2 = 1$. Note that convergence is exponential. \square

Example 10.4.4. *The random effects model.* We follow Example 3.2.4 and have

$$X_{ij} = \mu + \Delta_i + \varepsilon_{ij}, \quad 1 \leq i \leq I, \quad 1 \leq j \leq J,$$

with Δ_i and ε_{ij} i.i.d. and independent of each other, $N(0, \sigma_\Delta^2)$ and $N(0, \sigma_e^2)$, respectively. Then, by (3.2.17), if $\tau \equiv (\mu, \sigma_e^2, \sigma_\Delta^2)^T$, $\Delta \equiv (\Delta_1, \dots, \Delta_l)^T$, respectively,

$$p(\mathbf{x}|\Delta) = \prod_{i,j} \varphi_{\sigma_e}(x_{ij} - \mu - \Delta_i)$$

which depends on τ . Here the Δ_i are *latent* (unobserved) variables with joint density

$$p(\Delta|\mu, \sigma_e^2, \sigma_\Delta^2) = \prod_i \varphi_{\sigma_\Delta}(\Delta_i).$$

In the Bayesian framework we put a prior $\pi(\cdot)$ on τ and include Δ in our list of parameters. The full posterior distribution is then

$$\pi(\tau, \Delta|\mathbf{x}) \propto p(x|\Delta)p(\Delta|\tau)\pi(\tau),$$

where the proportionality constant $c(\mathbf{x})$ is such that

$$\int \pi(\tau, \Delta|\mathbf{x}) d\tau d\Delta = 1.$$

The value of the constant is irrelevant if we want the posterior mode but matters for credible Bayesian regions of τ . It is absolutely necessary if, say, we want the density of $\sigma_\Delta^2|\mathbf{x}$. We can, in principle, obtain this constant by the Metropolis-Hastings method finding an homogeneous Markov kernel $\tilde{K}(\tau_1, \tau_2)$ which has the posterior distribution of $\tau|\mathbf{x}$ as its stationary distribution in ways we have discussed. Note that since we are generating an empirical probability distribution, say that of $\tau_{b_0}, \tau_{b_0+t}, \dots, \tau_{b_0+nt}$, observations from the Markov chain after “burn in” spaced t apart (where t is “large”), this empirical distribution yields estimates of every plausible function of $\pi(\tau|\mathbf{x})$.

This approach also works for a non-Bayesian formulation of the model. Suppose $\mu, \sigma_\Delta^2, \sigma_e^2$ are simply unknown parameters. Then we are interested in good estimates $\hat{\mu}, \hat{\sigma}_e^2, \hat{\sigma}_\Delta^2$ and confidence regions for Δ just based on the conditional distribution of Δ given \mathbf{X} with parameter values $\hat{\mu}, \hat{\sigma}_e^2, \hat{\sigma}_\Delta^2$. We still need to calculate the conditional distributions. These are necessarily Gaussian in this case since (Δ, \mathbf{X}) have a joint Gaussian distribution which is easy to compute. This is not the case, however, if, for instance, we consider a robust alternative to model (3.2.16) and assume Δ_i have, say, logistic distributions. Neglecting what is now the difficulty of estimating $\sigma_\Delta^2, \sigma_e^2$, and μ , we are still left with a formidable integration problem finding the conditional density of Δ_i given \mathbf{X} . MCMC can come to the rescue as before. \square

The natural generalization of the simple Gibbs samplers appropriate to Bayesian problems of this structure is to the case where we want to simulate from p , the density of $\mathbf{U} = (U_1, \dots, U_k)$, where the U_j may themselves be vectors. We are given that the conditional distribution of U_i given $\{U_j : j \neq i\}$, call it $\mathcal{L}(U_i|\mathbf{U}_{-i})$, is easy to sample from.

Then, the idea is to cycle, initializing by picking $\mathbf{U} = \mathbf{u}^0$ arbitrarily, then $U_1 = u_1^1$ from $\mathcal{L}(U_1 | \mathbf{U}_{-i} = \mathbf{u}_{-i}^0)$, then U_2 from $\mathcal{L}(U_2 | \mathbf{U}_{-2} = (u_1^1, u_3^0, \dots, u_k^0))$, and continuing till the k th coordinate when we have $\mathbf{U}_1 = (u_1^1, \dots, u_k^1)$ complete, and then repeat. Although this chain does not obey detailed balance, we can show that its stationary distribution is p by arguing as we did for the case $k = 2$. An example is the random effects model (Example 10.4.4) where $U_i = \Delta_1$.

10.4.4 Speed of Convergence and Efficiency of MCMC

As we have noted, a reasonable measure of the desirability of a Metropolis or Gibbs sampler $\{X_1, \dots, X_n\}$ once simplicity has been taken into account is governed by the speed of convergence of the distribution of X_n to the desired stationary distribution with discrete or continuous case function p . But how is convergence to be measured in general? If the state space $\mathcal{X} = \{x_1, \dots, x_N\}$ is finite and the transition matrix is $\mathbf{K} = \|K(\cdot, \cdot)\|_{N \times N}$ then, as we note in Appendix D5, if a homogeneous Markov chain is irreducible, aperiodic, and satisfies detailed balance, then, as required in (10.4.1), for some $\rho \in (0, 1)$

$$\Delta_n(K) \equiv \sum_{k=1}^N |P[X_n = x_k] - p(x_k)| \asymp \rho^n$$

where “ $\asymp \rho^n$ ” means that, as $n \rightarrow \infty$, $n^{-1} \log \Delta_n(K) \rightarrow \log \rho$.

We next relate convergence to eigenanalysis. Consider an aperiodic irreducible reversible Markov Chain on the finite state space $\{1, 2, \dots, N\}$. Under these conditions it can be shown (Grimmett and Stirzaker (2001), Sections 6.6 and 6.14) that the transition matrix \mathbf{K} has real eigenvalues $\lambda_1, \dots, \lambda_N$ such that $\lambda_1 = 1$ and $|\lambda_j| < 1$ for $j = 2, \dots, N$. Moreover the entries v_{rj} in the corresponding eigenvectors \mathbf{v}_r are real. Define the inner product

$$(\mathbf{x}, \mathbf{y}) = \sum_{j=1}^N x_j y_j p_j$$

where p_j is the probability of observing j under the stationary distribution. We can take the eigenvectors $\mathbf{v}_1, \dots, \mathbf{v}_N$ to be an orthonormal basis with respect to (\cdot, \cdot) . It follows that $\mathbf{v}_1 = (1, \dots, 1)^T \equiv \mathbf{1}$. We can now give an exact expression for the n -step transition probability.

Proposition 10.4.3. *Set $p_{ij}(n) = P(X_{n+1} = j | X_1 = i)$. Under the conditions in the previous paragraph,*

$$p_{ij}(n) - p_j = p_j \sum_{r=2}^N \lambda_r^n v_{rj} v_{ri} .$$

Proof. We can write the unit vector $\mathbf{e}_j = (0, \dots, 1, \dots, 0)^T$ as

$$\mathbf{e}_j = \sum_{r=1}^N (\mathbf{e}_j, \mathbf{v}_r) \mathbf{v}_r = \sum_{r=1}^N v_{rj} p_j \mathbf{v}_r . \quad (10.4.15)$$

Next note that $\mathbf{K}^n \mathbf{e}_j = (p_{1j}(n), \dots, p_{Nj}(n))^T$ and $\mathbf{K}^n \mathbf{v}_r = \lambda_r^n \mathbf{v}_r$. Multiply (10.4.15) on the left by \mathbf{K}^n and find that

$$p_{ij}(n) = p_j \sum_{r=1}^N \lambda_r^n v_{rj} v_{ri}.$$

The result follows because $\lambda_1 = 1$ and $\mathbf{v}_1 = \mathbf{1}$; so the first term in the sum is p_j . \square

Now we can establish convergence rates:

Corollary 10.4.1. *Under the conditions of Theorem 10.4.3,*

$$|p_{ij}(n) - p_j| \leq p_j (|\lambda|_2)^n \sum_{r=2}^N v_{rj} v_{ri}$$

where $|\lambda|_2 = \max_{2 \leq j \leq N} |\lambda_j|$ is less than one. Moreover, the total variation norm satisfies

$$\sum_{j=1}^N |p_{ij}(n) - p_j| \leq |v_i|_2 (N-1) (|\lambda|_2)^n$$

where $|v_i|_2 = \max_{2 \leq r \leq N} |v_{ri}|$.

Proof. The first part follows from the definition of $|\lambda|_2$. The second part follows because $\sum_j p_j |v_{rj}| \leq 1$ by Cauchy-Schwarz and because $\sum_j p_j v_{rj}^2 = 1$ by the definition of the inner product for the eigenvectors.

Remark. For more on the convergence of MCMC, see Diaconis and Strook (1991), Grimmett and Stirzaker (2001), and Diaconis (2009).

Efficiency of MCMC

We will measure the effectiveness of MCMC by comparing it to the approximations we would have if we could obtain an i.i.d. sample from p . First we need the following result which shows that the second largest eigenvalue λ_2 can be interpreted as a maximal correlation.

Theorem 10.4.1. *Suppose the state space of a homogeneous chain is finite, and that $K(\cdot, \cdot)$ satisfies detailed balance. Then, if the eigenvalues of the transition matrix \mathbf{K} are ordered so that $1 = \lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_N$,*

$$\lambda_2 = \rho_+ \equiv \max\{\text{Corr}(g(X_1), g(X_2)) : g \text{ arbitrary}\}.$$

Proof. The pair (X_i, X_j) has discrete case density $K(x_i, x_j)p_i$, where $p_i = P(X_1 = x_i)$. The optimization problem whose solution we claim to be λ_2 is to maximize over vectors $\mathbf{g} = (g_1, \dots, g_n)^T$,

$$\sum_{i,j} g_i g_j K(x_i, x_j) p_i \tag{10.4.16}$$

subject to

$$(i) \sum_i g_i p_i = 0, \quad (ii) \sum_i g_i^2 p_i = 1.$$

Regard (10.4.16) as a sum over j . Fix j and maximize each term $g_j \sum_i g_i K(x_i, x_j) p_i$. The solution \mathbf{g}^0 , which necessarily exists, satisfies for all j , some γ_1, γ_2 ,

$$\sum_i g_i K(x_i, x_j) p_i = \gamma_2 p_j g_j + \gamma_1 p_j, \quad (10.4.17)$$

since the method of Lagrange multipliers applies, (Apostol (1957), Theorem 7.10, p.153; see Problem 10.4.9). The sum over j of the left side of (10.4.17) reduces to $\sum_i g_i p_i = 0$. Thus the sum over j of the right side is also 0. This implies that $\gamma_1 = 0$ by (i). By detailed balance,

$$K(x_i, x_j) \frac{p_i}{p_j} = K(x_j, x_i),$$

which implies that γ_2 is a right eigenvalue of $\mathbf{K} = (K(x_i, x_j))_{N \times N}$ and \mathbf{g}^0 is a right eigenvector. Since all eigenvectors are vectors of real numbers and $g \equiv 1$ is not a candidate (by (i)) we conclude by the Courant-Fischer theorem (Problem 8.3.24), that is, γ_2 is the second largest eigenvalues.

Moreover multiplying (10.4.17) by g_j , summing, and using (ii), we see that

$$\gamma_2 = \max\{\text{Corr}(g(X_1), g(X_2))\}.$$

□

Proposition 10.4.4. *Under the assumptions of Theorem 10.4.1,*

$$\max\{\text{Corr}(g(X_1), g(X_{k+1})) : g \text{ arbitrary}\} = (\rho_+)^k.$$

Proof. See Problem 10.4.10.

Remark. $\text{Corr}(X_1, X_{k+1})$ is called the *autocorrelation* of the chain, and

$$\max_g \{\text{Corr}(g(X_1), g(X_{k+1}))\},$$

the *maximal autocorrelation*.

Next we show that ρ_+ can be used to show the effectiveness of MCMC.

Example 10.4.5. Consider the problem of estimating $E_p g(X_1) = \sum_{j=1}^N g(x_j) p_j$ by $\bar{g} \equiv B^{-1} \sum_{b=1}^B g(X_b)$, where X_1 is initialized at some distribution and X_2, \dots, X_B are obtained via $K(\cdot, \cdot)$. Then

$$E(\bar{g} - E_p g(X_1))^2 = \left(\frac{1}{B} \sum_{b=1}^B (Eg(X_b) - E_p g(X_1)) \right)^2$$

$$+ \frac{1}{B} \left(\frac{1}{B} \sum_{b=1}^B \text{Var } g(X_b) + \frac{2}{B} \sum_{b=1}^{B-1} \sum_{a=b+1}^B \text{Cov}(g(X_a), g(X_b)) \right). \quad (10.4.18)$$

On the other hand, suppose that X'_1, \dots, X'_B are an i.i.d. sample from p and define $\bar{g}' \equiv B^{-1} \sum_{b=1}^B g(X'_b)$. It is reasonable to define the effectiveness or “efficiency” of the Markov chain sampling scheme by

$$e(K) \equiv \lim_{B \rightarrow \infty} \inf \left\{ \frac{E(\bar{g}' - E_p g(X_1))^2}{E(\bar{g} - E_p g(X_1))^2} : \text{all } g \text{ not constant} \right\}.$$

We claim that

Theorem 10.4.2. *Under the assumptions of Theorem 10.4.1*

$$e(K) = \frac{1 - \rho_+}{1 + \rho_+}.$$

The proof is sketched in Problems 10.4.10 and 10.4.11. It is quite possible that $\rho_+ < 0$ so that the Markov chain is *more* efficient than independent sampling — see Problem 10.4.12. Rosenthal (2003) shows, in the context of general chains, how to construct Metropolis samplers with good speed of convergence from samplers with ρ_+ close to -1 .

Speed of convergence results are also available for countable or, more importantly, continuous state spaces. However, more conditions are required and computation is even more challenging. An introductory treatment is in Tierney’s article in Gilks et al (1995) and a thorough treatment is given in Meyn and Tweedie (1993). See also Rosenthal (2003). Unfortunately, computing or even bounding λ_2 in cases of interest is difficult. A key question that remains is how to determine the length of the “burn in” period from initial simulations. We refer to Liu (2001), Diaconis (2009), and Brooks, Gelman, Jones, and Meng (2011), for instance, for more literature on MCMC.

Summary. We introduced Markov Chain Monte Carlo methods which generate X_1, X_2, \dots that are approximately i.i.d. from p , where p is a density known up to a constant, and from which generation of random variables is difficult. The method entails generating observations from a homogeneous Markov chain with p as stationary distribution. In particular, we presented the *Metropolis–Hastings* method for constructing such Markov chains. We introduced the Gibbs sampler which is an algorithm where knowledge of $\mathcal{L}(V|U = u)$ and $\mathcal{L}(U|V = v)$ is used to generate variables U and V from the marginal distributions $\mathcal{L}(U)$ and $\mathcal{L}(V)$ as well as from the joint distribution $\mathcal{L}(U, V)$. We showed, using an example, how in a Bayesian framework, the Gibbs sampler can be used to generate data from $\mathcal{L}(\theta_1|\mathbf{X})$, $\mathcal{L}(\theta_2|\mathbf{X})$, and $\mathcal{L}(\theta_1, \theta_2|\mathbf{X})$ when $\mathcal{L}(\theta_1|\mathbf{X}, \theta_2)$ and $\mathcal{L}(\theta_2|\mathbf{X}, \theta_1)$ are easy to generate from. Since variables which approximately have the desired p are only produced after the chain is run for some time — the so called “burn in” period, it is important to know how good the approximation is. We gave conditions under which the discrete case density p_k of X_k is close to p . In particular, when X has a finite state space \mathcal{X} , then under conditions on the Markov kernel K , the distance between p and p_k is of order ρ^k , where $\rho \in (0, 1)$. We finally discussed how MCMC can be used to approximate integrals of the form $\int g(x)p(x)dx$ and measured the effectiveness of such approximations by comparing them to Monte Carlo approximation based on (unavailable) i.i.d. samples from p .

10.5 Applications of MCMC to Bayesian and Frequentist Inference

Geyer makes the argument in his paper in Gilks et al (1996) that MCMC is as applicable in frequentist inference as it is in Bayesian inference, given that one is basically concerned with computing high dimensional integrals. We agree with this point of view but go even further. We argued in Section 5.5 that, at least from an asymptotic point of view, if the model we consider is regular parametric, the parameter θ belongs to R^d and the prior distribution put on θ is smooth, then optimal Bayes procedures coincide asymptotically with efficient frequentist procedures (Maximum Likelihood), from a frequentist point of view. This point of view runs into trouble when d is of the order of n or larger — see, for instance, Diaconis and Freedman (1986). These questions, as well as more basic issues of consistency of Bayes procedures continue to be vigorously investigated — see Ghosh and Ramamoorthi (2003).

We begin with a continuation of Example 10.4.2.

Example 10.5.1. *The Gaussian model.* We saw in Example 10.4.2 how we could use the Gibbs sampler to approximately generate observations $\theta_1^*, \dots, \theta_B^*$ from the posterior distribution of $\theta = (\mu, \sigma^{-2})$ when the prior distribution has μ and σ^{-2} independent Gaussian and Gamma, respectively. Assume for simplicity that the observations were obtained by running a sampler from independent starting points so that the X_i^* generated by the sampler are independent and that, although this is true only approximately, their distribution is correct.

We have seen earlier that the Bayes estimates of μ for quadratic loss or other symmetric loss functions if σ is assumed known is just the posterior mean

$$\left(\frac{n\bar{X}}{\sigma^2} + \frac{\mu_o}{\tau_o^2} \right) \left(\frac{n}{\sigma^2} + \frac{1}{\tau_o^2} \right)^{-1}.$$

However, if σ^2 is unknown and has a prior distribution as in Example 10.4.2, to get an estimate of μ , we have to compute the marginal posterior mean of μ ,

$$E \left\{ \left(\frac{n\bar{X}}{\sigma^2} + \frac{\mu_o}{\tau_o^2} \right) \left(\frac{n}{\sigma^2} + \frac{1}{\tau_o^2} \right)^{-1} \mid \mathbf{X} \right\},$$

for which we need

$$\int_0^\infty \left(n\bar{X}w + \frac{\mu_o}{\tau_o^2} \right) \left(nw + \frac{1}{\tau_o^2} \right)^{-1} q(w|\mathbf{X}) dw$$

where $q(w|\mathbf{X})$ is the posterior density of σ^{-2} , a quantity only determinable through numerical integration. On the other hand, we have, using $\theta_1^*, \dots, \theta_B^*$, the simple approximation $B^{-1} \sum_{b=1}^B \theta_{1b}^*$ of the marginal posterior mean of μ , where $\theta_b^* = (\theta_{1b}^*, \theta_{2b}^*)^T$. Similarly $B^{-1} \sum_{b=1}^B \theta_{2b}^*$ provides an estimate of the marginal posterior mean of σ^2 .

Next, suppose we want the shortest possible Bayes level $1 - \alpha$ credible interval for μ . We need to find $[\underline{\mu}, \bar{\mu}]$ such that

- (i) $\sum_{b=1}^B \mathbf{1}(\boldsymbol{\theta}_{1b}^* \in [\underline{\mu}, \bar{\mu}]) = [B(1 - \alpha)],$
- (ii) $\bar{\mu} - \underline{\mu}$ minimizes $\mu_2 - \mu_1$ among all pairs (μ_1, μ_2) , $\mu_1 < \mu_2$, satisfying (i).

Finding a minimum volume Bayes credible region for $\boldsymbol{\theta}$, a generalization of the problem stated above, can be done as follows. Since the posterior density of $\boldsymbol{\theta}$ is proportional to the joint density of $(\mathbf{X}, \boldsymbol{\theta})$, which is completely known, the regions which need to be considered are just all

$$\mathcal{S}_c = \{\boldsymbol{\theta} : p(\mathbf{X}|\boldsymbol{\theta}) q(\boldsymbol{\theta}) \geq c\}$$

where q is the prior density and $p(\mathbf{X}|\boldsymbol{\theta})$ is the conditional density of \mathbf{X} given $\boldsymbol{\theta} = \boldsymbol{\theta}$. Evidently we now choose c so that

$$\frac{1}{B} \sum_{b=1}^B \mathbf{1}(\boldsymbol{\theta}_b^* \in \mathcal{S}_c) \stackrel{\text{def}}{=} 1 - \alpha.$$

□

We pursue these ideas more generally. We observe $\mathbf{X} \sim p(\cdot|\boldsymbol{\theta})$, $\boldsymbol{\theta} \in \Theta$, where $p(\mathbf{x}|\boldsymbol{\theta})$ is a density function, $\boldsymbol{\theta} \in \Theta$, $\boldsymbol{\theta}$ has a prior density $\pi(\boldsymbol{\theta})$, $(\boldsymbol{\theta}, \mathbf{X})$ has joint density $\pi(\boldsymbol{\theta})p(\mathbf{x}|\boldsymbol{\theta})$, and we want to generate observations from the posterior density of $\boldsymbol{\theta}$ given \mathbf{X} . We took advantage of special features of Example 10.4.2 and we used the Gibbs sampler. How do we generate the $\boldsymbol{\theta}_b^*$ in general? The all purpose answer is a Metropolis algorithm since specification of $p(\mathbf{x}|\boldsymbol{\theta})$ and $q(\boldsymbol{\theta})$ is necessary to define the model. That is, we obtain the $\boldsymbol{\theta}_b^*$ by the usual algorithm: Choose a positive recurrent aperiodic Markov kernel $K(\cdot, \cdot)$ on $\Theta \times \Theta$ with stationary density $\pi(\boldsymbol{\theta}|\mathbf{x})$ and generate observations from it after a suitable burn-in from a number of starting points. As usual, one picks $K_0(\cdot, \cdot)$ positive recurrent, aperiodic Markov on $\Theta \times \Theta$ and then lets

$$K(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) = r(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) K_0(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$$

$$\text{if } \boldsymbol{\theta}_1 \neq \boldsymbol{\theta}_2 \text{ with } r(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) = \min \left(1, \frac{p(\mathbf{x}|\boldsymbol{\theta}_2)\pi(\boldsymbol{\theta}_2)K_0(\boldsymbol{\theta}_2, \boldsymbol{\theta}_1)}{p(\mathbf{x}|\boldsymbol{\theta}_1)\pi(\boldsymbol{\theta}_1)K_0(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)} \right). \quad (10.5.1)$$

All we are using here is that the posterior density of $\boldsymbol{\theta}$, $\pi(\boldsymbol{\theta}|\mathbf{X})$, is proportional (up to a data dependent constant) to $\pi(\boldsymbol{\theta})p(\mathbf{x}|\boldsymbol{\theta})$.

Let \mathcal{A} be an action space and $l : \Theta \times \mathcal{A} \rightarrow R^+$ be a loss function. The Bayes procedure for any such problem is, according to Section 3.2, given by

$$\delta^*(\mathbf{x}) = \arg \min_a E(l(\boldsymbol{\theta}, a)|\mathbf{x})$$

$$\text{where } E(l(\boldsymbol{\theta}, a)|\mathbf{x}) = \int l(\boldsymbol{\theta}, a) d\Pi(\boldsymbol{\theta}|\mathbf{x})$$

and $\Pi(\boldsymbol{\theta}|\mathbf{x})$ is the posterior probability distribution with density $\pi(\boldsymbol{\theta}|\mathbf{x})$. Suppose we are able to generate $\boldsymbol{\theta}_1^*, \dots, \boldsymbol{\theta}_B^*$ independent, each having (approximately) the posterior distribution of $\boldsymbol{\theta}$ given $\mathbf{X} = \mathbf{x}$. Then it is natural to use the approximation

$$\delta^{*(B)}(\mathbf{x}) = \arg \min_a \frac{1}{B} \sum_{b=1}^B l(\boldsymbol{\theta}_b^*, a). \quad (10.5.2)$$

Thus, for instance, in Example 10.5.1, in general, without our particular choice of $\pi(\mu, \sigma^{-2})$, the conditional distributions of μ and σ^{-2} , respectively, given each other and \mathbf{X} , would not have a closed form. But the Metropolis sampler is still implementable, although, of course, the usual questions of speed of convergence need to be asked.

The Gibbs and Metropolis sampler can be applied in frequentist inference in a number of ways. One we have already seen in the prediction context in Example 10.4.1. We pursue that in

Example 10.5.2. *Maximum likelihood and travel time prediction. EM-MCMC.* We assumed, in the prediction problem, that the travel time density was known to be $\mathcal{N}(0, 1)$. Suppose, more realistically, that the density is in fact unknown. We can approximate the density by a parametric model $f(\cdot, \boldsymbol{\theta})$, such as the gamma family or more generally finite mixtures of gammas. How do we carry out maximum likelihood estimation of $\boldsymbol{\theta}$? One possible approach is via the EM algorithm and MCMC. Represent Y_1, \dots, Y_n as $(Y_{(1)}, \dots, Y_{(n)}, R_1, \dots, R_n)$ with density

$$f(\mathbf{Y}, \boldsymbol{\theta} | \mathbf{Z}) = \prod_{i=1}^n f(Y_{(R_i)} - Z_i, \boldsymbol{\theta}).$$

Viewing R_1, \dots, R_n as missing, we see that we can alternate

$$\text{E step : Compute } \Lambda(\boldsymbol{\theta}, \widehat{\boldsymbol{\theta}}_{\text{OLD}}) \equiv \sum_{i=1}^n E_{\widehat{\boldsymbol{\theta}}_{\text{OLD}}} \{ \log f(Y_{(R_i)} - Z_i, \boldsymbol{\theta}) | Y_{(1)}, \dots, Y_{(n)} \}$$

M step : Maximize $\Lambda(\boldsymbol{\theta}, \widehat{\boldsymbol{\theta}}_{\text{OLD}})$ to get $\widehat{\boldsymbol{\theta}}_{\text{NEW}}$ by solving

$$\begin{aligned} \nabla \Lambda(\boldsymbol{\theta}, \widehat{\boldsymbol{\theta}}_{\text{OLD}}) &= \sum_{i=1}^n E_{\widehat{\boldsymbol{\theta}}_{\text{OLD}}} \{ \nabla \log f(Y_{(R_i)} - Z_i, \boldsymbol{\theta}) | Y_{(1)}, \dots, Y_{(n)} \} \\ &= 0 \end{aligned} \tag{10.5.3}$$

The first computation can be done, in principle, by MCMC since

$$\Lambda(\boldsymbol{\theta}, \widehat{\boldsymbol{\theta}}_{\text{OLD}}) = \sum_{i=1}^n \sum_{j=1}^n \log f(Y_{(R_i)} - Z_i, \boldsymbol{\theta}) P_{\widehat{\boldsymbol{\theta}}_{\text{OLD}}} [R_i = j | Y_{(1)}, \dots, Y_{(n)}] \tag{10.5.4}$$

which is a problem of the same type we considered in Example 10.4.2. (Here, everything is conditional on Z .) Thus, using the same notation as before on the M step, we maximize

$$\Lambda(\boldsymbol{\theta}, \widehat{\boldsymbol{\theta}}_{\text{OLD}}) \equiv \sum_{i=1}^n \sum_{j=1}^n \log f(Y_{(j)} - Z_i, \boldsymbol{\theta}) \widehat{P}_{\boldsymbol{\theta}_{\text{OLD}}} [R_i = j | Y_{(1)}, \dots, Y_{(n)}].$$

That is, we alternate, for fixed $\widehat{\boldsymbol{\theta}}_{\text{OLD}}$, running MCMC given $\widehat{\boldsymbol{\theta}}_{\text{OLD}}$, then maximize $\widehat{\Lambda}(\boldsymbol{\theta}, \widehat{\boldsymbol{\theta}}_{\text{OLD}})$ to get $\widehat{\boldsymbol{\theta}}_{\text{NEW}}$. Implementation of this kind of algorithm is discussed in Ostland (1999). \square

We can think of this implementation of EM and a more flexible approach in a general context, following Geyer — see Chapter 14 in Gilks et al (1995). We are given $f(\mathbf{x}, v, \boldsymbol{\theta})$, the joint density of $(\mathbf{X}, V, \boldsymbol{\theta})$ in an analytically tractable form. We observe \mathbf{X} so that the likelihood is,

$$g(\mathbf{X}, \boldsymbol{\theta}) = \int f(\mathbf{X}, v, \boldsymbol{\theta}) dv$$

and our goal is to maximize $g(\mathbf{X}, \boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$. According to EM, we need, for the E step, given $\hat{\boldsymbol{\theta}}_{\text{OLD}} = \boldsymbol{\theta}_0$

$$\Lambda(\boldsymbol{\theta}, \boldsymbol{\theta}_0) \equiv E_{\boldsymbol{\theta}_0}(\log f(\mathbf{X}, V, \boldsymbol{\theta}) | \mathbf{X})$$

For $\boldsymbol{\theta}_0$ fixed, we construct a Metropolis or other sampler for the conditional distribution of V given \mathbf{X} under $\boldsymbol{\theta}_0$ and get V_1, \dots, V_B . We estimate,

$$\widehat{\Lambda}(\boldsymbol{\theta}, \boldsymbol{\theta}_0) = B^{-1} \sum_{b=1}^B \log f(\mathbf{X}, V_b, \boldsymbol{\theta}_0)$$

and maximize $\widehat{\Lambda}$ in the M step.

A more basic alternative is to use formula (2.4.8) which gives

$$L(\boldsymbol{\theta}, \boldsymbol{\theta}_0) \equiv \frac{g(\mathbf{X}, \boldsymbol{\theta})}{g(\mathbf{X}, \boldsymbol{\theta}_0)} = E_{\boldsymbol{\theta}_0} \left(\frac{f(\mathbf{X}, V, \boldsymbol{\theta})}{f(\mathbf{X}, V, \boldsymbol{\theta}_0)} | \mathbf{X} \right).$$

Again, we generate V_1, \dots, V_B (approximately) i.i.d. from the conditional distribution of V given \mathbf{X} under $\boldsymbol{\theta}_0$ and estimate $L(\cdot, \boldsymbol{\theta}, \boldsymbol{\theta}_0)$ by

$$\widehat{L}(\boldsymbol{\theta}, \boldsymbol{\theta}_0) = B^{-1} \sum_{b=1}^B \frac{f(\mathbf{X}, V_b, \boldsymbol{\theta})}{f(\mathbf{X}, V_b, \boldsymbol{\theta}_0)}. \quad (10.5.5)$$

Naively, we can view \widehat{L} as the object to be maximized and thus we apparently do not need to run a new sampler for each $\boldsymbol{\theta}$. This is, however, illusory since the approximation of L by \widehat{L} is poor if $\boldsymbol{\theta}$ is not close to $\boldsymbol{\theta}_0$, as we illustrate in the next example.

But this approach does make it clear that we can vary strategies, for instance, by replacing EM steps by other methods of optimizing $\widehat{L}(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}_{\text{OLD}})$ such as the Newton–Raphson method. An important feature of these approaches is that since the sampler, and, hence the objective function, change at each iteration, we no longer have the guaranteed increase in the objective function that the classical EM and other good algorithms give us.

Here is an application which illustrates the risks of naive application of this method and indicates the correct way of proceeding.

Example 10.5.3. Transformation model. Many models including the Cox proportional hazard model, so called frailty models, and a semiparametric extension of the Box-Cox (1964) transformation model can be put in the following framework: We observe

$$(Z_i, Y_i), \quad 1 \leq i \leq n, \quad \text{i.i.d. as } (Z, Y), \quad Z \in R^d, \quad Y \in R.$$

We postulate that there is an unknown monotone strictly increasing transformation $a : R \rightarrow R$ such that for $\theta \in R^d$,

$$a(Y) = \theta^T Z + \varepsilon, \quad 1 \leq i \leq n \quad (10.5.6)$$

where ε is independent of Z and has *known* distribution F_0 with density f_0 . We can interpret this by saying that, on an unknown scale a , the responses Y follow an ordinary linear model. Thus, we have a semiparametric model $\{P_{(\theta,a)} : \theta \in R^d, a \text{ increasing}\}$. It may be shown (Problem 10.5.3) that if

$$f_0(t) = e^{-t} e^{-e^{-t}}, \quad t \geq 0,$$

a Gumbel density, then the model is a reparametrization of the Cox proportional hazard model. Other distributions correspond to frailty models (Vaupel, Manton and Stallard (1979)) where a latent variable ξ is postulated for each individual and a conditional hazard rate for Y given by

$$\lambda(y, \theta | \mathbf{z}, \xi) = \xi r(\theta, \mathbf{z}) \lambda(y). \quad (10.5.7)$$

If we assume ξ comes from a fixed known distribution, then this is again a transformation model. See Problem 10.5.6.

From the structure of these models we see that, to test $\theta = \theta_0$, we can invoke invariance of the model under the group of monotone transformations as in Chapter 8, to conclude that it is reasonable to base tests on the likelihood $p_\theta(\mathbf{r}, \mathbf{z})$ of (Z_i, R_i) , $1 \leq i \leq n$, where R_1, \dots, R_n are the ranks of Y_1, \dots, Y_n . Because $\text{Rank}(a(Y_i)) = \text{Rank}(Y_i)$, we can without loss of generality set $a(y) = y$ in this likelihood. Moreover, because $p_\theta(\mathbf{r}, \mathbf{z}) = p_\theta(\mathbf{r} | \mathbf{z}) h(\mathbf{z})$ and $h(\mathbf{z})$ does not depend on θ , it is enough to consider the conditional density function of \mathbf{R} given $Z_i = \mathbf{z}_i$, $1 \leq i \leq n$, which, for general $f_\theta(y | \mathbf{z})$, is

$$L(\theta) \equiv P_\theta[\mathbf{R} = \mathbf{r} | \mathbf{z}_1, \dots, \mathbf{z}_n] = E_{\theta'} \left(\prod_{i=1}^n \frac{f_\theta(Y_i | \mathbf{z}_i)}{f_{\theta'}(Y_i | \mathbf{z}_i)} | \mathbf{R} = \mathbf{r} \right) \quad (10.5.8)$$

where θ' is a reference value of θ which is selected in what follows, and $f_\theta(y | \mathbf{z})$ is the conditional density of Y given $Z = \mathbf{z}$.

In the linear model (10.5.6) a natural choice for θ' is $\mathbf{0}$, in which case,

$$L(\theta) = P_\theta[\mathbf{R} = \mathbf{r} | \mathbf{z}_1, \dots, \mathbf{z}_n] = \frac{1}{n!} E_0 \left\{ \prod_{i=1}^n \frac{f_\theta(Y_{(r_i)} | \mathbf{z}_i)}{f_0(Y_{(r_i)})} \right\}, \quad (10.5.9)$$

which is *Hoeffding's formula*. See Problem 9.1.23. Note that under $\theta = \mathbf{0}$, the Y_i are i.i.d. and $(R_1, \dots, R_n), (Z_1, \dots, Z_n)$ are independent with $P_0[\mathbf{R} = \mathbf{r}] \equiv 1/n!$. Example 10.5.2 also has this structure, save that (10.5.6) is replaced by a parametric model and we concentrated on $(Y_{(1)}, \dots, Y_{(n)})$ rather than (R_1, \dots, R_n) . Now it seems reasonable to estimate θ by maximizing a Monte Carlo approximation to $L(\theta)$. From (10.5.9), this appears possible even without MCMC, and was proposed by Doksum (1987) for θ satisfying $|\theta|/\sigma_0 = O(n^{-\frac{1}{2}})$ where σ_0^2 is the variance of ϵ . In this approach:

Generate sets of n independent observations, $Y_{1b}^*, \dots, Y_{nb}^*$, $1 \leq b \leq B$, with Y_{jb}^* having density $f_0(\cdot)$, $1 \leq j \leq n$, and, given $\mathbf{R} = \mathbf{r}$, $Z = \mathbf{z}$, as data, estimate $L(\boldsymbol{\theta})$ by

$$\widehat{L}(\boldsymbol{\theta}) \equiv \frac{1}{Bn!} \sum_{b=1}^B \prod_{i=1}^n \left\{ \frac{f_{\boldsymbol{\theta}}(Y_{(r_i)b}^* | \mathbf{z}_i)}{f_{\mathbf{0}}(Y_{(r_i)b}^*)} \right\} \quad (10.5.10)$$

where $Y_{(1)b}^* \leq \dots \leq Y_{(n)b}^*$ are the ordered Y_{ib}^* .

This can be viewed as importance sampling or as an implementation of (10.5.5). Here

$$\text{Var } \widehat{L}(\boldsymbol{\theta}) = \frac{1}{B(n!)^2} \text{Var}_{\mathbf{0}} \left(\prod_{i=1}^n \frac{f_{\boldsymbol{\theta}}(Y_{(r_i)} | \mathbf{z}_i)}{f_{\mathbf{0}}(Y_{(r_i)})} \right)$$

which for fixed n tends to zero as $B \rightarrow \infty$. However, we face the usual importance sampling issue. For the linear model (10.5.6),

$$\begin{aligned} \text{Var}_{\mathbf{0}} \left(\prod_{i=1}^n \frac{f_{\boldsymbol{\theta}}(Y_i | \mathbf{z}_i)}{f_{\mathbf{0}}(Y_i)} \right) &= \prod_{i=1}^n E_{\mathbf{0}} \left(\frac{f_{\boldsymbol{\theta}}^2(Y_i | \mathbf{z}_i)}{f_{\mathbf{0}}^2(Y_i)} \right) - 1 \quad (10.5.11) \\ &= \prod_{i=1}^n \left(\int \left\{ \frac{f_0^2(y - \boldsymbol{\theta}^T \mathbf{z}_i)}{f_0(y)} \right\} dy \right) - 1. \end{aligned}$$

Here, as $n \rightarrow \infty$ for any fixed $\boldsymbol{\theta} \neq \mathbf{0}$, the variance goes to ∞ exponentially (Problem 10.5.1), although it is $O(1)$ for the $|\boldsymbol{\theta}|/\sigma_0 = O(n^{-\frac{1}{2}})$ case considered by Doksum (1987) (Problem 10.5.2).

There are various alternative ways to proceed taking advantage of special features of particular models — see Murphy (1995) for gamma frailty for instance. However, for linear regression transformation models, a general way is to use (10.5.8) and MCMC ideas as follows:

1. Let $\widehat{\boldsymbol{\theta}}_0$ denote the MLE for Y_1, \dots, Y_n distributed as $\prod_{i=1}^n f_0(y_i - \boldsymbol{\theta}^T z_i)$.
2. At the m th step, $m \geq 0$, generate B i.i.d. vectors (Y_{1b}, \dots, Y_{nb}) , $1 \leq b \leq B$, where Y_{1b}, \dots, Y_{nb} are distributed as $\prod_{i=1}^n f_0(y_i - \widehat{\boldsymbol{\theta}}_m^T z_i)$. Let $Y_{(1)b}^*, \dots, Y_{(n)b}^*$ denote Y_{1b}, \dots, Y_{nb} ordered.
3. Estimate $L(\boldsymbol{\theta})$ by

$$L_m^*(\boldsymbol{\theta}) = \frac{1}{B} \sum_{b=1}^B \prod_{i=1}^n \frac{f_{\boldsymbol{\theta}}}{f_{\widehat{\boldsymbol{\theta}}_{\text{OLD}}}}(Y_{(r_i)b}^* | \mathbf{z}_i), \quad (10.5.12)$$

where $\widehat{\boldsymbol{\theta}}_{\text{OLD}} = \widehat{\boldsymbol{\theta}}_m$. Maximize $L^*(\boldsymbol{\theta})$ to obtain $\widehat{\boldsymbol{\theta}}_{m+1}$ and proceed to step $m + 1$.

We call this algorithm the *likelihood sampler*. It produces a semi-parametric likelihood estimate of $\boldsymbol{\theta}$ for the transformation model where $(a(Y)|\mathbf{z}) \sim f_{\boldsymbol{\theta}}(y|\mathbf{z})$ for unknown increasing $a(\cdot)$. When f_0 is $\mathcal{N}(0, 1)$, this is an extension of the Box-Cox model that has $a(t) = (t^\lambda - 1)/\lambda$, $\lambda \in R$. \square

Summary. We demonstrated how MCMC and Gibbs sampling algorithms can be used in Bayesian analysis to compute approximations to estimates and credible intervals by generating data approximately distributed as $\mathcal{L}(\theta_1, \theta_2 | \mathbf{X})$ when $\mathcal{L}(\theta_1 | \mathbf{X}, \theta_2)$ and $\mathcal{L}(\theta_2 | \mathbf{X}, \theta_1)$ are known. We next showed, in examples, how to use MCMC to approximate maximum likelihood estimates in difficult situations. In particular we showed how a combination of the EM algorithm and MCMC can be used to compute maximum likelihood estimates in the travel time model; and we showed how MCMC can be used to compute maximum likelihood estimates in a transformation model with unknown monotone transformation.

10.6 Problems and Complements

Problems for Section 10.1

1. In Example 10.1.1 (a), suppose both G_1 and G_2 are χ^2_2 under H . Let $m = n - m = 15$. Use R or other software to generate the Monte Carlo and permutation distributions of the two-sample t -statistic. Use $M = 1000$ Monte Carlo trials and 1000 random permutations. Summarize the results in histograms and compare the results to the student t_{28} density of Section B.3.1.
2. In Example 10.1.2, suppose that F_0 is known.
 - (a) Show that by conditioning on $\bar{X}_{([n\alpha]+1)}$ and $\bar{X}_{(n-[n\alpha])}$ it is possible to express $E_0(\bar{X}_\alpha^2)$ in terms of one and two dimensional integrals.
 - (b) Use (a) and MATLAB or some other software to evaluate $E_0(\bar{X}_\alpha^2)$ when $n = 10$, $\alpha = 0.25, 0.50, 0.75$ and (i) F_0 is $\mathcal{N}(0, 1)$, (ii) F_0 is Laplace, $f_0(x) = \frac{1}{2} \exp\{-|x|\}$.

Problems for Section 10.2

1. Let $r = m/2$ and $s = n/2$ for positive integers m and n . Describe how Theorem B.2.3 of Vol. I can be used to generate a random variable with a $\beta(r, s)$ distribution.
2. Show that in Theorem 10.2.1, V is minimized by choosing p_0 proportional to $|h|p$.
3. (a) Establish (10.2.6) by showing that for this example, as $s, r \rightarrow \infty$,

$$\frac{c(p, p_0)}{\sqrt{r+s}} \longrightarrow \frac{1}{\sqrt{2\pi t(1-t)}}.$$

- (b) Show that rejective sampling is not useful if p_0 is uniform and r or $s < 1$.
 4. Let p denote the $\beta(r, s)$ density with $r, s > 1$. Set $m = [2r]$ and $n = [2s]$ where $[.]$ is the greatest integer function. Let $r_0 = m/2$ and $s_0 = n/2$ and let p_0 be the $\beta(r_0, s_0)$ density. Discuss the behavior of $c(p, p_0)$ as s and r varies, including the case where $s, r \rightarrow \infty$.
 5. Show how one can simulate a random variable with the density (10.2.7).
- Hint.* Let Z be the Bernoulli ($\frac{1}{2}$). Generate a value of Z . If $Z = 0$, let X be generated by

the density rx^{r-1} , $0 < x < 1$ (show how). If $Z = 1$, let X be generated by the exponential, $\mathcal{E}(1)$ density.

- 6.** The logistic distribution has df $F_L(x) = [1 + \exp\{-x\}]^{-1}$ and density

$$f_L(x) = \frac{e^{-x}}{(1 + e^{-x})^2}, \quad -\infty < x < \infty.$$

- (a) Describe how the simple Monte Carlo technique can be used to generate X_1, \dots, X_n with density f_L .

Hint. Find $F_L^{-1}(\cdot)$.

- (b) Describe how rejective sampling and (a) preceding can be used to generate X_1, \dots, X_n with density

$$f(x) = \frac{af_L(x)}{1 + e^{-|x|}}, \quad -\infty < x < \infty, \quad a > 0.$$

- (c) Use MATLAB or other software to find a in (b) preceding. Then find $E(\tau)$ where τ is as in Theorem 10.2.2.

- (d) Carry out rejective sampling as in (b) preceding to obtain an observed sample x_1, \dots, x_{100} from f . You may use existing software. How long did the procedure take? Draw a histogram of your results.

- 7.** Suppose X has density

$$f(x) = \frac{ae^{-|x|}}{1 + e^{-|x|}}, \quad a > 0, \quad x \in R.$$

- (a) Describe how rejective sampling can be used to generate X .

- (b) Use MATLAB or other software to find the constant a .

- 8.** Let f and g be continuous case density functions. Assume there exists a finite M such that $f(y) \leq Mg(y)$ for all y . To obtain a random sample of size n from the density function $f(x)$ using the Rejection Method Algorithm, simulate Y_1, \dots, Y_N from the density $g(x)$. Then record the accepted sample $\{X_1, \dots, X_n\}$ and the rejected sample $\{Z_1, \dots, Z_{N-n}\}$. Let E_f denote expectation with respect to f .

- (a) Are $\delta_1 = n^{-1} \sum_{i=1}^n h(X_i)$ and $\delta_2 = N^{-1} \sum_{i=1}^N h(Y_i)f(Y_i)/g(Y_i)$ unbiased estimators of $E_f h(X)$?

- (b) Find the marginal density of Z_i .

- (c) Suppose $N > n$. Is the estimator

$$\delta_3 = \frac{1}{N-n} \sum_{j=1}^{N-n} \frac{(M-1)f(Z_j)h(Z_j)}{Mg(Z_j) - f(Z_j)}$$

and, unbiased estimator of $E_f h(X)$? Explain.

9. Suppose X_1, \dots, X_n are i.i.d. P where $P \in \mathcal{P} = \{P_\theta : \theta \in R\}$.

(a) Assume \mathcal{P} is regular in the sense that $\psi(x, \theta) \equiv \partial/\partial\theta \log p(x, \theta)$ satisfies conditions A1–A4, A6 in Section 5.4.2 of Volume I. Let $\theta_n = t/\sqrt{n} + \theta_0$ and $L(X_1, \dots, X_n) \equiv \sum_{i=1}^n (\log p(X_i, \theta_n) - \log p(X_i, \theta_0))$. Show that for X_1, \dots, X_n i.i.d.

$$\mathcal{L}_{\theta_0}(L(X_1, \dots, X_n)) \implies \mathcal{N}\left(-\frac{t^2}{2}I(\theta_0), t^2I(\theta_0)\right),$$

where $I(\cdot)$ denotes Fisher information.

Hint. See Examples 7.1.1 and 7.1.2.

(b) Suppose that $|\partial l/\partial\theta^j| \psi(X, \theta)|$ are uniformly bounded for θ in a compact set and $1 \leq j \leq 3$. Given a bounded function $g_n : \mathcal{X}^n \rightarrow R$, $|g_n(X_1, \dots, X_n)| \leq M < \infty$, we wish to estimate $E_n \equiv E_{\theta_n} g_n(X_1, \dots, X_n)$. Suppose it is easy to generate i.i.d. observations from P_{θ_0} and we use the importance sampling estimate

$$\hat{E}_n \equiv \frac{1}{B} \sum_{b=1}^B g_n(X_{1b}^*, \dots, X_{nb}^*) \exp\{L(X_{1b}^*, \dots, X_{nb}^*)\},$$

where $\{X_{ib}^* : 1 \leq i \leq n\}$ are i.i.d. from P_{θ_0} , $1 \leq b \leq B$.

Show that $\hat{E}_n = E_n \xrightarrow{P} 0$ as n and B tend to infinity.

Hint. $\text{Var}(\hat{E}_n) \leq C/B$ for some constant C independent of n . (In (b), if necessary, you may assume that the given boundedness condition implies the conditions for part (a).)

Problems for Section 10.3

1. Establish (10.3.6), (10.3.7), and Theorem 10.3.1 for the parameter $\theta(P) = g(\int x dP(x))$, where we assume g''' exists and $|g'''| \leq M$, $E|X_1|^4 \leq M$ for some constant $M > 0$. Formally show that if $B \asymp n^2$, then

$$\begin{aligned} \text{BIAS}_n^{(B)}(\hat{P}_n) &= \text{BIAS}_n(\hat{P}_n) + O_P((Bn)^{-\frac{1}{2}}) \\ \text{BIAS}_n^B(\hat{P}_n) &= \text{BIAS}_n(P) + O_P(n^{-\frac{3}{2}}) \\ \widehat{SD}_n^{(B)} &= SD_n(\hat{P}_n) + O_P(n^{-1}B^{-1}) \\ SD_n(\hat{P}_n) &= SD_n(P) + O_P(n^{-\frac{3}{2}}). \end{aligned}$$

Hint. For (10.3.6), use the definition of the bootstrap based on B replications. For (10.3.7), use the expansion

$$\theta(P^*) = \theta(\hat{P}_n) + g'(\bar{X})(\bar{X}^* - \bar{X}) + \frac{g''(\bar{X})}{2}(\bar{X}^* - \bar{X})^2 + \frac{g'''(\bar{X})}{6}(\bar{X}^* - \bar{X})^3 + O_P(n^{-\frac{3}{2}}).$$

$$\text{BIAS}_n^{(B)}(\hat{P}_n) = E^*[\theta(P^*) - \theta(\hat{P}_n)] + \frac{\hat{\sigma}^2}{n} + O_P(n^{-\frac{3}{2}}).$$

- 2.** Suppose $g : R^d \rightarrow R^p$ has total differential $g^{(1)}(\mathbf{x}) = \|(\partial g_i / \partial x_j)(\mathbf{x})\|_{p \times d}$ at $\mathbf{x} = \boldsymbol{\mu} \equiv E\mathbf{X}_1$. Generalize Theorem 10.3.2 (b) to

$$\mathcal{L}^*(\sqrt{n}(\mathbf{g}(\bar{\mathbf{X}}^*) - \mathbf{g}(\bar{\mathbf{X}}))) \rightarrow \mathcal{N}(\mathbf{0}, \mathbf{g}^{(1)}(\boldsymbol{\mu}) \text{Var}(\mathbf{X}_1) \mathbf{g}^{(1)}(\boldsymbol{\mu})) .$$

Hint. See Theorem 5.3.4.

- 3.** Assume X_1, \dots, X_n i.i.d. P . In Example 10.3.1, find the bias of $\hat{\sigma}_B^2 = \sigma^2(\hat{P}_n) - \text{BIAS}_n^{(B)}(\hat{P}_n)$. Assume that $0 < \sigma^2(P) < \infty$. Compare your result to the bias $-\sigma^2(p)/n$ of the empirically bias corrected estimate.

- 4.** Establish (10.3.4).

- 5.** Establish (10.3.5).

- 6.** In Section 10.3.3, show that if $\sqrt{n}[\hat{\mu}_n - \mu(P)]$ tends to a Gaussian distribution and the difference between the bootstrap quantiles and population quantiles are of order $O_P(n^{-\frac{1}{2}})$, then Efron's percentile bootstrap $\mu_{n[B(1-\alpha)+1]}^*$ with $B = \infty$ is an asymptotic $(1-\alpha)$ UCB.

- 7.** Let $K_4(P) = E_P X^4 - 3(E_P(X))^2$. Show that $EK_4(\hat{P}_n) = K_4(P) + O(n^{-1})$.

- 8.** *The Mallows' metric.* Let \mathcal{F} be the class of df's with $\int x dF(x) = 0$ and $0 < \int x^2 dF(x) < \infty$. The *Mallows' distance* ρ between $F, G \in \mathcal{F}$ is defined by

$$\rho^2(F, G) = \inf_P E_P(X - Z)^2 ,$$

where the inf is over bivariate probabilities P for (X, Z) with marginals F and G . A relevant result is *Fréchet's theorem*:

$$\rho^2(F, G) = \int_0^1 [F^{-1}(u) - G^{-1}(u)]^2 du ,$$

Let \xrightarrow{w} denote weak convergence.

Mallows' (1972) lemmas:

1. $\rho(F_k, G) \rightarrow 0$ iff $F \xrightarrow{w} G$ and $\int x^2 dF_k(x) \rightarrow \int x^2 dG(x)$.
 2. If $F_k = \mathcal{L}(\sum_{i=1}^k a_i X_i)$ where X_1, \dots, X_n are independent with $\mathcal{L}(X_i) = F_i \in \mathcal{F}$ and $\sum a_i^2 = 1$, then $F_k \in \mathcal{F}$ and $\rho^2(F^k, \Phi) \leq \sum a_i^2 \rho^2(F_i, \Phi)$, where Φ is the $\mathcal{N}(0, 1)$ df.
- (a) Show that ρ is a metric.
(b) Use Mallows' Lemma 2 above to show that if X_1, \dots, X_k and Z_1, \dots, Z_k are i.i.d. from F and G , respectively, and if F_k, G_k are the df's of

$$S_k = k^{-\frac{1}{2}} \sum_{i=1}^k [X_i - E(X_i)], \quad T_k = k^{-\frac{1}{2}} \sum_{i=1}^k [Z_i - E(Z_i)] ,$$

then $\rho(F_k, G_k) \leq \rho(F, G)$.

- (c) Suppose X_1, \dots, X_n are i.i.d. as $X \sim F$, where $\sigma^2 = \text{Var}(X) < \infty$. Show that $n^{-\frac{1}{2}}(\bar{X}^* - \bar{X})$ converge in law in probability to $\mathcal{N}(0, \sigma^2)$.

Hint: By (b), the distance between $\mathcal{L}^*(n^{\frac{1}{2}}(\bar{X}^* - \bar{X}))$ and $\mathcal{L}(n^{\frac{1}{2}}(\bar{X} - \mu))$ is bounded by $\rho(F, \hat{F}_n)$. We know that $\mathcal{L}(n^{\frac{1}{2}}(\bar{X} - \mu)) \rightarrow \mathcal{N}(0, \sigma^2)$ by the central limit theorem.

Remark. Let $|\cdot|$ denote the Euclidean metric. Bickel and Freedman (1981) define the Mallows' metric on the class \mathcal{P} of probabilities P on R^d with $E_P|\mathbf{X}|^2 < \infty$ as follows:

$$\rho_\alpha(P, Q) = \inf E\{|\mathbf{X} - \mathbf{Z}|^\alpha\}^{\frac{1}{\alpha}},$$

where $1 \leq \alpha < \infty$ and the inf is over all joint distributions of (\mathbf{X}, \mathbf{Z}) with marginals P and Q . They show that $\rho_2^2(P_k, P) \rightarrow 0$ is equivalent to “ $P_k \xrightarrow{w} P$ and $E_{P_k}|\mathbf{X}|^2 \rightarrow E_P|\mathbf{X}|^2$ ” and they show that if $d = 1$ and (X, Z) has marginals (F, G) , then $\rho_1(F, G) = \int |F(t) - G(t)| dt$.

9. Establish Theorem 10.3.2 for $d > 1$.

Hint: Use the uniform SLLN and the multivariate Lindeberg-Feller theorem. See Appendix D.1.

10. Suppose X_1, \dots, X_n are i.i.d. $\mathcal{N}(\mu, \sigma^2)$. Set $T = \sum_{i=1}^n (X_i - \bar{X})^2$. Use available software to plot the exact and bootstrap distribution of T when $n = 10$ and $\sigma = 1$. Use $B = 100$ and $B = 500$.

11. Suppose $B_n \rightarrow \infty$ as $n \rightarrow \infty$. In Example 10.3.3, show that $\bar{X} - \tilde{d}_{n\alpha}^{(B_n)}$ is an asymptotic $(1 - \alpha)$ UCB.

12. In Example 10.3.4, show that the simultaneous confidence region $\hat{F}(x) \pm c_n/\sqrt{n}$ has level $(1 - \alpha)$ for all distributions F .

Hint. By example 7.1.6, $\sqrt{n}|\hat{F}(x) - F(x)|$ converges in law to $W^0(F(x))$. Note that $\sup_x W^0(F(x)) \leq \sup_u W^0(u)$.

13. Establish (10.3.26).

14. For the model of Example 10.3.5, show that $n[\max(X_1, \dots, X_n) - \max(X_1^*, \dots, X_n^*)]$ does not converge in law in probability.

15. In Example 10.3.5 show how an appropriate choice of $m(n)$ with $\frac{m(n)}{n} \rightarrow 0$ and $m(n) \rightarrow \infty$ in the m out of n bootstrap makes $\lambda_{m(n)}(P) = \lambda(P) + o(1)$ and the “variance” $\lambda_n(P_n) - \lambda_n(P)$ arbitrarily small.

16. Let X_1, \dots, X_n be i.i.d. as F and let \hat{F}_n denote the empirical distribution. Let $X_n^* = (X_1^*, \dots, X_n^*)$ be a bootstrap sample drawn with replacement from \hat{F}_n .

(a) Suppose F has mean μ , unit variance, finite third absolute moment. Define

$$G_n(x, F) = P[\sqrt{n}(\bar{X} - \mu) \leq x].$$

Similarly, define the bootstrap counterpart $G_n^*(x, F_n) = P^*[\sqrt{n}(\bar{X}^* - \bar{X}) \leq x]$, where P^* denotes bootstrap probability given \hat{F}_n . Prove that as $n \rightarrow \infty$,

$$P\left(\sup_x |G_n^*(x, F_n) - G_n(x, F)| \rightarrow 0\right) = 1.$$

Hint: You may use the Berry-Esséen theorem. See Section A.15.

- 17.** Show that when $\hat{\theta} = \bar{X}$, the jackknife estimate of $\text{Var}(\hat{\theta})$ is unbiased.
- 18.** Show that when $\hat{\theta} = n^{-1} \sum_{i=1}^n (X_i - \bar{X})^2$, the expected value of the jackknife bias estimate is $E(\text{BIAS}_n^{(J)}) = -(n-1)^{-1}\sigma^2$.
- 19.** Show that for linear statistics of the form

$$T(\mathbf{X}) = a + n^{-1} \sum_{i=1}^n h(X_i),$$

The jackknife values $T(\mathbf{x}_{-i})$, $1 \leq i \leq n$, can be used to determine the value of $T(\mathbf{x}^*)$ for any bootstrap sample \mathbf{x}^* .

Hint: Set $t_i = T(\mathbf{x}_{-i})$ and solve the equations

$$t_i = a + n^{-1} \sum_{j \neq i} h_j, \quad 1 \leq i \leq n,$$

for h_1, \dots, h_n . Use this to find the value of $T(\mathbf{x}^*)$ for any bootstrap sample $\mathbf{x}^* = (x_1^*, \dots, x_n^*)^T$.

- 20.** Establish (10.3.12) formally.

Problems for Section 10.4

- 1.** Assume the conditions to Appendix D.5.D. Show that X_1, \dots, X_n of Section 10.4.1 will be approximately distributed as p for $M \rightarrow \infty$ provided (a) and (b) of Section 10.4.1 hold.

Hint: See Appendix D.5.

- 2.** Show that a symmetric kernel K_0 has the uniform distribution as its stationary distribution.

- 3.** Under the conditions of Proposition 10.4.2, show that the joint distribution of (U_m, V_m) in a step of Gibbs sampler obeys detailed balance.

- 4.** Consider the Markov chain with state space $\{1, 2\}$ and kernel

$$K = \begin{pmatrix} a & 1-a \\ 1-b & b \end{pmatrix}.$$

Compute the upper bounds in Corollary 10.4.1.

- 5.** In Example 10.4.2, show that the conditional distribution of σ^{-2} given μ and X_1, \dots, X_n is $\Gamma(p + \frac{1}{2}n, \lambda)$, where λ is given by (10.4.14).

- 6. (a)** Assume the conditions of Theorem 10.4.1. Suppose (X_1, X_2) is drawn from the distribution with joint probability $K(x_1, x_2)p(x_1)$. Let $\lambda_1, \dots, \lambda_N$ denote the eigenvalues of \mathbf{K} ordered so that $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_N$. Show that

$$\min\{\text{Corr}(g(X_1), g(X_2)) : g \text{ arbitrary}\} = \min_{j \geq 2} \lambda_j = \lambda_N \equiv \rho_-.$$

(b) Deduce that

$$\max\{\text{Corr}(g(X_1), h(X_2)) : g, h \text{ arbitrary}\} = \max\{\rho_+, -\rho_-\}.$$

Hint: $E[g(X_1)h(X_2)] = E[g(X_1)E(h(X_2)|X_1)].$

7. Show that if the state space is finite and the transition matrix \mathbf{K} satisfies detailed balance, then \mathbf{K} has equal right and left eigenvalues, all being real.

8. In Example 10.4.3, show that $X^{(k+1)} \sim \mathcal{N}(0, \rho^4 \sigma_k^2 + (1 - \rho^4))$.

9. Establish (10.4.17) using Lagrange multipliers.

10. Assume the conditions of Theorem 10.4.1. Show that

$$\max \{\text{corr}(g(X_1), g(X_{k+1}))\} = (\rho_+)^k.$$

Hint. Note that by Theorem B.1.1,

$$E[g(X_1)g(X_{k+1})] = E[g(X_1)E(g(X_{k+1})|X_1)]. \quad (10.6.1)$$

Let \mathbf{K} be the transition matrix of a chain with state space $\{1, 2, \dots, N\}$ and let $p_{ij}(n)$ be the (i, j) element of \mathbf{K}^n . Then use (10.6.1) to show that the eigenvalues of \mathbf{K}^n are the n th powers of eigenvalues of \mathbf{K} .

11. Prove Theorem 10.4.2.

Hint. Using Problem 10.4.10,

$$\begin{aligned} \text{Var}\left(\sum_{b=1}^B g(X_b)\right) &= \sum_{b=1}^B \text{Var } g(X_b) + 2 \sum_{a=1}^B \sum_{b < a} \text{Cov}(g(X_a), g(X_b)) \\ &= B \text{Var } g(X_1) + 2 \sum_{a=1}^B \sum_{b=1}^{a-1} \rho_+^{a-b} \asymp B \text{Var } g(X_1) \frac{1 + \rho_+}{1 - \rho_+}. \end{aligned}$$

12. Show that the Markov Chain transition matrix $\mathbf{K} = \begin{pmatrix} a & 1-a \\ 1-b & b \end{pmatrix}$ has second largest eigenvalue $\lambda_2 < 0$ iff $(a+b) < 1$.

Problems for Section 10.5

1. Show that the expression (10.5.11) tends to infinity at an exponential rate as $n \rightarrow \infty$. That is, let Δ_n denote the right hand side of (10.5.11). Then show that $n^{-1} \log \Delta_n \rightarrow c$ for some $c > 0$.

2. In the transformation model (10.5.6) with f_0 the $\mathcal{N}(0, \sigma_0^2)$ density, set $\mu_i = \boldsymbol{\theta}^T z_i$, $\sigma_0^2 = \text{Var}(\epsilon)$. Assume that $\max |\mu_i|/\sigma_0 \rightarrow 0$ as $n \rightarrow \infty$ and that $\sigma_0^{-2} \sum_{i=1}^n \mu_i^2 \leq M$ for some fixed $M > 0$. Show that the variance (10.5.11) is $O(1)$.

3. Show that the model (10.5.6) with f_0 a Gumbel density is equivalent to the Cox proportional hazard model.

Hint. See Problem 1.1.13.

4. Implement the likelihood sampler (10.5.12) of Example 10.5.3 on the computer for the transformation model (10.5.6) when f_0 is the $\mathcal{N}(0, 1)$ density. Describe the algorithm. Generate Y_1, \dots, Y_{50} from the model where $a(t) = \log t$, $E(a(Y)) = 1 + Z\theta$, $\theta = 1$, $Z \sim U(0, 1)$. Compute $L_1(\theta)$ and $L_2(\theta)$ for $\theta = .5 + \delta$, $\delta = 0, 0.1, 0.2, \dots, 1$. Plot $L_1(\theta)$ and $L_2(\theta)$ at these points.

5. Use the EM-MCMC approach of Example 10.5.2 to estimate θ in the transformation model of Example 10.5.3. Describe an algorithm.

6. Suppose that Y and ξ are independent continuous random variables, and that (10.5.7) holds. Show that if $\xi \sim G$, then the model for Y can be written as (10.5.6) for some F_0 .

Hint. $\Lambda(y|\mathbf{z}, \xi) = \xi r(\theta, \mathbf{z})\Lambda(y)$ where $\Lambda(y|\mathbf{z}, \xi) = -\log[1 - P(Y \leq y|\mathbf{z}, \xi)]$ and $\Lambda(y) = -\log[1 - H(y)]$ for some continuous df H . Solve for $P(Y \leq y|\mathbf{z}, \xi)$.

10.7 Notes

Note for Section 10.1

(1) We speak of generating independent observations by machine. This is not possible since computers are deterministic. But many methods have been devised for generating “pseudo random” numbers ($\mathcal{U}(0, 1)$ observations) which behave as if they were random for most usual purposes. For more on this topic see, for instance, Ripley (1987).

Note for Section 10.3

(1) Efron’s presentation in Efron (1979) is much less abstract and focussed on the estimation of bias and variance, while this presentation follows more that of Bickel and Freedman (1981). But the basic idea is in Efron’s paper. See also Efron and Tibshirani (1993) and Bickel, Götze and van Zwet (1997).

Note for Section 10.4

Another generalization of rejective sampling which uses (Y_1, Y_2, \dots) to generate X_1, X_2, \dots, X_B , which, though not necessarily independent, does have distribution p , called *perfect sampling*, is also coming into use.

Chapter 11

NONPARAMETRIC INFERENCE FOR FUNCTIONS OF ONE VARIABLE

11.1 Introduction

We saw in Chapter 9 that to construct estimates of low dimensional Euclidean parameters in semiparametric models we need to estimate infinite dimensional parameters such as curves. Examples are densities and derivatives of densities on the one hand and conditional expectations on the other. These quantities, which will be considered in this chapter and the next, are not defined for discrete distributions and, so, as we saw in Section I.5 we need to consider regularized approximations to these parameters before we can construct empirical plug-in estimates. This is not a consequence of our needing to estimate functions but rather of the irregular nature of these parameters. For instance, we can estimate distribution functions using plug-in estimates without regularizing.

To indicate the usefulness of the estimates in this chapter, we cite Example 9.3.1 where efficiency (and even consistency for arbitrary underlying distributions) requires us to estimate the derivatives of the logs of the densities. Similarly, in Example 9.1.10 we need estimates of the regression functions $u \rightarrow E(Y|U = u)$ and $u \rightarrow E(Z|U = u)$. A very important class of situations where we are led to function estimation is nonparametric classification or, more generally, prediction. This class of problems is also loosely categorized under the topic “machine learning.” Such topics, including the estimation of the nonparametric regression $\mu(\mathbf{z}) = E(Y|\mathbf{Z} = \mathbf{z})$ and some of its semiparametric versions when \mathbf{z} is d -dimensional, will be covered in Chapter 12.

We introduce and analyze in this chapter nonparametric estimates of a continuous case density $f(x)$, $x \in R$, and estimates of the nonparametric regression curve $\mu(x) = E(Y|X = x)$ for a continuous bivariate pair $(X, Y) \in R^2$. Key methods and asymptotic expressions are introduced. We assume all distributions are continuous since discrete cases are “easy.”

11.2 Convolution Kernel Estimates on R

There are many uses for density estimates in visualization and, more generally, exploratory data analysis. For instance suppose we are contemplating using a gamma distribution to model the distribution of the failure times of a new type of computer chip. Comparing a nonparametric density estimate visually with a fitted gamma density is one way of exploring possible models. For more details, see Silverman (1986), Scott (1992), Wand and Jones (1995), and Loader (1999). The methods of this chapter can also be used to estimate the hazard rate discussed in Section 9.1.

We turn to the simplest methods of density estimation for real valued random variables and more generally vector valued variables: histogram and kernel density estimation.

Consider first the problem we discussed in Example 7.1.5, estimation of the density (in the continuous sense) f of a real valued random variable X on the basis of X_1, \dots, X_n , which are i.i.d. as X . The models \mathcal{P} we shall consider are nonparametric, in the sense that any probability distribution may be approximated weakly by members of \mathcal{P} , but members of \mathcal{P} have densities which range over a set \mathcal{F} which is restricted in some way. For instance, we may consider $\mathcal{F}_r = \{f : \|f^{(r)}\|_\infty < \infty\}$ for $r \geq 1$. More generally we can consider Sobolev spaces,

$$\mathcal{F} = \{f : \int |f^{(j)}(t)|dt < \infty, \quad 0 \leq j \leq r\},$$

and, more generally still, spaces which permit discontinuities (Besov spaces; see Daubechies (1993) for instance). When f is not continuous it is not unambiguously defined. For simplicity we shall assume in this section that f is continuous and positive on intervals of the form $(-\infty, \infty)$, $(-\infty, b]$, $[a, \infty)$ or a closed interval $[a, b]$ with $f(x) = 0$ for $x \notin [a, b]$. This interval $S(f)$ is called the *support* of f and the upper and lower bounds are called *boundary points*. It is relatively difficult to find good estimates of $f(x)$ when x is at or near one of the boundary points.

Within this context we consider local problems such as estimating $f(x)$ at points x and global ones such as estimating $f(\cdot)$ as a function. We will occasionally use the notation $f(P; x)$ for $f(x)$ and $f(P; \cdot)$ for the function to remind ourselves that these are parameters defined on \mathcal{P} . Recall that the histogram estimator of Section I.5 is obtained by approximating $f(P; x)$ by $h^{-1}P(I(x))$ where for integers j , $I_j = (jh, (j+1)h]$, $-\infty < j < \infty$, and $I(x) = I_{j(x)}$ is the unique interval containing x ; and then plugging in the empirical probability \widehat{P} for P . Now $f_h(\widehat{P}; \cdot)$ is a discontinuous function and intuitively should be improvable if we assume f is smooth. To demonstrate this we introduce a famous easily analyzed family of estimates called the *kernel estimates*. These can be approached by taking histogram estimates and recentering them at each x . Specifically, approximate $f(x)$ by the continuous function

$$f_h(x) \equiv f_h(P; x) \equiv \frac{1}{2h}P(x-h, x+h] \tag{11.2.1}$$

and plug-in \widehat{P} to get the density estimate $\widehat{f}_h(x) = f_h(\widehat{P}; x)$.

Let F denote the distribution function of X and set $J_h(u) = h^{-1}J(u/h)$ with $J(u) = \frac{1}{2}\mathbf{1}[-1 \leq u \leq 1]$. Then it is easy to see that, if $S(f) = R$ (Problem 11.2.1),

$$f_h(P; x) = \int J_h(x - z)dF(z), \quad \hat{f}_h(x) = f_h(\hat{P}; x) = \int J_h(x - z)d\hat{F}(z).$$

That is, $f_h(\cdot)$ is the convolution of a uniform density $J_h(\cdot)$ on $[-h, h]$ with $f(\cdot)$ and $\hat{f}_h(x)$ is the empirical plug-in estimate of $f_h(x)$.

Convoluting $f(\cdot)$ with a density K gives a function with at least the smoothness of K , and the resulting estimate can be expected to be better if f itself is as smooth as K . More generally, for any function $K(\cdot)$ with $\int K(u)du = 1$, we consider regularizations of f of the form

$$f_h(x) = f_h(F; x) = \int K_h(x - z)dF(z) = EK_h(x - Z) \quad (11.2.2)$$

where $Z \sim F$, $K_h(u) = h^{-1}K(u/h)$ and we identify P with the distribution function F . If $K(\cdot)$ is a density and $S(f) = R$, then (11.2.2) is the density of $X + hV$ where $V \sim K(\cdot)$ and V is independent of X . See Problem 11.2.2. The plug-in estimate of f is then

$$\hat{f}_h(x) = f_h(\hat{F}; x) = n^{-1} \sum_{i=1}^n K_h(x - X_i) \quad (11.2.3)$$

where \hat{F} is the empirical df based on X_1, \dots, X_n i.i.d. from F .

A function $K(u)$ with $\int K(u)du = 1$ is called a *kernel function*. Thus all densities are kernels, but kernels need not be non-negative. We shall sometimes refer to the estimates $f_h(\hat{F}, x)$ with $f_h(F; x)$ of the form (11.2.2) as *convolution kernel estimates* to distinguish them from generalized *kernel estimates* of the form

$$f(\hat{P}, x) = \int K(x, z)d\hat{P}(z). \quad (11.2.4)$$

The condition

$$\int K(x, z)dx = 1$$

and $K(x, z) \geq 0$ for all z is required for $f(\hat{P}, \cdot)$ to be a density.

In addition to the uniform kernel $J(u)$, common kernels are the $\mathcal{N}(0, 1)$ density $\varphi(u)$, the *Epanechnikov kernel*,

$$K_E(u) = \frac{3}{4}(1 - u^2)\mathbf{1}[-1 \leq u \leq 1],$$

and the *quartic kernel*,

$$K_Q(u) = 15(1 - u^2)^2\mathbf{1}[-1 \leq u \leq 1]/16.$$

This can be put in the general regularization framework of Chapter 9 by considering the sequence of parameters $f_h(P; x)$ which tend to $f(x)$ as $h \downarrow 0$ but which for fixed h are estimated by the ‘‘plug-in’’ method where P is replaced by the empirical probability \widehat{P} .

Let the *support* of K be the closure of the set $\{u : K(u) \neq 0\}$, then

Lemma 11.2.1. *Let $f_h(\cdot)$ be as defined in (11.2.2). If either of the following holds:*

- (a) *f is continuous at x , x is in interior of $S(f)$, and K has compact support,*
- (b) *f is continuous at x and bounded on R ,*

then, as $h \rightarrow 0$,

$$f_h(x) \rightarrow f(x).$$

Proof. Note that $\int K(u)du = 1$ permits us to write

$$\begin{aligned} f_h(x) - f(x) &= \frac{1}{h} \int K\left(\frac{x-z}{h}\right) [f(z) - f(x)] dz \\ &= \int K(u)[f(x-hu) - f(x)] du \end{aligned} \quad (11.2.5)$$

by changing variable to $u = (x - z)/h$. Since, for all M ,

$$\sup\{|f(x-hu) - f(x)| : |u| \leq M\} \rightarrow 0$$

as $h \rightarrow 0$ and because, under (a), $K(u) = 0$, $|u| > M$, for some $M > 0$, our claim follows for (a). For (b), see Problem 11.2.3. □

We proceed to study the plug-in estimate $\widehat{f}_h(x)$ given by

$$\widehat{f}_h(x) \equiv \int K_h(x-z)d\widehat{F}(z) = \frac{1}{nh} \sum_{i=1}^n K\left(\frac{x-X_i}{h}\right). \quad (11.2.6)$$

Lemma 11.2.2. *If K is bounded and has compact support, then if $h = h_n \rightarrow 0$ and $nh_n \rightarrow \infty$ as $n \rightarrow \infty$,*

$$\widehat{f}_h(x) \xrightarrow{P} f(x)$$

at all x which are points of continuity of f .

Proof. We can show that $E\widehat{f}_h(x) \rightarrow f(x)$ and $\text{Var}\widehat{f}_h(x) \rightarrow 0$ as $n \rightarrow \infty$. Consistency follows. See Problem 11.2.4.

11.2.1 Uniform Local Behavior of Kernel Density Estimates

We analyze the bias and variance of $\hat{f}_h(x)$ at a point x under the assumption of a bounded third derivative f''' . This strong assumption gives uniformity of our convergence statements in $f \in \mathcal{F}_0 \subset \mathcal{F}$ and in x restricted to some finite interval $[c, d]$ contained in the support of f . Without such uniformity our asymptotic analyses comparing the performance of estimates are suspect since depending on f , the n for which an approximation to precision ε is valid might be arbitrarily great. Let

$$m_j(K) = \int u^j K(u) du, \quad (11.2.7)$$

and for some constant $M \in (0, \infty)$, let

$$\mathcal{F}_3 = \{f : S(f) = [a, b], f \text{ three times differentiable on } (a, b), \|f'''\|_\infty \leq M\}.$$

Proposition 11.2.1. *Let K be a kernel with $m_1(K) = 0$, $m_j(K)$ finite, $1 \leq j \leq 3$. Then, as $h \rightarrow 0$, uniformly in $x \in [c, d] \subset (a, b)$, $f \in \mathcal{F}_3$, and n ,*

$$E\hat{f}_h(x) = f_h(x) = f(x) + \frac{1}{2}m_2(K)f''(x)h^2 + O(h^3). \quad (11.2.8)$$

Proof. Evidently, $E\hat{f}_h(x) = f_h(x)$. For $x \in [c, d]$, Taylor expand $f(x - hu)$ about $h = 0$, to get, if $|u| \leq M_1$ for some constant $M_1 \varepsilon (0, \infty)$,

$$f(x - hu) = f(x) - huf'(x) + \frac{1}{2}h^2u^2f''(x) - \frac{1}{6}h^3u^3f'''(x^*)$$

for $|x - x^*| \leq |hu| \leq hM_1$. Next substitute this expression into (11.2.5) to obtain the result. \square

Note that $m_j(K)$ is finite for all j if K has compact support, and that $m_1(K) = 0$ if K is symmetric about 0. Also note that the bias does not depend on n , but the result holds if, in particular, we let $h = h_n \rightarrow 0$ as $n \rightarrow \infty$. As with histogram estimates, the bias tends to 0 as $h \rightarrow 0$, but we now have a faster rate and uniformity of convergence of the bias based on the assumption of a bounded third derivative. If we assume more derivatives we can control the bias further if we do not restrict to nonnegative K . See Problem 11.2.5.

Next we turn to the variance of \hat{f}_h . Let

$$\nu_j(K) = \int u^j K^2(u) du, \quad \nu(K) = \nu_0(K), \quad (11.2.9)$$

and for $M > 0$, let $\mathcal{F}_1 = \{f : S(f) = [a, b], f \text{ is differentiable on } (a, b) \text{ and } \|f'\|_\infty \leq M\}$.

Proposition 11.2.2. Assume that $\nu_j(K) < \infty$, $j = 0, 1$. Then, as $n \rightarrow \infty$, uniformly for $x \in [c, d] \subset (a, b)$, $f \in \mathcal{F}_1$,

$$\text{Var } \widehat{f}_h(x) = (nh)^{-1} \nu(K) f(x) + O(n^{-1}). \quad (11.2.10)$$

Proof. By (11.2.6),

$$\begin{aligned} \text{Var } \widehat{f}_h(x) &= (nh)^{-2} n \text{Var } K\left(\frac{x-X}{h}\right) \\ &= n^{-1} h^{-2} \left[\int K^2\left(\frac{x-z}{h}\right) f(z) dz - h^2 f_h^2(x) \right]. \end{aligned}$$

Change variable to $u = (x-z)/h$ and rewrite the first term as

$$\begin{aligned} (nh)^{-1} \int K^2(u) f(x-hu) du &= (nh)^{-1} f(x) \nu(K) \\ &\quad + (nh)^{-1} \int K^2(u) (f(x-hu) - f(x)) du. \end{aligned} \quad (11.2.11)$$

Argue as in Proposition 11.2.1 to complete the proof (Problem 11.2.6). □

As with histogram estimates, $\text{Var } \widehat{f}_h(x)$ is proportional to $f(x)$ and is small if the density is small. Further, the variance is proportional to $(nh)^{-1}$ and tends to 0 as $n \rightarrow \infty$ for fixed h and to ∞ as $h \rightarrow 0$ for fixed n .

Since, by Proposition 1.3.1, $\text{MSE} = \text{VAR} + (\text{BIAS})^2$, we have established

Corollary 11.2.1. Under the conditions of Propositions 11.2.1 and 11.2.2, uniformly for $x \in [c, d] \subset (a, b)$, $f \in \mathcal{F}_3$, as $n \rightarrow \infty$ and $h \rightarrow 0$

$$\text{MSE}[\widehat{f}_h(x)] = (nh)^{-1} \nu(K) f(x) + \frac{1}{4} h^4 m_2^2(K) [f''(x)]^2 + O(n^{-1} + h^5). \quad (11.2.12)$$

The bias-variance tradeoff is clear: Small h leads to small bias and large variance while large h has the opposite effect. By choosing $h = h_n$ such that $h_n \rightarrow 0$ and $nh_n \rightarrow \infty$, we have consistency of $\widehat{f}_h(x)$. The sum of the first two terms is sometimes called the asymptotic MSE $[\widehat{f}_h(x)]$ (AMSE $[\widehat{f}_h(x)]$). Set $A(h) = \text{AMSE}[\widehat{f}_h(x)]$. Then by solving $A'(h) = 0$ for h we find that when $f(x) > 0$ and $0 < [f''(x)]^2 < \infty$, the minimizer of $A(h)$ is of the form $h_{\text{opt}} = cn^{-\frac{1}{5}}$ for some $c(f) > 0$ (Problem 11.2.10). Substituting h_{opt} in (11.2.12) gives

$$\inf_h \text{MSE}[\widehat{f}_h(x)] \asymp n^{-\frac{4}{5}},$$

a rate that is slower than the rate n^{-1} for estimation of regular parameters.

11.2.2 Global Behavior of Convolution Kernel Estimates

We can measure the distance, $\rho(f, g)$, between two curves f and g in many ways, e.g. $\|f - g\|_2 \equiv (\int (f - g)^2)^{\frac{1}{2}}$, $\|f - g\|_p \equiv (\int |f - g|^p)^{\frac{1}{p}}$, $p \geq 1$, and $\|f - g\|_\infty \equiv \sup |f(x) - g(x)|$. For most types of density estimates, all of these, other than $\|f - g\|_\infty$, behave in the same way qualitatively in terms of rate of convergence to 0. We focus on $\|f - g\|_2^2$. Then the loss, using \hat{f} , is the Integrated Squared Error $\int_{-\infty}^{\infty} (\hat{f} - f)^2(x)dx$ and the risk is the Integrated MSE,

$$\begin{aligned}\text{IMSE}(\hat{f}, f) &= \int_{-\infty}^{\infty} E(\hat{f}_h(x) - f(x))^2 dx \\ &= \int_{-\infty}^{\infty} [E\hat{f}_h(x) - f(x)]^2 dx + \int_{-\infty}^{\infty} \text{Var}\hat{f}_h(x) dx.\end{aligned}\quad (11.2.13)$$

For IMSE we invoke slightly stronger conditions than for MSE. For some constant $M > 0$, let

$$\mathcal{F}_3^I = \left\{ f_3 \in \mathcal{F}_3 : \int_a^b [f'']^2(x) dx < \infty, \int_a^b |f'''|^2(x) dx \leq M \right\}.$$

Theorem 11.2.1. Suppose the conditions of Corollary 11.2.1 hold save that \mathcal{F}_3 is replaced by \mathcal{F}_3^I . Then, uniformly for $f \in \mathcal{F}_3^I$,

$$\text{IMSE}(\hat{f}_h, f) = (nh)^{-1}\nu(K) + \frac{h^4}{4}m_2^2(K) \int_{-\infty}^{\infty} [f'']^2(x) dx + O(n^{-1} + h^5). \quad (11.2.14)$$

Proof. We use Laplace's form of the remainder in the Taylor expansion of $f(x - hu)$:

$$f(x - hu) = f(x) - huf'(x) + \frac{(hu)^2}{2}f''(x) - \frac{(hu)^3}{2} \int_0^1 (1 - \lambda)^2 f'''(x - \lambda hu) d\lambda.$$

Then, by (11.2.5) and $\int uK(u) = 0$,

$$\begin{aligned}E\hat{f}_h(x) - f(x) &= \frac{1}{2}h^2m_2(K)f''(x) \\ &\quad - \frac{h^3}{2} \int_{-\infty}^{\infty} \int_0^1 u^3 K(u)(1 - \lambda)^2 f'''(x - \lambda hu) d\lambda du.\end{aligned}\quad (11.2.15)$$

The square of the integral term in (11.2.15) is bounded by

$$\left\{ \int_{-\infty}^{\infty} |u^3 K(u)| \int_0^1 (1 - \lambda)^2 |f'''(x - \lambda hu)|^2 d\lambda du \right\}^2$$

which, by the inequality $EU \leq E^{\frac{1}{2}}(U^2)$ applied to $U = |f'''(x - \Lambda hu)|$, with Λ having density $3(1 - \lambda)^2$ on $(0, 1)$, is bounded by

$$\frac{1}{3} \left\{ \int_{-\infty}^{\infty} |u^3 K(u)| \left[\int_0^1 3(1 - \lambda)^2 [f'''(x - \lambda hu)]^2 d\lambda \right]^{\frac{1}{2}} du \right\}^2 \quad (11.2.16)$$

which, because $\int_0^1 (1-\lambda)^2 |f'''(x - \lambda hu)|^2 d\lambda \leq \int_{-\infty}^{\infty} |f'''(t)|^2 dt$, in turn is bounded by

$$\frac{1}{3} \left(\int_{-\infty}^{\infty} |u^3 K(u)| du \right)^2 \left\{ \int_{-\infty}^{\infty} |f'''(x)|^2 dx \right\}. \quad (11.2.17)$$

It follows from (11.2.15) that the squared bias is $\frac{1}{4} h^4 m_2 \int |f''(x)|^2 dx + O(h^5)$. Next, using (11.2.11) (Problem 11.2.11),

$$\int_{-\infty}^{\infty} \text{Var} \widehat{f}_h(x) dx = (nh)^{-1} \int_{-\infty}^{\infty} K^2(u) du + O(n^{-1}) \quad (11.2.18)$$

and the result follows. \square

Note that the bias-variance tradeoff for IMSE is the same as for MSE and that the first two terms of (11.2.14) are of the form $A(h) = c_1(nh)^{-1} + c_2 h^4$. To minimize $A(h)$, we set the derivative equal to zero and find (Problem 11.2.12)

$$h_{\text{opt}} = \{\nu(K)/m_2(K)\nu(f'')\}^{\frac{1}{5}} n^{-\frac{1}{5}} \quad (11.2.19)$$

provided $0 < \nu(f'') = \int [f''(x)]^2 dx < \infty$. Substituting this in (11.2.14) gives

$$\inf_{h>0} \text{IMSE}[\widehat{f}(\cdot)] = \frac{5}{4} \{m_2(K)\nu(K)\nu(f'')\}^{\frac{1}{5}} n^{-\frac{4}{5}} + o(n^{-\frac{4}{5}}). \quad (11.2.20)$$

That is, as with MSE, assuming a well behaved f''' , the order of uniform convergence over the family \mathcal{F}_3 is $n^{-\frac{4}{5}}$.

11.2.3 Performance and Bandwidth Choice

As we have noted, an immediate observation from the preceding results is that the order of convergence of MSE and IMSE is strictly worse than the n^{-1} we have seen in the case of regular parameters. This is not a feature of a local vs. global focus nor of curve vs. Euclidean parameter estimation since the distribution function is a regular parameter.

In fact, performance depends quite specifically on the type of f we consider. We have already seen (Section I.5) that with histogram estimates if we only assume $0 < |f'(z)| \leq M < \infty$ for z in a neighborhood of x the fastest pointwise convergence locally is $n^{-2/3}$ and this continues to be true locally with kernel estimates. Moreover the optimal bandwidth is of the form $cn^{-1/3}$. On the other hand, if we specify greater smoothness for f , for instance, $0 < |f^{(2k+1)}(z)| \leq M < \infty$, $k > 2$, for z in a neighborhood of x we can use K which are sometimes negative to achieve the rate $n^{-2k/(2k+1)}$ using bandwidth $cn^{-1/(2k+1)}$. See Problem 11.2.9.

The major difficulty that has to be faced is the choice of the bandwidth h_n . The trouble is that even if we are willing to accept the order of magnitude $n^{-1/5}$ (see (11.2.19)) dictated by the amount of smoothness for f that we are willing to assume, the optimal c in $h_{\text{opt}} = cn^{-\frac{1}{5}}$ for both MSE and IMSE depends on f'' . For instance, from (11.2.19), h_{opt}

in this case depends on $\int [f''(x)]^2 dx$ which, unfortunately, is harder to estimate than f and requires more smoothness assumptions. We are caught in a no win situation. If we believe that more than two derivatives of f are available we should use bandwidth of larger order than $n^{-1/5}$ and obtain a final rate better than $n^{-4/5}$, but that requires the estimation of $\int [f^{(4)}(x)]^2 dx$. However, if $f^{(4)}(\cdot)$ exists, we should use an even larger order bandwidth, and so on. There are a number of ways out. The one we consider soundest is cross validation which does not require a smoothness assumption. We discuss this method further in Section 12.5.

A simple bandwidth choice if we are willing to assume a bounded f'' is to employ the h_{opt} appropriate for a *reference distribution* (Bickel and Doksum (1977)), such as the Gaussian distribution and replacing σ by the sample standard deviation s . For $F = \mathcal{N}(\mu, \sigma^2)$, we find $\nu(f'') = 3/(8\sqrt{\pi}\sigma^5)$ and h_{opt} is $c(K)n^{-1/5}\sigma$ with $c(K)$ equal to 1.95, 2.34, and 1.06 for the $\mathcal{U}[-1, 1]$, Epanechnikov, and $\mathcal{N}(0, 1)$ kernels, respectively. With this choice $\hat{f}_h(x)$ will achieve the optimal rate of convergence for $f \in \mathcal{F}_3$ provided $0 < f''(x) < c$ and $0 < \sigma^2 < \infty$.

11.2.4 Discussion of Convolution Kernel Estimates

Perhaps the strongest point in favor of convolution kernel estimates is that they are easy to analyze theoretically and that they can easily be extended to the multivariate case. However, the convolution kernel estimate (11.2.6) has a weakness if f has compact support $[a, b]$ because for x near the boundary points a and b , (11.2.6) with a symmetric kernel will lead to biases by trying to account for impossible observations. A skew kernel which reduces the bias near a will worsen the bias near b , and vice versa. A possible solution is to permit the shape of the kernel to vary with x , that is, use a generalized kernel $K(x, z)$ as introduced in (11.2.4). Put another way, in replacing the density f of X with that of $X + hV$, we will not limit ourselves to X, V independent. A simple method leading to such solutions is based on using a local parametric approximation to $f(x)$ and using the plug-in method to estimate the parameters in this approximation. We consider this approach in Section 11.3.1. Other methods such as penalization can also lead to estimates insensitive to the support.

Summary. We consider estimation of a continuous case density $f(\cdot)$ and because this is an irregular parameter where empirical plug-in estimates are not directly available, we first approximate $f(\cdot)$ by a regular parameter $f_h(\cdot)$, $h > 0$, that converges to $f(\cdot)$ as $h \downarrow 0$. This regular parameter is the convolution of $f(\cdot)$ with a function $K_h(u) = h^{-1}K(u/h)$ where the *kernel* K satisfies $\int K(t)dt = 1$. The constant h is called the *bandwidth* and is an example of a *tuning parameter*. By applying the empirical plug-in principle to $f_h(\cdot)$ we obtain the convolution kernel estimate $\hat{f}_h(\cdot)$. We develop uniform asymptotic properties of the bias, variance, and mean squared error (MSE) of $\hat{f}_h(x)$ first for fixed x and next for the mean integrated squared error (MISE). We find that the MSE and IMSE tend to zero and $\hat{f}_h(\cdot)$ is consistent provided $nh \rightarrow \infty$ as $n \rightarrow \infty$. Assuming a bounded third derivative for MSE and a bounded integrated absolute third derivative for IMSE, we find that the fastest possible rate for uniform convergence of MSE and IMSE to zero is $n^{-4/5}$, which is obtained

when $h = O(n^{-\frac{1}{5}})$. In this section we discuss a simple approach which asks for optimality at a reference density. We postpone an effective data based method for selecting h called cross validation until Section 12.5.

11.3 Minimum Contrast Estimates: Reducing Boundary Bias

In this section we extend the minimum contrast approach of Volume I to density estimation. This extended contrast approach also applies to other curves and to a variety of classes of regular approximating functions to these curves. Here we show that this approach produces density estimates with desirable properties.

Let \mathcal{F} be a nonparametric class of densities (e.g. densities that satisfy the conditions of Corollary 11.2.1), and let \mathcal{F}_s denote a class of regular approximating functions $f_s(\cdot, \beta)$, $\beta \in R^p$, which contains the constant functions. Here $s \in S$ is a tuning constant such as h whose choice was discussed above and we will return to later. We introduce a *discrepancy*

$$D_{t,x}(f(\cdot), f_s(\cdot; \beta)) > 0$$

between $f(x)$ and $f_s(x, \beta)$ where $t \in T$ is another tuning constant. Typically, either S or T is empty. For fixed s and t , we assume that there is a unique minimizer $\beta_{s,t}(F, x)$ of D , and we assume that for each x in the interior (a, b) of the support $S(f)$ of f

$$\inf\{D_{t,x}(f(\cdot), f_s(\cdot; \beta_{s,t}(F, x))) : s \in S, t \in T\} = 0.$$

We also assume that $\beta_{s,t}(F, x)$ depends on F only, that is, not on f . Our estimate of β is $\hat{\beta} = \beta_{s,t}(\hat{F}, x)$ and the estimate of $f(x)$ is $f_s(x, \hat{\beta})$. Note that we will get the same solution if we use the global discrepancy

$$D_t(f(\cdot), f_s(\cdot; \beta)) = E_F D_{t,X}(f(\cdot), f_s(\cdot; \beta))$$

because local minimization implies global minimization, cf. Proposition 3.2.1.

Example 11.3.1. *Linear kernel estimates as minimum contrast estimates.* We first consider x fixed and assuming f is continuous we want an approximation $f(z; \beta)$ that is close to $f(z)$ for z in a neighborhood of x , where $\beta = \beta(x)$ depends on x . One way is to select a kernel $K \geq 0$ and to find $\beta(x)$ that minimizes the local discrepancy

$$\begin{aligned} D_{h,x}(f(\cdot), f(\cdot; \beta)) &= \frac{1}{W(x)} \int_{S(f)} \{[f(z) - f(z; \beta)]^2 K_h(z - x)\} dz \quad (11.3.1) \\ &= E_{Q_h} \{[f(Z) - f(Z; \beta(X))]^2 | X = x\} \end{aligned}$$

where (X, Z) has joint df Q_h with density $q_h(x, z) = f(x)q_h(z|x)$, for $x, z \in S(f)$, and

$$q_h(z|x) = \frac{K_h(z - x)1[a \leq z \leq b]}{W(x)}$$

with

$$W(x) = \int_a^b K_h(z-x)dz = \int_{(a-x)/h}^{(b-x)/h} K(u)du.$$

That is, $q_h(z|x)$ is the conditional density of $Z = X + hV$ given that $X = x$ and that $X + hV$ is in $[a, b]$, where $V \sim K$. Note that the global D_h measures how well $f(Z; \beta(X))$ predicts $f(Z)$.

By equating the derivative of $D_{h,x}$ with respect to β to zero, we formally obtain a minimizer $\beta_h(F, x)$ and then the estimators $\hat{\beta}_h(x) = \beta_h(\hat{F}, x)$ and $\hat{f}_h(x) = f(x; \hat{\beta}_h(x))$. We can use the empirical plug-in principle because minimizing $D_{h,x}$ is equivalent to minimizing

$$R(F; \beta) \equiv -2 \int_{S(f)} f(z; \beta) K_h(z-x)dF(z) + \int_{S(f)} f^2(z; \beta) K_h(z-x)dz.$$

We call $R(\hat{F}; \beta)$ a *contrast function* and $\hat{f}_h(x)$ a *minimum contrast estimate*, cf. Section 2.1.1.

As a simple illustration, take $\{f(z; \beta), \beta \in R^p\}$ to be the class of constants $\{\beta, \beta \in R\}$. In this case we find, provided $E_F K_h(X - x) > 0$ and $x \pm h \in S(f)$, that the unique minimizer of D is

$$\beta_h(F, x) = E_F K_h(X - x), \quad x \pm h \in S(f).$$

When K is symmetric, this coincides with the convolution kernel approach (11.2.2) when $x \pm h \in S(f)$, but when $x \pm h$ is not in $S(f)$, that is, x is in the boundary regions of $S(f)$, then the *minimum contrast estimate* is consistent (Problem 11.3.2) and the convolution estimate is not (Problem 11.3.1). A key feature of the minimum discrepancy approach is that it automatically leads to boundary corrections. We next elaborate on this remark. \square

Boundary points

Suppose $S(f) = [a, b]$ with a and b finite and $2h < b-a$. We say that x is in the *interior region* of $S(f)$ if $x-h$ and $x+h$ both are in $[a, b]$. In the asymptotic theory of Section 11.2, we could argue that for any x in (a, b) we can choose $h = h_n$ small enough so that this condition holds. However, for finite n as well as when $n \rightarrow \infty$ the kernel estimator (11.2.6) may be severely biased for x in the *boundary regions* $[a, a+h]$ and $(b-h, b]$. In fact for x_h of the form $a + \lambda h$ or $b - \lambda h$, $0 < \lambda < 1$, the kernel estimator (11.2.6) is asymptotically biased when K has compact support (Problem 11.3.1).

However, if we use the minimizer of (11.3.1) with $f(z; \beta)$ a constant, and we assume that K is a density with support $[-1, 1]$, then this asymptotic bias of the plug-in estimate disappears. We can write the unique solution to $dD/d\beta = 0$ for $x \in (a, b)$ as (Problem 11.3.3)

$$f_h(x) = E_Q[f(Z)|X = x] = \int_a^b K_h(z-x)dF(z) / \int_{(a-x)/h}^{(b-x)/h} K(u)du, \quad (11.3.2)$$

which leads to a generalized kernel estimate of the form (11.2.4). For symmetric K , it corresponds to the density of a random variable of the form $X + hV$, $V \sim K$, where V depends on X , i.e., we no longer have a convolution as in (11.2.2).

Let $\hat{f}_h(x)$ be the plug-in estimate using (11.3.2). It can be shown (Problem 11.3.2) that for x_h in the boundary region, $\text{Bias } \hat{f}_h(x_h) = O(h)$. The rate $O(h)$ at which the bias tends to zero is slower than the rate $O(h^2)$ obtained for interior points. A remedy is to use a locally linear approximation.

Example 11.3.1 continued: *Locally linear and polynomial estimates.* Suppose $S(f) = [a, b]$ with a and b finite. Let $K(u)$ be a nonnegative symmetric kernel with support $[-1, 1]$. We find $\alpha = \alpha(x, F)$ and $\beta = \beta(x, F)$ that minimize the local discrepancy

$$\int_a^b \{f(z) - [\alpha + \beta(z-x)]\}^2 K_h(z-x) dz \quad (11.3.3)$$

from the locally linear fit $g(z-x) = \alpha + \beta(z-x)$ to $f(z)$. The locally linear approximation to $f(x)$ is

$$f_h(x) = g(0) = \alpha(x, F) + \beta(x, F)(x-x) = \alpha(x, F),$$

Setting the derivative of the quadratic in (11.3.3) with respect to α and β equal to zero and simplifying, $\alpha(x, F)$ and $\beta(x, F)$ will, for $x \in (a, b)$, be the solutions to the normal equations

$$\begin{aligned} \int_a^b K_h(z-x) dF(z) &= \alpha m_0(x, h) + \beta h m_1(x, h) \\ \int_a^b (z-x) K_h(z-x) dF(z) &= \alpha h m_1(x, h) + \beta h^2 m_2(x, h) \end{aligned} \quad (11.3.4)$$

where

$$m_j(x, h) \equiv \int_{(a-x)/h}^{(b-x)/h} u^j K(u) du, \quad j = 0, 1, 2.$$

By the argument above $f_h(x, \hat{F}) \equiv \alpha(x, \hat{F})$ is an estimate of $f(x)$. If $a+h \leq x \leq b-h$, or, equivalently, $a-x \leq -h, b-x \geq h$, then $m_1(x, h) = 0$ by the symmetry of K and K vanishing outside $[-1, 1]$. Thus, on this range,

$$f_h(x, \hat{F}) = \int K_h(x-z) d\hat{F}(z), \quad (11.3.5)$$

the usual convolution kernel estimate. However, for $a \leq x < a+h$ and $b-h < x \leq b$ we obtain, solving (11.3.4) (see Problem 11.3.5),

$$f_h(x, \hat{F}) = \int K_h(x, z) d\hat{F}(z) \quad (11.3.6)$$

where

$$K_h(x, z) = \frac{m_2(x, h) - m_1(x, h)(z - x)/h}{m_0(x, h)m_2(x, h) - m_1^2(x, h)} K_h(z - x). \quad (11.3.7)$$

Note that $K_h(x, z)$ is not a convolution kernel, and is sometimes negative.

For $M > 0$, let $f \in \mathcal{F}_2 \equiv \{f \text{ on } [a, b] : |f''(x)| \leq M \text{ for all } x \in [a, b]\}$, where f'' is one sided at a and b .

Proposition 11.3.1. *Suppose that the conditions of Propositions 11.2.1 and 11.2.2 hold. Then, uniformly for $f \in \mathcal{F}_2$ and $x \in [c, d] \subset (a, b)$,*

$$E f_h(x, \hat{F}) = f(x) + O(h^2), \text{ as } h \rightarrow 0 \quad (11.3.8)$$

$$\text{Var} f_h(x, \hat{F}) = O((nh)^{-1}), \text{ as } n \rightarrow \infty. \quad (11.3.9)$$

The proof of these assertions and the asymptotic MSE of $f_h(x, \hat{F})$ are left to Problem 11.3.5. Note also that since $\beta(x) = f'(x)$, $\beta(x, \hat{F})$ gives us an estimate of $f'(x)$ which is given in Problem 11.3.5 as well. We give extensions to estimates based on local polynomial approximations in Problem 11.3.6.

Remark 11.3.1 The conditional expectation version of (11.3.1), and Theorem 1.4.3 with $(f(Z), Z - X)$ in place of (Y, Z) and with $\mathcal{L}(f(Z), Z - X|X = x)$ in place of $\mathcal{L}(Y, Z)$, give,

$$\begin{aligned} \alpha(x, F) &= E[f(Z)|X = x] - \beta(x, F)E(Z - X|X = x), \\ \beta(x, F) &= \frac{\text{Cov}(f(Z), (Z - X)|X = x)}{\text{Var}(Z - X|X = x)}. \end{aligned}$$

That is, $\alpha(x; F)$ is the value of the optimal MSPE linear predictor of $f(Z)$ at $z = x$ when $\mathcal{L}(Z|X = x)$ is $q_h(z|x)$. This representation gives a different derivation of $f_h(x; \hat{F})$ (Problem 11.3.7) and shows the connection to prediction.

Remark 11.3.2. The locally linear estimate of Example 11.3.1 can be computed as follows: Let $\{x^{(1)}, \dots, x^{(g)}\}$ denote a set of grid points in $[\hat{a}, \hat{b}]$ where $\hat{a} = x_{(1)}$ and $\hat{b} = x_{(n)}$. For a given $x^{(k)}$ determine whether it is in the left or right boundary region or the interior region of $[\hat{a}, \hat{b}]$. Thus if $x^{(k)}$ is in the left boundary region, $x^{(k)} = \hat{a} + \lambda h, 0 < \lambda < 1$, determine $\lambda^{(k)} = (x^{(k)} - \hat{a})/h$, and use $K_h(x, z)$ with

$$m_j(x, h) = \int_{-1}^{\lambda^{(k)}} u^j K(u) du.$$

A point $x^{(k)}$ in the central part uses the kernel $K(u)$, and the right boundary point $x^{(k)}$ uses $K_h(x, z)$ with $m_j(x, h) = \int_{-\gamma^{(k)}}^1 u^j K(u) du$ and $\gamma^{(k)} = (\hat{b} - x^{(k)})/h$. This procedure

yields $\{(x^{(k)}, \hat{f}(x^{(k)})); k = 1, \dots, g\}$. These points are then connected using a smoothing routine and the final output is a smooth curve.

Example 11.3.2. *Series expansions.* Suppose f is in $L_2(a, b)$, for $-\infty \leq a \leq b \leq \infty$. Thus, $\int_a^b f^2(x)dx < \infty$. Let $\{B_j(\cdot) : j \geq 1\}$ be a complete orthonormal basis of L_2 ,

$$\int_a^b B_j(x)B_k(x)dx = 1(j = k),$$

and let

$$\beta_j(F) = \int_{-\infty}^{\infty} f(x)B_j(x) dx.$$

Then,

$$f(x) = \sum_{j=1}^{\infty} \beta_j(F)B_j(x) \quad (11.3.10)$$

in the sense that $\int_a^b (f(x) - \sum_{j=1}^m \beta_j(F)B_j(x))^2 dx \rightarrow 0$ as $m \rightarrow \infty$ (L_2 convergence). For example, if $a = -\pi$, $b = \pi$, we can consider the Fourier basis,

$$\begin{aligned} B_k(x) &= \pi^{-\frac{1}{2}} \cos kx, & k > 0 \\ &= (2\pi)^{-\frac{1}{2}}, & k = 0 \\ &= \pi^{-\frac{1}{2}} \sin kx, & k < 0 \end{aligned}$$

since $\int_{-\pi}^{\pi} \cos^2 kx dx = \pi$. Another example is the Legendre polynomials. If $a = -\infty$, $b = \infty$, we can consider the Hermite functions defined by

$$B_k(x) = \varphi(x)H_k(x) = \frac{(-1)^k}{c_k} \frac{d^k}{dx^k} \varphi(x).$$

Here φ is the $\mathcal{N}(0, 1)$ density, H_k are the Hermite polynomials, and c_k makes

$$\int H_k^2(x)\varphi^2(x)dx = 1$$

(see also Section 5.3). Other important bases are the wavelet bases — see Donoho and Johnstone (1995) for instance. For all these cases, we can estimate $\beta_k(F)$ by empirical plug-in,

$$\hat{\beta}_k = \beta_k(\hat{F}) = \frac{1}{n} \sum_{i=1}^n B_k(X_i). \quad (11.3.11)$$

Unfortunately, while, by Plancharel's theorem,

$$\int f^2(x)dx = \sum_{k=1}^{\infty} \beta_k^2(F) < \infty,$$

$\beta_k(\widehat{F}) \not\rightarrow 0$ as $k \rightarrow \infty$ in general. Thus, the direct plug-in approach for $f(x)$ as in (11.3.10) is impossible. On the other hand, minimizing the discrepancy

$$D(f(\cdot), f_m(\cdot, \beta)) = \int [f(x) - \sum_{j=0}^m \beta_j B_j(x)]^2 dx$$

yields $\beta_k(\widehat{F})$ as in (11.3.11) and the preceding plug-in estimates $\widehat{\beta}_1, \widehat{\beta}_2, \dots$, as well as the density estimate

$$\widehat{f}_m(x) = \sum_{j=0}^m \widehat{\beta}_j B_j(x) = \int K_m(x, z) d\widehat{F}(z)$$

where

$$K_m(x, z) = \sum_{j=0}^m B_j(x) B_j(z).$$

It may be shown that if $[a, b]$ is compact and B_0 is constant, then $\int K_m(x, z) dx = 1$ (Problem 11.3.8). Analysis of such estimates is not as simple as that of convolution kernels and they are necessarily sometimes negative. We refer to Silverman (1986) and Viollaz (1989) for further discussion. This idea becomes much more important and attractive in connection with regression. Note that, as with convolution kernel estimates, a crucial “tuning parameter,” m in this instance, needs to be chosen.

□

Kernel estimates, despite their formal simplicity, have been computationally unattractive since a separate summation has to be carried out for each x in a grid of x -values. However, advances in computer technology and the availability of software (Loader, 1999, Chapter 5) have reduced this disadvantage. Orthogonal series estimates do not share this problem since the $B_j(\cdot)$ can be stored once and for all and $m << n$ but, as we have indicated, are hard to analyze and can be negative. There are many nonlinear (nongeneralized kernel) methods of density estimation which are competitive with kernel density estimates. We consider these in a general discussion of regularization in the next section.

Summary. We consider a systematic approach to constructing regular approximations to $f(\cdot)$ which consists of minimizing a *discrepancy* between $f(\cdot)$ and a parametric approximation $f(\cdot; \beta)$, $\beta \in R^p$, to $f(\cdot)$. The minimizer will be regular if a local quadratic discrepancy is used. The empirical version of the minimizer is called a minimum *contrast estimate*. We illustrate this approach using $f(\cdot; \beta) \equiv \beta$, $\beta \in R$, and linear approximations $f(\cdot; \beta)$ to $f(\cdot)$ over local neighborhoods of x . We show that this approach yields estimates of f with reduced bias for boundary points x_h of the form $a + \lambda h$ and $b - \lambda h$ where $0 < \lambda < 1$ and a and b are the left and right boundaries of the support $S(f)$ of f . In particular, by plugging the empirical distribution into the linear minimizer of a local quadratic discrepancy based on bandwidth h , we obtain an estimate with bias of order $O(h^2)$ for all $x \in (a, b)$ including boundary points. The variance of this estimate is of order $O((nh)^{-1})$. We also

briefly consider estimates based on approximating $f(\cdot)$ by a truncated series expansion and show how minimizing a quadratic discrepancy leads to a natural series expansion estimate of $f(\cdot)$.

11.4 Regularization and Nonlinear Density Estimates

11.4.1 Regularization and Roughness Penalties

What we call regularization is related in spirit to the notion of regularization introduced by Tikhonov (1963) in applied mathematics. His problem was, essentially, given an integral equation of the form $Kf = g$ with $g \in K(\mathcal{F}_0)$ and \mathcal{F}_0 a “nice” set of functions (not necessarily densities), to find numerical approximations $f_N \in \mathcal{F}_0$ to f which converged to f as $N \rightarrow \infty$. \mathcal{F}_0 being “nice” corresponded to smoothness, for instance f twice differentiable on $(0, 1)$ with $J(f) \equiv \int_0^1 [f''(x)]^2 dx < \infty$. Simply discretizing K to a matrix K_N and g_N to a vector \mathbf{g} , defining $\mathbf{f}_N = K_N \mathbf{g}_N$, and extrapolating \mathbf{f}_N to $f_N \in \mathcal{F}_0$ would not work because the matrix K_N was too ill conditioned. Tikhonov proposed minimizing $|K_N \mathbf{f} - \mathbf{g}|^2 + \lambda J_N(\mathbf{f})$ for $\lambda = \lambda_N \rightarrow 0$ as $N \rightarrow \infty$, where $J_N(\mathbf{f})$ penalizes \mathbf{f} for an extrapolation with $J(f)$ too large. In particular, Tikhonov considered the natural numerical approximation to $J(f)$, $N^{-1} \sum_{i=1}^N (f(\frac{i+2}{N}) - 2f(\frac{i+1}{N}) + f(\frac{i}{N}))^2$. Passing to the limit, as $N \rightarrow \infty$, for fixed $\lambda > 0$ we see that we have defined f_λ as the solution of the problem: Minimize $\int_0^1 (Kf - g)^2 + \lambda \int_0^1 [f''(x)]^2 dx$. Under weak conditions on K , discretizations f_{λ_N} of this problem converge to f_λ in strong senses for fixed $\lambda > 0$, and $f_\lambda \rightarrow f$ as $\lambda \rightarrow 0$. By letting $\lambda_N \rightarrow 0$ sufficiently slowly, we have $f_{\lambda_N} \rightarrow f$.

As we indicated in Sections I.5 and 9.1, we think of regularization as follows: We are given a parameter $\theta(P)$ such as $\theta(P) \equiv f(\cdot)$, the continuous case density, which is defined on a model \mathcal{P}_0 , for instance, all P with twice differentiable densities f such that $\int |f''(x)|^2 dx < \infty$, but not on $\mathcal{M} \equiv$ all distributions. Since we cannot plug the empirical distribution \hat{P} into $\theta(P)$, we construct a sequence $\theta_s(P)$, $s \in R$ such that

- (i) θ_s is defined on \mathcal{M} .
- (ii) $\theta_s(P) \rightarrow \theta(P)$ on \mathcal{P}_0 as $s \rightarrow \infty$.

Then use $\theta_{\hat{s}}(\hat{P})$ with the tuning parameter \hat{s} chosen using the data.

We have already noted that if $\theta(P) = f(\cdot)$, then the parameter corresponding to convolution kernel estimation is

$$f_s(x) \equiv s \int K(s[z-x]) dP(z)$$

where $s \equiv 1/h$, $h \rightarrow 0$. As we also noted the log likelihood $\int \log f(x) d\hat{P}(x)$ is not maximized for $P \in \mathcal{P}_0$ since the supremum is infinite and is formally achieved as $f(x)$ looks more and more like $n^{-1} \sum_{i=1}^n \delta(x - X_i)$, where δ is the Dirac function. What Tikhonov regularization, as he introduced it, naturally corresponds to in this case is penalized maximum likelihood introduced by Good and Gaskins (1978) and Tapia and Thompson (1978).

Penalized maximum likelihood (PLM)

Let \mathcal{F} be a space of densities and let $J : \mathcal{F} \rightarrow R^+$ be a “roughness penalty.” In analogy to Tikhonov we define a population parameter $f(\cdot, F)$ as the minimizer f_λ of

$$\int \log f_\lambda(x) dF(x) + \lambda J(f_\lambda). \quad (11.4.1)$$

The corresponding estimate is just $f_\lambda(\cdot, \hat{F})$. Good and Gaskins (1978) took, for f on R , $J(f) = \int \{[f'(x)]^2/f(x)\} ds$. A computationally simpler proposal, to take $J(f) = \int [f''(x)]^2 dx$, consistent with the Tikhonov penalty and with proposals made by Wahba (1969) for nonparametric regression, was made by de Montricher, Tapia and Thompson (1975). See also Wahba (1990). The solution to (11.4.1) for $J(f) = \int [f''(x)]^2 dx$, if we restrict f to densities on $[a, b]$ with $f^{(j)}(a) = f^{(j)}(b) = 0, j = 0, 1$, is a quadratic spline with one continuous derivative and knots at $a, x_{(1)} < \dots < x_{(n)}, b$. Here a *spline* is a continuous piecewise polynomial function with the *knots* $t_1 < \dots < t_m$ being the points where the polynomial changes. That is, the spline f is a different polynomial of the same degree over each $(t_k, t_{k+1}]$, and $f^{(j)}(t_k^-) = f^{(j)}(t_k^+), k = 1, \dots, m$, some $j \geq 0$. See de Boor (1978) for an extensive discussion. For proofs of these claims see Tapia and Thompson (1978), p.106. We do not pursue this subject further here.

11.4.2 Sieves. Machine Learning. Log Density Estimation

A general method of regularization, introduced in Section 9.1, is the following: Let \mathcal{F} be a general (NP) class of functions. Then,

- (i) Find a sequence of parametric models (usually but not necessarily nested), $\mathcal{F}_k \equiv \{f_k(\cdot, \boldsymbol{\theta}_k) : \boldsymbol{\theta}_k \in R^{d_k}\}$ such that every $f \in \mathcal{F}$ is the limit as $k \rightarrow \infty$ in some norm of $f_k(\cdot, \boldsymbol{\theta}_k(f))$ (for some $\boldsymbol{\theta}_k(f) \in R^{d_k}$), where $f_k \in \mathcal{F}_k$.
- (ii) Select a method of estimating $\boldsymbol{\theta}_k$ from the data, under the assumption that $f_k \in \mathcal{F}_k$ holds. Typically this method is maximum likelihood. If specifying f_k does not completely specify a model, such as a regression framework where the distribution of the error is not assumed to be of specified parametric form, least squares or some other contrast function criteria may be used.
- (iii) Determine, using a data based criterion or otherwise, a suitable sequence k_n which tends to ∞ as $n \rightarrow \infty$.
- (iv) Act as if \mathcal{F}_{k_n} is true. That is, estimate $f(\cdot)$ by $f_k(\cdot, \hat{\boldsymbol{\theta}}_k)$, where $\hat{\boldsymbol{\theta}}_k$ is the estimate from (ii).

When f_k determines a model and the estimation method is maximum likelihood then this method, for fixed x , reduces to replacing $f(x)$ by the regular parameter $f_k(x, \boldsymbol{\theta}_k(f))$ where

$$\boldsymbol{\theta}_k(f) = \arg \max_{\boldsymbol{\theta}} \int \log f_k(x, \boldsymbol{\theta}) dF(x). \quad (11.4.2)$$

Recall Section 2.2.2 where we showed that in the i.i.d. case the MLE is the plug-in estimate of the θ that minimizes the Kullback-Leibler divergence between $f(x)$ and $f_k(\cdot, \theta)$. In our new terminology, $f_k(x, \hat{\theta}_k)$ is a minimum discrepancy estimate in the sense of Section 11.3.

There are two critical choices involved, that of $\{\mathcal{F}_k\}$ and that of k_n . We shall discuss the latter choice further in Section 12.5. The choice of $\{\mathcal{F}_k\}$ is clearly problem dependent. We want \mathcal{F}_k to be a good approximation for the type of f we expect. The method of estimation is also dependent on the problem, and two other considerations.

(a) *Ease of computation*

Methods which are in closed form or involve optimization of convex functions over convex sets are preferred for obvious reasons.

(b) *Efficiency for the family \mathcal{F}_k .*

The effectiveness of this criterion depends on how closely f can be approximated by a member of \mathcal{F}_k since estimates may have bias of order much greater than $\frac{1}{\sqrt{n}}$ for $f \notin \mathcal{F}_k$. Since no density of interest typically belongs to any \mathcal{F}_k this criterion is secondary.

Vapnik (1999) has argued for a third consideration.

(c) *Empirical optimization of a loss function: Machine Learning*

Vapnik (1998,1999), as did Wald (1950) before him, has argued that one should start with a loss function, attached to some feature of P , such as classification with 0-1 loss or estimation of $E(Y|X)$ with L_2 loss. Vapnik then requires us to specify a class of decision procedures. Given this framework, he insists one should minimize a natural empirical estimate of the risk. This point of view can be viewed as a modification of the method of sieves with model selection and fitting method combined. It differs from the classical Wald (1950) framework in two ways. First no model is specified but only a class of decision procedures. Second, empirical risk minimization has to be followed. For instance, having our loss function for choosing g when f is true as the Kullback-Leibler divergence and specifying \mathcal{F}_k would, to Vapnik, be equivalent to choosing the class of procedures in which f is estimated by $f(\cdot, \theta_k)$, $\theta_k \in R^k$. The risk would be just $-\int \log [f(x, \theta_k)/f(x)] f(x) dx$. The empirical risk which is to be minimized is $-\int \log [f(x, \theta_k)/f(x)] d\hat{P}(x)$, which is equivalent to maximum likelihood since $\int \log f(x) d\hat{P}(x)$ is fixed. If, however, we specify loss as

$$\int (f(x) - f(x, \theta_k))^2 dx = \int f^2(x) dx - 2 \int f(x, \theta_k) f(x) dx + \int f^2(x, \theta_k) dx$$

we would be led to minimize $-2 \int f(x, \theta_k) d\hat{P}(x) + \int f^2(x, \theta_k) dx$, a method suggested by Rudemo (1982). Note that Vapnik's criterion contradicts (b) unless Kullback-Leibler loss is used because the theory of Section 6.2 tells us that if $P \in \mathcal{F}_k$ (or presumably if it is close), then maximum likelihood should be strictly better than the Rudemo method even using Rudemo's loss! See Problem 11.4.1.

Log density estimation

One sieve $\{\mathcal{F}_k\}$ that satisfies both efficiency and relative computational ease is based on expanding the log of the density. Let B_1, B_2, \dots , be a sequence of functions on $(a, b) \subset R$. Suppose that the B_j are such that finite linear combinations can approximate an “arbitrary” function. Examples of B_1, B_2, \dots include orthogonal polynomials with respect to different measures such as the Hermite, Laguerre, or Legendre systems, splines, trigonometric polynomials, and wavelets.

Define

$$f_k(x, \boldsymbol{\eta}) = \exp \left\{ \sum_{j=1}^k \eta_j B_j(x) - A(\boldsymbol{\eta}) \right\} h(x) \quad (11.4.3)$$

where $\boldsymbol{\eta} \in \mathcal{E}_k$, the natural parameter space of this canonical exponential family, which we assume open. Assume that we can approximate any $f \in \mathcal{F}$, a “nonparametric” set of densities on (a, b) as follows,

$$\inf \{ \rho(f(\cdot, \boldsymbol{\eta}), f) : \boldsymbol{\eta} \in \mathcal{E}_k \} \rightarrow 0 \quad (11.4.4)$$

as $k \rightarrow \infty$, where ρ is a “distance” such as the Kullback-Leibler discrepancy.

Let X_1, \dots, X_n be i.i.d. as $X \sim F$, $f = F'$, and set $\bar{B}_j(\mathbf{X}) = n^{-1} \sum_{i=1}^n B_j(X_i)$,

$$p_k(\mathbf{x}, \boldsymbol{\eta}) = \prod_{i=1}^n f_k(x_i, \boldsymbol{\eta}).$$

By Corollary 1.6.2 and Theorem 2.3.1, we have

Proposition 11.4.1. *If \mathcal{E}_k is open and $1, B_1(\mathbf{x}), \dots, B_k(\mathbf{x})$ are linearly independent, then $\log p_k(\mathbf{x}, \boldsymbol{\eta})$ is strictly concave. A unique $\hat{\boldsymbol{\eta}}$ which maximizes $\log p_k(\mathbf{x}, \boldsymbol{\eta})$ exists and satisfies $\dot{A}(\hat{\boldsymbol{\eta}}) = \bar{\mathbf{B}}(\mathbf{x})$, where $\bar{\mathbf{B}}(\mathbf{x}) = (\bar{B}_1(\mathbf{x}), \dots, \bar{B}_k(\mathbf{x}))^T$.*

Recall from Section 2.2.2 that this $\hat{\boldsymbol{\eta}}$ is the plug-in estimate of the minimizer of the Kullback-Leibler divergence between $f(\cdot)$ and $f_k(\cdot, \boldsymbol{\eta})$. The estimate of $f(x)$ is now $\hat{f}_k(x) = f_k(x, \hat{\boldsymbol{\eta}})$.

A key property of \hat{f}_k not satisfied by many of the methods we have considered so far is that $\hat{f}_k \geq 0$ by definition. Here $\boldsymbol{\eta}$ can be computed using standard optimization algorithms; see Section 2.6, for example. Also note that $[\log \hat{f}_k(x)]'$, which has a simple expression, is an estimate of $[\log f_k(x)]'$ as required in Example 9.3.1.

If the B_j are a splines basis Stone and Koo (1986) and Stone, Hansen, Kooperberg and Truong (2007) have established further properties of $\hat{f}_k(x)$. We are left here too with a choice of tuning parameters, in this case the number and location of the knots. We turn to this in Section 12.5.

11.4.3 Nearest Neighbor Density Estimates

We close with a simple nonlinear method that plays an important role in classification. Consider the convolution kernel estimate with K corresponding to $\mathcal{U}(-1, 1)$. Then

$$\hat{f}_h(x) = \frac{\hat{F}(x+h) - \hat{F}(x-h)}{2h}. \quad (11.4.5)$$

It seems reasonable to use larger h if $f(x)$ is small. One way of achieving this is to note that for h close to 0 and n large,

$$\hat{F}(x+h) - \hat{F}(x-h) \simeq 2f(x)h,$$

and then choose h so that $\hat{F}(x+h) - \hat{F}((x-h)^-)$ is the same for each x . This notion corresponds to finding h such that

$$\hat{F}(x+h) - \hat{F}((x-h)^-) = k$$

and choosing $\hat{h} = \hat{h}(k, x)$ as the smallest such h . Let $|x - X|_{(1)} < \dots < |x - X|_{(n)}$ denote $|x - X_i|$, $1 \leq i \leq n$, ordered, then

$$\hat{h}(k, x) = |x - X_{i_h}|$$

where i_k is defined by

$$|x - X_{i_h}| = |x - X|_{(k)}.$$

The k^{th} -nearest neighbour (kNN) density estimate is

$$\tilde{f}_k(x) = \frac{\hat{F}(x + \hat{h}(k, x)) - \hat{F}((x - \hat{h}(k, x))^-) }{2\hat{h}(k, x)} = \frac{k}{2\hat{h}(k, x)}. \quad (11.4.6)$$

Evidently, $\{\hat{h}(k, x)\}$ can also be used as a family of bandwidth choices for the other kernels. Unfortunately, it is easy to see that if $S(f) = R$, $p \lim_{|x| \rightarrow \infty} \tilde{f}_k(x) \neq 0$ and thus kNN estimates are properly defined and renormalizable as densities only when $S(f)$ is a finite interval (Problems 11.4.2 and 11.4.3). Using the arguments used to establish Corollary 11.2.1, we can show (see Problem 11.4.4)

Theorem 11.4.1. If f has support $[a, b]$ for finite a and b , if $k = k_n \rightarrow \infty$, $n^{-1}k_n \rightarrow 0$ as $n \rightarrow \infty$, if $x \in (a, b)$, and if

$$\sup |f''(x) : x \in [a, b]| \leq M$$

for some $M > 0$, then, uniformly in f , $MSE(\tilde{f}_k(x)) = O(k^{-1})$. Thus \tilde{f}_k is consistent.

Summary. We compare our approach to regularization with regularization based on roughness penalties. Our approach involves approximating a parameter $\theta(P)$ where $\theta(\hat{P})$ is undefined with a parameter $\theta_s(P)$ where $\theta_s(\hat{P})$ is defined, and using $\theta_s(\hat{P})$ to estimate $\theta(P)$.

This approach applies in particular when $\theta(P)$ is a smooth function such as $f(\cdot)$. In statistics, the roughness penalty approach to estimating a smooth function $f(\cdot)$ that does not allow a plug-in estimate involves finding the minimizer $f_\lambda(\cdot) = f_\lambda(\cdot; \lambda)$ of an expression of the form

$$\int \rho(f_\lambda(x), f(x)) dx + \lambda J(f_\lambda)$$

where the first term measures the discrepancy between f , and $J(f_\lambda)$ is a roughness penalty such as $\int |f''_\lambda(x)|^2 dx$ that increases with the “roughness” of f_λ , and $\lambda > 0$ is a tuning parameter. The estimate $f_\lambda(\cdot; \hat{F})$ of $f(\cdot)$ is typically a piecewise polynomial called a *spline*. This approach includes the *penalized maximum likelihood* approach obtained by setting $\rho(f_\lambda(x), f(x)) = f(x) \log f_\lambda(x)$. Next we discuss the application of the method of sieves to the regularization and estimation of an irregular function $f(\cdot)$ and discuss its properties in terms of Wald decision theory and Vapnik learning theory. We apply the sieve approach to log density estimation and find that by using results from Section 1.6 on convexity in exponential family models, we obtain simple estimates of the log density and its derivative. Finally we introduce k th *nearest neighbor estimates* that are obtained from convolution kernel estimates with $K(u) = \frac{1}{2}1[|u| \leq h]$ by choosing h so that there are k x ’s in the interval $[-h, h]$.

11.5 Confidence Regions

We have until now focussed on estimation. For testing the hypothesis that f is a certain parametric shape it is, in some ways (see Section 9.5), better to use goodness-of-fit tests based on the empirical df \hat{F} , rather than on \hat{f}_h . However, by plotting confidence regions for f based on \hat{f}_h , we get insights into the possible shapes for f provided the sample size is large. We will indicate how to find such regions using an asymptotic result of Bickel and Rosenblatt (1973) and the bootstrap. We motivate this approach by developing the regions from student t -type intervals.

First we consider confidence intervals for the density $f(x)$ with x a fixed point of interest, then we show how to widen these intervals to obtain confidence regions valid for an interval of x ’s. We use ideas from Section 5.3. Set

$$Z_i = K_h(x - X_i), \theta = E(Z_i), \tau = \text{Var}(Z_i).$$

Then $\hat{f}_h(x) = \bar{Z}$, $\theta = f_h(x)$, and to get a confidence interval for $f_h(x)$ we can use the pivot

$$\frac{\sqrt{n}(\bar{Z} - \theta)}{s_z} \tag{11.5.1}$$

where s_z^2 is the sample variance of Z_1, \dots, Z_n . The asymptotic distribution of this pivot is $\mathcal{N}(0, 1)$ (see (5.3.18)). Moreover, Examples 4.4.1 and 5.3.3 suggest that its distribution can be closely approximated by the \mathcal{T}_{n-1} distribution. Thus

$$\hat{f}_h(x) \pm |t_{n-1}|_{1-\alpha} s_z / \sqrt{n} \tag{11.5.2}$$

defines an approximate level $(1 - \alpha)$ confidence interval for $f_h(x)$. If $h = cn^{-a}$ with $\frac{1}{5} < a < 1$, then (11.2.8) implies that the interval is asymptotically valid for $f(x)$ as well under the conditions of Corollary 11.2.1, because

$$\sqrt{nh}[\hat{f}_h(x) - f(x)] = \sqrt{nh}\{[\hat{f}_h(x) - f_h(x)] + [f_h(x) - f(x)]\}$$

and $\sqrt{nh}[f_h(x) - f(x)] \rightarrow 0$ by (11.2.8) as $n \rightarrow \infty$. However, $h = cn^{-a}$, $\frac{1}{5} < a < 1$, excludes the asymptotically optimal choice $h = cn^{-\frac{1}{5}}$.

If instead of estimating τ we approximate it using (11.2.10), we obtain the pivot

$$T(f_h(x), \hat{f}_h(x)) = \frac{\sqrt{nh}[\hat{f}_h(x) - f_h(x)]}{\nu(K)\sqrt{f_h(x)}} \quad (11.5.3)$$

which by the Central limit theorem and Slutsky's theorem has an asymptotic $\mathcal{N}(0, 1)$ distribution (Problem 11.5.1). Solving

$$|T(f_h(x), \hat{f}_h(x))| \leq |z|_{1-\alpha}$$

for $f_h(x)$ is equivalent to solving $\theta^2 - (2\hat{\theta} + c_\alpha^2)\theta + \hat{\theta}^2 \leq 0$ for θ where $\theta = f_h(x)$, $\hat{\theta} = \hat{f}_h(x)$, and $c_\alpha = |z|_{1-\alpha}\nu(K)/\sqrt{nh}$. The solution (see Example 4.4.3) gives the approximate level $(1 - \alpha)$ confidence interval $\hat{f}^-(x) \leq f_h(x) \leq \hat{f}^+(x)$ where

$$\hat{f}^\pm(x) = \hat{f}_h(x) + c_\alpha^2/2 \pm c_\alpha\sqrt{\hat{f}_h(x) + c_\alpha^2/4}. \quad (11.5.4)$$

Under the conditions of Proposition 11.2.2, this band is valid asymptotically for $f_h(x)$, and for $f(x)$ if $h = cn^{-a}$, $\frac{1}{5} < a < 1$.

Asymptotic regions

To modify the above intervals to have probability $(1 - \alpha)$ of containing $f(x)$ for all x in an interval $I = [r, s]$, one approach is to widen the preceding intervals to attain asymptotic level $(1 - \alpha)$. Thus we could use the α quantile k_α of the asymptotic distribution (Bickel and Rosenblatt (1973)) of

$$T = \sup_{x \in I} |T(f_h(x), \hat{f}_h(x))|.$$

This would lead to a simultaneous confidence region $\hat{f}^\pm(x), x \in [r, s]$, that covers $f_h(x)$ (and, under certain conditions, $f(x)$) with probability tending to $(1 - \alpha)$ as $n \rightarrow \infty$ when c_α in (11.5.4) is replaced by $\tilde{c}_\alpha = \tilde{k}_\alpha\nu(K)/\sqrt{nh}$, where \tilde{k}_α is the asymptotic Bickel-Rosenblatt (1973) approximation to k_α . Here \tilde{k}_α can be a poor approximation to k_α for small and moderate n . Unfortunately, the ordinary bootstrap is, we believe, not an alternative but the m out of n bootstrap may be — see Bickel and Sakov (2008).

Summary. Confidence regions for a curve provide a visualization of the possible shapes of that curve provided the sample size is large. We first give approximate confidence

intervals for $f_h(x)$ and $f(x)$ for a fixed x by using pivots based on treating $\hat{f}_h(x)$ as a sample average. Then we show how to construct $(1 - \alpha)$ confidence regions for the curves $f_h(\cdot)$ and $f(\cdot)$ by taking the maximum over x of the fixed x pivot and using the asymptotic distribution of this new pivot to construct the boundaries of the $(1 - \alpha)$ confidence region.

11.6 Nonparametric Regression for One Covariate

11.6.1 Estimation Principles

This book has considered many experiments where the task is to model and analyze statistically the relationship between a response variable Y and a vector of covariates \mathbf{X} . In this section we consider nonparametric regression in the case of a one-dimensional covariate X . The one dimensional procedures are often used as building blocks for the d -dimensional case, and some of the one dimensional methods we consider will have straightforward d -dimensional extensions. Moreover in some applications there is one variable X of principle interest while the others are confounding variables. In this case Y will be the response after the effect of the other variables have been subtracted out or controlled for.

Throughout the book we have considered various models connecting Y with X . In Section 2.2.1 we argued that for x restricted to a narrow range, a model linear in x may provide a good fit to $\mu(x) = E(Y|X = x)$. We now will allow $\mu(x)$ to take a general shape over the range \mathcal{X} of X . One very effective technique is the one suggested in Section 2.2.1: Do linear fits for X restricted to local neighborhoods of multiple x 's, then piece together the results from the multiple fits. The result is local linear regression.

Principles for construction of estimated nonparametric regression include local averaging, local parametric modelling such as local linear regression, global modelling, and penalized modelling. Or we can make a somewhat different breakdown:

- (i) Methods based on plugging the empirical probability \hat{P} into a regular approximation to $\mu(x)$ in the formula,

$$\mu(x) = E(Y|X = x) = \int y f(y|x) dy$$

where $f(\cdot | \cdot)$ is the density of $(Y|X = x)$.

- (ii) Methods based on viewing $E(Y|X)$ as the minimizer of $E(Y - g(X))^2$ over $g(\cdot)$ (see Section 1.4) and obtaining a regular approximation to $\mu(\cdot)$ by restricting g to a locally parametric class of functions and replacing quadratic loss with locally weighted quadratic loss.

Under each of these, there are local methods and global methods.

We begin with two examples of paradigm (i) approaches. We assume that we observe (X_i, Y_i) , $1 \leq i \leq n$, i.i.d. as (X, Y) with (X, Y) having a continuous case density $f \in \mathcal{F}$. We let $\tilde{P}_X(\cdot)$ and $\hat{P}(\cdot, \cdot)$ denote the empirical distributions of X_1, \dots, X_n and (X_i, Y_i) , $1 \leq i \leq n$.

Example 11.6.1. *The regressogram and Nadaraya-Watson estimates.* If we divide the x axis into intervals $I_j = (jh, (j+1)h]$, $-\infty < j < \infty$, the natural approximation to the parameter $f(\cdot|x)$ is the conditional density $f_h(\cdot|x)$ of Y given $X \in I(x)$, the I_j to which x belongs. This leads to approximating the parameter $\mu(x)$ by

$$E_h(Y|X=x) = \sum_{j=-\infty}^{\infty} \int_y f(y|x) 1(x \in I_j) dy .$$

The empirical plug-in estimate is

$$\hat{\mu}_h^R(x) \equiv \sum_{j=-\infty}^{\infty} \bar{Y}_{jh} 1(x \in I_j)$$

where \bar{Y}_{jh} is the average of the Y_i with $X_i \in I_j$. This is called the *regressogram* and like the histogram density estimate doesn't take enough advantage of possible smoothness of $\mu(x)$.

Let $K_h(t) = h^{-1}K(t/h)$ where K is a kernel as in Section 11.2. A smooth weighted average is

$$\hat{\mu}_{\text{NW}}(x) = \frac{\sum_{i=1}^n Y_i K_h(X_i - x)}{\sum_{i=1}^n K_h(X_i - x)}, \quad (11.6.1)$$

the *Nadaraya-Watson (NW) estimate*. By selecting a kernel with enough derivatives we can intuitively obtain a version of NW which is as smooth as we please. \square

We can obtain the NW estimate by viewing it from the optimization viewpoint as well and, in that way, also generalize it to estimate smoother $\mu(x)$. Both are examples of *local averaging estimates*, $\hat{\mu}(x) = \sum_{i=1}^n Y_i w_n(x, X_i)$ where the data-determined weight, $w_n(x, X_i)$, decrease as $|X_i - x|$ increases.

Example 11.6.2. *Minimum contrast and local polynomial estimates.* Here is a class of procedures which falls under paradigm (ii). Let x be fixed and let $\mu(z, \beta)$ be a parametric family of functions such that for suitable $\beta(x)$, $\mu(x, \beta(x)) = \mu(x)$ and $\mu(z, \beta(x))$ approximates $\mu(x)$ well locally for z close to x . Mainly we consider $\mu(z, \beta(x))$ a polynomial in $z - x$. Suppose $K \geq 0$. We obtain a regular parameter by replacing minimization of $E(Y - g(X))^2$ by that of the discrepancy

$$D(\beta(\cdot)) = \int (y - \mu(z, \beta(x)))^2 K_h(z - x) dz f(x, y) dx dy \quad (11.6.2)$$

with respect to $\beta = \beta(x)$, where $\beta(x)$ is in R^p for some $p \geq 0$, and then approximating $\mu(x)$ by $\mu(x, \beta(x))$. As $h \rightarrow 0$ we approach paradigm (ii). To simplify our notation, we introduce (X', Y, Z) with density proportional to $K_h(z - x)f(x, y)1(x \in S(f))$. Conditioning on $X' = x$, we obtain (Problem 11.6.1)

$$\beta(x) = \arg \min_{\beta} \int (y - \mu(z, \beta))^2 q(z, y|x) dz dy$$

where $q(z, y|x)$ is the conditional density of $(Z, Y)|X' = x$, that is

$$q(z, y|x) = \frac{f(z, y)K_h(z - x)1(x \in S(f))}{f_h(x)}$$

with $f_h(\cdot)$ the marginal density of X' given by

$$\begin{aligned} f_h(x) &= \int K_h(z - x)f(z, y)dzdy1(x \in S(f)) \\ &= \int K_h(z - x)f_X(z)dz1(x \in S(f)). \end{aligned}$$

Note that

$$\beta(x) = \beta(P; x) \equiv \arg \min_{\beta} \frac{\int [y - \mu(z, \beta)]^2 K_h(z - x)1(x \in S(f))dP(z, y)}{\int K_h(z - x)1(x \in S(f))dP_X(z)}. \quad (11.6.3)$$

The empirical plug-in estimate $\beta(\hat{P}; x)$ is a *minimum contrast* and *local modelling estimate* in the same spirit as (11.3.1); we approximate $\mu(x)$ locally by $\mu(z, \beta)$, where β can be expressed as $\beta(P; x)$.

If we specialize to $\mu(z, \beta) = \beta \in R$ we obtain the Nadaraya-Watson estimate (Problem 11.6.2). If we take $\mu(z, \beta) = \beta_0 + \beta_1(z - x)$ we get

$$\beta_1(x, P) = \frac{E_Q(Y(Z - E_Q Z))}{\text{Var}_Q Z}, \quad \beta_0(x, P) = E_Q(Y) - \beta_1(x, P)(E_Q(Z) - x) \quad (11.6.4)$$

where Q is the distribution corresponding to $q(z, y|x)$. See Remark 11.3.1.

We can use empirical plug in because dependence on P is purely through $dP(z, y)$ and $dP_X(z)$. Specifically,

$$\beta_1(x, \hat{P}) = \frac{\sum_{i=1}^n Y_i(X_i - \bar{X}(x))K_h(X_i - x)}{\sum_{i=1}^n (X_i - \bar{X}(x))^2 K_h(X_i - x)} \quad (11.6.5)$$

where

$$\bar{X}(x) = \frac{\sum_{i=1}^n X_i K_h(X_i - x)}{\sum_{i=1}^n K_h(X_i - x)}. \quad (11.6.6)$$

Set $\bar{Y}(x) = \hat{\mu}_{\text{NW}}(x)$, then the estimate of $\mu(x)$ is the *locally linear estimate*

$$\hat{\mu}_h(x) = \beta_0(x, \hat{P}) = \bar{Y}(x) - \beta_1(x, \hat{P})(\bar{X}(x) - x). \quad (11.6.7)$$

Evidently, we may replace $\beta_0 + \beta_1(z - x)$ by a higher order polynomial in $(z - x)$ and use Theorem 1.4.4 to obtain locally polynomial estimates. \square

Remark 11.6.1. We obtain a robust minimum contrast estimate if we replace the squared error e^2 , $e = y - u$, in (11.6.2) by the Huber function $\rho(e) = e^2$, $|e| \leq k$, $\rho(e) = k|e|$, $|e| > k$. With this alteration, our estimates resemble the popular LOESS estimate of Cleveland (1979) and Cleveland and Devlin (1988).

11.6.2 Asymptotic Bias and Variance Calculations

Our calculations are kept simple by first computing the bias and variance conditional on $\mathbf{X} = (X_1, \dots, X_n)^T$ and then computing limits in probability as $n \rightarrow \infty$. Thus, because $E(Y_i|\mathbf{X}) = \mu(X_i)$,

$$E\{\widehat{\mu}_{\text{NW}}(x)|\mathbf{X}\} = \frac{n^{-1} \sum_{i=1}^n K_h(X_i - x)\mu(X_i)}{n^{-1} \sum_{i=1}^n K_h(X_i - x)}. \quad (11.6.8)$$

Moreover, if $\|\mu'''\|_\infty \leq M < \infty$, then uniformly in x ,

$$\mu(X_i) = \mu(x) + \mu'(x)(X_i - x) + \frac{1}{2}\mu''(x)(X_i - x)^2 + O_P(X_i - x)^3 \quad (11.6.9)$$

which makes $E[\widehat{\mu}_{\text{NW}}(x)|\mathbf{X}]$ a simple function of the i.i.d. sums

$$S_{nj} = \sum_{i=1}^n (X_i - x)^j K_h(X_i - x), \quad j = 0, 1, 2.$$

Recall that $[a, b]$ is the support of the density $f(x)$ of X and recall the notation

$$m_j = m_j(K) = \int u^j K(u) du, \quad \nu = \nu(K) = \nu_0(K), \quad \nu_j = \nu_j(K) = \int u^j K^2(u) du.$$

Theorem 11.6.1. Suppose that $[c, d] \subset (a, b)$, $\inf\{f(x) : x \in [c, d]\} > 0$, and that $\mu'''(\cdot)$ and $f''(\cdot)$ are bounded on $[c, d]$. If K is symmetric, $m_2(K) < \infty$, and $\nu_5(K)$ exists, then, for the estimate (11.6.1), uniformly for $x \in [c, d]$ as $n \rightarrow \infty$, $h = h_n \rightarrow 0$,

$$\text{Bias}[\widehat{\mu}_{\text{NW}}(x)|\mathbf{X}] = \frac{1}{2} \left[\mu''(x) + \frac{2\mu'(x)f'(x)}{f(x)} \right] m_2(K)h^2 + O_P(h^3). \quad (11.6.10)$$

Proof. Using (11.6.8) and (11.6.9), we have

$$\text{Bias}[\widehat{\mu}_{\text{NW}}(x)|\mathbf{X}] = \frac{\mu'(x)S_{n1} + \frac{1}{2}\mu''(x)S_{n2} + O_P(S_{n3})}{S_{no}}. \quad (11.6.11)$$

The rest of the proof uses Chebychev's inequality, the law of large numbers, and Taylor expansions. See Appendix D.6.

Remark 11.6.2. We see from the expansion used to derive (11.6.10) that under the conditions of Theorem 11.6.1, if $m_2(K) = 0$, and if $m_4(K) < \infty$, then the conditional bias of $\widehat{\mu}_{\text{NW}}$ is of the order $O_P(h^4)$. In general, if r is even, f'' and $\mu^{(r+3)}$ are bounded, K is symmetric, $m_j(K) = 0$, $j = 1, \dots, r$, and $m_{r+3}(K) < \infty$, then the conditional bias of $\widehat{\mu}_{\text{NW}}(K)$ is of order $O_P(h^{r+2})$. Kernels with $m_j(K) = 0$, $j = 1, \dots, r$, are called *higher order kernels of order r*. See Problem 11.2.8 for an example. If r is odd, the preceding with r replaced by $r - 1$ applies. \square

We turn to

$$\text{Var}(\hat{\mu}_{\text{NW}}(x)|\mathbf{X}) = \sum_{i=1}^n w_i^2(x) \sigma^2(X_i) \quad (11.6.12)$$

where

$$\sigma^2(x) = \text{Var}(Y|X=x), \quad w_i(x) = K_h(X_i - x) / \sum K_h(X_i - x). \quad (11.6.13)$$

Theorem 11.6.2. Suppose that for $[c, d] \subset (a, b)$, $\inf \{f(x) : x \in [c, d]\} > 0$ and that $f'(x)$ and $[\sigma^2(x)]'$ are bounded on $[c, d]$. If $\nu_j(K) < \infty$, $j = 0, 1$, then, uniformly for $x \in [c, d]$, as $n \rightarrow \infty$, $h \rightarrow 0$, $nh \rightarrow \infty$,

$$\text{Var}(\hat{\mu}_{\text{NW}}(x)|\mathbf{X}) = \frac{\nu(K)\sigma^2(x)}{nhf(x)} + o_P\left(\frac{1}{nh}\right). \quad (11.6.14)$$

Proof. See Appendix D.6.

Remark 11.6.3 The bias-variance tradeoff is again clear. Let $B(x)h^4$ and $V(x)(nh)^{-1}$ denote the leading terms of the squared conditional bias (11.6.10) and variance (11.6.14), then the asymptotic conditional MSE is

$$A_h(x) = B(x)h^4 + V(x)(nh)^{-1}.$$

By solving $(d/dh)A_h(x) = 0$ for h , we find the minimizer

$$h = h_{\text{optimal}} = \left(\frac{V(x)}{4B(x)}\right)^{\frac{1}{5}} n^{-\frac{1}{5}}.$$

It follows that the conditional MSE of $\hat{\mu}_{\text{NW}}(x)$ satisfies

$$\inf_{h>0} \text{MSE}(\hat{\mu}_{\text{NW}}(x)|\mathbf{X}) = 5 \cdot 4^{-\frac{4}{5}} B^{\frac{1}{5}}(x) V^{\frac{4}{5}}(x) n^{-\frac{4}{5}} + o_P(n^{-\frac{4}{5}}).$$

Thus $\hat{\mu}_{\text{NW}}(x)$ with h of the form $cn^{-\frac{1}{5}}$ converges to $\mu(x)$ at the rate $O_P(n^{-\frac{4}{5}})$. \square

If $S(f) = [a, b]$ and $x_\ell = a + \lambda h$, $x_r = b - \lambda h$ are left and right boundary points, then $\hat{\mu}_{\text{NW}}(x_\ell)$ is still consistent, but the bias at these points is of order $O_P(h)$ rather than the $O_P(h^2)$ rate of Theorem 11.6.1. To improve on this rate we turn to other estimates.

Local polynomial estimates

In Example 11.6.2 consider the polynomial

$$\mu(z; \boldsymbol{\beta}) = \sum_{k=0}^p \beta_k (z - x)^k.$$

The partial derivative with respect to β_j is $\nabla_j \mu(z - x, \boldsymbol{\beta}) = (z - x)^j$ and the minimizer of the discrepancy (11.6.2) satisfies the set of linear equations ($j = 0, \dots, p$)

$$E_P\{Y(X - x)^j K_h(X - x)\} = \sum_{k=0}^p E_P\{(X - x)^j K_h(X - x)(X - x)^k\} \beta_k. \quad (11.6.15)$$

Let $\mathbf{L}(X - x) = (1, X - x, \dots, (X - x)^p)^T$, then the solution to (11.6.15) is

$$\boldsymbol{\beta} = \boldsymbol{\beta}(x; P) = [E_P\{\mathbf{L}(X - x)K_h(X - x)\mathbf{L}^T(X - x)\}]^{-1} E_P\{YK_h(X - x)\mathbf{L}(X - x)\}$$

provided the indicated matrix inverse exists. Thus the empirical plug-in estimate of $\boldsymbol{\beta}$ is

$$\hat{\boldsymbol{\beta}}(x) = (\mathbf{Z}^T \mathbf{W} \mathbf{Z})^{-1} \mathbf{Z}^T \mathbf{W} \mathbf{Y} \quad (11.6.16)$$

where $\mathbf{W} = \text{Diagonal}(K_h(X_1 - x), \dots, K_h(X_n - x))$, $\mathbf{Y} = (Y_1, \dots, Y_n)^T \mathbf{i}$, and $\mathbf{Z} = ((X_i - x)^j), i = 1, \dots, n, j = 0, \dots, p$. Note that (11.6.16) is also the solution to the weighted least squares problem with weight matrix \mathbf{W} ; see (2.2.20). Using the notation

$$\hat{\boldsymbol{\beta}}(x) = (\hat{\beta}_0(x), \hat{\beta}_1(x), \dots, \hat{\beta}_p(x))^T$$

the estimate of $\mu(x)$ is

$$\hat{\mu}(x) = \mu(z; \hat{\boldsymbol{\beta}}(x))|_{z=x} = \hat{\beta}_0(x).$$

For the local linear case where $p = 1$, this reduces to (11.6.7).

We next show that when μ is smooth the rate at which the bias tends to zero as $h \rightarrow 0$ can be made arbitrarily fast by selecting p large. Note that (11.6.16) and $E(Y_i | \mathbf{X}_i) = \mu(X_i)$ yields

$$E(\hat{\boldsymbol{\beta}}(x) | \mathbf{X}) = (\mathbf{Z}^T \mathbf{W} \mathbf{Z})^{-1} \mathbf{Z}^T \mathbf{W} \boldsymbol{\mu}$$

where $\boldsymbol{\mu} = (\mu(X_1), \dots, \mu(X_n))^T$. Set $\boldsymbol{\beta} = (\mu(x), \mu'(x), \dots, \mu^{(p)}(x)/p!)^T$ and $\mathbf{r} = \boldsymbol{\mu} - \mathbf{Z}\boldsymbol{\beta}$; then the conditional bias of $\hat{\boldsymbol{\beta}}(x)$ as an estimate of $\boldsymbol{\beta}$ is

$$\text{Bias}(\hat{\boldsymbol{\beta}}(x) | \mathbf{X}) = (\mathbf{Z}^T \mathbf{W} \mathbf{Z})^{-1} \mathbf{Z}^T \mathbf{W} \mathbf{r}. \quad (11.6.17)$$

Here is a key lemma.

Lemma 11.6.1. Assume that $\inf\{f(x) : x \in [c, d]\} > 0$ for $[c, d] \subset (a, b)$. Also assume that $f'(\cdot)$ and $\mu^{(p+2)}(\cdot)$ are bounded on $[c, d]$. Then, uniformly for $x \in [c, d]$,

$$\mathbf{r} = (\beta_{p+1}(X_i - x)^{p+1} + o_P(X_i - x)^{p+1})_{1 \leq i \leq n}. \quad (11.6.18)$$

Proof. The result follows because by Taylor expansion

$$\mu(X_i) = \sum_{j=0}^p \beta_j (X_i - x)^j + \beta_{p+1}(X_i - x)^{p+1} + o_P(X_i - x)^{p+1}, \quad (11.6.19)$$

that is, the first $(p + 1)$ terms in the Taylor expansion of \mathbf{r} are identically zero! \square

It follows that

Proposition 11.6.1. *If $\hat{\mu}(x)$ is a polynomial of order p , then $\hat{\beta}(x)$ is unbiased. In particular, $\hat{\mu}(x) = \hat{\beta}_0(x)$ is unbiased.*

Lemma 11.6.1 also yields the asymptotic conditional bias of $\hat{\mu}(x)$ in general: For the proof and the conditional bias of the vector $\hat{\beta}(x)$, see Appendix D.6.

Theorem 11.6.3. *Suppose that K is symmetric with $m_{2p+1}(K) < \infty$. Under the conditions of Lemma 11.6.1, uniformly for $x \in [c, d]$,*

$$\text{Bias}[\hat{\mu}(x)|\mathbf{X}] = \frac{\eta_{p+1}(K)}{(p+1)!} \mu^{(p+1)}(x)h^{p+1} + o_P(h^{p+1}) \quad (11.6.20)$$

where $\eta_{p+1}(K)$ is a constant defined in Appendix D.6.

When p is even, then $\eta_{p+1}(K) = 0$ (Appendix D.6), so the leading term in (11.6.20) is zero. In the case of $p = 0$, we have shown that the next term in the expansion gives (11.6.10) as the leading term, which is seen to be $O_P(h^2)$. For the leading term when $p \geq 2$ is even, see Fan and Gijbels (1996, Theorem 3.1). When $p = 1$, we show in Appendix D.6 that $\eta_2(K) = m_2(K)$, thus

$$\text{Bias}(\hat{\mu}(x)|\mathbf{X}) = \frac{1}{2}m_2(K)\mu''(x)h^2 + o_P(h^2), \quad p = 1.$$

This bias is of same order, $O_P(h^2)$, as the bias (11.6.10) of the local constant estimate. However, at boundary points we can show that the local linear estimate retains the order $O_P(h^2)$ while the bias of the local constant estimator converges to zero at the slower rate $O_P(h)$ (Problem D.6.5). Let $m_{\ell,\lambda}(K) = \int_{-\lambda}^1 u^\ell K(u)du$ and

$$Q_\lambda(K) = \frac{m_{2\lambda}^2(K) - m_{1\lambda}(K)m_{3\lambda}(K)}{m_{2\lambda}(K)m_{0\lambda}(K) - m_{1\lambda}^2(K)}.$$

Proposition 11.6.2. *Suppose the support of f is $S(f) = (a, b)$ with a and b finite, K is symmetric with $K(u) = 0$ for $|u| > 1$, $\mu^{(p+1)}(a^+)$ exists, $f(a^+) > 0$, and $f(\cdot)$ and $\mu^{(p+1)}(\cdot)$ are right continuous at a , then for $x = a + \lambda h$,*

$$(a) \quad \text{for } p = 0, \text{Bias}[\hat{\mu}(x)|\mathbf{X}] = m_{2\lambda}(K)\mu'(a^+)h + o_P(h); \quad (11.6.21)$$

$$(b) \quad \text{for } p = 1, \text{Bias}[\hat{\mu}(x)|\mathbf{X}] = \frac{1}{2}Q_\lambda(K)\mu''(a^+)h^2 + o_P(h^2). \quad (11.6.22)$$

The variance

Theorem 11.6.4. Suppose K is a symmetric density with $\nu_{2p+1}(K) < \infty$, $\inf\{f(x) : x \in [c, d]\} > 0$ for $[c, d] \subset (a, b)$ and that $f'(x)$, $\mu'_{p+1}(x)$ and $[\sigma^2(x)]'$ exist and are bounded on $[c, d]$. If $h \rightarrow 0$ and $nh \rightarrow \infty$ as $n \rightarrow \infty$, then, uniformly for $x \in [c, d]$,

$$\text{Var}(\hat{\mu}(x)|\mathbf{X}) = \nu_p(K) \frac{\sigma^2(x)}{f(x)} (nh)^{-1} \{1 + o_p(1)\},$$

where $\nu_p(K)$ is a constant given in Appendix D.6.

It follows from Appendix D.6 that $\nu_1(K) = \nu_0(K)$, and the asymptotic variances for the $p = 0$ and $p = 1$ cases coincide. Also note that the order $O_P\{(nh)^{-1}\}$ of $\text{Var}\hat{\mu}(x)|\mathbf{X}$ does not depend on p . Thus by increasing p , assuming smoothness, we can make the asymptotic conditional bias arbitrarily small without increasing the order of the asymptotic conditional variance.

We next turn to conditional MSE, its asymptotic and integrated versions, and its minimizers.

MSE and IMSE

Corollary 11.6.1. Under the conditions of Theorems 11.6.3 and 11.6.4, uniformly for $x \in [c, d]$,

$$\begin{aligned} \text{MSE}(\hat{\mu}(x)|\mathbf{X}) &= \left[\frac{\eta_{p+1}(K)}{(p+1)!} \mu^{(p+1)}(x) h^{p+1} \right]^2 + \nu_p(K) \frac{\sigma^2(x)}{f(x)} (nh)^{-1} \\ &\quad o_P\{h^{p+1} + (nh)^{-1}\}. \end{aligned} \quad (11.6.23)$$

We denote the sum of the first two terms in (11.6.23) by $A(x, h)$. Note that $A(x, h)$ is of the form $B(x)h^{2(p+1)} + V(x)n^{-1}h^{-1}$, where $B(x)$ and $V(x)$ are given in (11.6.23). By solving $(d/dh)A(x, h) = 0$ for h we find the minimizer of $A(x, h)$

$$h_0(x) = \arg \min_{h>0} A(x, h) = \left[\frac{V(x)}{2(p+1)B(x)} \right]^{\frac{1}{2p+3}} n^{-\frac{1}{2p+3}}.$$

By substituting $h_0(x)$ in $A(x, h)$ we find

$$\min_{h>0} A(x, h) = c'n^{-\frac{2p+2}{2p+3}} \quad (11.6.24)$$

where c' does not depend on n . The same order of convergence holds for the IMSE.

Remark 11.6.4. Note that when $\mu(x)$ is infinitely differentiable the rate of convergence of the MSE and IMSE to zero can be made arbitrarily close to the regular parametric rate n^{-1} . \square

The (asymptotic) integrated mean squared error (IMSE) is

$$\text{IMSE}(\hat{\mu}) = \int A(x, h) w(x) dx$$

for an appropriate weight function $w(x)$. If $f(\cdot)$ has finite support $[a, b]$, then $w(x) = f(x)$ is a natural choice. In this case, if $\sigma^2(x)$ is constant and $E\{[\mu^{(p+1)}(X)]^2\} < \infty$, the optimal h is (Problem 11.6.2.1)

$$h_0 = C_p \left\{ \frac{[b-a]\sigma^2}{E\{[\mu^{(p+1)}(X)]^2\}} \right\}^{\frac{1}{2p+3}} n^{-\frac{1}{2p+3}}. \quad (11.6.25)$$

With this h_0 , IMSE tends to zero at the rate $n^{-\frac{2p+2}{2p+3}}$.

Regression curves

So far our main focus has been on estimating the *regression level* $\mu(x) = E(Y|X=x)$. However, in parametric analysis the emphasis is on estimating regression slopes (often called regression coefficients) because they give the expected change in the response Y as the covariate x is increased (or decreased). In a nonparametric framework the regression curve $b(x) \equiv \mu'(x)$ has a similar interpretation, and is also of interest. Recall that our local polynomial estimate $\hat{\beta} = \hat{\beta}(x) = (\hat{\beta}_0(x), \dots, \hat{\beta}_p(x))^T$ provides an estimate $\hat{\mu}(z) = \sum_{j=0}^p \hat{\beta}_j(z)(z-x)^j$ of $\hat{\mu}(z)$ for z close to x . It follows that

$$\hat{b} = \hat{b}(x) \equiv \hat{\mu}'(z)|_{z=x} = \hat{\beta}_1(x)$$

is an estimate of $b(x)$. The asymptotic properties of \hat{b} are contained in the results in Appendix D.6. We state these results when $p = 2$, which is a good choice of p because it gives conditional bias of order h^2 at boundary as well as interior points.

Theorem 11.6.5. Suppose the conditions of Theorems 11.6.3 and 11.6.4 are satisfied with $p = 2$. Then

$$\begin{aligned} MSE(\hat{b}(x)|\mathbf{X}) &= \left\{ \frac{1}{6} \frac{m_4(K)}{m_2(K)} \mu^{(3)}(x) h^2 \right\}^2 + \frac{\nu_2(K)}{m_2(K)} \frac{\sigma^2(x)}{f(x)} n^{-1} h^{-3} + o_P(h^2 + n^{-1} h^{-3}) \\ &= B(x) h^4 + V(x) n^{-1} h^{-3} + R_n, \text{ say.} \end{aligned}$$

The asymptotically optimal bandwidth h is obtained by setting the derivative of $Bh^4 + Vn^{-1}h^{-3}$ equal to zero, which with $C(K) = \{27m_2(K)\nu_2(K)/m_4^2(K)\}^{\frac{1}{7}}$ (Problem 11.6.7) yields

$$h_0 = \left(\frac{3V}{4B} \right)^{\frac{1}{7}} n^{-\frac{1}{7}} = C(K) \left\{ \frac{\sigma^2(x)}{[\mu^{(3)}(x)]^2 f(x)} \right\}^{\frac{1}{7}} n^{-\frac{1}{7}}. \quad (11.6.26)$$

Smoothing, variance estimation, and confidence regions

Our estimates of $\mu(\cdot)$ are often referred to as *linear smoothers* of the data (X_i, Y_i) , $1 \leq i \leq n$. More precisely, let $\boldsymbol{\mu} = (\mu(X_1), \dots, \mu(X_n))'$ and $\hat{\boldsymbol{\mu}} = (\hat{\mu}(X_1), \dots, \hat{\mu}(X_n))'$ be an estimator (predictor) of $\boldsymbol{\mu}$. Then we can write

$$\hat{\boldsymbol{\mu}} = \mathbf{S}\mathbf{Y} \quad (11.6.27)$$

for some $n \times n$ smoother matrix \mathbf{S} . In the case of the NW estimate, (11.6.1), $\mathbf{S} = (w_i(X_j))_{n \times n}$ (see (11.6.13)). Estimates of the form \mathbf{SY} include regression splines; see e.g. Hastie and Tibshirani (1990).

For variance estimation, we consider the model

$$Y_i = \mu(X_i) + \varepsilon_i, \quad i = 1, \dots, n$$

where X_i and ε_i are independent and $v(X_i) = \text{Var}(Y|X_i) = \text{Var}(\varepsilon_i)$. An estimator of $v = (v(X_1), \dots, v(X_n))^T$ is obtained by considering \mathbf{TR}^2 for some $n \times n$ smoother matrix \mathbf{T} , where $\mathbf{R} = (\mathbf{I} - \mathbf{S})\mathbf{Y}$ is the vector of residuals and \mathbf{R}^2 is the vector of squared residuals. To obtain a bias correction for \mathbf{TR}^2 , we note that when $v(x) \equiv \sigma^2$,

$$E(\mathbf{R}^2|\mathbf{X}) = [E(\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}|\mathbf{X})]^2(\mathbf{1} + \mathbf{A})\sigma^2$$

where $\mathbf{A} = \text{diag}(\mathbf{SS}^T - 2\mathbf{S})$ and $\mathbf{1} = (1, \dots, 1)_{n \times 1}^T$. To reduce bias, we write $A = (a_{ij})$ and set

$$r_i = \frac{R_i}{\sqrt{1 + a_{ii}}}, \quad \mathbf{r}^2 = (r_1^2, \dots, r_n^2)^T.$$

Our estimate of \mathbf{v} is (see Ruppert, Wand, Holst and Hössjer (1997))

$$\hat{\mathbf{v}} = \mathbf{Tr}^2.$$

When $\hat{\boldsymbol{\mu}}$ is conditionally unbiased (see Proposition 11.6.1) and $v(x) \equiv \sigma^2$, $\hat{\mathbf{v}}$ is conditionally unbiased. Moreover, for local polynomial estimates, by Theorem 11.6.3, the conditional bias is of order $o_P(h^{2(p+1)})$. In the homeoscedastic case when $v(x) \equiv \sigma^2$, we use

$$\hat{\sigma}^2 = n^{-1} \sum_{i=1}^n \hat{v}(X_i). \tag{11.6.28}$$

We now turn to the estimation of the variances of our estimates. The estimates of $\mu^{(j)}(\cdot)$, $j = 0, \dots, p$, can be written in the form $\sum_{i=1}^n w_{ij}(x)Y_i$ for appropriate $w_{ij}(x)$; see (11.6.1), (11.6.5), and (11.6.7). It follows that in the homeoscedastic case

$$\text{Var}(\hat{\mu}^{(j)}(x)|\mathbf{X}) = \sigma^2 \sum_{i=1}^n w_{ij}^2(x),$$

where we use $\hat{\sigma}^2$ from (11.6.28) to estimate σ^2 . In this case, under conditions on h where $\hat{\mu}^{(j)}(x)$ has bias of lower order than the order of its standard deviation, the pivot

$$Z_j(x) = \frac{\hat{\mu}^{(j)}(x) - \mu^{(j)}(x)}{\hat{\sigma}\{\sum_{i=1}^n w_{ij}^2(x)\}^{\frac{1}{2}}}$$

will approximately have a $\mathcal{N}(0, 1)$ distribution and yield approximate $(1 - \alpha)$ confidence intervals for $\mu^{(j)}(x)$ with x in a grid $\mathcal{G} = \{x^{(1)}, \dots, x^{(g)}\}$. We expect $100(1-\alpha)\%$ of these intervals to contain the true values of $\mu^{(j)}(x)$, $x \in \mathcal{G}$. Simultaneous level $(1 - \alpha)$ intervals can be obtained from the asymptotic distribution of $\sup_x |Z_j(x)|$ (Nadaraya (1989)).

Multiple testing in matched pair experiments

In Section 4.9.2 we considered experiments where subjects serve as their own control and the difference Y between a treatment response and a control response is recorded. Now in addition to Y we include a covariate such as age or BMI (body mass index). In this case we may want to investigate for which covariate values the treatment has an effect and turn to multiple testing where we test $H_j : \mu(x_j) = 0$, $j = 1, \dots, g$, vs one-sided or two-sided alternatives using $Z_j(x)$ or $|Z_j(x)|$. The simultaneous confidence intervals discussed above provide possible solutions to the multiple testing problem. If we test $H_0 : \mu(x) = 0$ for all x , $\hat{\mu}^{(j)}(x)$ will not have a bias problem under H_0 .

Bandwidth choice

The discussion of Section 11.2.3 also applies to the local polynomial estimates of $\mu(\cdot)$. Again the reference distribution approach to bandwidth selection yields the optimal rate of convergence but not the optimal constant when $p = 1$ if we assume a well behaved μ'' but are not willing to estimate $\int [\mu''(x)]^2 f(x)dx$ nonparametrically. For instance, we can use the parametric model where $(Y|X = x)$ is $\mathcal{N}(\alpha_0 + \alpha_1 x + \alpha_2 x^2, \sigma^2)$ and substitute the parametric estimates of σ^2 and $\mu''(x) = \alpha_2$ in (11.6.25) with $(b - a)$ replaced by $(X_{(n)} - X_{(1)})$. However, the soundest approach is cross validation, which adapts to the degree of smoothness present. See Section 12.5.

Summary. Let $(X_1, Y_1), \dots, (X_n, Y_n)$ be i.i.d. as $(X, Y) \sim P$ where (X, Y) is continuous. We consider kernel estimates $\hat{\mu}_h(x)$ of the curve $\mu(x) = E_P(Y|X = x)$ based on weighted local fits to $\mu(x)$ over intervals $[x - h, x + h]$. The Nadaraya-Watson (NW) estimate $\hat{\mu}_{\text{NW}}(x)$ is a kernel estimate of the form $\sum w_i(x)Y_i$, where

$$w_i(x) = K_h(X_i - x) / \sum_{i=1}^n K_h(X_i - x),$$

and can be viewed as giving a locally constant fit to $\mu(x)$ over $[x - h, x + h]$. We also consider locally linear and locally polynomial fits and give their bias and variance properties for $n \rightarrow \infty$ and $h \rightarrow 0$. The NW estimate has conditional bias of order $O_P(h^2)$ for interior points x and order $O_P(h)$ for boundary points while the locally linear estimate has conditional bias of order $O_P(h^2)$ for all points x in the support of X . These estimates have the same asymptotic variance.

11.7 Problems and Complements

Problems for Section 11.2

1. Let X_1, \dots, X_n be i.i.d. with density f , where f has support $S(f)$.
 - (a) Suppose $S(f) = R$. Show that $(2h)^{-1}\hat{P}(x - h, x + h] = \int J_h(x - z) \, dz$ where $J_h(u) = h^{-1}J(u/k)$ and $J(t) = \frac{1}{2}1[-1 \leq t \leq 1]$.

(b) Suppose that $S(f) = [a, b]$ and $0 < 2h < b - a$. Specify $K_h(x, z)$ so that

$$\frac{\widehat{P}(x-h, x+h)}{2h} = \int K_h(x, z) d\widehat{P}(z).$$

Consider (i) $x \in [a, a+h]$, (ii) $x \in [a+h, b-h]$, and (iii) $x \in (b-h, b]$

2. Suppose K is a kernel with $K \geq 0$ and support $S(K) = [-c, c]$, $0 < c \leq \infty$. Let $X_h = Z + hV$ where $Z \sim f$, $V \sim K$, Z and V are independent, $S(f) = [a, b]$, $-\infty \leq a < b \leq \infty$, $2h < b - a$.

(a) Show that X_h has density

$$\begin{aligned} f_h(x) &= \int_a^b K_h(x-z)f(z)dz, \quad x \in [a-ch, b+ch] \\ &= 0, \quad \text{otherwise}. \end{aligned}$$

Hint. Z and V have joint density $f(z)K(v)$, $z \in [a, b]$, $v \in [-c, c]$. By the Jacobian transformation theorem (Theorem B.2.2), X_h and Z have joint density

$$f(z)h^{-1}K((x-z)/h), \quad z \in [a, b], \quad x \in [a-ch, b+ch].$$

(b) It follows from (a) that $\int_{a-ch}^{b+ch} f_h(x)dx = 1$. Thus $\int_a^b f_h(x)dx$ may not be one and $f_h(x)$ may not be a density on $[a, b]$. Find $\int_a^b f_h(x)dx$ when $V = U \sim \mathcal{U}[-1, 1]$ and (i) f is $\mathcal{N}(0, 1)$, (ii) f is $\mathcal{U}[0, 1]$.

(c) We can think of $f_h(x) = \int_a^b K_h(x-z)f(z)dz$ as a weighted average of $f(z)$ with weights $K_h(x-z)$, $z \in [a, b]$. However, the weights may not integrate to one. Let $W_h(x) = \int_a^b K_h(x-z)dz$. Show that $W_h(x) = L((x-a)/h) - L((x-b)/h)$, where $L(t) = \int_{-\infty}^t K(t)dt$.

(d) Suppose $-\infty < a < b < \infty$. For $x = a + \lambda h$ and $x = b - \lambda h$, $0 < \lambda < 1$, find $W_h(x)$ as defined in (c) when (i) K is $\mathcal{U}[-1, 1]$ and (ii) $K = K_E$.

(e) We can get a closer approximation to $f(x)$ than $f_h(x)$ by rescaling the weights $K_h(x-z)$; that is, use $f(x; h) = \int K_h(x, z)f(z)dz$ with $K_h(x, z) = K_h(x-z)/W_h(x)$ and $W_h(x)$ as in (c). Show that $\int_a^b f(x; h)dx = 1$.

3. Show that part (b) of Lemma 11.2.1 holds using the dominated convergence theorem.

4. Establish Lemma 11.2.2 using Lemma 11.2.1.

Hint. By (A.15.2),

$$P(|\widehat{f}_h(x) - f(x)| \geq \varepsilon) \varepsilon^2 \leq \text{MSE}(\widehat{f}_h(x)) = \{\text{Bias}[\widehat{f}_h(x)]\}^2 + \text{Var}[\widehat{f}_h(x)].$$

- 5.** Show that a symmetric kernel with $m_2(K) = 0$ will yield a bias term of uniform order h^4 for $\hat{f}_h(x)$ if $\|f^{(5)}\| < \infty$ and $m_5(K) < \infty$. Show that $K(t) = \frac{1}{2}(3 - x^2)\phi(x)$ is such a kernel when ϕ is the $\mathcal{N}(0, 1)$ density. Do the same for $K(t) = (15/32)(7t^4 - 10t^2 + 3)1(|t| \leq 1)$.

- 6.** Complete the proof of Proposition 11.2.2.

- 7.** Show that the conclusions of Propositions 11.2.2 and Corollary 11.2.1 remain valid if K doesn't have compact support.

- 8. Kernels K such that**

$$\int_{-\infty}^{\infty} K^{(j)}(x) dx = 0 \quad \text{for } 2 \leq j \leq r \quad (11.7.1)$$

are called *higher order kernels of order r* . Show that $K(x) = H_j(x) \exp(-x^2/2)$, where H_j is a Hermite polynomial (Section 5.3), satisfies (11.7.1).

- 9.** By repeating the arguments of Propositions 11.2.1 and 11.2.2, show that if we assume that $\|f^{(r+3)}\|_{\infty} < \infty$, $m_{r+3}(K) < \infty$, and r is even, then, by using K as in (11.3.7) and bandwidth $cn^{-1/(2r+1)}$, we achieve the uniform rate of convergence $n^{-\frac{2r}{2r+1}}$ of $\text{MSE}[\hat{f}(x)]$ to zero. Show that if K is symmetric and r is odd, the result holds with r replaced by $r - 1$.

- 10.** Assume the conditions of Corollary 11.2.1, that $f(x) > 0$ and $0 < [f''(x)]^2 < \infty$.

- (a) Show that the sum $\text{AMSE}(\hat{f}_h(x))$ of the first two terms in the expansion (11.2.12) of $\text{MSE}[\hat{f}(x)]$ is minimized when $h = cn^{-\frac{1}{5}}$ for some $c \neq 0$. Find the expression for c .

Hint. Note that the expression is of the form $a(nh)^{-1} + bh^2$. Set the derivative equal to zero and solve for h . Check that this gives the minimum.

- (b) Show that using $h = cn^{-\frac{1}{2}}$ leads to the convergence rate $n^{-\frac{4}{5}}$ for $\text{MSE}[\hat{f}_h(x)]$.

- 11.** Establish (11.2.18).

- 12.** Establish (11.2.19) and (11.2.20). Verify that when $F = N(\mu, \sigma^2)$ and $K = U[-1, 1]$, then $f_{\text{opt}} = 1.95\sigma n^{-\frac{1}{5}}$. See also Chapter 12.

- 13.** Generate a sample of size $n = 100$ from a chi-square distribution with 5 degrees of freedom. Plot the density of f and the estimate $\hat{f}_h(\cdot)$ given in (11.2.6) based on

- (a) $\mathcal{N}(0, 1)$ kernel and bandwidth $1.06n^{-\frac{1}{5}}s$ based on a $\mathcal{N}(0, 1)$ reference distribution.

- (b) $U[-1, 1]$ kernel and $h = 1.95n^{-\frac{1}{5}}s$ based on a $\mathcal{N}(0, 1)$ reference distribution.

- 14. Smooth distribution function estimates.** When F is a continuous distribution function, we may want a continuous estimate of F . Consider the two estimates (see Problem 11.2.1).

$$\hat{F}_h(x) = \int_{-\infty}^x \hat{f}_h(t) dt = \int_{-\infty}^x \int_a^b K_h(t-z) d\hat{F}(z) dt$$

$$\widehat{F}(x; h) = \int_{-\infty}^x \widehat{f}(t; h) dt = \int_{-\infty}^x \int_a^b K(t - z; h) d\widehat{F}(z) dt$$

where $K(t - z; h) = K_h(t - z)/w_h(t)$, $w_h(t) = \int_a^b K_h(t - z) dz$. Suppose $S(f) = [a, b]$.

- (a) Let $L(x) = \int_{-\infty}^x K(t) dt$. Show that $\widehat{F}_h(x) = n^{-1} \sum_{i=1}^n \left[L((x - X_i)/h) - L((a - X_i)/h) \right]$.
- (b) When $K = \mathcal{U}[-1, 1]$ and a and b are finite, find $\widehat{F}(x; h)$ for $x \in [a + h, b - h]$, $x = a + \lambda h$ and $x = b - \lambda h$, $0 < \lambda < 1$.
- (c) Show that

$$\begin{aligned} E\widehat{f}_h(x) &= L((x - b)/h) + \int_{(x+b)/h}^{(x-a)/h} F(x - uh) K(u) du \\ &\quad - \left[L((a - b)/h) + \int_{(a-b)/h}^0 F(a - uh) K(u) du \right]. \end{aligned}$$

Hint. Use integration by parts: $\int L dF = LF - \int F dL$.

- (d) When $K = \mathcal{U}[-1, 1]$ and a and b are finite, find $E\widehat{F}(x; h)$ for x as in (b) preceding. Compare your answer to $E\widehat{F}_h(x)$ with $K = \mathcal{U}[-1, 1]$.
- (e) Find $\text{Var}\widehat{F}_h(x)$. Show that we can choose $h = h_n$ so that $\text{MSE}(\widehat{F}_h(x)) = O(n^{-1})$. What smoothness conditions are needed on F ? What conditions are needed on K ?

Hint. Use (c) preceding.

- (f) A good choice for L is the logistic df L_0 . For $K = L_0$, write a program that computes $\widehat{F}_h(x)$, $\widehat{F}(x; h)$, and $\widehat{F}(x)$ and plot the graphs for a sample of size $n = 20$ from $F(x) = 1 - e^{-x}$, $x > 0$. Use $h = 0.5s$. Repeat with $h = 0.25s$ and $h = s$. Include the true $f(x)$ on the figures.

Hint. See Remark 11.2.2.

Problems for Section 11.3

1. Suppose that K has support $[-1, 1]$. Under the conditions of Proposition 11.2.1, show that for $\widehat{f}_h(x)$ defined by (11.2.6), if $S(f) = [a, b]$ with a and b finite and $x = a + \lambda h$, $0 < \lambda < 1$, then

$$E\widehat{f}_h(x) = \mu_{0,\lambda}(K)f(x) - h\mu_{1,\lambda}(K)f'(x) + \frac{1}{2}h^2 f''(x)\mu_{2,\lambda}(K) + O(h^3)$$

where $\mu_{j,\lambda}(K) = \int_{-1}^{\lambda} u^j K(u) du$.

2. Suppose $K(\cdot)$ is a symmetric density on $[-1, 1]$. Assume the conditions of Problem 11.3.1. Show that the estimate $\hat{f}_h(x)$ based on (11.3.2) has bias $O(h)$ for the boundary points $a + \lambda h$ and $b - \lambda h$, $0 < \lambda < 1$.

3. Establish (11.3.2).

4. Let $\hat{f}_h(x)$ be defined by (11.2.6). Suppose f is the $\mathcal{U}[0, 1]$ density, $0 < h < \frac{1}{2}$, and $h \leq x \leq 1 - h$. When $K = \mathcal{U}[-1, 1]$ find

(a) $\text{MSE}[\hat{f}_h(x)]$ and give the rate at which $\text{MSE}[\hat{f}_h(x)]$ tends to zero as $n \rightarrow \infty$.

(b) h_{opt} which minimize $\text{MSE}[\hat{f}_h(x)]$.

(c) Solve (a) and (b) above when $K(u) = \frac{3}{4}(1 - u^2)1(|u| \leq 1)$.

(d) Solve (c) when $x = \lambda h$, $0 < \lambda < 1$.

5. In Example 11.3.1, (see (11.3.4)) set $m_j = m_j(x, h)$, $j = 0, 1, 2$, and

$$s_j = \int_a^b (z - x)^j K_h(z - x) dF(z), \quad j = 0, 1.$$

(a) Show that the solution to (11.3.4) is

$$(\alpha, \beta)^T = \begin{pmatrix} s_0 h^2 m_2 - s_1 h m_1 \\ s_1 m_0 - s_0 h m_0 \end{pmatrix} \frac{1}{h^2(m_0 m_2 - m_1^2)}.$$

(b) Use (a) preceding to show that an estimate of $f'(x)$ is

$$f'(x, \hat{F}) \equiv \beta(x, \hat{F}) = \int_a^b K_h(x, z) d\hat{F}(z)$$

where

$$K_h(x, z) = \frac{m_0(x, h)[(z - x)/h - 1]/h}{m_0(x, h)m_2(x, h) - m_1^2(x, h)} K_h(z - x).$$

(c) Establish (11.3.8) and (11.3.9).

Hint: For $x \in [a + h, b - h]$, by (11.3.5), this follows from Propositions 11.2.1 and 11.2.2. For $x_h = a + \lambda h$, $0 < \lambda < 1$, follow the proofs of these propositions.

6. Local polynomial estimates. Suppose that for z close to x , in Example 11.2.1 continued, we use a local polynomial approximation

$$f(z - x; \boldsymbol{\beta}) \cong \beta_0 + \sum_{j=1}^p \beta_j (z - x)^j. \quad (11.7.2)$$

That is, we approximate $f(x)$ by $f_h(x) = f(0; \boldsymbol{\beta}(x)) = \beta_0(x)$ where

$$\boldsymbol{\beta}(x) = \boldsymbol{\beta}(F; x) = \arg \min_{\boldsymbol{\beta}} \int [f(z) - f(z - x; \boldsymbol{\beta})]^2 K_h(z - x) dz.$$

(a) Show that $\beta(x) = \beta(F; x)$ is the functional

$$\beta(F; x) = H^{-1}C^{-1}a(F)$$

where $H = \text{diag}(1, h, \dots, h^p)$, $C = (s_{jk})$, $a(F) = (a_0(F), \dots, a_p(F))^T$,

$$s_{jk} = h^{[1-(j+k)]} \int_{(x-a)/h}^{(b-x)/h} u^{j+k} K(u) du, \quad j, k = 0, 1, \dots, p,$$

$$a_j(F) = \int \left(\frac{z-x}{h} \right)^j K_h(z-x) dF(z), \quad j = 0, \dots, p.$$

Hint. This $\beta(x)$ satisfies the linear equations, for $k = 0, \dots, p$,

$$\int (z-x)^k K_h(z-x) dF(z) = \sum_{j=0}^p \beta_j \int (z-x)^{j+k} K_h(z-x) dz.$$

Our approximation to $f(x)$ is the functional $\beta_0(F; x)$ with plug-in estimate $\hat{f}_h(x) = \beta_0(\hat{F}; x)$.

(b) Show that if $f(x)$ is a polynomial of order p and if $\hat{f}_h(x)$ is the empirical plug-in estimate of $f(x)$ obtained from the local polynomial approximation to $f(x)$ of order p , then $\hat{f}_h(x)$ is an unbiased estimate of $f(x)$.

If $f(x)$ has a continuous $(p+2)$ th derivative at the point x in its support, if p is odd, and if K is continuous and symmetric with support $[-1, 1]$, then it can be shown (Jiang and Doksum (2003)) that this $\hat{f}_h(x)$ converges in probability to $f(x)$ at the rate $O_P(n^{-a})$ with $a = (p+1)/(2p+3)$ provided the optimal (the one minimizing asymptotic MSE) $h \asymp (n^{(-2p+3)^{-1}})$ is employed. This holds for boundary as well as interior points. Note that if $f(x)$ is infinitely differentiable, the rate of convergence of $\hat{f}_h(x)$ to $f(x)$ can be made arbitrarily close to the regular rate $n^{-\frac{1}{2}}$ by choosing a polynomial approximation with p large.

Suppose we assume that $f(x)$ has a continuous $(p+2)$ th derivative. Are there estimates that achieve better rates of convergence than $O_P(n^{-(p+1)/(2p+1)})$? The answer is no; see e.g. van der Vaart (1998).

7. Show that $\alpha(x; F)$ as given in Remark 11.3.1 reduces to $f_h(x; F) = \int K_h(x, z) dF(z)$ where $K_h(\cdot, \cdot)$ is defined by (11.3.7).

Hint. $E[(Z-x)f(Z)|X=x] = \frac{\int (z-x)K_h(z-x) dF(z)}{m_0(x; F)}$,

$$E(Z-x|X=x) = \frac{\int (z-x)K_h(z-x) dz}{m_0(x; h)} = \frac{hm_1(x, h)}{m_0(x, h)}, \text{ etc.}$$

8. In Example 11.3.2, show that $\int K_m(x, z) dx = 1$ when B_0 is a constant and a and b are finite.

9. Let X_1, \dots, X_{200} be a sample from a beta (2,3) distribution. Consider the grid $\{0, 0.02, \dots, 0.98, 1\}$.

- (a) Plot the locally linear density estimate (11.3.6) using the algorithm in Remark 11.3.2 using $h = 0.1$. Also plot the beta (2,3) density on the same figure.
- (b) For comparison, repeat (a) with (11.3.6) replaced by (11.2.6).

Problems for Section 11.4

1. Consider a sample X_1, \dots, X_n from $f_2(x, \boldsymbol{\eta})$, $x \in R$, as given in (11.4.3) with $h(x) \equiv 1$, $B_1(x) = x$ and $B_2(x) = x^2 - 1$, the first two Hermite polynomials.

(a) Give the asymptotic covariance matrix of the MLE $\hat{\boldsymbol{\eta}}$ of $\boldsymbol{\eta}$.

(b) Suppose we estimate $\boldsymbol{\eta}$ using the Rudemo criteria. We obtain

$$\hat{\boldsymbol{\eta}}_R = \arg \min \left\{ -2 \int f_2(x, \boldsymbol{\eta}) d\hat{P}(x) + \int f_2^2(x, \boldsymbol{\eta}) dx \right\} .$$

Give the asymptotic covariace matrix of $\hat{\boldsymbol{\eta}}_R$.

- (c) Compare your answers to (a) and (b). Because the MLE of a function of a parameter is the function of the MLE of this parameter, this comparison has implications for the comparison of $f_2(x, \hat{\boldsymbol{\eta}})$ and $f_2(x, \hat{\boldsymbol{\eta}}_R)$. Give this implication.

Hint. $f_2(x, \boldsymbol{\eta})$ is a normal density.

Remark. If X_1, \dots, X_n is from $f(\cdot) \neq f_2(\cdot, \boldsymbol{\eta})$, it could be that $f(x, \hat{\boldsymbol{\eta}}_R)$ has smaller IMSE than $f(x, \hat{\boldsymbol{\eta}})$.

2. Show that if $f(x) > 0$, $x \in R$, then for the nearest neighbor estimate (11.4.6), $p \lim_{|x| \rightarrow \infty} \tilde{f}_k(x) > 0$.

Hint. $2\hat{h}(k, x)$ is the difference $X_{(l)} - X_{(j)}$ of two order statistics where $(X_{(j)}, X_{(l)})$ contains k order statistics. Let $r = [nF(x)]$; then $l \leq r + 2k$, $j \geq r - 2k$, $(l/n) \rightarrow F(x)$, and $(j/n) \rightarrow F(x)$.

3. Suppose $S(f) = [a, b]$ with a and b finite. Show that $\lim_{|x| \rightarrow \infty} \tilde{f}_k(x) = 0$.

Hint. See Problem 11.4.2. To contain k order statistics, the length of the interval $(X_{(j)}, X_{(l)})$ must tend to ∞ as $|x| \rightarrow \infty$.

4. Let $\tilde{f}_k(x)$ be the k th nearest neighbor density estimate. Assume the conditions of Theorem 11.4.3.1.

(a) Derive an expression of the form (11.2.12) for $MSE[\tilde{f}_k(x)]$.

(b) Compare the solution to (a) preceding to $MSE[\tilde{f}_k(x)]$ as given by (11.2.12) for the case $K(w) = 1[|u| \leq 1]/2$.

(c) Derive an expression of the form (11.2.14) for $IMSE(\tilde{f}_k, f)$.

(d) Compare the solution to (c) preceding to $IMSE(\tilde{f}_h, f)$ as given by (11.2.14) for $K(u) = 1[|u| \leq 1]/2$.

- (e) Justify informally the approximation $k \cong 2nhf(x)$ for n large.
- (f) Use (e) to informally find the k that minimizes IMSE (see (11.2.19)).
- (g) Use (f) to informally find the k that minimizes IMSE for the $\mathcal{N}(\mu, \sigma^2)$ reference distribution (see Section 11.2.3).

Problems for Section 11.5

1. Argue that T as defined by (11.5.3) has an asymptotic $\mathcal{N}(0, 1)$ distribution.
2. Let X_1, \dots, X_{1000} be a sample from a χ^2_5 distribution. Plot \hat{f}_h and the confidence intervals (11.5.2) at the grid points $\{0.5, 1.0, 1.5, 2, \dots, 8\}$ using the $U(-1, 1)$ kernel and $h = 0.2$.

Problems for Section 11.6

1. Let

$$\begin{aligned} D_x(\mu(\cdot), \mu(\cdot; \beta)) &= \frac{1}{f_h(x)} \int [\mu(z) - \mu(z; \beta)]^2 K_h(z - x) dF(z) \\ &= E\{[\mu(Z) - \mu(Z; \beta(X))]^2 | X = x\} \end{aligned} \quad (11.7.3)$$

where (X, Z) has joint density $f(x)q_h(z|x)$, $(x, z) \in [a, b]$, with

$$q_h(z|x) = \frac{f(z)K_h(z - x)1(x \in [a, b])}{f_h(x)}.$$

Show that minimizing (11.7.3) is equivalent to minimizing (11.6.2).

2. Show that if we select $\mu(z, \beta) \equiv \beta \in R$ in (12.6.13), then (12.6.13) yields $\hat{\mu}_{\text{NW}}(\cdot)$.
3. Establish (11.6.4).
4. Establish (11.6.25).
5. In Remark 11.6.1, give a type (11.6.10) expansion of $\text{Bias}(\hat{\mu}_{\text{NW}}(x)|\mathbf{X})$ when K is a kernel of order 3.
6. In Remark 11.6.2, justify the formula for (a) h_{optimal} ; (b) the conditional MSE of $\hat{\mu}_{\text{NW}}(x)$; (c) the optimal convergence rate $n^{-\frac{4}{5}}$ for the MSE in (b).
7. Establish (11.6.26).
8. Consider the model $Y_{ni} = \mu(x_{ni}) + \varepsilon_{ni}$, where $x_{n1} \leq \dots \leq x_{nn}$, $E(\varepsilon_{ni}) = 0$, $\text{Var}(\varepsilon_{ni}) = \sigma^2(x_{ni})$. Let $I_{nk}(x)$ denote the indices of the k values of x_{n1}, \dots, x_{nn} closest to x , where ties are broken by inducing only the smallest index. Define the *nearest neighbor estimate*

$$\hat{\mu}(x) = \sum_{i \in I_{nk}(x)} Y_i/k.$$

(a) Show that

$$P\left(\sup_x |\widehat{\mu}(x) - E\widehat{\mu}(x)| \leq t\right) \geq 1 - 8(kt)^{-2} \left[\sum_{i=1}^{n-k} \sigma^2(x_{ni}) + \sum_{i=1}^n \sigma^2(x_{ni}) \right].$$

Hint: Use Kolmogorov's inequality.

- (b) Find bounds on $\sup_x |E\widehat{\mu}(x) - \mu(x)|$. (Bjerve, Doksum and Yandell (1985) bound $\sup_x |\widehat{\mu}(x) - \mu(x)|$ in probability.)

9. Nonparametric binary regression. Suppose we want to study the effect of a certain treatment. Consider the following approach. Dosages $x_1 < \dots < x_n$ are fixed. Dosage x_i is administered to the i th member of this sample and a binary response Y_i is recorded; e.g., $Y_i = 0$ if the treatment does not have a beneficial effect, $Y_i = 1$ if the treatment has a beneficial effect. Assume that the Y_i are independent and that Y_i has the Bernoulli distribution with success probability $\theta(x_i)$, $i = 1, \dots, n$. Let x be some fixed dosage level.

Consider the following nearest neighbor estimate of $\theta(x)$:

$$\widehat{\theta}(x) = \frac{1}{k} \sum_{j=i+1}^{i+k} Y_j, \quad x \in I_i, \quad i = 0, \dots, n-k$$

where

$$\begin{aligned} I_0 &= \left(-\infty, \frac{1}{2}(x_1 + x_{1+k})\right] \\ I_i &= \left(\frac{1}{2}(x_i + x_{i+k}), \frac{1}{2}(x_{i+1} + x_{i+k+1})\right], \quad i = 1, \dots, n-k-1 \\ I_{n-k} &= \left(\frac{1}{2}(x_{n-k} + x_n), \infty\right). \end{aligned}$$

Show that

$$Pr\left(\sup_{-\infty < x < \infty} |\widehat{\theta}(x) - E(\widehat{\theta}(x))| \leq \varepsilon\right) \geq 1 - \frac{n}{\varepsilon^2 k^2}.$$

Hint: Use Kolmogorov's inequality.

Chapter 12

PREDICTION AND MACHINE LEARNING

12.1 Introduction

In this chapter we face the major challenge of modern statistics, devising and analyzing methods of inference for high dimensional complex data. The issues we discuss in this chapter are:

1. *Prediction with high dimensional covariates*

Under this heading we will consider classification, regression and prediction problems now associated with machine learning, and statistical learning. The methods we shall briefly discuss include ones easily related to statistical methods such as the univariate curve estimation in Chapter 11 and the logistic regression in Sections 6.4.3 and 6.5; and ones more closely related to machine learning such as support vector machines and boosting.

2. *Regularization*

A central theme as in the last chapter is regularization. How do we control overcomplexification of models and procedures which both theoretically and practically lead to poor performance? We will discuss major semiparametric and parametric methods such as those based on sieves (Section 9.1.4), shrinkage (Section I.6), and penalized least squares.

3. *Dimension reduction and model selection*

While prediction and classification are the primary goals in many engineering and marketing applications, the goal in science is usually to try to isolate key factors which can be interpreted. Factors with subject matter interpretations, if available, can lead to the construction of better predictors and scientific insights. Isolating key factors is usually identified in statistics with *model selection* and *variable selection* — see Section I.7.

4. *Nonparametric prediction and classification. The Bayes classifier.*

We were presented in Section 1.4 with the classical prediction problem: Given a d dimensional vector $\mathbf{Z} = (Z_1, \dots, Z_d)^T$ of covariates, predict the value of a response Y . In that section we showed that for quadratic loss, the best predictor is

$$\mu(\mathbf{Z}) = E(Y|\mathbf{Z}).$$

In the ‘‘known P ’’ case, the joint distribution of (\mathbf{Z}, Y) is known and μ can be computed. In practice, when P is unknown, we can construct an estimate $\hat{\mu}_n(\cdot)$ based on a *sample* $(\mathbf{Z}_1, Y_1), \dots, (\mathbf{Z}_n, Y_n)$ i.i.d. as $(\mathbf{Z}, Y) \sim P$. When $k > n$ and \mathbf{Z}_k is known but Y_k is unknown, we can use $\hat{\mu}(\mathbf{Z}_k)$ to predict Y_k from \mathbf{Z}_k where $(\mathbf{Z}_k, Y_k) \sim P$. In this context, $(Z_1, Y_1), \dots, (Z_n, Y_n)$ is referred to as a *training sample*.

We may try to estimate $\mu(\mathbf{z}) \equiv E(Y|\mathbf{Z} = \mathbf{z})$ by $\hat{\mu}(\mathbf{z}) \equiv \hat{E}(Y|\mathbf{Z} = \mathbf{z})$ where \hat{E} is the expected value under the empirical probability \hat{P} which assign probability n^{-1} to each observed data vector (\mathbf{z}_i, y_i) , and 0 elsewhere. If \mathbf{Z} is discrete with a small number of values, this is a parametric problem and our choice of $\hat{\mu}(\cdot)$ is reasonable if we observe all possible values of \mathbf{Z} . However, if as is typically the case, \mathbf{Z} is continuous, then the estimate $\hat{\mu}(\mathbf{z})$ is undefined for unobserved \mathbf{z} . To go further we need to assume some special properties of $\mu(\cdot)$, most naturally, smoothness.

In some parts of this chapter we will consider generalizations of the problem of estimating a nonparametric regression that we discussed in Section 11.6. The major differences are:

i) The Y to be predicted can be categorical so that 0 – 1 loss as discussed in Section 1.3 is appropriate. That is, we have a problem of *classification* where Y can take on values $1, \dots, C$ corresponding to C distinct categories. Throughout we consider the loss function

$$\begin{aligned} l(j, d) &= 0 \text{ if } d = j \\ &= 1 \text{ otherwise} \end{aligned}$$

where d is the predicted value (class) and j is the actual value. Then, if $p_j(\mathbf{z})$ is the conditional density of \mathbf{Z} given $Y = j$, and the ‘‘prior’’ probability is $P[Y = j] = \pi_j$, the conditional Bayes risk of a classifier d given $\mathbf{Z} = \mathbf{z}$ is

$$E(l(Y, d)|\mathbf{z}) = 1 - P(Y = d|\mathbf{z}) = 1 - \pi_d p_d(\mathbf{z})/q(\mathbf{z})$$

where $q(\mathbf{z})$ is the density of \mathbf{Z} . Thus the conditional Bayes risk is minimized by

$$\begin{aligned} \delta_B(\mathbf{Z}) &= \arg \max_j P[Y = j|\mathbf{Z}] \\ &= \arg \max_j \pi_j p_j(\mathbf{Z}). \end{aligned} \tag{12.1.1}$$

This δ_B is called the *Bayes classifier*. We also see that the minimum Bayes risk is

$$R_B \equiv 1 - E\left[\max_j \{\pi_j p_j(\mathbf{Z})\}/q(\mathbf{Z})\right]. \tag{12.1.2}$$

Since $P[Y = j|\mathbf{Z}] = E(l(Y, d)|\mathbf{Z})$, we see that, in some ways, the classification problem can be viewed as the same as estimating the regression of $1(Y = j)$ on \mathbf{Z} from a sample, where a *sample* $(\mathbf{Z}_1, Y_1), \dots, (\mathbf{Z}_n, Y_n)$ is i.i.d. as $(\mathbf{Z}, Y) \sim P$ with $Y \in \{1, \dots, C\}$. The sample is used to construct a method that can be used to classify Y_k from \mathbf{Z}_k for $k > n$, where $(\mathbf{Z}_k, Y_k) \sim P$. This version of classification as estimation is somewhat misleading in some frameworks, as we shall see in Section 12.2.4 when we consider boosting.

ii) The predictor \mathbf{Z} is potentially very high dimensional. The coordinates Z_1, \dots, Z_d can be categorical or ordinal as well as real but what matters is that there are a lot of them. The high dimensional case is important because of new technologies that generate vast amounts of data.

It is worth noting that high dimensionality is implicit in the case of one dimensional regression as well. As is well known in prediction, covariates can always be added to a regression. Thus, if $Z \in [0, 1]$, say, then $E(Y|Z) = E(Y|Z, Z^2, \dots, Z^p)$, and estimation of the second form leads naturally to linear regression of Y on the $p+1$ vector $(1, Z, \dots, Z^p)^T$, a method we considered in Section 11.6.2, for smooth $\mu(Z)$. Similarly, if Z takes on a large number of known categories, the categories can, for instance, be coded by $\mathbf{e}_1, \dots, \mathbf{e}_d$, the standard basis vectors of R^d . That is, $\mathbf{e}_1 = (1, 0, 0, \dots, 0)^T$, $\mathbf{e}_2 = (0, 1, 0, \dots, 0)^T, \dots, \mathbf{e}_k = (0, 0, \dots, 0, 1)$. Now if C_j denotes the j th category, $E(Y|Z \in C_j) = E(Y|\mathbf{e}_j)$.

iii) In addition to estimation of the regression $E(Y|\mathbf{Z})$, the prediction of Y is of major concern. In particular, prediction error plays a key role in model and tuning parameter selection based on cross validation.

□

Remark 12.1.1. *Prediction versus estimation.* This chapter examines ways of constructing good predictors; however the discussion is sometimes in terms of estimation. This is because the problem of selecting a good predictor is sometimes equivalent to selecting a good estimator. For squared error loss this is clear because if $\hat{\mu}(\cdot)$ is based on $\mathbf{X} = (\mathbf{Z}_1, Y_1), \dots, (\mathbf{Z}_n, Y_n)$ and we are trying to predict Y_{n+1} from \mathbf{Z}_{n+1} , where $(\mathbf{Z}_{n+1}, Y_{n+1})$ is independent of X , then, for $\mu(\mathbf{Z}) = E(Y|\mathbf{Z})$,

$$\begin{aligned}\text{Prediction MSE} &= E\{[Y_{n+1} - \hat{\mu}(\mathbf{Z}_{n+1})]^2\} \\ &= E\{[Y_{n+1} - \mu(\mathbf{Z}_{n+1})]^2\} + E\{[\mu(\mathbf{Z}_{n+1}) - \hat{\mu}(\mathbf{Z}_{n+1})]^2\} \\ &= \text{constant} + \text{estimation IMSE}\end{aligned}$$

because the cross product term is zero by the iterated expectation theorem. Thus optimal prediction is equivalent to optimal estimation in this case. (Here the constant does not depend on the procedures $\hat{\mu}$, but it contributes a substantial amount to the prediction MSE of any predictor. Thus while estimation IMSE typically tends to zero, prediction MSE does not.) □

The formulation given so far is the center of our discussion in this chapter which examines most of the major approaches to regularization, dimension reduction, model selection and to the construction of prediction rules, regression procedures, and classifiers. We next add more introductory details.

12.1.1 Statistical Approaches to Modeling and Analyzing Multidimensional data. Sieves

As in Section 12.1, let $(\mathbf{Z}_1, Y_1), \dots, (\mathbf{Z}_n, Y_n)$ be i.i.d. as $(\mathbf{Z}, Y) \sim P$ with $\mathbf{Z} = (Z_1, \dots, Z_d)^T$ and $Y \in R$. Consider the problems of estimating $\mu(\mathbf{z}) = E(Y|\mathbf{Z} = \mathbf{z})$, predicting Y from \mathbf{Z} , or using \mathbf{Z} to classify Y . Two methods are:

a) Generalizations of kernel regression methods. These methods for the estimation of $E(Y|\mathbf{Z})$ use product kernels to replace the univariate ones of Section 11.6.2. See Section 12.2.1.

b) Regularization methods based on sieves. We pursue the regularization ideas introduced in Section 9.1.4. Thus for P in a general class of models, we approximate $\mu(\mathbf{Z}) = E_P(Y|\mathbf{Z})$ by postulating a sequence of parametric models $\{\mathcal{P}_j\}$ for (\mathbf{Z}, Y) with members P_j of \mathcal{P}_j identified by a k_j dimensional parameter $\boldsymbol{\theta}^{(j)}(P)$ which is defined for all P . We assume that each P_j is regular as defined in Section 1.1.3. Define $\mu_j(\mathbf{Z}) = E_{P_j}(\mathbf{Z})$. We require that for any P in the nonparametric class we consider, as $j \rightarrow \infty$,

$$\mu_j(\mathbf{Z}) \rightarrow \mu(\mathbf{Z}),$$

where “ \rightarrow ” denotes convergence in an appropriate metric. The parameter $\boldsymbol{\theta}^{(j)}(P)$ is defined by minimum contrast; e.g. if p_j denotes the density of P_j , we could use the Kullback-Leibler contrast and set

$$\boldsymbol{\theta}^{(j)}(P) = \arg \min_{\boldsymbol{\theta}^{(j)}} \int \log p_j(\mathbf{z}, y) dP(\mathbf{z}, y).$$

We can then generate a sequence $\{\hat{\mu}_j(\cdot)\}$ of estimates or predictors with $\hat{\mu}_j(\cdot) \equiv E_{\hat{P}_j}(Y|\mathbf{Z})$ and $\hat{P}_j \in \mathcal{P}_j$ corresponding to $\boldsymbol{\theta}^{(j)}(\hat{P})$, where \hat{P} denotes the empirical distribution function based on $(\mathbf{Z}_1, Y_1), \dots, (\mathbf{Z}_n, Y_n)$. Here regularization is achieved by using a sieve made up of regular parametric models. The index j is called a *regularization* or *tuning parameter* and is chosen to minimize an estimate of risk. We postpone the issue of choosing $\hat{j} \equiv j(\hat{P})$ to obtain our final estimate $\hat{\mu}_{\hat{j}}(\cdot)$. See Section 12.5. There is a huge literature on regularization; see Bickel and Li (2006), for discussions and a bibliography. We now turn to several special cases of sieves.

(1) Gaussian Sieves

A natural special case, if, say, $\mathbf{Z} \in I^d$ where $I \equiv [0, 1]$, is to choose an orthonormal basis $\{f_k(\cdot)\}_{k \geq 1}$ with $f_1 \equiv 1$,

$$\int_{I^d} f_a(\mathbf{z}) f_b(\mathbf{z}) d\mathbf{z} = 1(a = b)$$

and postulate as model \mathcal{P}_j with dimension k_j

$$Y = \sum_{k=1}^{k_j} \theta_k^{(j)} f_k(\mathbf{Z}) + \varepsilon \tag{12.1.3}$$

where $\varepsilon \sim \mathcal{N}(0, \sigma^2)$ independent of \mathbf{Z} , and $\boldsymbol{\theta}^{(j)} = (\theta_1^{(j)}, \dots, \theta_{k_j}^{(j)})^T$ is defined as the minimizer of mean squared prediction error as in Theorem 1.4.4.

(2) Penalized Least Squares and Ridge Regression

Having chosen $\{f_k\}$ as in (12.1.3), it is natural to consider a Lagrangian version of the method of sieves which simplifies computation and makes the need for regularization explicit. Choose a function $J(\boldsymbol{\theta})$, $J : R \times R^{k_j} \rightarrow R^+$ and $\lambda > 0$; and in model \mathcal{P}_j minimize, over $\boldsymbol{\theta}^{(j)}$,

$$\frac{1}{n} \sum_{i=1}^n \left(Y_i - \sum_{k=1}^{k_j} \theta_k^{(j)} f_k(\mathbf{Z}_i) \right)^2 + \lambda J(\boldsymbol{\theta}^{(j)}) .$$

The population parameter $\boldsymbol{\theta}_\lambda^{(j)}(P)$ in model \mathcal{P}_j is defined by

$$\boldsymbol{\theta}_\lambda^{(j)}(P) = \arg \min \left\{ \int (y - [\boldsymbol{\theta}^{(j)}]^T \mathbf{f}^{(j)}(\mathbf{z}))^2 dP(\mathbf{z}, y) + \lambda J(\boldsymbol{\theta}^{(j)}) : \boldsymbol{\theta}^{(j)} \right\} ,$$

where $\mathbf{f}^{(j)}(\cdot) = (f_1(\cdot), \dots, f_{k_j}(\cdot))^T$ and $P(\cdot, \cdot)$ is the probability distribution of (\mathbf{Z}, Y) . The empirical plug-in estimate is now $\widehat{\boldsymbol{\theta}}_\lambda^{(j)} = \boldsymbol{\theta}_\lambda^{(j)}(\widehat{P})$. Note that this approach still falls under our general formulation of empirical plug-in estimation.

A choice we pursue further is *ridge regression* which corresponds to

$$J(\boldsymbol{\theta}^{(j)}) = \frac{1}{2} \sum_{k=1}^{k_j} [\theta_k^{(j)}]^2 = \frac{1}{2} |\boldsymbol{\theta}^{(j)}|^2 .$$

Another interesting choice is

$$J(\boldsymbol{\theta}^{(j)}) = \sum_{k=1}^{k_j} |\theta_k^{(j)}|$$

which corresponds to the *Lasso*.

(3) Bayes and Penalized Least Squares

Ridge regression has a natural Bayesian interpretation. Consider the sequence of models $\{\mathcal{P}_j\}$ as a dense subset of the grand model,

$$Y = \sum_{k=1}^{\infty} \theta_k f_k(\mathbf{Z}) + \varepsilon \tag{12.1.4}$$

where $\varepsilon \sim \mathcal{N}(0, \sigma^2)$. Now assume that $\theta_1, \dots, \theta_j$ are i.i.d. $\mathcal{N}(0, 1/\lambda)$, $\theta_k = 0$, $k > j$. Then, for fixed j and λ , the ridge regression estimate $\widehat{\boldsymbol{\theta}}_\lambda^{(j)}$ is just the mode of the posterior distribution of $\boldsymbol{\theta}^{(j)}$ given the data.

This formulation suggests that other choices of prior distribution might also give interesting results. A class of possible penalties corresponding to priors in this way is

$$J_r(\boldsymbol{\theta}) = \sum_{k=1}^j |\theta_k|^r, \quad 0 < r < \infty ,$$

with

$$J_0(\boldsymbol{\theta}) = \sum_{k=1}^j 1(|\theta_k| \neq 0)$$

defined by a limiting process as $r \rightarrow 0$ and not corresponding to a proper prior. Among $J_r(\boldsymbol{\theta})$, $r > 0$, the penalty $J_1(\boldsymbol{\theta})$ which corresponds to $\theta_1, \dots, \theta_j$ i.i.d. with prior $\pi(\theta) = \frac{1}{2}e^{-|\theta|}$, the *Laplace* or *double exponential distribution*, has particularly interesting properties. See Section 12.2.3.

(4) Logistic Sieves. Linear Classifiers

Another important sieve suitable for classification problems is based on logistic regression. Here Y takes on the values $1, \dots, C$. The functions to be estimated in the Bayes classifier (12.1.1) are $\pi_a p_a(\mathbf{Z})$, $a = 1, \dots, C$, where $\pi_a = P(Y = a)$ and $p_a(Z)$ is the density of Z given $Y = a$. Of these, if we have a training sample, π_a is estimated by $n^{-1} \sum_{i=1}^n 1(Y_i = a)$, the fraction of class a in the training sample, while estimating $p_a(\mathbf{Z})$ is the problem of estimating a general multivariate density from $\{\mathbf{Z}_i : Y_i = a\}$.

Alternatively we can think of, from the beginning, estimating $(\pi(1|\mathbf{Z}), \dots, \pi(C|\mathbf{Z}))$, where

$$\pi(a|\mathbf{Z}) \equiv P[Y = a|\mathbf{Z}] = \pi_a p_a(\mathbf{Z}) / \sum_{k=1}^C \pi_k p_k(\mathbf{Z}),$$

using a generalized linear logistic regression model (see Sections 1.6.2, 6.4.3, 6.5 and Hastie, Tibshirani and Friedman (2001, 2009)). That is, given functions $f_1(\mathbf{Z}), \dots, f_{k_j}(\mathbf{Z})$ and $1 \leq a \leq C$, specify conditionally, with $\boldsymbol{\pi} \equiv (\pi_1, \dots, \pi_C)$, $\boldsymbol{\theta} \equiv \{\theta_{a,k}^{(j)} : 1 \leq k \leq k_j, 1 \leq a \leq C\}$, the j th model as

$$\pi(a|\mathbf{Z}, \boldsymbol{\theta}, \boldsymbol{\pi}) = \frac{\pi_a \exp\{\sum_{k=1}^{k_j} \theta_{a,k}^{(j)} f_k(\mathbf{Z})\}}{\sum_{l=1}^C \pi_l \exp\{\sum_{k=1}^{k_j} \theta_{l,k}^{(j)} f_k(\mathbf{Z})\}}, \quad (12.1.5)$$

the *logistic regression model*.

In the j th model \mathcal{P}_j , the classifier using training data based on (12.1.5) classifies an unknown Y corresponding to covariate vector \mathbf{Z} by

$$\delta(\mathbf{Z}, \widehat{\boldsymbol{\theta}}) = a \text{ if } \sum_{k=1}^{k_j} (\widehat{\theta}_{a,k}^{(j)} - \widehat{\theta}_{l,k}^{(j)}) f_k(\mathbf{Z}) \geq \log(\widehat{\pi}_l / \widehat{\pi}_a) \quad (12.1.6)$$

for all l , with ties broken arbitrarily. Here $\widehat{\boldsymbol{\theta}}$ and $\widehat{\boldsymbol{\pi}}$ are obtained by maximum likelihood from

$$L(\boldsymbol{\theta}, \boldsymbol{\pi}) = \prod_{i=1}^n \prod_{a=1}^C \pi^{\varepsilon_{ia}}(a|\mathbf{Z}_i, \boldsymbol{\theta}, \boldsymbol{\pi}),$$

where $\varepsilon_{ia} = 1(Y_i = a)$. Thus $\widehat{\pi}_a = \sum_i \varepsilon_{ia}/n$ and $\widehat{\boldsymbol{\theta}}$ is the MLE for a canonical exponential family — see Example 1.6.7 and Section 2.3, for instance. The classifier (12.1.6) is an example of a *linear classifier*, that is, one such that $\log[\widehat{\pi}(a|\mathbf{Z})/\widehat{\pi}(l|\mathbf{Z})]$ is a linear combination of functions $f_k(\mathbf{Z})$, $1 \leq k \leq k_j$. Geometrically, we can think of such classifiers as

classifying a new point Y with given \mathbf{Z} in class l if $(f_1(\mathbf{Z}), \dots, f_k(\mathbf{Z}))^T$ is on one side of a hyperplane in R^k and “not l ” if it is on the other side. We have already seen one example of such a classifier, Fisher’s linear discriminant function in Section 4.2.

12.1.2 Machine Learning Approaches

The machine learning point of view as discussed in Section 11.4.2 is to specify a parametric family \mathcal{D} of possible decision procedures, based on predictors in our case, and select a member by some optimization criteria which may or may not include a penalty for complexity (overfitting). The focus is, in particular, on classification.

a) Neural Nets

The motivation behind neural nets is a primitive notion of how human stimulus response learning proceeds. Abstracted it is relatively simple. If $\mathbf{Z}_{d \times 1}$ is the vector of predictors, let the classes that a corresponding Y can fall in be $\{1, \dots, C\}$. By definition, a *neural net* with one hidden layer of M interior nodes, based on $\{(\mathbf{Z}_1, Y_1), \dots, (\mathbf{Z}_n, Y_n)\}$, outputs a vector, $\{0 \leq \hat{\pi}(j|\mathbf{Z}) : 1 \leq j \leq C, \sum_{k=1}^C \hat{\pi}(k|\mathbf{Z}) = 1\}$, of estimated posterior probabilities $j, 1 \leq j \leq n$, satisfying

$$\log \frac{\hat{\pi}(j|\mathbf{Z})}{\hat{\pi}(1|\mathbf{Z})} = (\hat{\beta}_j - \hat{\beta}_1)^T \boldsymbol{\sigma}(\hat{A}\mathbf{Z}^*) . \quad (12.1.7)$$

Here the $\hat{\beta}_j$ are $(M+1) \times 1$, \hat{A} is $(M+1) \times (d+1)$, $\mathbf{Z}^* = (1, \mathbf{Z}^T)^T$, and $\boldsymbol{\sigma}(\omega) = (\sigma(\omega_1), \dots, \sigma(\omega_{M+1}))^T$, where $\sigma(\omega) = (1 + e^{-\omega})^{-1}$ and $\omega^T = (\omega_1, \dots, \omega_{M+1})$. Essentially the classifier is obtained by a fitting process from all classifiers based on

$$\arg \max_j \pi(j|\mathbf{Z}, A, B)$$

where A is $(M+1) \times (d+1)$, $B = (\beta_1, \dots, \beta_C)$ is $(M+1) \times C$, and $\pi(j|\mathbf{Z}, A, B)$ is given by (12.1.7) with the “hats” removed.

Neural nets essentially use logistic regression applied to a nonlinear coordinatewise transformation of an affine transformation of \mathbf{Z} into R^{M+1} . The second optimization problem is not convex. The tuning parameter here is M , the number of nodes in the hidden layer parametrized by A . It may be shown that as $M \rightarrow \infty$ the functions

$$\log [\pi(j|\mathbf{Z}, A, B)/\pi(1|\mathbf{Z}, A, B)]$$

can approximate any $\log [\pi(j, \mathbf{Z})/\pi(1, \mathbf{Z})]$. Neural nets may also be applied to the estimation of the regression mean $\mu(\mathbf{z}) = E(Y|\mathbf{Z} = \mathbf{z})$. See Problem 2.3.41. We do not pursue this method. A full discussion may be found, for instance, in Ripley (1996) and Hastie et al (2001, 2009).

b) Support Vector Machines

This method introduced by Vapnik (1996) is limited to classification. We discuss it initially for the case of $C = 2$ classes here denoted by “−1” and “1”. These are, in fact, linear

classifiers, but as with neural nets, follow a nonlinear transformation of \mathbf{Z} from R^d to a typically much higher dimensional space R^M . As we remarked earlier, given any \mathbf{Z} , we can always choose an orthonormal basis in $L_2(P)$, $\{f_k(\mathbf{Z}), k \geq 1\}$, approximating any $L_2(P)$ function g of \mathbf{Z} in the L_2 sense. That is, we can formally write

$$g(\mathbf{Z}) = \sum_{k=1}^{\infty} a_k(g) f_k(\mathbf{Z}) \quad (12.1.8)$$

where $a_k(g) = E g(\mathbf{Z}) f_k(\mathbf{Z})$.

Now consider a two class $\{-1, 1\}$ classification problem, where the Bayes rule is $\delta(\mathbf{Z}) = 21(\pi(\mathbf{Z}) \geq \frac{1}{2}) - 1$ with $\pi(\mathbf{Z}) \equiv P[Y = 1|\mathbf{Z}]$. Then, if we represent $\log \pi(\cdot)$ as in (12.1.8), the Bayes rule is a linear classifier in R^∞ determined by the hyperplane $\{(a_1, a_2, \dots) : \sum_k a_k f_k(\mathbf{Z}) = \log \frac{1}{2}\}$. All points on one side of the hyperplane are classified as $Y = 1$ and the others as $Y = -1$. In fact, it is possible to show (Problem 12.1.2) that if we map $\mathbf{Z}_1, \dots, \mathbf{Z}_n, \mathbf{Z}_j \in R^d$, in a 1-1 fashion to R^M , $M >> \max(d, n)$, and if \mathbf{Z}_i^* are the R^M images of the \mathbf{Z}_i , then $\{\mathbf{Z}_i^* : Y_i = 1\}$ and $\{\mathbf{Z}_i^* : Y_i = -1\}$ can be separated by a hyperplane. That is, we can construct a linear classifier that classifies the Y 's in a *training sample* $(\mathbf{Z}_1, Y_1), \dots, (\mathbf{Z}_n, Y_n)$, $1 \leq i \leq n$, perfectly. There are evidently many such hyperplanes if one exists. Vapnik (1996) proposed using that separating hyperplane which maximizes the minimum distance between points of the training sample and the hyperplane. This problem it turns out can be cast as one in quadratic programming and is then rapidly and uniquely solvable.

The original form of support vector machines, for $C = 2$, is valid only in a context where there is a perfect separation of the two classes in the training sample. However, the introduction of suitable slack variables in the program permit regularization, i.e. lead to the choice of a hyperplane which permits a specified minimal number of misclassification of the training sample. We shall go further into the theory of this linear classifier in Section 12.2.4.

c) Boosting

Schapire (1990), Freund and Schapire (1997), and others proposed a method of classifying using larger and larger linear combinations of so called weak classifiers. This method has since been identified (Friedman, Hastie and Tibshirani (2000)) as an algorithm of a known type (Gauss-Southwell) for minimizing a data-based convex objective function. See Section 12.2.4.

d) Tree-based Methods

These are hybrid methods introduced independently by Breiman, Olshen, Friedman, Stone (1984), Quinlan (1987), and earlier by Morgan and Sonquist (1963). They may be viewed as constructing iteratively a partition of the covariate space R^d into blocks and then, for regression, estimating conditional expectations of Y in each block or, for classification, estimating the conditional probabilities of belonging to class $j \in \{1, \dots, C\}$ given that \mathbf{Z} belongs in a block. Again we shall discuss this method in Section 12.2.4.

12.1.3 Outline

Here is the rest of our outline for this chapter. In Section 12.2, we describe methods mentioned in Section 12.1 in more detail and start to give their properties. In Sections 12.3 and 12.4 we shall focus on asymptotic theory both from a statistical (minimax theorems) and machine learning (oracle inequalities) point of view and focus on the critical implication that “sparsity” in some sense is needed for any statistical successes when the covariate space R^d is high dimensional. Sparsity in this context means roughly that either the population regression function depends on a small subset $\{Z_j : j \in \mathcal{S}\}$ of covariates, or that the function takes on a simple form, e.g. is a sum of functions of one variable Z_j only. In other words, the regression function is much less complex than a general function on R^d .

In Section 12.5 we introduce cross validation as a tool for tuning parameter choice and in Section 12.6, we will discuss model selection and dimension reduction. In this context we shall discuss Bayesian model selection and principal component analysis. Finally, Section 12.7 discusses multiple testing and has pointers to work not covered in this book.

Summary. In this section, we have introduced the prediction framework both in the context of classification and regression, and

- i) Statistical approaches to these problems, in particular, sieves, penalization, and empirical Bayes methods.
- ii) Machine learning approaches to these problems, in particular, support vector machines, boosting, and tree-based methods.

12.2 Classification and Prediction

In this section, we give more detailed descriptions and properties of the approaches outlined in Section 12.1. However, first we introduce, in Section 12.2.1, tools to be used for classification and prediction. These include a multivariate density estimate that is used in classification procedures and a nonparametric statistical method based on multivariate kernel regression that can be used to predict a response Y from a vector \mathbf{x} of observed covariate values. In Sections 12.2.2, 12.2.3, and 12.2.4 we describe classification procedures based on nonparametric methods, sieves, and machine learning approaches.

12.2.1 Multivariate Density and Regression Estimation

The statistical methods in Chapter 11 for univariate nonparametric estimation will be generalized straightforwardly in this section to the multivariate case with continuous \mathbf{X} .

(a) Density Estimates

Let $\mathbf{X}_1, \dots, \mathbf{X}_n$ be i.i.d. as $\mathbf{X} \sim F$, where $\mathbf{X} \in R^d$. We assume that \mathbf{X} has a density $f(\mathbf{x})$ with convex support $S(f)$ equal to the closure of $\{\mathbf{x} : f(\mathbf{x}) > 0\}$. All the methods we have considered for estimating f when $d = 1$ generalize naturally. Here are two of the generalizations.

(1) Convolution kernel estimates

Let $K(\mathbf{u})$ be a multivariate kernel, that is, a map $R^d \rightarrow R$ with

$$\int_{R^d} K(\mathbf{u}) d\mathbf{u} = 1.$$

A kernel is said to be of *order* r if

$$\int_{R^d} \prod_{j=1}^d u_j^{i_j} K(\mathbf{u}) d\mathbf{u} = 0, \text{ for all } i_j \text{ with } 1 \leq i_1 + \dots + i_d \leq r.$$

For simplicity, we consider a *product kernel*

$$K(\mathbf{u}) = \prod_{j=1}^d K_j(u_j),$$

where each K_j is a 1 dimensional kernel. A product kernel is of order r iff each of its components is. Since scale may vary by coordinate, we specify a vector of bandwidths $\mathbf{h} = (h_1, \dots, h_d)^T$. Thus, as for $d = 1$, we define

$$K_{\mathbf{h}}(\mathbf{u}) = \prod_{j=1}^d h_j^{-1} K_j(u_j/h_j).$$

Often $K_j(u) = K_1(u)$, $j = 1, \dots, d$, and $K_1(u)$ is chosen as the normal density φ , or if compact support is desired, $K_1(u) = \frac{1}{2}1[|u| \leq 1]$ or $K_1(u) = \frac{3}{4}(1 - u^2)1(|u| \leq 1)$.

The empirical distribution is

$$\widehat{F}(\mathbf{x}) = n^{-1} \sum_{i=1}^n 1[\mathbf{X}_i \leq \mathbf{x}]$$

where $\mathbf{X} \leq \mathbf{x}$ is componentwise inequality, $\mathbf{x} \in R^d$. We define

$$f_{\mathbf{h}}(\mathbf{x}, F) = \int K_{\mathbf{h}}(\mathbf{x} - z) dF(z) \tag{12.2.1}$$

and the (plug-in) *convolution kernel density estimate*,

$$\widehat{f}_{\mathbf{h}}(\mathbf{x}) \equiv f_{\mathbf{h}}(\mathbf{x}, \widehat{F}) = \int K_{\mathbf{h}}(\mathbf{x} - \mathbf{z}) d\widehat{F}(\mathbf{z}) = \frac{1}{n} \sum_{i=1}^n K_{\mathbf{h}}(\mathbf{x} - \mathbf{X}_i). \tag{12.2.2}$$

Bias, variance, MSE, and IMSE results are completely analogous to the $d = 1$ case; see Silverman (1986), Scott (1992), and the problems. Briefly, let

$$D^2 f(\mathbf{x}) = \left| \frac{\partial^2 f(\mathbf{x})}{\partial x_i \partial x_j} \right|_{d \times d}, \quad \mathcal{F} = \{f : D^2 f(\mathbf{x}) \leq M \text{ all } \mathbf{x} \in R^d\}.$$

Then, if K has compact support and is of order 1,

$$E_F \hat{f}_h(\mathbf{x}) = f(\mathbf{x}) + O(|\mathbf{h}|^2) \quad (12.2.3)$$

$$\text{Var}_F \hat{f}_h(\mathbf{x}) = (n|\mathbf{h}|^d)^{-1} f(\mathbf{x}) \int K^2(\mathbf{u}) d\mathbf{u} (1 + o(1)) \quad (12.2.4)$$

as $n \rightarrow \infty$, $\mathbf{h} \rightarrow \mathbf{0}$; uniformly for $f \in \mathcal{F}$. Note that the dimension d inflates the variance by the factor $|\mathbf{h}|^{-d}$. If $K_j \equiv K_1, j = 1, \dots, d$, with K_1 of order 1, if $0 < \int |D^2 f(\mathbf{x})|^2 d\mathbf{x} < \infty$, and we select $\mathbf{h} = h(1, \dots, 1)^T$ to minimize the approximate IMSE

$$\text{AIMSE} \hat{f}_h(\mathbf{x}) \simeq \frac{1}{4} \left(\int |\mathbf{t}|^2 K(\mathbf{t}) d\mathbf{t} \right)^2 \left(\int |D^2 f(\mathbf{x})|^2 d\mathbf{x} \right) h^4 + \left(\int K^2(\mathbf{u}) d\mathbf{u} \right) (nh^d)^{-1} \quad (12.2.5)$$

then we obtain the optimal asymptotic rate

$$h \asymp n^{-\frac{1}{d+4}}.$$

For this h , the minimum IMSE is of order $n^{-4/(d+4)}$ (Problem 12.2.2).

Remark 12.2.1. For d large, say $d \geq 4$, the rate of convergence of $\hat{f}_h(\mathbf{x})$ to $f(\mathbf{x})$ is no better than $n^{-1/4}$. Thus if it takes 100 observations to achieve desired precision in a parametric setting this suggests it will take at least 10,000 in this high dimensional nonparametric setting. We can do no better with any other method (van der Vaart (1998), Section 24.3) unless we use stronger regularity conditions. As in the one dimensional case, the optimal rates become better if one assumes f has bounded derivatives of order $p > 2$, but there is no data-dependent way of checking such assumptions. This gloomy picture is relieved by the observation that data distributions tend to stay close to low dimensional manifolds and thus the worst case analyses (nonparametric for d large) are much too conservative.

Cross validation methods to estimate \mathbf{h} are given in Section 12.5 and the reference distribution approach is given in Problem 12.2.2.

As in the one dimensional case, we can eliminate difficulties in estimation at the boundary of f if $S(f)$ is a convex set by using locally polynomial estimates or more generally plugging into the minimizer of $D_{h,x}(f(\cdot), f(\cdot, \beta))$ defined as in Example 11.3.1, with the univariate integral over (a, b) replaced by integrating over $S(f) \subset R^d$, and K_h now a multivariate kernel. Details are given in Problem 12.2.3.

(2) Multivariate nearest neighbor estimates

Let $K(\mathbf{u}) = 1(|\mathbf{u}| \leq 1)/V(d)$ where $V(d)$ is the volume of the unit sphere in R^d (*not* a product kernel) and set

$$K_h(\mathbf{u}) = h^{-d} K\left(\frac{\mathbf{u}}{h}\right).$$

Then, if $\hat{f}_h(\mathbf{x})$ is defined by (12.2.2),

$$\hat{f}_h(\mathbf{x}) = \hat{F}(\mathbf{z} : |\mathbf{z} - \mathbf{x}| \leq h)/V(d)h^d.$$

If we take $h = \hat{h}(k, \mathbf{x})$ to be the smallest h such that there are k sample points \mathbf{x}_i in the set $\{\mathbf{z} : |\mathbf{z} - \mathbf{x}| \leq h\}$ then we obtain the k th *nearest neighbour estimate*

$$\tilde{f}_k(\mathbf{x}) = \frac{k}{V(d)[\hat{h}(k, \mathbf{x})]^d} = \frac{k}{V(d)[|\mathbf{x} - \mathbf{X}|_{(k)}]^d} \quad (12.2.6)$$

with $|\mathbf{x} - \mathbf{X}|_{(1)} < \dots < |\mathbf{x} - \mathbf{X}|_{(n)}$ denoting $|\mathbf{x} - \mathbf{X}_i|$, $1 \leq i \leq n$, ordered.

(b) Kernel Regression Estimates

We introduce a nonparametric product kernel estimate of $E(Y|\mathbf{X} = \mathbf{x})$ and give mean squared error properties of this estimate. Suppose the observations (\mathbf{X}_i, Y_i) , $i = 1, \dots, n$ are, i.i.d. as (\mathbf{X}, Y) , where $\mathbf{X}_i \equiv (X_{i1}, \dots, X_{id})^T$, $Y_i \in R$, with the \mathbf{X}_i having continuous case density f and loss is mean square error. We assume that $\mathcal{S}^0 \equiv \{\mathbf{x} : f(\mathbf{x}) > 0\}$, the interior of the support \mathcal{S}_f of $f(\cdot)$ is an ellipse or a rectangle with possibly one or more infinite boundaries. We want to estimate $\mu(\mathbf{x}) = E(Y|\mathbf{X} = \mathbf{x})$, where $\mathcal{L}(Y|\mathbf{X} = \mathbf{x})$ is defined to be pointmass at zero for $\mathbf{x} \notin \mathcal{S}^0$, so that $\mu(\mathbf{x}) = 0$, $\mathbf{x} \notin \mathcal{S}^0$. Suppose $K : R^d \rightarrow R$ is a kernel, that is, a function such that $\int_{R^d} K(\mathbf{t}) d\mathbf{t} = 1$. Then, define

$$K_{\mathbf{h}}(\mathbf{t}) \equiv (h_1 h_2 \dots h_d)^{-1} K\left(\frac{t_1}{h_1}, \dots, \frac{t_d}{h_d}\right),$$

with the t_j, h_j being the coordinates of \mathbf{t}, \mathbf{h} , where $h_j > 0$. Evidently $K_{\mathbf{h}}$ is also a kernel. Assume that K is symmetric, that is $K(-\mathbf{t}) = K(\mathbf{t})$. Then

$$\hat{f}_{\mathbf{h}}(\mathbf{x}) = n^{-1} \sum_{i=1}^n K_{\mathbf{h}}(\mathbf{X}_i - \mathbf{x})$$

is the multivariate convolution kernel density estimate of $f(x)$ of part (a) of this section. The generalized *multivariate Nadaraya-Watson estimate* of the nonparametric regression function $\mu(\mathbf{x}) \equiv E(Y|\mathbf{X} = \mathbf{x})$ is $\hat{g}_{\mathbf{h}}(\mathbf{x})/\hat{f}_{\mathbf{h}}(\mathbf{x})$ where $\hat{g}_{\mathbf{h}}(\mathbf{x}) = n^{-1} \sum Y_i K_{\mathbf{h}}(\mathbf{X}_i - \mathbf{x})$, that is,

$$\hat{\mu}_{\text{NW}}(\mathbf{x}) = \frac{\sum_{i=1}^n Y_i K_{\mathbf{h}}(\mathbf{X}_i - \mathbf{x})}{\sum_{i=1}^n K_{\mathbf{h}}(\mathbf{X}_i - \mathbf{x})}, \quad (12.2.7)$$

where $(0/0) = 0$. In the asymptotic analysis, we assume the *product kernel*

$$K(\mathbf{t}) \equiv \prod_{j=1}^d K_0(t_j), \quad h_j = h, \quad 1 \leq j \leq d,$$

where K_0 is a univariate, non-negative, symmetric kernel.

We will develop asymptotic results for the multivariate $\hat{\mu}_{\text{NW}}$: Let \mathcal{S} be a compact set contained in $\mathcal{S}^0 = \{\mathbf{x} : f(\mathbf{x}) > 0\}$. We select \mathcal{S} so that it has a “simple” and “smooth” boundary which we for simplicity take to be an ellipse. Let $|\cdot|$ be the Euclidean norm, let G denote the distribution of (\mathbf{X}, Y) , and let $\sigma^2(\mathbf{x}) = \text{Var}(Y|\mathbf{X} = \mathbf{x})$. Moreover, let $M > 0$ and $\varepsilon > 0$ be generic constants. We will use the assumptions:

A_1 : $\mu(\mathbf{x}) = E_G(Y|\mathbf{x})$ is Lipschitz; $|\mu(\mathbf{x}) - \mu(\mathbf{z})| \leq M|\mathbf{x} - \mathbf{z}|$, all $\mathbf{x}, \mathbf{z} \in \mathcal{S}$, and $\mu(\mathbf{x}) \leq M < \infty$ all $\mathbf{x} \in \mathcal{S}$.

A_2 : $f(\mathbf{x})$ is Lipschitz; $|f(\mathbf{x}) - f(\mathbf{z})| \leq M|\mathbf{x} - \mathbf{z}|$ all $\mathbf{x}, \mathbf{z} \in \mathcal{S}$, and $f(\mathbf{x}) \leq M$, all $\mathbf{x} \in \mathcal{S}$.

A_3 : $K_0 \geq 0$ has compact support, is continuous, symmetric, and is bounded above; $K_0(t) \leq M$ for all $t \in R$.

A_4 : $f(\mathbf{x})$ is bounded below on \mathcal{S} ; $f(\mathbf{x}) \geq \varepsilon$ all $\mathbf{x} \in \mathcal{S}$.

A_5 : The conditional distribution of $Y|\mathbf{X} = \mathbf{x}$ has a bounded moment generating function; $E_G\{e^{tY}|\mathbf{X} = \mathbf{x}\} \leq M$ for $|t| \leq \delta < \infty$, some $\delta > 0$, all $\mathbf{x} \in \mathcal{S}$.

Let \mathcal{G} be the class of G where A_1, \dots, A_5 hold uniformly in G and \mathbf{x} . Note that by A_5 , $\sigma^2(\mathbf{x}) \leq M < \infty$ for all \mathbf{x} . Let $\text{IMSE}(\hat{\boldsymbol{\mu}}_{\text{nw}}) = E[\text{MSE}(\hat{\boldsymbol{\mu}}_{\text{nw}}(\mathbf{X}))1(\mathbf{X} \in \mathcal{S})]$, then rates of convergence of the Nadaraya-Watson estimates are

Theorem 12.2.1. Under assumptions A_1, \dots, A_5 ,

- (i) $\inf_{h>0} \text{MSE}(\hat{\boldsymbol{\mu}}_{\text{nw}}(\mathbf{x})) = O(n^{-\frac{2}{2+d}})$ uniformly for $\mathbf{x} \in \mathcal{S}$ and $G \in \mathcal{G}$.
- (ii) $\inf_{h>0} \text{IMSE}(\hat{\boldsymbol{\mu}}_{\text{nw}}) = O(n^{-\frac{2}{2+d}})$ uniformly for $G \in \mathcal{G}$, for \mathcal{S} with smooth boundaries.
- (iii) The minimizers in (i) and (ii) are of the form $h = cn^{-1/(d+2)}$.

Proof. The proof is in Appendix D.7.

Remark 12.2.2. (1) Assumptions A_1, A_2 , and A_3 can be considerably weakened. Moreover A_4 can be replaced by a condition on how quickly $f(\mathbf{x}) \rightarrow 0$ as $\mathbf{x} \rightarrow \partial\mathcal{S}$ and A_5 can be replaced by moment conditions. The technical cost of weakening the assumptions does not seem worthwhile.

(2) If \mathcal{G} is defined by Lipschitz assumptions on partial derivatives of μ or more generally, $\mu(\cdot)$ lies in a Sobolev ball, the rate $n^{-2/(2+d)}$ can be replaced by $n^{-2s/(2s+d)}$, where s is some measure of smoothness determined by \mathcal{G} provided that we use local polynomial regression estimates of a sufficiently high order, such as linear, quadratic, etc.

(3) For other results of this type see Stone (1984), Ruppert and Wand (1994), Fan and Gijbels (1996), Jiang and Doksum (2003), and Problem 12.2.4. \square

Remark 12.2.3. (1) As in the case of density estimation (Remark 12.2.1), we face slow convergence rates unless the distribution of \mathbf{X} concentrates near a low dimensional manifold (sparsity).

(2) We also face the choice of bandwidth h , or number of nearest neighbours k , if we use a kernel with support $[-1, 1]^d$ and use the h that yields a fixed number k of \mathbf{x}_i 's in $[\mathbf{x} - h, \mathbf{x} + h]$. The asymptotically optimal choice of h given in Problem 12.2.4 depends on an unknown constant c . Prediction suggests the use of cross validatory choice, that is, fitting of the rule using a fraction of the sample and selection of the regularization parameter h by optimizing prediction or classification performance on the rest of the sample. We shall discuss these issues further in Section 12.5.

12.2.2 Bayes Rule and Nonparametric Classification

We begin with a result relating the Bayes classification rule to nonparametric density estimation. We observe a training sample $(X_1, I_1), \dots, (X_n, I_n)$ i.i.d. as $(X, I) \sim P$, where $I \in \{1, \dots, C\}$, $X \in \mathcal{X}$, $P \in \mathcal{P}$. Let $(X_{n+1}, I_{n+1}) \sim (X, I)$ independent of the other (X_i, I_i) . On the basis of X_{n+1} we are to predict I_{n+1} using 0-1 loss: $l(I, a) = 0$ if $a = I$, 1 otherwise.

Suppose $\hat{p}_j(\cdot)$, $1 \leq j \leq C$, are estimates of $p_j(\cdot)$, the conditional densities with respect to $\nu(x)$ of X given $I = j$, and $\hat{\pi}_{jn}$ are estimates of π_j , $1 \leq j \leq C$. The Bayes rule is, from (12.1.1),

$$\delta_B(x) = j \text{ iff } \pi_j p_j(x) = \max\{\pi_k p_k(x) : 1 \leq k \leq C\},$$

with ties broken arbitrarily if more than one j achieves the max. Let

$$\hat{\delta}_n(x : X_1, \dots, X_n) = j \text{ iff } \hat{\pi}_j \hat{p}_j(x) = \max\{\hat{\pi}_k \hat{p}_k(x) : 1 \leq k \leq C\} \quad (12.2.8)$$

with the same rule for ties. We show that $\hat{\delta}_n$ has risk close to the minimum Bayes risk. This minimum Bayes risk is from (12.1.2)

$$R_B = 1 - \int \max\{\pi_j p_j(x) : 1 \leq j \leq C\} d\nu(x).$$

Let \mathcal{P} be a general class of P 's satisfying the conditions specified in what follows.

Theorem 12.2.2. *Let $\hat{\delta}_n$ be as defined in (12.2.8). Suppose that as $n \rightarrow \infty$,*

$$\int |\hat{p}_j(x) - p_j(x)| d\nu(x) \xrightarrow{P} 0 \text{ and } \hat{\pi}_j \xrightarrow{P} \pi_j, \quad 1 \leq j \leq C. \quad (12.2.9)$$

Then, as $n \rightarrow \infty$, the risk of $\hat{\delta}_n$ for 0-1 loss satisfies

$$R(P, \hat{\delta}_n) \equiv P[\hat{\delta}_n(X_{n+1} : X_1, \dots, X_n) \neq I_{n+1}] = R_B + o(1). \quad (12.2.10)$$

If (12.2.9) holds uniformly for $P \in \mathcal{P}$, then $o(1)$ in (12.2.10) is also uniform.

Proof. For $\delta(\cdot) : \mathcal{X} \rightarrow \{1, \dots, C\}$, we note the following useful formula,

$$R(P, \delta) - R_B = \int (\max_j \pi_j p_j(x) - \pi_{\delta(x)} p_{\delta(x)}(x)) d\nu(x). \quad (12.2.11)$$

Therefore, for $\hat{a} = \arg \max \hat{\pi}_j \hat{p}_j(x)$,

$$\begin{aligned} R(P, \hat{\delta}_n) - R_B &= E \left\{ \int (\max_j \pi_j p_j(x) - \pi_{\hat{a}} p_{\hat{a}}(x)) d\nu(x) \right\} \\ &\leq E \int (\max_j \pi_j p_j(x) - \min_j \pi_j p_j(x)) \mathbf{1}(\bigcup_{a=1}^C A_a) d\nu(x) \end{aligned}$$

where $A_a = \{x : \widehat{\pi}_a \widehat{p}_a(x) > \widehat{\pi}_a \widehat{p}_a(x), \pi_a p_a(x) < \pi_a p_a(x)\}$. By hypothesis, $P[A_a] \rightarrow 0$, since $\max_a |\widehat{\pi}_a \widehat{p}_a(X) - \pi_a p_a(X)| \xrightarrow{P} 0$. The result follows from the dominated convergence theorem (Theorem B.7.5). \square

Definition 12.2.1. A classifier $\widehat{\delta}_n$ that satisfies (12.2.10) is called *Bayes consistent*.

We next give examples of rules that satisfy the conditions of Theorem 12.2.2.

Example 12.2.1. Kernel and nearest neighbour rules.

(a) Kernels. It is evident that we can plug in kernel estimates $\widehat{p}_j(\cdot)$ for $p_j(\cdot)$ when X is continuous. For instance, suppose $\mathcal{X} = R$, $K(u) = \frac{1}{2}1(-1, 1)$, we use the same bandwidth h for all \widehat{p}_j , and the $\widehat{\pi}_j = n^{-1} \sum_{i=1}^n 1(I_i = j)$ in the sample are unbiased and consistent estimates of the population π_j . Then, a reasonable classification rule is

$$\widehat{\delta}_n(X_{n+1} : X_1, \dots, X_n) = j \text{ if } \widehat{p}_j(X_{n+1}) > \frac{\widehat{\pi}_m}{\widehat{\pi}_j} \widehat{p}_m(X_{n+1}) \text{ for all } m$$

where $\widehat{p}_j(\cdot)$ is the kernel density estimate based on $\{X_i : I_i = j\}$. This rule is equivalent to preferring j over m if the ratio of the number $N_j(h)$ of X observations from category j in the training set at distance $\leq h$ from X_{n+1} to the corresponding number $N_m(h)$ from category m is larger than $\widehat{\pi}_m/\widehat{\pi}_j$.

(b) Nearest Neighbours. If we use the k th nearest neighbour density estimate of Sections 11.4.3, and 12.2.1(a), and if π_j is uniform, $\pi_j = C^{-1}$, we order $|X_{n+1} - X_j|$ to get $|X_{n+1} - X_j|_{(1)} < \dots < |X_{n+1} - X_j|_{(n)}$, the ordered distances between the X_j , $1 \leq j \leq n$, and X_{n+1} . The $\arg \max_j \{\widehat{p}_j(x_{n+1})\}$ rule leads to (Problem 12.2.5): Predict $I_{n+1} = I_{\widehat{j}_k}$, where \widehat{j}_k is defined by

$$|X_{n+1} - X_{\widehat{j}_k}| = |X_{n+1} - X|_{(k)}.$$

That is, \widehat{j}_k is the subscript of the k th nearest neighbour to X_{n+1} . If $k = 1$, this leads to the highly plausible rule:

Predict $I_{n+1} = I_{\widehat{i}}$, where

$$|X_{n+1} - X_{\widehat{i}}| = \min\{|X_{n+1} - X_m| : 1 \leq m \leq n\}.$$

That is, \widehat{i} is the subscript of the X_i , $1 \leq i \leq n$, closest to X_{n+1} . We call this the *nearest neighbour classifier*.

Corollary 12.2.1. Assume the framework of Example 12.2.2 (a). If $h_n \rightarrow 0$, $nh_n \rightarrow \infty$, if $C = 2$, if $p_j(\cdot)$ concentrates on $(0, 1)$, if $\|p_j\|_\infty \leq M$ for $j = 0, 1$, and if δ_B is the Bayes rule, then

$$P[\widehat{\delta}_n(X_{n+1} : X_1, \dots, X_n) \neq I_{n+1}] = P[\delta_B(X_{n+1}) \neq I_{n+1}] + o(1). \quad (12.2.12)$$

Moreover, (12.2.12) holds uniformly over $\mathcal{F} = \{(p_0, p_1) : \|p_j''\|_\infty \leq M\}$.

Proof. Note that, using Theorem 11.2.1,

$$\left(E \int_0^1 |\widehat{p}_j(x) - p_j(x)| dx \right)^2 \leq E \int_0^1 (\widehat{p}_j(x) - p_j(x))^2 dx = O((nh)^{-1}) + O(h^4)$$

uniformly on \mathcal{F} . The result follows from Theorem 12.2.2. \square

Remark 12.2.4 (1) The conditions of Corollary 12.2.1 are much too strong. If no uniformity is required, then (12.2.12) holds for kernel density estimates for any f — see Devroye and Lugosi (1996) for instance.

(2) If $d > 1$, the condition $nh_n \rightarrow \infty$ is replaced by $nh_n^d \rightarrow \infty$.

(3) The k th nearest neighbour classifier is consistent under regularity conditions when $d = 1$ if $k = k_n \rightarrow \infty$, $n^{-1}k_n \rightarrow 0$ as $n \rightarrow \infty$. See Theorem 11.4.1 and Problem 12.2.6.

(4) Define a (poor) estimate based on the training sample of the misclassification probability by

$$\hat{R} = n^{-1} \sum_{j=1}^C \sum_{i=1}^n \mathbb{1}[I_i = j, \hat{\delta}_n(X_i : X_1, \dots, X_n) \neq j].$$

Note that the nearest neighbour classifier has $\hat{R} = 0$, a clear underestimate because the nearest neighbour classifier, which corresponds to a kernel estimate with bandwidth $h = O_P(n^{-1})$, has a bandwidth that is too small to give a good density estimate. To obtain Bayes consistency we have seen that we have to take $k = k_n \rightarrow \infty$, $k_n/n \rightarrow 0$. On the other hand, if $\hat{\delta}_n(\cdot)$ is the 1-nearest neighbour rule and $R(P, \hat{\delta}_n)$ is its classification risk corresponding to the 0–1 loss function, then

$$\overline{\lim}_n \frac{R(P, \hat{\delta}_n)}{R_B(P)} \leq 2$$

where $R_B(P)$ is the minimum Bayes risk (Cover and Hart (1967)). Thus, a terrible density estimate can lead to a perfectly reasonable, though in general, suboptimal classification rule.

(5) Kernel and k th nearest neighbour methods were first advocated by Fix and Hodges (1951). The latter particularly are still in wide use.

(6) Properties of classifiers will be discussed in more detail in Section 12.2.4.

12.2.3 Sieve Methods

Logistic regression (LR)

As we have discussed one of the most widely used classification method is based on logistic regression. Fitting is particularly easy since it involves just the calculation of the MLE for a canonical exponential family. The issue of regularization appears in this context through the choice of the number of functions $f_i(\mathbf{Z})$ appearing in the logistic regression. Of even greater importance is the nature of the sieve determined by the class $\mathcal{F} = \{f_1, f_2, \dots\}$. If $f_j(\mathbf{Z}) = Z_j \in R$, then logistic regression may be viewed as a competitor to the rule based on Fisher's linear discriminant analysis (LDA) which is constructed on the assumption that \mathbf{Z} has an $\mathcal{N}_d(\mu_j, \Sigma)$ distribution for $j = 1, \dots, C$. It can be shown (Problem 12.2.8 and 12.2.9) that LDA can never do better than LR asymptotically at least to 0th order, that is converge to the Bayes rule.

Penalized least squares

Recall the framework of Section 12.1.1. Again the choice of sieve is of first importance. Consider $\{f_j(\mathbf{Z}) : 1 \leq j \leq k\}$ as the k th member of a sieve, that is,

$$Y = \mathbf{F}^T \boldsymbol{\theta} + e, \quad (12.2.13)$$

where $F = (f_1(\mathbf{Z}), \dots, f_k(\mathbf{Z}))^T$, $\boldsymbol{\theta} = (\theta_1, \dots, \theta_k)^T \in R^k$. Let e_j be i.i.d. $\mathcal{N}(0, \sigma^2)$, so that the model for n observations is

$$\mathbf{Y}_{n \times 1} = F_{n \times k} \boldsymbol{\theta} + \mathbf{e}. \quad (12.2.14)$$

$F \equiv [f_j(\mathbf{Z}_i)]$, $1 \leq j \leq k$, $1 \leq i \leq n$, $\mathbf{e} \equiv (e_1, \dots, e_n)^T$, $\mathbf{Y} \equiv (Y_1, \dots, Y_n)^T$. Then, penalizing by $J(\boldsymbol{\theta}) = \frac{\lambda}{2} |\boldsymbol{\theta}|^2$ is equivalent to minimizing

$$\begin{aligned} |\mathbf{Y} - F\boldsymbol{\theta}|^2 &+ \frac{\lambda}{2} |\boldsymbol{\theta}|^2 \\ &= \mathbf{Y}^T \mathbf{Y} - 2\boldsymbol{\theta}^T F^T \mathbf{Y} + \boldsymbol{\theta}^T (F^T F + \frac{\lambda}{2} \mathbf{I}) \boldsymbol{\theta} \end{aligned} \quad (12.2.15)$$

where I is the $k \times k$ identity matrix. The minimizer is easily seen to be

$$\hat{\boldsymbol{\theta}}_R = (F^T F + \lambda \mathbf{I})^{-1} F^T \mathbf{Y}. \quad (12.2.16)$$

The minimizer is always uniquely defined if $\lambda > 0$ even if $\text{rank}(F) < k$. The *ridge regression* estimator $\hat{\boldsymbol{\theta}}_R$ leads to the estimate of the regression $\mu(\mathbf{Z})$,

$$\hat{\mu}(\mathbf{Z}) = F \hat{\boldsymbol{\theta}}_R. \quad (12.2.17)$$

Ridge regression was first advanced (Hoerl and Kennard (1970)) for numerical stability purposes in situations where F was ill-conditioned, but has important statistical properties as well. As is noted in Section 12.1.1, $\hat{\boldsymbol{\theta}}_R$ can be interpreted as a Bayes estimate. As such, it shrinks the (unpenalized) least squares predictor $\hat{\mu}_{LS}(\mathbf{Z}) = F[F^T F]^{-1} F^T \mathbf{Y}$ towards the mean $\mathbf{0}$ of the prior distribution $\mathcal{N}_n(\mathbf{0}, FF^T/\lambda)$ for $F\boldsymbol{\theta}$.

The choice of regularization parameter λ will be discussed further in Section 12.5. However, it is straightforward to make a connection to Stein shrinkage as in Section I.6. Suppose $F^T F = \mathbf{I}$, so that the vectors $\{f_j(\mathbf{Z}_i) : i = 1, \dots, n\}$ are orthonormal. Then the MLE of $\boldsymbol{\theta}$ is

$$\hat{\boldsymbol{\theta}} \equiv F^T \mathbf{Y} \sim \boldsymbol{\theta} + \mathbf{e}_{k \times 1}$$

where $\mathbf{e}_{k+1} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ and $\boldsymbol{\theta} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}/\lambda)$ are independent under our (empirical) Bayesian model. Thus, marginally, $\hat{\boldsymbol{\theta}} \sim \mathcal{N}(\mathbf{0}, (\lambda + 1)\mathbf{I}/\lambda)$ and the MLE of

$$(1 + \lambda)^{-1} = 1 - \left(\frac{\lambda + 1}{\lambda}\right)^{-1}$$

is

$$\widehat{(1 + \lambda)}^{-1} = \left(1 - \frac{d}{|\hat{\boldsymbol{\theta}}|^2}\right)_+, \quad (12.2.18)$$

which is almost the Stein positive part estimate in which $|\hat{\theta}|^2/d$ is replaced by $|\hat{\theta}|^2/(d-2)$. It can also be interpreted as the estimate which for some $c(\lambda)$ minimizes

$$|\mathbf{Y} - F\boldsymbol{\theta}|^2$$

subject to $|\boldsymbol{\theta}|^2 \leq c(\lambda)$.

Recall the notation $J_r(\boldsymbol{\theta}) = \sum_{l=1}^k |\theta_l|^r$. The penalty $J_1(\boldsymbol{\theta})$ is the closest convex penalty of the form J_r to J_0 . It is also the only *convex* member of the J_r family which permits making some coordinates of the optimizing $\boldsymbol{\theta}$ to be 0 (Problem 12.2.10). Thus if $\hat{\theta}_j$ is a regression coefficient and is set to zero, this corresponds to leaving the j th predictor Z_j out of the model. As mentioned earlier, this method is known as the “Lasso” (least absolute shrinkage and selection operator) (Tibshirani (1996)). The properties of the Lasso are discussed extensively by Bühlmann and van de Geer (2011), and Hastie, Tibshirani and Wainwright (2015). See also Problem 12.2.10. Properties of extensions of the Lasso to grouped variables (*the grouped Lasso*) can be found in Yuan and Lin (2007) and Zhao, Rocha and Yu (2009).

12.2.4 Machine Learning Approaches

We return to the methods discussed in Section 12.1.2 but now with $Y \in \{-1, 1\}$. Recall that (\mathbf{Z}_i, Y_i) , $1 \leq i \leq n$, are i.i.d. as (\mathbf{Z}, Y) , where (\mathbf{Z}, Y) is independent of $\{(\mathbf{Z}, Y_i), 1 \leq i \leq n\}$. The goal is to build a classifier $\delta_n(\mathbf{Z})$ based on a training sample (\mathbf{Z}_i, Y_i) , $1 \leq i \leq n$, that can be used to classify an individual or object with unknown Y and known $\mathbf{Z} \equiv (Z_1, \dots, Z_d)^T$ as being either in the “ $Y = -1$ ” or “ $Y = 1$ ” category.

a) Neural Nets

We do not further discuss these methods but refer to Hastie et al (2001, 2009) and Ripley (1996).

b) Support Vector Machines

As we indicated, support vector machines, were proposed by Vapnik as a classification algorithm for 2 classes which finds the hyperplane of “maximum margin” separating the members of the training sample perfectly according to class. Of course, there may be no hyperplane which can separate at all — for instance, imagine one class distributed uniformly on a spherical shell and the other limited to the interior of the sphere. But if separation is possible then we can define the *margin* for any separating hyperplane H by

$$M(H) = \min_i |\mathbf{Z}_i - \Pi(\mathbf{Z}_i | H)|$$

where $|\cdot|$ is the Euclidean norm and Π is projection on H . This is just the minimum distance of the training set from H . The hyperplane sought is the one which maximizes $M(H)$. Formally, for the hyperplane

$$H = \{\mathbf{z} : \boldsymbol{\beta}^T \mathbf{z} + \beta_0 = 0\}, |\boldsymbol{\beta}| = 1,$$

the problem may be stated as (Problem 12.2.13)

$$\max_{\beta, \beta_0} \min_i \left\{ |\beta^T \mathbf{Z}_i + \beta_0| : |\beta| = 1, \quad Y_i(\beta^T \mathbf{Z}_i + \beta_0) > 0, \quad 1 \leq i \leq n \right\}. \quad (12.2.19)$$

We refer to Hastie et al (2009), Section 12.2, for a discussion showing how this convex optimization problem may be transformed upon dropping the restriction that $|\beta| = 1$ into

$$\min_{\beta, \beta_0} |\beta| \quad (12.2.20)$$

subject to

$$Y_i(\mathbf{Z}_i^T \beta + \beta_0) \geq 1 \quad (12.2.21)$$

for all i . Here the maximum margin is $1/|\beta^*|$, where β^* is the minimizer in (12.2.20).

If separation is impossible we can modify (12.2.21) to

$$Y_i(\mathbf{Z}_i^T \beta + \beta_0) \geq 1 - \xi_i \quad (12.2.22)$$

for all i where $\xi_i \geq 0$. Note that if $\xi_{i_0} > 1$, then any hyperplane misclassifying Y_{i_0} but correctly classifying all Y_i such that $\xi_i = 0$ becomes a candidate for separating all $\{\mathbf{Z}_i : \xi_i = 0\}$. The reason is that we can write (12.2.22) as

$$Y_i \left(\frac{\mathbf{Z}_i^T \beta}{|\beta|} + \frac{\beta_0}{|\beta|} \right) \geq \frac{1 - \xi_i}{|\beta|}.$$

Then, if $\xi_{i_0} > 1$, and there is a hyperplane $H = \{\mathbf{z} : \mathbf{z}^T \beta^* + \beta_0 = 0\}$ which misclassifies only Y_{i_0} and $Y_{i_0}(\mathbf{Z}_{i_0}^T \beta^* + \beta_0) = -\delta < 0$, we see that, for some $0 < C < \infty$, $-\delta/C \geq 1 - \xi_{i_0}$ and so $|\beta^*|/C$ is feasible, i.e. a candidate for the minimizing $|\beta|$ as in (12.2.20) subject to (12.2.22). Evidently, if we permit $\xi_i \geq 0$ to be arbitrary we can make the inf in (12.2.20) to be 0. Thus, the restriction $\sum_{i=1}^n \xi_i \leq K$ is added, and we arrive at

$$\min \frac{1}{2} |\beta|^2$$

subject to

$$\xi_i \geq 0, \quad Y_i(\mathbf{Z}_i^T \beta + \beta_0) \geq 1 - \xi_i, \quad i = 1, \dots, n, \quad \sum_{i=1}^n \xi_i \leq K. \quad (12.2.23)$$

By Lagrangian duality (see, for instance, Boyd and Vandenberghe (2004)), we note that the problem is equivalent to

$$\max \left\{ \sum_{i=1}^n \alpha_i - \frac{1}{2} \left| \sum_{i=1}^n \alpha_i Y_i \mathbf{Z}_i \right| \right\} \quad (12.2.24)$$

subject to

$$0 \leq \alpha_i \leq \gamma, \quad \sum_{i=1}^n \alpha_i Y_i = 0$$

for some γ depending on K and with the relation between the optimizing $\boldsymbol{\alpha}^*$ and β^*, β_0^* , $\xi = (\xi_1, \dots, \xi_n)$, and γ , given by

$$\boldsymbol{\alpha}_i^* [Y_i(\mathbf{Z}_i^T \boldsymbol{\beta}^* + \beta_0^*) - (1 - \xi_i)] = 0, \quad (12.2.25)$$

$$\xi_i(\boldsymbol{\alpha}_i^* - \gamma) = 0, \quad (12.2.26)$$

for $i = 1, \dots, n$ and

$$\boldsymbol{\beta}^* = \sum_{i=1}^n \alpha_i^* Y_i \mathbf{Z}_i. \quad (12.2.27)$$

Note that $\alpha_i^* > 0$ iff $Y_i(\mathbf{Z}_i^T \boldsymbol{\beta}^* + \beta_0^*) = 1 - \xi_i$. The corresponding \mathbf{Z}_i are called *support vectors*. Note also that all α_i^* with $\xi_i > 0$ equal γ . Any $0 < \alpha_i^* < \gamma$ will correspond to a correctly classified \mathbf{Z}_i at minimum distance from the hyperplane corresponding to $(\boldsymbol{\beta}^*, \beta_0^*)$. For further discussion of the implementation of these methods via reproducing kernel Hilbert spaces, we refer to Hastie et al (2009).

c) Boosting

The initial form of this method, as limited to classification into 2 classes was in terms of reweighting. We are given

- (i) A set of “weak” classifiers, $h_k : \mathbf{Z} \rightarrow \{-1, 1\}$, $k = 1, 2, \dots, M$, $M \leq \infty$, where $\{-1, 1\}$ correspond to the classification decisions. “Weak” loosely means that the classifiers are expected to have $P[\text{Misclassification}] < \frac{1}{2}$, but barely so.
- (ii) A training sample (\mathbf{Z}_i, Y_i) , $i = 1, \dots, n$, where $\mathbf{Z}_i \in R^d$ and $Y_i \in \{-1, 1\}$.

Boosting then produces a sequences of linear classifiers of the form

$$G_j(\mathbf{Z}) = \text{sgn}\left(\sum_{k=1}^M \alpha_{kj} h_k(\mathbf{Z})\right), \quad j = 1, 2, \dots,$$

by selecting the weights so as to emphasize the classifiers doing particularly well on the hard to classify observations.

There are many flavors of boosting; see Meir and Rätsch (2003) for an overview. One division can be made between the algorithms in which (1) the h_k are considered in a pre-specified order, so that $\alpha_{kj} = 0$, $k > \min(j, M)$, and (2) the other more common approach where $\alpha_{kj} \neq 0$ for at most $\min(j, M)$ indices, but the h to be considered at stage j is chosen optimally.

The basic form of the original AdaBoost algorithm describes a process of generating the α_{kj} by iterative reweighting of the training sample. Specifically, suppose we introduce h_m at step m ,

AdaBoost 1.

1. Initialize $G_1(\mathbf{Z}) = h_1(\mathbf{Z})$
- $w_{i1} = \frac{1}{n}$, $i = 1, \dots, n$
- $\text{err}_1 = \sum_{i=1}^n w_{i1} 1(Y_i \neq G_1(\mathbf{Z}_i))$
- $\alpha_1 = \log \left[(1 - \text{err}_1) / \text{err}_1 \right]$.

$$\begin{aligned}
2. \text{ Given } G_m &= \operatorname{sgn}(\sum_{k=1}^m \alpha_k h_k) \\
w_{i(m+1)} &= w_{im} \exp [\alpha_m 1(Y_i \neq h_m(\mathbf{Z}_i))] \quad i = 1, \dots, n \\
\operatorname{err}_{m+1} &= \frac{\sum_{i=1}^n w_{i(m+1)} 1(Y_i \neq h_{m+1}(\mathbf{Z}_i))}{\sum_{i=1}^n w_{i(m+1)}} \\
\alpha_{m+1} &= \log [(1 - \operatorname{err}_{m+1}) / \operatorname{err}_{m+1}].
\end{aligned}$$

define $G_{m+1} = \operatorname{sgn}(\sum_{k=1}^{m+1} \alpha_k h_k)$.

3. Iterate for $m = 1, \dots, M, \dots$,

Note. Iterations need not stop once h_M has been considered but we can restart with G_M replacing G_1 .

AdaBoost 2. Instead of h_k being presented in the original order, in the more common version of the algorithm, h_{i_1}, h_{i_2}, \dots are determined in data driven order. Thus, set $h^{(k)} \equiv h_{i_k}$ and initialize as before. At stage m

$$G_m = \operatorname{sgn}\left(\sum_{k=1}^m \alpha_k h^{(k)}\right),$$

then at stage $m + 1$,

$$h_{i_{m+1}} = \arg \min_k \frac{\sum_{i=1}^n w_{i(m+1)} 1(Y_i \neq h^{(k)}(\mathbf{Z}_i))}{\sum_{i=1}^n w_{i(m+1)}}$$

where

$$w_{i(m+1)} \equiv w_{im} \exp \{ \alpha_m 1(Y_i \neq h^{(m)}(\mathbf{Z}_i)) \}.$$

Then

$$G_{m+1} = \operatorname{sgn}\left(\sum_{k=1}^{m+1} \alpha_k h^{(k)}\right)$$

with

$$\alpha_{m+1} = \log [(1 - \operatorname{err}_{m+1}) / \operatorname{err}_{m+1}]$$

and

$$\operatorname{err}_{m+1} = \sum_{i=1}^n w_{i(m+1)} 1(Y_i \neq h^{(m+1)}(\mathbf{Z}_i)) \tag{12.2.28}$$

□

As was discovered by a number of authors (see Meir and Rätsch (2003)) the first of these algorithms is just coordinate ascent as described in Section 2.4.2 and the more common version is a variant usually called the Gauss-Southwell algorithm for a particular optimization problem. We next establish this result. Given P , a distribution of (\mathbf{Z}, Y) , let

$$Q(\boldsymbol{\alpha}, P) \equiv \int \exp - \left\{ \left[\sum_{j=1}^M \alpha_j h_j(\mathbf{z}) \right] y \right\} dP(\mathbf{z}, y).$$

This is evidently a convex function of $\alpha \in R^M$ and if \widehat{P}_n is the empirical distribution of the training sample, we note that

$$Q_n(\alpha) \equiv Q(\alpha, \widehat{P}_n) = \frac{1}{n} \sum_{i=1}^n \exp \left\{ - \sum_{j=1}^M \alpha_j h_j(\mathbf{Z}_i) \right\}.$$

If we fix $\alpha_{jm} = \alpha_{j(m-1)}$, $1 \leq j \leq m-1$, and $\alpha_{jm} = 0$, $j > m$, then,

$$\begin{aligned} \frac{\partial}{\partial \alpha_m} Q_n(\alpha) &= \frac{1}{n} \sum_{i=1}^n \exp \left[-G_{m-1}(\mathbf{Z}) Y_i \right] \left[e^{-\alpha_m} 1(h_m(\mathbf{Z}) Y_i = 1) \right. \\ &\quad \left. - e^{\alpha_m} 1(h_m(\mathbf{Z}) Y_i = -1) \right] = 0 \end{aligned}$$

iff

$$(e^{2\alpha_m} + 1) = \frac{\sum_{i=1}^n e^{-G_{m-1}(\mathbf{Z}) Y_i}}{\sum_{i=1}^n e^{-G_{m-1}(\mathbf{Z}) Y_i} 1(h_m(\mathbf{Z}) Y_i = 1)}$$

or

$$\alpha_m = \frac{1}{2} \log \frac{(1 - \text{err}_{m-1})}{\text{err}_{m-1}}.$$

The correspondence to the first version of AdaBoost we described is complete when we notice that (Problem 12.2.12)

$$\begin{aligned} e^{-G_{m-1}(\mathbf{Z}) Y_i} &= \prod_{j=1}^{m-1} e^{-\alpha_j h_j(\mathbf{Z}) Y_i} \\ &= \left(\prod_{j=1}^m e^{2\alpha_j 1(h_j(\mathbf{Z}) Y_i = -1)} \right) e^{-\sum_{j=1}^m \alpha_j}. \end{aligned} \tag{12.2.29}$$

Define

$$\text{logit}(u) = \log[u/(1-u)], 0 < u < 1,$$

and suppose that

$$\text{logit}(P(Y = 1 | \mathbf{Z})) = \sum_{j=1}^M \alpha_j h_j(\mathbf{Z}) \tag{12.2.30}$$

for some $\alpha \equiv (\alpha_1, \dots, \alpha_M)^T$. If M is fixed and if the sample optimizer $\widehat{\alpha}$ of $Q(\alpha, \widehat{P}_n)$ exists, then if we iterate either version of AdaBoost, the algorithms converge to

$$\delta(\mathbf{Z}) = 1 \text{ iff } \sum_{j=1}^M \widehat{\alpha}_j h_j(\mathbf{Z}) > 0$$

with probability tending to 1 for fixed $n > M$ (Problem 12.2.14). In turn, $\widehat{\alpha}$ converges as $n \rightarrow \infty$ in probability to the minimizer of $Q(\alpha, P)$ (Problem 12.2.15). However, suppose $\text{logit}(P(Y = 1 | \mathbf{Z}))$ is of the form (12.2.30) with $M = \infty$ and it is approximable by functions of the form $\sum_{j=1}^m \alpha_j h_j(\mathbf{Z})$. This is, for instance, the case if the $[h_j(\mathbf{Z}) + 1]/2$

are taken as indicators of arbitrary hyper rectangles in R^d . Then, it may be shown that the population version of the algorithm of type 2 where the “best” direction h_j is chosen on each iteration converges to the Bayes classifier, but the sample version may not unless it is regularized in some way by stopping the algorithm early. See Problem 12.2.16.

Other convex objective functions than $Q(\alpha, P)$ can also be considered and these ideas can also be applied to nonparametric regression and even density estimation — see Bickel, Ritov, Zakkai (2001), Friedman, Hastie, Tibshirani (2000), and references therein for further discussions of boosting from a statistical point of view. See also Meir and Rätsch (2003) and Rudin, Daubechies and Schapire (2004) for a discussion from a machine learning/dynamical systems point of view.

We note that support vector machines, as defined by (12.2.24), may be put in the boosting framework (Problem 12.2.17). In particular, solving (12.2.24) is equivalent to minimizing the convex objective function,

$$R(\alpha, P) = \int W((\beta_0 + \sum_{j=1}^n \alpha_j h_j(\mathbf{z}))y) dP(\mathbf{z}, y) + \lambda |\alpha|^2 \quad (12.2.31)$$

where $\lambda = (2\gamma)^{-1}$ and the function W is given by $W(t) = (1-t)_+$, with $x_+ = x1(x \geq 0)$. W may be thought of as the closest convex approximation to $1(t \leq 1)$. Note (Problem 12.2.17) that, if $\lambda = 0$, and the Bayes rule is of the form

$$\delta_B(\mathbf{Z}) = \text{sgn}\left(\sum_{j=1}^M \alpha_j^* h_j(\mathbf{Z})\right),$$

then δ_B minimizes $R(\alpha, P)$, as well as the Bayes risk,

$$P\left[Y \sum_{j=1}^M \alpha_j^* h_j(\mathbf{Z}) < 0\right].$$

For a fuller discussion of these relations, see Hastie et al (2001, 2009).

d) Tree-structured Methods

We define trees informally. We ask that the reader examine the captions of Figure 12.1 for the definition of the terms we use in our discussion. We consider the two-class case, that is, $Y = 0$ or 1 , say. The general multiclass case is developed in Problem 12.2.17.

We will consider tree-based classifiers based on classification regions derived from a training sample $(Z_1, Y_1), \dots, (Z_n, Y_n)$. At level 1, a classification rule $\delta^{(1)}$ based on a predictor $Z \in \mathcal{Z}$ can be viewed as specifying a “critical” region $C_1 = \{z : \delta^{(1)}(z) = 1\}$, where Y is classified as 1, and its complement, $C_0 \equiv \{z : \delta^{(1)}(z) = 0\}$, where Y is classified as 0. We can represent $\delta^{(1)}$ as a two-branch tree as follows. We ask the question: Does the observed Z belong to C_1 ? If so, send Z to node $[1]$, on level 1 in Figure 12.1, and predict $Y = 1$, or, if not, then send Z to node $[0]$ and predict $Y = 0$. For level 2, we view C_1 and C_0 as separate universes with classifiers $\delta^{(2)}$ and $\bar{\delta}^{(2)}$ where $\delta^{(2)}$ takes values

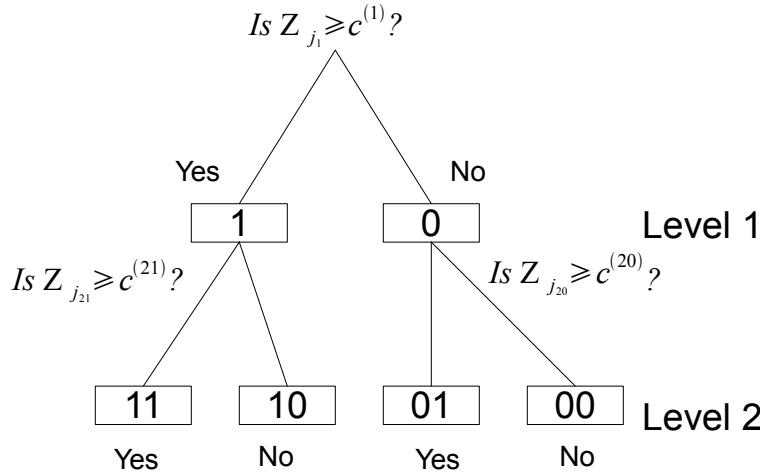


Figure 12.1. An example of a classification tree. The nodes at the final level are called terminal nodes. Y is classified to equal the last digit in the terminal node.

0 and 1 on C_1 while $\bar{\delta}^{(2)}$ does so on C_0 . Let $C_{11}, C_{10}, C_{01}, C_{00}$ be the sets defined by $\{\delta^{(1)} = 1, \delta^{(2)} = 1\}, \{\delta^{(1)} = 1, \delta^{(2)} = 0\}, \{\delta^{(1)} = 0, \bar{\delta}^{(2)} = 1\}, \{\delta^{(1)} = 0, \bar{\delta}^{(2)} = 0\}$. These form a partition of \mathcal{Z} , even as C_0 and C_1 did. Classify as follows: If $\delta^{(1)}$ sends Z to C_1 , ask if $Z \in C_{11}$ and if yes, send Z to node 11 on level 2 and classify Y as 1; if not send Z to node 10 and classify Y as 0, etc. We have in this way constructed a more complex classifier taking values 1 on $C_{11} \cup C_{01}$ and 0 on $C_{01} \cup C_{00}$. Given a sequence of such classifiers, we can evidently grow as large a tree as we wish. The classification we make for a given $Z = z$ always corresponds to the value of the last rule, for the terminal node we send z to.

We are immediately faced with the questions:

- 1) How do we select the consecutive rules for the different levels of the tree?
- 2) How deeply do we grow the tree?

Our goals in 1) are to have each member of the sequence easily computable but exhibiting the feature that the complex rules corresponding to growing the tree to a large depth can approximate the Bayes rule. Question 2) is more complex and can be viewed as a regularization question as discussed in Section 12.5.

To understand how Question 1 is answered, we begin with the unrealistic assumption that we know P , that is, π, p_0 , and p_1 . Then,

A: We specify a class of simple rules. If Z is real the class of rules for classification and

regression trees (CART)(Breiman, Friedman, Olshen and Stone (1984)) is

$$\{\delta_c, \bar{\delta}_c : \delta_c \equiv 1[c, \infty), \bar{\delta}_c \equiv 1 - \delta_c\}.$$

That is, $\delta(z) = 1$ if $z \geq c$, 0 otherwise; and $\bar{\delta}_c(z) = 1$ if $z < c$, 0 otherwise. This means that we consider all possible ways of classifying Y according as z is larger or smaller than a threshold.

B: Construct level 1. We begin by examining how well the populations would be split if we used δ_c or $\bar{\delta}_c$. A perfect or “pure” split would occur if

$$P[Y = 1|Z \geq c] = 1 - P[Y = 0|Z < c]$$

or conversely. A way to measure the “goodness of split,” proposed by Breiman et al. (1984), is to use

$$\tau^2(c) \equiv E \operatorname{Var}(Y|\delta_c(Z))$$

which vanishes iff the split is “pure” as an *impurity measure* of the split. In the two-class case this means:

Choose c so that, if $A_c \equiv \{\delta_c(Z) = 1\}$,

$$\tau^2(c) \equiv P(A_c)P[Y = 1|A_c]P[Y = 0|A_c] + P(\bar{A}_c)P[Y = 1|\bar{A}_c]P[Y = 0|\bar{A}_c]$$

is minimal. Once the optimal c , c_{opt} is chosen we use δ_c or $\bar{\delta}_c$ accordingly as

$$\pi P[Y = 1|A_c] > (1 - \pi)P[Y = 1|\bar{A}_c].$$

An alternative to τ^2 , proposed by Quinlan (1993), is to use the entropy (see Problem 12.2.18)

$$-E(E(\log P[Y = 1|\delta_c(Z)])).$$

Here is an algorithm.

The T algorithm. Suppose $\mathbf{Z} = (Z_1, \dots, Z_k)^T \in R^k$ is the vector of predictors. For each, $1 \leq j \leq k$, compute $c_{\text{opt}}(j)$, the split of variable j which gives highest purity, i.e. smallest $\tau^2(c)$. Then, for the first level of the tree, choose that variable X_j which minimizes $\tau^2(c_{\text{opt}}(j))$, the impurity, over all possible variable choices. Call the index on this variable j_1 and the optimal splitting point $c^{(1)}$. This gives a rule $\delta^{(1)}$ for the first level of the tree. For the second level, consider splitting separately $[c^{(1)}, \infty)$ and $(-\infty, c^{(1)})$ into $[c^{(1)}, c']$, $[c', \infty)$ on the right and $(-\infty, c)$, $[c, c^{(1)}]$ on the left for $c < c^{(1)}$ and $c' \geq c^{(1)}$, respectively. That is, use $\delta^{(2)}(z) = \delta_c(z)$ or $\bar{\delta}_c(z)$ defined for $z \geq c^{(1)}$, and $c \geq c^{(1)}$, and similarly, $\bar{\delta}^{(2)}(z) = \delta_{c'}(z)$ or $\bar{\delta}_{c'}(z)$ for $z < c^{(1)}$, $c < c^{(1)}$, and proceed as in level 1 to obtain $(c^{(21)}, j_{21})$, $(c^{(20)}, j_{20})$. Here, $c^{(21)}, c^{(20)}$ are the optimal cut points for variables $Z_{j_{21}}, Z_{j_{20}}$. Two nodes now arise from each of the nodes of level 1 with predictions carried out as described earlier — see Figure 1. We can continue in this fashion indefinitely if $P \leftrightarrow (\pi, p_0, p_1)$ where p_0, p_1 are densities. \square

It is intuitively clear and can be shown rigorously (Problem 12.2.24), that if each member of the partition defined by level m of the tree and has positive probability and

$C_m(\mathbf{z})$ is the member of the partition containing \mathbf{z} , then $C_m(\mathbf{z}) \supset C_{m+1}(\mathbf{z})$ for all m and $\bigcap_m C_m(\mathbf{z}) = \{\mathbf{z}\}$. Hence, if $|C_m(\mathbf{z})|$ denotes the length of $C_m(\mathbf{z})$

$$P[\mathbf{Z} \in C_m(\mathbf{z}) | Y = y] / |C_m(\mathbf{z})| \rightarrow p_y(\mathbf{z}) \text{ as } m \rightarrow \infty$$

where $p_y(\mathbf{z})$ is the conditional density of \mathbf{Z} given $Y = y$. Thus, the decision rule

$$\delta_m(\mathbf{z}) = 1 \text{ iff } \pi P[Y = 1 | \mathbf{Z} \in C_m(\mathbf{z})] \geq (1 - \pi) P[Y = 0 | \mathbf{Z} \in C_m(\mathbf{z})]$$

converges to the Bayes rule for each \mathbf{z} , as $m \rightarrow \infty$.

If we have a training sample $\mathbf{Z}_1, \dots, \mathbf{Z}_n$ and replace P by the empirical probability \hat{P} , C_m with the version \hat{C}_m based on \hat{P} , and π with $\hat{\pi} = \hat{P}[Y = 1]$, we have a way of growing empirical classification trees. We claim that if we stop at any level m and define $\hat{p}_{mj}(\mathbf{z}) = \hat{P}[\mathbf{Z} \in \hat{C}_m(\mathbf{z}) | Y = j]$, the ratio of the number of $Y_i = j$ which appear in the node, to which \mathbf{z} belongs, to the total number of $Y_i = j$. Then, the $\hat{p}_{mj}(z)$ are estimates of the $p_j(z)$ and the rule

$$\hat{\delta}(z) = 1 \text{ iff } \hat{\pi} \hat{p}_{m1}(z) \geq (1 - \hat{\pi}) \hat{p}_{m0}(z)$$

converges to the Bayes rule provided that $n \rightarrow \infty$, $m_n \rightarrow \infty$ slowly — see Breiman, Friedman, Olshen and Stone (1984) for rules for selecting the depth m .

CART and Sieves

It should be clear that, if we stop at stage m , the estimates of p_{mj} are just histograms but with bin size adapted to the data. They can be related to a sieve as follows. Take without loss of generality the support of P to be the cube I^k . Let $\mathcal{P}_m = \{\text{All histograms on } I^k \text{ with up to } m \text{ break points in each coordinate}\}$. That is, if $p_j \leftrightarrow P \in \mathcal{P}$, $j = 0, 1$

$$p_j(\mathbf{z}) = \sum_{i=1}^{2^m} a_{ij} \mathbf{1}(\mathbf{z} \in A_{m_{ij}})$$

where $\{A_{m_{ij}}\}$ is a partition of I^k into k -cubes $\{\mathbf{z} : b_i \leq z_i \leq c_i, 0 \leq i \leq m + 1, b_0 = 0, b_{m+1} = 1\}$. So \mathcal{P}_m is parametrizable by $\{a_{ij}, 1 \leq i \leq 2^m - 1, j = 0, 1\}$, the 2^{m-1} left hand vertices of k -cubes, and the indices (j_1, \dots, j_m) of the variables. If we call this parameter θ , then the tree gives us an estimate $\hat{\nu}_m$ of a parameter $\nu_m(P)$ such that $\nu_m(P_\theta) = \theta$. However, the method of estimation is not maximum likelihood or more generally of minimum contrast form. As we have noted, if $P \notin \mathcal{P}$, then $P_{\nu_m(P)}$, the member of \mathcal{P} which is “closest” in the appropriate sense to P converges to P as $m \rightarrow \infty$. Then, again, regularization, which in this case is choice of the tuning parameter m , is called for. For an extensive discussion of tree-structured methods, see Breiman et al (1984). \square

Summary. In this section we describe and analyze some of the procedures we mentioned previously in detail. In Section 12.2.1 we establish consistency and determine the rates of uniform convergence of the Nadaraya-Watson over suitable nonparametric models in regression. In Section 12.2.2 we study consistency of k th nearest neighbour rules in the classification. In Section 12.2.3 we also touch on logistic regressions, ridge regressions as examples of sieve methods, and their connections to empirical Bayes. Finally, in Section 12.2.4 we describe support vector machines, boosting, and the tree-based method CART.

12.3 Asymptotic Risk Criteria

In what sense are the various procedures we have discussed best or even good? The classical decision theory point of view we have discussed in Section 1.3 of Volume I is to first consider a probability model \mathcal{P} for $X \in \mathcal{X}$, specify a decision space D , a loss function $l : \mathcal{P} \times D \rightarrow R^+$, and the class of all possible (possibly randomized) decision procedures $\delta : \mathcal{X} \rightarrow D$. We measure the performance of a decision procedure δ by

$$R(P, \delta) \equiv E_P l(P, \delta(X)), \quad P \in \mathcal{P}. \quad (12.3.1)$$

The principle focussed on in the analysis of nonparametric classification and regression procedures is minimax. That is, we compute

$$R(\delta) \equiv \max_{\mathcal{P}} R(P, \delta),$$

and consider δ^* minimax optimal if

$$R(\delta^*) = \arg \min_{\mathcal{D}} R(\delta).$$

A key result which we extend is Theorem 3.3.3 which states that δ^* is minimax if we can find a sequence of priors π_k such that

$$\inf_{\delta} r(\pi_k, \delta) \rightarrow R(\delta^*),$$

where $r(\pi_k, \delta)$ denotes the Bayes risk of δ .

Recall from Section 12.1 that classification, regression, and other prediction problems are naturally split into two parts:

(i) “ P known”: Here we envisage observing a vector of predictors $\mathbf{Z} \in \mathcal{Z}$ and know P , where P is the distribution of $\mathbf{X} = (\mathbf{Z}, Y)$. Y is unknown and we want to predict Y for each given $\mathbf{Z} = \mathbf{z}$. A decision rule is a prediction rule $d(\cdot) : \mathcal{Z} \rightarrow \mathcal{Y}$ where \mathcal{Y} is the set of values or classes to be predicted. The risk of $d(\cdot)$ is

$$R(P, d(\cdot)) = \int l(y, d(\mathbf{z})) dP(\mathbf{z}, y) = E_P l(Y, d(\mathbf{Z})),$$

where $l(y, d)$ is the loss associated with predicting y to be $d(\mathbf{z})$. If we think of Y as a random “parameter,” there is, for any such problem, an optimal rule, the Bayes rule of Proposition 3.2.1,

$$\delta_P^*(\mathbf{z}) = \arg \min_d E_P \{l(Y, d(\mathbf{z})) | \mathbf{Z} = \mathbf{z}\}$$

which yields the minimum Bayes risk,

$$R_P^* \equiv E_P R(P, \delta_P^*(\mathbf{Z})).$$

(ii) *The “real” problem.* P unknown. We are given training data $\mathbf{X}_i \equiv (\mathbf{Z}_i, Y_i), i = 1, \dots, n$, i.i.d. as $\mathbf{X} \sim P$. A decision procedure is a mapping

$$\delta(\cdot : \mathbf{X}_i, 1 \leq i \leq n) : \mathcal{Z} \rightarrow D$$

and has risk

$$R(P, \delta) = E_P \int l(y, \delta(\mathbf{z} : \mathbf{X}_i, 1 \leq i \leq n)) dP(\mathbf{z}, y),$$

where l is the loss when predicting y to be $\delta(\mathbf{z} : \mathbf{X}_i, 1 \leq i \leq n)$.

It's useful in this framework to equivalently measure performance of δ as its *regret*

$$\tilde{R}(P, \delta) \equiv R(P, \delta) - R_P^*, \quad (12.3.2)$$

and ask first that, asymptotically, we have, for a sequence $\{\delta_n\}$,

Consistency: $\tilde{R}(P, \delta_n) \rightarrow 0$ as $n \rightarrow \infty$ for all P , uniformly on \mathcal{P} .

And second, given consistency, that we have

Asymptotic minimax regret on \mathcal{P} :

$$\tilde{R}(\delta_n) \equiv \sup \{ \tilde{R}(P, \delta_n) : P \in \mathcal{P} \} \asymp \inf_{\delta} \sup \{ \tilde{R}(P, \delta) : P \in \mathcal{P} \} \equiv \tilde{R}_n. \quad (12.3.3)$$

Here \asymp can mean *strongly*.

$$\frac{\tilde{R}(\delta_n)}{\tilde{R}_n} \rightarrow 1 \quad (12.3.4)$$

or *weakly*

$$\tilde{R}(\delta_n) = O(\tilde{R}_n) \text{ and } \tilde{R}_n = O(\tilde{R}(\delta_n)), \quad (12.3.5)$$

noting that we are in a situation where $\tilde{R}_n \rightarrow 0$. Typically, unless \mathcal{P} is regular parametric, (12.3.5) is the best that can be hoped for.

When proving minimaxity the essential argument always has the same nature. Exhibit, or at least show, the existence of a sequence of Bayes priors and corresponding Bayes rules whose *Bayes regret*, that is, the regret of the Bayes rules, is asymptotically the same or no better than the maximum regret $\tilde{R}(\delta_n)$ of the candidate minimax procedures. We begin with the parametric case.

12.3.1 Optimal Prediction in Parametric Regression Models

We establish the consistency and strong asymptotic minimax regret property of an empirical plug-in Bayes predictor for squared error loss when \mathcal{P} is a regular parametric model. Let $\mathbf{X} = (\mathbf{Z}, Y)$ be distributed according to $P_{\boldsymbol{\theta}}$ with density $p(\mathbf{z}, y | \boldsymbol{\theta}) : \boldsymbol{\theta} \in R^p$ such that the model $\{P_{\boldsymbol{\theta}} : \boldsymbol{\theta} \in \Theta\}$, $\Theta \subset R^p$, is regular and obeys the conditions of Theorem 6.2.2. Suppose also that, as usual, the marginal density $q(\mathbf{z})$ of \mathbf{Z} does not depend on $\boldsymbol{\theta}$, and that

- (i) Y is bounded above: $|Y| \leq M$ for some $M > 0$.

(ii) $q(\cdot)$ is bounded below: $P(q(\mathbf{Z}) \geq \delta) = 1$ for some $\delta > 0$.

$$(iii) \max_{\boldsymbol{\theta} \in \Theta} \left\{ \int \nabla_{\boldsymbol{\theta}}^T e(\mathbf{z}, \boldsymbol{\theta}) I^{-1}(\boldsymbol{\theta}) \nabla_{\boldsymbol{\theta}} e(\mathbf{z}, \boldsymbol{\theta}) q(\mathbf{z}) d\mathbf{z} \right\} < \infty,$$

where, by (1.4.5), the Bayes regret of δ at $P_{\boldsymbol{\theta}}$ for squared error loss is

$$R_{\boldsymbol{\theta}}(\delta) \equiv E_{\boldsymbol{\theta}} \left[\int (\delta(\mathbf{z} : \mathbf{X}_1, \dots, \mathbf{X}_n) - e(\mathbf{z}, \boldsymbol{\theta}))^2 q(\mathbf{z}) d\mathbf{z} \right].$$

Here, $E_{\boldsymbol{\theta}}$ is expectation with respect to $P_{\boldsymbol{\theta}}$, and by (1.4.4), the optimal predictor of Y given $\boldsymbol{\theta}$ and $\mathbf{Z} = \mathbf{z}$ is

$$e(\mathbf{z}, \boldsymbol{\theta}) = E_{\boldsymbol{\theta}}(Y|\mathbf{z}) = \int_{-\infty}^{\infty} y p(\mathbf{z}, y|\boldsymbol{\theta}) dy / q(\mathbf{z}). \quad (12.3.6)$$

When $\boldsymbol{\theta}$ is unknown, the natural estimate of $e(\mathbf{z}, \boldsymbol{\theta})$ is the plug-in estimate $e(\mathbf{z}, \hat{\boldsymbol{\theta}})$ where $\hat{\boldsymbol{\theta}}$ is the MLE of $\boldsymbol{\theta}$, which we assume behaves regularly. Thus, we take

$$\delta^*(\mathbf{z} : \mathbf{X}_1, \dots, \mathbf{X}_n) \equiv e(\mathbf{z}, \hat{\boldsymbol{\theta}}).$$

By the expansion

$$e(\mathbf{z}, \hat{\boldsymbol{\theta}}) = e(\mathbf{z}, \boldsymbol{\theta}) + \nabla_{\boldsymbol{\theta}}^T e(\mathbf{z}, \boldsymbol{\theta})(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) + o_P(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}),$$

formally,

$$\begin{aligned} R_{\boldsymbol{\theta}}(\delta^*) &= E_{\boldsymbol{\theta}} \int (e(\mathbf{z}, \hat{\boldsymbol{\theta}}) - e(\mathbf{z}, \boldsymbol{\theta}))^2 q(\mathbf{z}) d\mathbf{z} \\ &= E_{\boldsymbol{\theta}} \int [\nabla_{\boldsymbol{\theta}}^T e(\mathbf{z}, \boldsymbol{\theta})(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})]^2 q(\mathbf{z}) d\mathbf{z} (1 + o_P(1)) \\ &= n^{-1}(1 + o(1)) \int \nabla_{\boldsymbol{\theta}}^T e(\mathbf{z}, \boldsymbol{\theta}) I^{-1}(\boldsymbol{\theta}) \nabla_{\boldsymbol{\theta}} e(\mathbf{z}, \boldsymbol{\theta}) q(\mathbf{z}) d\mathbf{z} \end{aligned} \quad (12.3.7)$$

where $I(\boldsymbol{\theta})$ is the Fisher information matrix. These formal calculations can be justified under our conditions (Problem 12.3.1) and establish Bayes consistency of the decision rule δ^* .

Proposition 12.3.1. *Under (i), (ii), (iii), and the assumptions of Theorem 6.2.2, the predictor $\delta^* = e(\mathbf{z}, \hat{\boldsymbol{\theta}})$ is asymptotically Bayes consistent for squared error loss.*

To show that the asymptotic minimax regret property (12.3.4) holds we will examine what the Bayes solution of this problem is if we assign to $\boldsymbol{\theta}$ a positive continuous bounded prior density π . Then, for \mathbf{Z} independent of $\mathbf{X}_1, \dots, \mathbf{X}_n$ and $\boldsymbol{\theta}$, the Bayes risk of a decision rule δ is

$$\begin{aligned} \int R_{\boldsymbol{\theta}}(\delta) \pi(\boldsymbol{\theta}) d\boldsymbol{\theta} &= E_{\pi} \{ E_{\boldsymbol{\theta}} (\delta(\mathbf{Z} : \mathbf{X}_1, \dots, \mathbf{X}_n) - e(\mathbf{Z}, \boldsymbol{\theta}))^2 | \boldsymbol{\theta} \} \\ &= E_{\boldsymbol{\theta}} \{ E_{\pi} (\delta(\mathbf{Z} : \mathbf{X}_1, \dots, \mathbf{X}_n) - e(\mathbf{Z}, \boldsymbol{\theta}))^2 | \mathbf{Z}, \mathbf{X}_1, \dots, \mathbf{X}_n \}, \end{aligned} \quad (12.3.8)$$

and Theorem 1.4.1 leads to the Bayes rule,

$$\delta_\pi(\mathbf{Z} : \mathbf{X}_1, \dots, \mathbf{X}_n) = E_\pi[e(\mathbf{Z}, \boldsymbol{\theta}) | \mathbf{Z}, \mathbf{X}_1, \dots, \mathbf{X}_n] = E_\pi[e(\mathbf{Z}, \boldsymbol{\theta}) | \mathbf{X}],$$

where now \mathbf{X} denotes $(\mathbf{X}_1, \dots, \mathbf{X}_n)$, and the Bayes regret for δ_π , the minimizer of Bayes risk, is

$$R_\pi = \int E_{\boldsymbol{\theta}}(E_\pi(e(\mathbf{Z}, \boldsymbol{\theta}) | \mathbf{X}) - e(\mathbf{Z}, \boldsymbol{\theta}))^2 \pi(\boldsymbol{\theta}) d\boldsymbol{\theta}. \quad (12.3.9)$$

Next we argue that we get a close approximation to R_π if in (12.3.9) we replace $E_\pi[e(\mathbf{Z}, \boldsymbol{\theta}) | \mathbf{X}]$ with $e(\mathbf{Z}, \hat{\boldsymbol{\theta}})$. By assumptions (i), (ii), and (iii),

$$e(\mathbf{z}, \hat{\boldsymbol{\theta}} + \mathbf{v}n^{-\frac{1}{2}}) - e(\mathbf{z}, \hat{\boldsymbol{\theta}}) = \nabla_{\boldsymbol{\theta}}^T e(\mathbf{z}, \hat{\boldsymbol{\theta}}) \mathbf{v}n^{-\frac{1}{2}} + O_P(n^{-1}), \quad (12.3.10)$$

uniformly for \mathbf{v} in a compact set. Writing $e(\mathbf{z}, \boldsymbol{\theta}) \cong e(\mathbf{z}, \hat{\boldsymbol{\theta}}) + \nabla_{\boldsymbol{\theta}} e(\mathbf{z}, \hat{\boldsymbol{\theta}})(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})$, we have

$$\mathcal{L}_{\boldsymbol{\theta}}\{E_\pi[e(\mathbf{Z}, \boldsymbol{\theta}) | \mathbf{X}] - e(\mathbf{z}, \hat{\boldsymbol{\theta}})\} \cong \mathcal{L}_{\boldsymbol{\theta}}\{\nabla_{\boldsymbol{\theta}}(e(\mathbf{Z}, \hat{\boldsymbol{\theta}})) E_\pi((\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})) | \mathbf{X}\}.$$

Noting that e is uniformly bounded, and applying the Bernstein–von Mises theorem to $\sqrt{n}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}) | \mathbf{X}$ (Theorems 5.5.2 and 6.2.3), we get that

$$\begin{aligned} & \int E_{\boldsymbol{\theta}}(E_\pi(e(\mathbf{Z}, \boldsymbol{\theta}) | \mathbf{X}) - e(\mathbf{Z}, \hat{\boldsymbol{\theta}))^2 \pi(\boldsymbol{\theta}) d\boldsymbol{\theta} \\ &= \frac{1}{n} \int E_{\boldsymbol{\theta}} \left[\int \nabla_{\boldsymbol{\theta}}^T(e(\mathbf{Z}, \hat{\boldsymbol{\theta}})) \mathbf{v}^T \varphi(\mathbf{v}, I(\hat{\boldsymbol{\theta}})) d\mathbf{v} \right]^2 \pi(\boldsymbol{\theta}) d\boldsymbol{\theta} (1 + o(1)), \end{aligned} \quad (12.3.11)$$

where $\varphi(\mathbf{v}, \Sigma)$ is the $\mathcal{N}(\mathbf{0}, \Sigma)$ density. By the symmetry of the $\mathcal{N}(\mathbf{0}, \Sigma)$ distribution the inside integral is zero. Then, (12.3.11) is $o(n^{-1})$ and using (1.4.5), we can conclude that

$$R_\pi = \int E_{\boldsymbol{\theta}}(e(\mathbf{Z}, \hat{\boldsymbol{\theta}}) - e(\mathbf{Z}, \boldsymbol{\theta}))^2 \pi(\boldsymbol{\theta}) d\boldsymbol{\theta} + o(n^{-1}). \quad (12.3.12)$$

An application of the argument leading to the Bernstein–von Mises theorem similarly yields

$$\int R_{\boldsymbol{\theta}}(\delta^*) \pi(\boldsymbol{\theta}) d\boldsymbol{\theta} = R_\pi(1 + o(1)). \quad (12.3.13)$$

What we have shown is that, by (12.3.12) and the expansion leading to (12.3.7),

Proposition 12.3.2. *Under the conditions of Proposition 12.3.1,*

(i) *the Bayes regret is*

$$R_\pi = \frac{c(\pi)}{n}(1 + o(1))$$

where

$$c(\pi) = \int E_{\boldsymbol{\theta}}[\nabla_{\boldsymbol{\theta}}^T e(\mathbf{Z}, \boldsymbol{\theta}) I^{-1}(\boldsymbol{\theta}) \nabla_{\boldsymbol{\theta}} e(\mathbf{Z}, \boldsymbol{\theta})] \pi(\boldsymbol{\theta}) d\boldsymbol{\theta}$$

and thus

(ii) $e(\cdot, \hat{\theta})$ is an asymptotically efficient estimate of $e(\cdot, \theta)$.

We next show that δ^* has the strong asymptotic minimax property.

Proposition 12.3.3. *Under the assumptions of Proposition 12.3.1, δ^* is asymptotically minimax regret in the sense of (12.3.4).*

Proof. Take $\pi = \pi_t$ with π_t converging weakly to point mass at θ^* , as $t \rightarrow 0$, where

$$\theta^* \equiv \arg \max E_{\theta} [\nabla_{\theta}^T e(\mathbf{Z}, \theta) I^{-1}(\theta) \nabla_{\theta} e(\mathbf{Z}, \theta)].$$

Since the integral of a function is less than its maximum,

$$\max_{\pi} c(\pi) = \max_{\theta} E_{\theta} [\nabla_{\theta}^T e(\mathbf{Z}, \theta) I^{-1}(\theta) \nabla_{\theta} e(\mathbf{Z}, \theta)].$$

It follows that the minimum Bayes regret is

$$\frac{1}{n} \max_{\theta} E_{\theta} [\nabla_{\theta}^T e(\mathbf{Z}, \theta) I^{-1}(\theta) \nabla_{\theta} e(\mathbf{Z}, \theta)] (1 + o(1))$$

and hence we conclude that δ^* is asymptotically minimax regret in the strong sense of (12.3.4) by the arguments establishing Theorem 3.3.3. That is, we have shown that there are priors whose Bayes regret has the same asymptotic behavior as the maximum regret of our candidate rule δ^* . \square

12.3.2 Optimal Rates of Convergence for Estimation and Prediction in Nonparametric Models

In this section we will develop methods for showing weak regret minimality in general nonparametric frameworks. Again, this involves introducing a candidate procedure and finding a “least favorable” prior where the Bayes risk converges to the maximum risk of the candidate rule. Our main application is

Example 12.3.1. *Nonparametric regression.* We show that with proper bandwidth choice, the Nadaraya-Watson kernel of Section 12.2.1 is asymptotically minimax regret in the weak sense for the nonparametric problem of predicting $\mu(\mathbf{z})$ where $\mu(\mathbf{z}) \equiv E(Y|\mathbf{Z} = \mathbf{z})$ satisfies a first order Lipschitz condition,

$$\mathcal{P} = \{P : |\mu(\mathbf{z}) - \mu(\mathbf{z}')| \leq M|\mathbf{z} - \mathbf{z}'| \text{ for all } \mathbf{z}, \mathbf{z}'\}. \quad (12.3.14)$$

We recall first that the rate $n^{-2/(2+d)}$ is achievable in nonparametric regression for the problem of predicting $\mu(\mathbf{Z}) = E(Y|\mathbf{Z})$ with squared error loss on the basis of (\mathbf{Z}, Y) , $\mathbf{Z} \in R^d$, when the joint distribution P of (\mathbf{Z}, Y) belongs to \mathcal{P} given by (12.3.14). The following theorem, which is easily generalized to more general \mathbf{Z} , Y , follows from the proof of Theorem 12.2.1 in Appendix D.7.

Theorem 12.3.1. *Let $\mathcal{P}_0 = \{P \in \mathcal{P} : P \text{ on } I^d\}$, \mathbf{Z} uniform on I^d . Let $\hat{\mu}_{NW}$ be the NW estimate given in (12.2.7) with product kernel $\prod_{j=1}^d K_0(t_j)$, where K_0 is non-negative, continuous, bounded, and has compact support. Then, using bandwidth $h = cn^{-1/(2+d)}$, $c > 0$,*

- (i) For fixed \mathbf{z} in the interior of I^d , $\sup \{ |\widehat{\mu}_{NW}(\mathbf{z}) - \mu(\mathbf{z})| : P \in \mathcal{P}_0 \} = O(n^{-\frac{2}{2+d}})$.
- (ii) $\sup \{ E_P |\widehat{\mu}_{NW}(\mathbf{Z}) - \mu(\mathbf{Z})|^2 : P \in \mathcal{P}_0 \} = O(n^{-\frac{2}{2+d}})$. (12.3.15)

Remark 12.3.1(a). In the problems we will show that rate $n^{-2s/(2s+d)}$ can be obtained if we generalize (12.3.14) to the *Hölder class*

$$\mathcal{P}_s \equiv \{P : |D^{s-1}(\mu(\mathbf{z}) - \mu(\mathbf{z}'))| \leq M|\mathbf{z} - \mathbf{z}'| \text{ for all } \mathbf{z}, \mathbf{z}'\}$$

where D^{s-1} is the $s-1$ differential of μ , $D^0\mu = \mu$, provided we replace Nadaraya-Watson by a suitable local polynomial estimate.

(b) These rates for Hölder classes or similar Sobolev classes can be achieved by a variety of predictors — see Tsybakov (2008) and Györfi et al (2002) for examples. This is also true for classes which permit discontinuous functions. In this case an example of appropriate procedures is methods based on wavelet expansions of the members of \mathcal{P} in (12.3.14). See Donoho and Johnstone (1994) for basic results in this direction. \square

Lower bounds

Using lower bounds on risk, we will show that for P as in (12.3.14), indeed, $n^{-\frac{2}{2+d}}$ is the best minimax rate one can hope for. We also reveal something equally significant. For IMSE, the “least favorable” prior distributions are quite unlike those in Section 12.3.1. The former can be thought of as smooth densities concentrating around a least favorable point θ^* . The latter concentrate around probabilities P whose corresponding $\mu(\cdot)$ is as wiggly as possible but still obeys the constraints on P . Such μ seem rather implausible in actual practice and the minimax theorems are not as compelling as in the low dimensional parametric case. As Einstein said (in English translation), “God is subtle but not malevolent.” \square

In general, to prove lower bound results, we necessarily want to use asymptotic versions of the minimax theory of Chapter 8. There are a number of approaches to lower bounds. Some of the ones we present and others are discussed in full detail in Tsybakov (2008), Chapter 2.

The testing approach to lower risk bounds

Suppose that our decision problem is of the form: We observe $X \in \mathcal{X}$, $X \sim P$, $P \in \mathcal{P}$ and want to estimate a parameter $\theta : \mathcal{P} \rightarrow \mathcal{T}$ where we are given a metric ρ on \mathcal{T} and our goal is to construct $\tilde{\theta} : \mathcal{X} \rightarrow \mathcal{T}$ so that $\tilde{\theta}$ has the minimax property

$$\max_{\mathcal{P}} E_P \rho(\tilde{\theta}(X), \theta(P)) = \min_{\tilde{\theta}} \max_{\mathcal{P}} E_P \rho(\tilde{\theta}(X), \theta(P)).$$

Evidently prediction in regression is a problem of this type with $\theta(P)(\cdot) = E_P(Y|\cdot) = \mu(\cdot)$. Here $X = (\mathbf{Z}, Y)$, \mathcal{T} is a collection of functions of \mathbf{Z} and a possible metric to use as a loss function is

$$\rho(t_1, t_2) \equiv \left(\int (t_1(\mathbf{z}) - t_2(\mathbf{z}))^2 d\nu(\mathbf{z}) \right)^{\frac{1}{2}}$$

for some probability ν and $t_1, t_2 \in \mathcal{T}$. For this $\theta(P)$ and ρ , if $\nu = Q$ is the distribution of a \mathbf{Z} independent of the data vector \mathbf{X} ,

$$E[\rho^2(\hat{\theta}(\mathbf{X}), \theta(P))] = E_Q[MSE(\hat{\mu}(\mathbf{Z}))] = IMSE(\hat{\mu})$$

by Fubini's theorem. Another example is $\nu =$ point mass at a given point \mathbf{z} .

Returning to the general case, let $\theta_0 \equiv \theta(P_0)(\cdot)$ where P_0 is fixed and for $r > 0$ let

$$S_r \equiv \{P : \rho(\theta(P), \theta_0) \geq r\}.$$

Our strategy is to select a subset

$$\Omega_r = \{Q_\lambda \in S(r) : \lambda \in \Lambda \subset R\}$$

of $S(r)$ parametrized by a parameter $\lambda \in \Lambda$ and to select a prior $\pi_r(\lambda)$ on Λ such that the Bayes test generates lower bounds in risk. We will identify $P \in \Omega_r$ with λ and write $\Pi_r(dP)$ for $\Pi_r(\lambda)d\lambda$ as well as $\Pi_r(\Omega_r)$ for $\Pi_r(\Lambda)$.

Theorem 12.3.2. Suppose all $P \in \mathcal{P}_r \equiv \{P_0\} \cup \Omega_r$ have discrete or continuous case density functions p . Consider the testing problem, $H : P = P_0$ vs $K : P \in \Omega_r$ with $0 - 1$ loss. Let Π_r be a prior distribution on \mathcal{P}_r such that $\frac{1}{2} = \Pi_r(\{P_0\}) = 1 - \Pi_r(\Omega_r)$ and let φ_B be a corresponding Bayes test,

$$\begin{aligned} \varphi_B(x) &= 1 \quad \text{if } L \equiv \int_{\Omega_r} [p(x)/p_0(x)]\Pi_r(dP) > c \\ &= 0 \quad \text{if } L < c. \end{aligned}$$

Let

$$R_B(\Pi_r) \equiv r(\Pi_r, \varphi_B) = \frac{1}{2} \left(E_0 \varphi_B(X) + \int E_P(1 - \varphi_B(X))\Pi_r(dP) \right) \quad (12.3.16)$$

be the minimum Bayes risk, and assume $0 < R_B(\Pi_r) < 1$. Then, for any $\hat{\theta}$ and any metric ρ ,

$$\max_{\mathcal{P}_r} P[\rho(\hat{\theta}, \theta(P)) \geq \frac{r}{2}] \geq R_B(\Pi_r).$$

Proof: Consider the test

$$\begin{aligned} \varphi(x) &= 1 \quad \text{if } \rho(\hat{\theta}, \theta_0) > \frac{r}{2} \\ &= 0 \quad \text{if } \rho(\hat{\theta}, \theta_0) \leq \frac{r}{2}. \end{aligned}$$

Then φ has Bayes risk no lower than the Bayes test, thus

$$\frac{1}{2} \left(E_0 \varphi(X) + \max_{\Omega_r} E_P(1 - \varphi(X)) \right) \geq R_B(\Pi_r). \quad (12.3.17)$$

So, either

$$P_0\left[\rho(\hat{\theta}, \theta_0) > \frac{r}{2}\right] \geq R_B(\Pi_r)$$

or

$$\max_{\Omega_r} P\left[\rho(\hat{\theta}, \theta_0) \leq \frac{r}{2}\right] \geq R_B(\Pi_r). \quad (12.3.18)$$

In the second case, writing θ for $\theta(P)$ we have, by the triangle inequality,

$$\rho(\hat{\theta}, \theta_0) \geq \rho(\theta, \theta_0) - \rho(\hat{\theta}, \theta) \geq r - \rho(\hat{\theta}, \theta).$$

Then,

$$\rho(\hat{\theta}, \theta_0) \leq \frac{r}{2} \implies \rho(\hat{\theta}, \theta) \geq \frac{r}{2}.$$

Using (12.3.18), we see there exists $P \in \Omega_r$ such that

$$R_B(\Pi_r) = P\left[\rho(\hat{\theta}, \theta_0) \leq \frac{r}{2}\right] \leq P\left[\rho(\hat{\theta}, \theta) \geq \frac{r}{2}\right],$$

and the result follows. \square

Further, by Markov's inequality we obtain a minmax risk bound,

Corollary 12.3.1. *Under the conditions of Theorem 12.3.2*

$$\min_{\hat{\theta}} \max_{\mathcal{P}_r} E_P \rho(\hat{\theta}, \theta(P)) \geq \frac{2R_B(\Pi_r)}{r}.$$

Bounds from asymptotic likelihoods in the Gaussian contiguous case.

The main weakness of the bound based on Theorem 12.3.2 is that we are left with the problem of lower bounding the Bayes risk $R_B(\Pi_r)$ itself. However, in the case that $(X_1, \dots, X_n)^T$ is a vector of independent random variables with $p(x) = \prod_{j=1}^n p^{(j)}(x_j)$ densities, there is a natural approach to asymptotic bounds using the contiguity ideas of Section 9.5. Let $r = r_n$ depend on n and write Π_n for Π_{r_n} . We are looking for $\pi_n(\cdot)$ on $S(r_n) \cup \{P_0\}$ which make the testing problem as hard as possible, i.e. an (approximately) least favorable prior distribution π_n . Let, Ω_n denote Ω_{r_n} and

$$L_n(P) = \log \prod_{j=1}^n \frac{p^{(j)}}{p_0^{(j)}}(X_j),$$

where $P \in \Omega_n$. Suppose that we can choose Ω_n and Π_n such that $\pi_n(\Omega_n) = \frac{1}{2} = \pi_n(P_0)$ and, as $n \rightarrow \infty$,

$$\pi_n\left\{P_0 : L_n \xrightarrow{P_0} \mathcal{N}\left(-\frac{\sigma^2}{2}, \sigma^2\right)\right\} \rightarrow 1.$$

It readily follows that, asymptotically, we should take $c = \frac{1}{2}$ in the Bayes test φ_B and then (Problem 12.3.7), as $n \rightarrow \infty$,

$$R_B(\Pi_r) \rightarrow 1 - \Phi\left(\frac{\sigma}{2}\right).$$

Thus, a general approach to getting lower bounds is to construct Ω_n and Π_n as above. We can then conclude by Corollary 12.3.1 that it will not be possible to estimate $\theta(P)(\cdot)$ at a rate faster than $r_n \downarrow 0$.

Example 12.3.1. (Continued) Nonparametric regression.

We next apply the likelihood idea to the Lipschitz model \mathcal{P}_0 of Theorem 12.3.1 and show minimaxity of the Nadaraya-Watson estimate. As before, for simplicity, assume \mathbf{Z} is uniform on the unit cube $I^d = [0, 1]^d$.

(a) Estimation of $\mu(\cdot)$ at a Point

We begin with the easier problem of estimating $\mu(\mathbf{z})$ at a point, say, $\mathbf{z}_0 \equiv (\frac{1}{2}, \dots, \frac{1}{2})^T$ with loss function $|\hat{\mu}(\mathbf{z}_0) - \mu(\mathbf{z}_0)|$. For our submodel, we consider

$$Y_i = \mu_h(\mathbf{Z}_i) + N_i, \quad i = 1, \dots, n \quad (12.3.19)$$

where N_1, \dots, N_n are i.i.d. $\mathcal{N}(0, 1)$ and we assume, without loss of generality, that

$$\mu_0(\cdot) \equiv \mu_{P_0}(\cdot) = 0.$$

Given that $\hat{\mu}_{NW}$ is the candidate rule, it is natural to consider a prior π_n putting all its mass on $\mu_h(\cdot)$ given by

$$\mu_h(\mathbf{z}) \equiv hg\left(\frac{\mathbf{z} - \mathbf{z}_0}{h}\right)1(|\mathbf{z} - \mathbf{z}_0| \leq \frac{h}{2}) \quad (12.3.20)$$

where $h = h_n$ depends on n , and the kernel g is of the form

$$g(\mathbf{z}) \equiv \prod_{j=1}^d g_1(z_j)$$

with g_1 differentiable and, for some $M_0 > 0$,

$$g_1 : [0, 1] \rightarrow R, \quad g_1 \neq 0, \quad g_1(0) = g_1(1) = 0, \quad |g_1'| \leq M_0. \quad (12.3.21)$$

Note that, for $\mathbf{z}^{(1)}, \mathbf{z}^{(2)} \in I^d$,

$$\begin{aligned} \text{(i)} \quad & |g(\mathbf{z}^{(k)})| \leq M_0^d, \quad k = 1, 2. \\ \text{(ii)} \quad & |g(\mathbf{z}^{(1)}) - g(\mathbf{z}^{(2)})| \leq M_0^{d-1} \sum_{j=1}^d |g_1(\mathbf{z}_j^{(1)}) - g_1(\mathbf{z}_j^{(2)})| \\ & \leq M_0^d \sum_{j=1}^d |\mathbf{z}_j^{(1)} - \mathbf{z}_j^{(2)}| \leq M_0^d d^{\frac{1}{2}} |\mathbf{z}^{(1)} - \mathbf{z}^{(2)}|. \end{aligned} \quad (12.3.22)$$

Thus if $M_0 \equiv (Md^{-\frac{1}{2}})^{-d}$, for some $M > 0$ and $|\mathbf{z}^{(k)} - \mathbf{z}_0| \leq \frac{1}{2}h$, $|\mathbf{z}^{(k)} - \mathbf{z}_0| \leq \frac{1}{2}h$, $k = 1, 2$, then

$$|\mu_h(\mathbf{z}^{(1)}) - \mu_h(\mathbf{z}^{(2)})| \leq h \left| g\left(\frac{\mathbf{z}^{(1)} - \mathbf{z}_0}{h}\right) - g\left(\frac{\mathbf{z}^{(2)} - \mathbf{z}_0}{h}\right) \right| \leq M |\mathbf{z}^{(1)} - \mathbf{z}^{(2)}|. \quad (12.3.23)$$

In view of (12.3.21), $hg_1[(\mathbf{z} - \frac{1}{2})/h]1(|\mathbf{z} - \frac{1}{2}| \leq h)$ is differentiable on $(0, 1)$ with derivative bounded by M_0 in absolute value. Thus the first order Lipschitz condition (12.3.14) holds for all $\mathbf{z}_1, \mathbf{z}_2$. Moreover, by construction,

$$|\mu_h(\mathbf{z})| \leq hM_0^d, \quad (12.3.24)$$

and

$$L_n(P) = \sum_{i=1}^n \left(\mu_h(\mathbf{Z}_i) N_i - \frac{\mu_h^2(\mathbf{Z}_i)}{2} \right), \quad (12.3.25)$$

where N_i are i.i.d. $\mathcal{N}(0, 1)$. So, given $\mathbf{Z}_1, \dots, \mathbf{Z}_n$,

$$L_n \sim \mathcal{N}\left(-\frac{1}{2} \sum_{i=1}^n \mu_h^2(\mathbf{Z}_i), \sum_{i=1}^n \mu_h^2(\mathbf{Z}_i)\right).$$

Now,

$$\begin{aligned} \mu_h(\mathbf{Z}_i) &= 0 \text{ if } |\mathbf{Z}_i - \mathbf{z}_0| > h \\ &= hg\left(\frac{\mathbf{Z}_i - \mathbf{z}_0}{h}\right) \text{ otherwise.} \end{aligned}$$

Evidently,

$$P[|\mathbf{Z}_i - \mathbf{z}_0| \leq h] \asymp h^d$$

and given $|\mathbf{Z}_i - \mathbf{z}_0| \leq h$, $\frac{\mathbf{Z}_i - \mathbf{z}_0}{h}$ is uniform on I^d . Thus,

$$\begin{aligned} E \sum_{i=1}^n \mu_h^2(\mathbf{Z}_i) &\asymp nh^{d+2} \\ \text{Var} \sum_{i=1}^n \mu_h^2(\mathbf{Z}_i) &\leq nE\mu_h^4(\mathbf{Z}) \asymp nh^{d+4} \end{aligned}$$

and if we take $h = n^{-\frac{1}{d+2}}$, then $\sum_{i=1}^n \mu_h^2(\mathbf{Z}_i) \xrightarrow{P} \sigma^2$. Hence,

$$L_n(P) \implies \mathcal{N}\left(-\frac{1}{2}\sigma^2, \sigma^2\right), \quad (12.3.26)$$

where $\sigma^2 = (\int_0^1 g^2(z)dz)^d$. In view of (12.3.23), (12.3.24) and (12.3.26) we can apply Corollary 12.3.1 and obtain that the lower bound on the risk has the rate of $n^{-\frac{2}{d+2}}$ we got for the Nadaraya-Watson estimate in Theorem 12.2.1. That is, the Nadaraya-Watson estimate achieves the optimal rate of convergence.

(b) Estimating $\mu(\cdot)$ as a Function. We approach this problem from the estimation point of view with a prior suggested by our analysis of estimation at a point. Let

$$\mu_h(\mathbf{z}, \boldsymbol{\varepsilon}) \equiv h \sum_{j=1}^N \varepsilon_j g\left(\frac{\mathbf{Z} - \mathbf{z}_j}{h}\right) 1(\mathbf{z} \in C_j), \quad (12.3.27)$$

where the ε_j are i.i.d. $P[\varepsilon_j = 0] = P[\varepsilon_j = 1] = \frac{1}{2}$, $\boldsymbol{\varepsilon} \equiv (\varepsilon_1, \dots, \varepsilon_N)^T$, and

$$g(\mathbf{z}) = \prod_{j=1}^d g_1(z_j)$$

where $g_1(z) = z1(0 \leq z \leq \frac{1}{2}) + (1-z)1(\frac{1}{2} < z \leq 1)$ for $0 \leq z \leq 1$, C_1, \dots, C_N are a partition of I^d into $N = h^{-d}$ cubes of side length h , and the \mathbf{z}_j are the centers of the C_j . It can be shown (Problem 12.3.4) that if N_1, \dots, N_n are i.i.d. $\mathcal{N}(0, 1)$ and

$$Y_i = \mu_h(\mathbf{Z}_i, \boldsymbol{\varepsilon}) + N_i, \quad i = 1, \dots, n, \quad (12.3.28)$$

then, for $M = (d^{-\frac{1}{4}})$, (12.3.27) and (12.3.28) define a prior distribution π on P with

$$|\mu(\mathbf{z}_1, \boldsymbol{\varepsilon}) - \mu(\mathbf{z}_2, \boldsymbol{\varepsilon})| \leq M|\mathbf{z}_1 - \mathbf{z}_2|.$$

We now apply (12.3.6) and (12.3.9) treating $\boldsymbol{\varepsilon} \equiv (\varepsilon_1, \dots, \varepsilon_N)^T$ as θ . Then, $e(\mathbf{z}, \boldsymbol{\varepsilon}) = \mu_h(\mathbf{z}, \boldsymbol{\varepsilon})$ and

$$E\{e(\mathbf{z}, \boldsymbol{\varepsilon})|\mathbf{X}_1, \dots, \mathbf{X}_n\} = h \sum_{j=1}^N P[\varepsilon_j = 1|\mathbf{X}_1, \dots, \mathbf{X}_n] g\left(\frac{\mathbf{z} - \mathbf{z}_j}{h}\right) 1(\mathbf{z} \in C_j). \quad (12.3.29)$$

Using the fact that the $1_j \equiv 1(\mathbf{z} \in C_j)$ are orthogonal we see from (12.3.9) and (12.3.29) that the Bayes risk is

$$\begin{aligned} R_\pi &= E\left\{h^2 \int \left(\sum_{j=1}^N (P[\varepsilon_j = 1|\mathbf{X}_1, \dots, \mathbf{X}_n] - \varepsilon_j) g\left(\frac{\mathbf{z} - \mathbf{z}_j}{h}\right) 1_j(\mathbf{z})\right)^2 d\mathbf{z}\right\} \\ &= h^2 E\left\{\sum_{j=1}^N [E_{\boldsymbol{\varepsilon}}(P[\varepsilon_j = 1|\mathbf{X}_1, \dots, \mathbf{X}_n] - \varepsilon_j)^2] \int g^2\left(\frac{\mathbf{z} - \mathbf{z}_j}{h}\right) 1_j(\mathbf{z}) d\mathbf{z}\right\}, \end{aligned} \quad (12.3.30)$$

where $E_{\boldsymbol{\varepsilon}}$ indicates expectation over $\boldsymbol{\varepsilon}$, while the second E is over $\mathbf{X}_1, \dots, \mathbf{X}_n$. But,

$$\int g^2\left(\frac{\mathbf{z} - \mathbf{z}_j}{h}\right) 1_j(\mathbf{z}) d\mathbf{z} = \left(\int g^2(\mathbf{z}) d\mathbf{z}\right) h^d, \quad (12.3.31)$$

and

$$\begin{aligned} &E_{\boldsymbol{\varepsilon}}(P[\varepsilon_j = 1|\mathbf{X}_1, \dots, \mathbf{X}_n] - \varepsilon_j)^2 \\ &= \frac{1}{2}(1 - P[\varepsilon_j = 1|\mathbf{X}_1, \dots, \mathbf{X}_n])^2 + \frac{1}{2}P^2[\varepsilon_j = 1|\mathbf{X}_1, \dots, \mathbf{X}_n] \\ &= \frac{1}{2} - P[\varepsilon_j = 1|\mathbf{X}_1, \dots, \mathbf{X}_n](1 - P[\varepsilon_j = 1|\mathbf{X}_1, \dots, \mathbf{X}_n]) \\ &\geq \frac{1}{2} - \frac{1}{4} = \frac{1}{4}. \end{aligned} \quad (12.3.32)$$

Put $h = n^{-\frac{1}{d+2}}$ and combine (12.3.30)–(12.3.32) to get that, for some universal C ,

$$R_\pi \geq \frac{h^2}{4} M_0^2 \int_{I^d} g^2(\mathbf{z}) d\mathbf{z} = C n^{-\frac{2}{2+d}}$$

since $Nh^d = 1$.

Thus the rate of convergence to zero of the lower bound on risk is no faster than the rate of convergence of $\widehat{\mu}_{\text{NW}}(\cdot)$, and we have established the asymptotic rate optimality of the Nadaraya-Watson estimate for IMSE. \square

As we have seen in our examples establishing the required lower bound on Bayes risks even with plausible least favorable priors can be subtle. For instance, computing the natural bound on integrated absolute error is not easy using the method we developed for IMSE in the second part of Example 12.3.2.

The multiple decision approach to lower bounds

There are a number of general purpose lower bounds for the Bayes risk in problems with finite parameter spaces and corresponding finite multiple decision problems which generalize Lemma 12.3.1 and can be similarly applied. Chief among these are lemmas due to Fano (1952) and Assouad (1983). See also Tsybakov (2008), Theorem 2.4 for a more direct generalization of Lemma 12.3.1 to multiple decision procedures rather than just tests as we did in Corollary 12.3.1. We discuss Fano's Lemma and refer to Tsybakov (2008) for a general form of Assouad's and other lemmas.

Let $\mathcal{P} = \{P_\theta : \theta \in \{1, \dots, m\}\}$ be probability distributions on some \mathcal{X} and consider the decision problem of deciding, given an observation X , which is the true P_j with 0-1 loss. Identify P_j with j . A possible rule $\delta : \mathcal{X} \rightarrow \{1, \dots, m\}$ has risk

$$R(j, \delta) = P_j[\delta(X) \neq j], \quad 1 \leq j \leq m. \quad (12.3.33)$$

We introduce two quantities familiar in information theory; see Cover and Thomas (1991), Chapter 2. For (X, Y) , a pair of random variables, with X taking on m values $\{1, \dots, m\}$ only, define

$$H(X) = - \sum_{j=1}^m P[X = j] \log P[X = j],$$

the *entropy* of X . It is immediate that

$$H(X) = \log m - K(P_X, U), \quad (12.3.34)$$

where P_X is the distribution of X , U is the uniform distribution on $\{1, \dots, m\}$, and K is the Kulback-Leibler divergence defined in (2.2.24). Next define the *mutual information* of (X, Y) as

$$\begin{aligned} I(Y, X) &= I(X, Y) \equiv K(P_{(X,Y)}, P_X P_Y) \\ &= \int \int p_X(x) p_{Y|X}(y|x) \log \left(\frac{p(x,y)}{p_X(x)p_Y(y)} \right) dx dy, \end{aligned} \quad (12.3.35)$$

the Kullback–Leibler divergence between the joint distribution of (X, Y) and the distribution with independent coordinates having p_X, p_Y , as marginals. Identifying Θ with $Y \sim U$, $\mathcal{L}(X|\Theta = j)$ with P_j as above, Fano’s inequality states

Theorem (Fano). *For any δ as in (12.3.33), $m \geq 2$,*

$$\max_j R(j, \delta) \geq \frac{1}{m} \sum_{j=1}^m R(j, \delta) \geq 1 - \frac{I(\Theta, X)}{\log m} \quad (12.3.36)$$

$$\geq 1 - \max_{k,l} \left[\frac{K(P_k, P_l) + \log 2}{\log m} \right] \frac{(m-1)}{m}. \quad (12.3.37)$$

The interpretation of (12.3.36) is natural in information theory; see Cover and Thomas (1991) for a discussion and proof of (12.3.36). The modification (12.3.37) is not the best possible — see Le Cam (1986) for a better but somewhat harder to prove version.

To prove (12.3.37) we argue as follows. If $Y \sim U$ and X takes values in $\{1, \dots, m\}$,

$$\begin{aligned} \log \left[\frac{p(x, y)}{p_X(x)p_Y(y)} \right] &= \log [p(x|y)(m^{-1} \sum_{k=1}^m p(x|k))^{-1}] \\ &= \log p(x|y) - \log \left[m^{-1} \sum_{k=1}^m p(x|k) \right] \\ &\leq \log p(x|y) - \frac{1}{m} \sum_{k=1}^m \log p(x|k) = \frac{1}{m} \sum_{k=1}^m \log \frac{p(x|y)}{p(x|k)}, \end{aligned} \quad (12.3.38)$$

since $-\log z$ is convex in z . Substituting (12.3.38) back in (12.3.35) we get

$$I(\Theta, X) \leq \frac{1}{m} \sum_{j=1}^m \int \left(p(x|j) \frac{1}{m} \sum_{k=1}^m \log \frac{p(x|j)}{p(x|k)} \right) dx \leq \frac{m-1}{m} \max_{j,k} K(P_j, P_k)$$

and (12.3.37) follows. \square

Fano’s inequality can be utilized as a generalization of the testing argument of Corollary 12.3.1. We look for m points P_1, \dots, P_m in \mathcal{P} such that $\rho(\theta(P_k), \theta(P_l)) \geq r_n$ for all $k \neq l$, and $\cup_{k=1}^m S(\theta(P_k), \frac{r_n}{2}) = \mathcal{P}$, where $S(\theta, r)$ is the ρ sphere of radius r around θ , and define the rule $\hat{\delta}(x) = j$ iff $\rho(\hat{\theta}, \theta(P_j)) < r_n/2$. Note that j is uniquely defined (Problem 12.3.8). Then, by Markov’s inequality, we have

$$E_P \rho(\hat{\theta}, \theta(P_0)) \geq r_n P[\rho(\hat{\theta}, \theta(P_0)) \geq r_n] \geq r_n \left(1 - \frac{\max_{l \neq k} K(P_l, P_k)}{\log m} \right).$$

Therefore

$$\max_{\mathcal{P}} E_P \rho(\hat{\theta}, \theta(P)) \geq r_n \delta \quad (12.3.39)$$

for some $\delta > 0$, all n , provided that $\max_{l \neq k} K(P_l, P_k)/\log m$ is bounded above by $1 - \delta$.

We apply this method to Example 12.3.1 with integrated squared error loss. It is natural to consider the $m = 2^N$ functions

$$\mu_{\varepsilon}(\mathbf{z}) = \sum_{j=1}^N \varepsilon_j h g\left(\frac{\mathbf{z} - \mathbf{z}_j}{h}\right)$$

as ε varies. Any two such functions $\mu_{\varepsilon_1}, \mu_{\varepsilon_2}$ differ on at least one of the C_j and, by the calculation made for part (i) of the example, if $\rho^2(\mu_1, \mu_2) = \int_{I^d} (\mu_1(\mathbf{z}) - \mu_2(\mathbf{z}))^2 d\mathbf{z}$, then

$$\rho^2(\mu_{\varepsilon_1}, \mu_{\varepsilon_2}) \geq h^2 \int_{I^d} g^2(\mathbf{z}) d\mathbf{z}. \quad (12.3.40)$$

On the other hand, the maximum divergence corresponds to the functions

$$\mu_2(\mathbf{z}) = \sum_{j=1}^N h g\left(\frac{\mathbf{z} - \mathbf{z}_j}{h}\right) 1(\mathbf{z} \in C_j), \quad \mu_1(\mathbf{z}) \equiv 0.$$

If P_{1n}, P_{2n} are two product probabilities for which (\mathbf{Z}_i, Y_i) , $1 \leq i \leq n$ are i.i.d., then

$$K(P_{1n}, P_{2n}) = n K(P_{11}, P_{21}), \quad (12.3.41)$$

and if we let $W \sim \mathcal{N}(0, 1)$, then

$$\begin{aligned} K(P_{11}, P_{21}) &= -E_1 \log \left\{ \frac{\varphi(Y_1 - \mu_2(\mathbf{Z}))}{\varphi(Y_1)} \right\} = E(\mu_2(\mathbf{Z}W)) + \frac{1}{2} E\mu_2^2(\mathbf{Z}) \\ &= \frac{1}{2} h^2 \int_{I^d} g^2(\mathbf{z}) d\mathbf{z}. \end{aligned} \quad (12.3.42)$$

If we put $r_n = n^{-\frac{1}{2+d}}$, $h = r_n$ so that $N = n^{\frac{d}{2+d}}$, we see from (12.3.40)–(12.3.42) that (12.3.39) with $\delta = \frac{1}{2}$ yields

$$\max_{\mathcal{P}} E_P (\hat{\mu}(\mathbf{Z}) - \mu(\mathbf{Z}))^2 \geq \frac{r_n^2}{2}$$

in accordance with our results in part (ii) of Example 12.3.1.

Remark 12.3.2. (a) This method based on Fano's inequality permits us to consider other loss functions ρ such as $\rho(\mu_1, \mu_2) = \int_{I^d} |\mu_1(\mathbf{z}) - \mu_2(\mathbf{z})| d\mathbf{z}$.

(b) We can also consider other \mathcal{P} than the ones corresponding to Lipschitz μ . For instance, suppose \mathcal{P} corresponds to $\{\mu : |D^s \mu| \leq M < \infty\}$ where D^s is the collection of s th partial derivatives of μ , $s \geq 1$. Then, if we use integrated squared error, we can show (Problem 12.3.9) that the minmax risk of the Nadaraya-Watson estimate is no smaller in order than $n^{-\frac{2s}{2s+d}}$. It is also fairly easy to show (Problem 12.3.10) that this rate can be achieved. This type of result was first obtained by Stone (1982).

12.3.3 The Gaussian White Noise (GWN) Model

Donoho and Johnstone (1994) and Kerkyacharian and Picard (1995) initiated the study of behaviour we expect in nonparametric formulations generally.

Let $\mu \equiv (\mu_1, \dots, \mu_p, \dots)^T$ be an infinite dimensional vector lying in $l_2 \equiv \{\mu : \sum_j \mu_j^2 < \infty\}$. Suppose $\sigma^2 = \sigma_0^2$ is known and that we observe $\mathbf{Y}_1, \dots, \mathbf{Y}_n$ i.i.d. as \mathbf{Y} with

$$\mathbf{Y} = \mu + \varepsilon,$$

where $\mu \in l_2$ and ε is a vector of independent $\mathcal{N}(0, \sigma_0^2)$ variables. Note that $\mathbf{Y} \notin l_2$ since $\sum \varepsilon_i^2 = \infty$ with probability 1. By sufficiency of \bar{Y} this model reduces to the model

$$\bar{\mathbf{Y}} = \mu + \bar{\varepsilon}, \quad (12.3.43)$$

where $\bar{\varepsilon}$ is a vector of i.i.d. $\mathcal{N}(0, \sigma_0^2/n)$ variables. This, (12.3.43), is the *Gaussian white noise model* (GWN).

Note that if the vector were p rather than ∞ dimensional this would be a known variance linear model. Just as we did in Remark 9.5.4, we will argue that, at least formally, the GWN is the limit of nonparametric models in analyzing decision procedures just as the canonical linear model is the limit of regular parametric models. As we shall see doing so not only suggests convergence rates and the difficulty of problems, but also procedures. We illustrate by considering again

Example 12.3.1 (Continued). Nonparametric regression. Using the same notation as in Example 11.3.2, suppose that we estimate by orthogonal series. For instance, consider first one dimensional Fourier series on $I \equiv [0, 1]$. That is,

$$\mu(x) = \sum_{k=1}^{\infty} \theta_k \varphi_k(x)$$

where

$$\varphi_1(x) \equiv 1, \varphi_{2k}(x) = \sqrt{2} \cos(2\pi kx), \varphi_{2k+1}(x) = \sqrt{2} \sin(2\pi kx),$$

is the Fourier basis. Note that this basis is orthonormal in $L_2(0, 1)$ because

$$\varphi_{2k}(x) + i\varphi_{2k+1}(x) = \sqrt{2}e^{2\pi ikx}.$$

We can promote these functions by tensor products, to orthogonal series for $\mu(\cdot)$ on I^d . We consider

$$\mathcal{H} = \{h : h(\mathbf{x}) = \prod_{j=1}^d \varphi_{i_j}(x_j) : 1 \leq i_j < \infty\}.$$

Then, if $\mu \in L_2(I^d)$, we can write, for $h_l, h_k \in \mathcal{H}$,

$$\mu(\mathbf{Z}) = \sum_{l=1}^{\infty} \theta_l h_l(\mathbf{Z}), \quad \theta_k(P) = \int_{I^d} \mu(\mathbf{z}) h_k(\mathbf{z}) d\mathbf{z}. \quad (12.3.44)$$

Consider the sieve of approximating regression models,

$$Y_i = \sum_{k=1}^p \theta_k h_k(\mathbf{Z}_i) + \varepsilon_i \quad i = 1, \dots, n$$

where the ε_i are i.i.d. $\mathcal{N}(0, \sigma_0^2)$. By Theorem 6.1.4, the MLE of $\boldsymbol{\theta}^{(p)} \equiv (\theta_1(P), \dots, \theta_p(P))^T$ is

$$\hat{\boldsymbol{\theta}}^{(p)} = G_n^{-1}(\hat{\boldsymbol{\theta}}_1^*, \dots, \hat{\boldsymbol{\theta}}_p^*)^T,$$

where

$$G_n \equiv \left\| \frac{1}{n} \sum_{i=1}^n h_j(\mathbf{Z}_i) h_k(\mathbf{Z}_i) \right\|_{d \times d}$$

and

$$\hat{\boldsymbol{\theta}}_j^* = \frac{1}{n} \sum_{i=1}^n Y_i h_j(\mathbf{Z}_i), \quad 1 \leq j \leq p.$$

Given $\mathbf{Z}_1, \dots, \mathbf{Z}_n$, by Section 6.1,

$$\hat{\boldsymbol{\theta}}^{(p)} \sim \mathcal{N}_p(\boldsymbol{\theta}^{(p)}, \sigma_0^2 G_n^{-1}/n). \quad (12.3.45)$$

As $n \rightarrow \infty$, for fixed p ,

$$\sqrt{n}(\hat{\boldsymbol{\theta}}^{(p)} - \boldsymbol{\theta}^{(p)}) \implies \mathcal{N}(0, \sigma_0^2 J_p)$$

since $G_n \xrightarrow{P} J_p$ by our choice of h_1, \dots, h_p, \dots . If we now let $p \rightarrow \infty$ we see that we have arrived at the Gaussian white noise model (12.3.43). Then, estimation of $\mu(\cdot)$ and estimation of $\boldsymbol{\theta}(P) = (\theta_1(P), \dots, \theta_p(P), \dots) \in l_2$ are formally equivalent, and for IMSE and squared error on $\boldsymbol{\theta}$ rigorously so, since

$$\int_{I^d} (\hat{\mu}(\mathbf{z}) - \mu(\mathbf{z}))^2 d(\mathbf{z}) = \sum_j (\hat{\theta}_j - \theta_j)^2$$

by the Parseval identity. That suggests that if we can carry over descriptions in $\mu(\cdot)$ space into ones on $\boldsymbol{\theta}$ in l_2 we can gain some insight into the behaviour of procedures, even as we did from the correspondence between ‘‘limits’’ of finite dimensional parametric models and the canonical Gaussian linear model in Chapter 6. We will only proceed formally but want to note that the rigorous theory of asymptotic approximation of experiments, developed by Le Cam, which fully justifies the equivalence of results such as Wilks’ theorem in the parametric case has been developed by Brown et al, Nussbaum, and others in an important series of papers (1996–2004) to apply to regression, density estimation, and other nonparametric methods. \square

12.3.4 Minimax Bounds on IMSE for Subsets of the GWN Model

In Example 12.3.1 (continued) we restrict θ to subsets of l_2 of the form,

$$\Theta_\beta = \left\{ \theta : \sum_{k=1}^{\infty} (k^\beta \theta_k)^2 \leq M \right\}, \quad \beta > 0. \quad (12.3.46)$$

These sets are, if θ_k are Fourier coefficients, sets that specify some degrees of smoothness on $\mu(\cdot)$. For instance, Θ_1 corresponds to the set, $\{\mu(\cdot) : \mu(\cdot) \text{ is absolutely continuous, and } \int_0^1 \mu'(s)^2 ds \leq M\}$. To see this we can use Parseval's identity since the Fourier coefficients of μ' are (Problem 12.3.5)

$$\tilde{\theta}_1 = 0, \quad \tilde{\theta}_{2k} = 2\pi k, \quad \tilde{\theta}_{2k+1} = -2\pi k. \quad (12.3.47)$$

We consider estimation of θ by $\hat{\theta} \in R^\infty$ with loss

$$l(\hat{\theta}, \theta) \equiv \sum_{j=1}^{\infty} (\hat{\theta}_j - \theta_j)^2 \quad (12.3.48)$$

and return to the problem of setting lower bounds in the GWN model with IMSE risk. We use a direct approach. Note that because $\Sigma k^{-\alpha}$ converges iff $\alpha > 1$,

$$\Theta_\beta \supset \tilde{\Theta}_\lambda \equiv \left\{ \theta : |\theta_k| \leq M_\lambda k^{-\lambda}, \lambda > \beta + \frac{1}{2}, k \geq 1 \right\}$$

for suitable M_λ .

We restrict, without loss of generality, to priors making the θ_k independent. See Problem 12.3.6. It turns out that such priors are asymptotically least favorable. In that case, it seems reasonable to consider $\theta_k \sim F_k$ where F_k is least favorable for the univariate model $\{\theta_k : |\theta_k| \leq M k^{-\lambda}\}$. It has been shown (see Casella and Strawderman (1981), Bickel (1981), and Levit (1981)), that the least favorable priors, π_δ , for estimating μ , given $X \sim \mathcal{N}(\mu, 1)$, $|\mu| \leq m > 0$, have finite support and as $m \downarrow 0$ are symmetric on points $\{-c, 0, c\}$ and in the limit are point mass at 0. That is, there exists $m_1 > 0$ such that the least favorable prior is of the above form for $m \leq m_1$. Replacing Y_i by $\sqrt{n}Y_i \sim \mathcal{N}(\mu_i, 1)$ with $\mu_i \equiv n^{\frac{1}{2}}\theta_i$ we see that $\tilde{\Theta}_\lambda$ corresponds to $\{\mu : |\mu_k| \leq M_\lambda k^{-\lambda} n^{\frac{1}{2}}\}$ so that for $k \geq [M_\lambda n^{\frac{1}{2}} m_o]^{\frac{1}{\lambda}} \equiv k_0$ the least favorable prior is pointmass at 0.

In general, let $\delta_{\{t\}}$ denote point mass at t , consider the prior

$$\pi_\delta = \varepsilon \delta_{\{0\}} + (1 - \varepsilon) \left[\frac{\delta_{\{-c\}} + \delta_{\{c\}}}{2} \right].$$

The Bayes estimate is (Problem 12.3.12)

$$\begin{aligned}\delta_m(t) &= (1 - \varepsilon)m \frac{[\varphi((t - m)\sqrt{n}) - \varphi((t + m)\sqrt{n})]}{\varphi(t - m\sqrt{n}) + \varphi((t + m)\sqrt{n})} \\ &= (1 - \varepsilon)m \frac{(1 - e^{-2m\sqrt{n}t})}{(1 + e^{-2m\sqrt{n}t})}\end{aligned}\tag{12.3.49}$$

and its Bayes risk is

$$\begin{aligned}R_m &\equiv (1 - \varepsilon)E_m(\delta_m(Y) - m)^2 + \varepsilon m^2 \\ &= m^2[(1 - \varepsilon)E(e^{-4m\sqrt{n}Y}(1 + e^{-2m\sqrt{n}Y})^{-2}) + \varepsilon]\end{aligned}\tag{12.3.50}$$

Put $m = n^{-\frac{1}{2}}$, $\varepsilon = 0$, and $Y = Z + m$ where $Z \sim \mathcal{N}(0, 1)$ and get

$$R_m = \frac{4}{n}[Ee^{-4mZ}(1 + e^{-2mZ})^{-2}](1 + o(1)) = \frac{4}{n}(1 + o(1)).$$

It is easy to see that putting $\theta_k \sim \pi_m$, $m = n^{-\frac{1}{2}}$, $k < k_0$, and $m = 0$, $k \geq k_0$, yields a Bayes risk for the loss (12.3.48) that is bounded by

$$\frac{4}{n}(k_0 - 1) + \sum_{k=k_0}^{\infty} \theta_k^2 = \frac{4k_0}{n} + M_\lambda^2 n \sum_{k=k_0}^{\infty} k^{-2\lambda} \asymp n^{-(1-\frac{1}{2\lambda})}.$$

Since this holds for all $\lambda > \beta + \frac{1}{2}$ we expect that $n^{-2\beta(1+2\beta)^{-1}}$ is a correct minmax order bound for Θ_β . In particular, $\beta = 1$ give us the familiar Lipschitz $n^{-\frac{2}{3}}$ rate. This is, in fact, established in Theorem 2.7 of Tsybakov (2008). The Bayes estimate achieves minimaxity in terms of rate but the simpler estimate,

$$\begin{aligned}\hat{\theta}_k &= Y_k, & 1 \leq k \leq k_0 \\ &= 0, & k > k_0,\end{aligned}\tag{12.3.51}$$

does as well. One can, in this and related cases, do better and determine the actual value of the minmax bound asymptotically by using *Pinsker's theorem* (1980). See also Theorem 3.1 in Tsybakov (2008). By taking $\hat{\theta}_k$ of the form $(1 - \alpha_k)Y_k \mathbf{1}(k \leq k_0)$, it is possible to obtain the bound asymptotically.

12.3.5 Sparse Submodels

The GWN also accommodates subsets which are “sparse” but where the small $\{\theta_k\}$ can occur anywhere, and not just for k large. These types of models arise quite generally. An example arises with covariance matrices of vectors $(X_1, \dots, X_d)^T$ which represent gene expression in a microarray. Then we expect that most genes are independent corresponding to 0’s in the off diagonal part of the covariance matrix, but, if the genes are dependent, we

expect the dependence is strong. Donoho and Johnstone (1994) suggested a class of GWN subsets exhibiting this type of sparsity

$$\Theta_{q,r} \equiv \{\boldsymbol{\theta} : \theta_k = 0, k > r, \sum_{k=1}^r |\theta_k|^q \leq M\} \quad (12.3.52)$$

for $0 \leq q < 1$. The case $q = 0$ where the sum is simply counting nonzero elements in $\{\theta_k : 1 \leq k \leq r\}$ is the prototype. The minmax risk of squared error estimation for such models is discussed extensively in Donoho and Johnstone (1994), Kerkyacharian and Picard (1995), and more recently in Abramovich, Benjamini, Donoho and Johnstone (2006) and Johnstone and Silverman (2005). A key emphasis is on adaptive procedures, procedures defined here as achieving uniform minimax rates. We will only derive minmax rates for $\Theta_{0,n}$ and exhibit procedures which achieve them.

The principle that we should consider priors with independent components is still legitimate, as is the principle that least favorable priors asymptotically concentrate on $\{-c, 0, c\}$. However, we do not know which are likely to be the 0 coordinates other than $\theta_j, j > r$. On the other hand there is no reason to behave differently for different nonzero coordinates.

Fano's inequality suggests considering the points $\boldsymbol{\theta}_S$ where S ranges over all subsets of size M of $\{1, \dots, r\}$ given by $\theta_j = 0, j \notin S, \theta_j = \pm c, j \in S$. The l_2 distance between two such points is $\geq c$. The Kullback-Leibler divergence between the probability distribution corresponding to two such points is

$$K(P_{\theta_1}, P_{\theta_2}) = n \frac{c^2}{2} \sum_{i=1}^r 1(\theta_{1j} \neq \theta_{2j}),$$

so that

$$\max K(P_{\theta_1}, P_{\theta_2}) = n M c^2.$$

There are $m = 2^M \binom{r}{M}$ such sets. Since $\binom{r}{M} \left(\frac{M}{r}\right)^M \left(1 - \frac{M}{r}\right)^{r-M} \leq 1$, we obtain from (12.3.39), for $M/r \rightarrow 0$, the bound

$$\min_{\widehat{\boldsymbol{\theta}}} \max_{\boldsymbol{\theta} \in \Theta_{0,r}} E_{\boldsymbol{\theta}} l(\boldsymbol{\theta}, \widehat{\boldsymbol{\theta}}) \geq c^2$$

if

$$\frac{n M c^2}{M \log \frac{r}{M}} \geq \delta.$$

For $r = n$ and M bounded we obtain the bound $M \delta n^{-1} \log(n/M)$ for c^2 . We have shown that

Proposition 12.3.4. *For squared error loss, the asymptotic minimax rate for $\Theta_{0,n}$ is $n^{-1} \log n$.*

The natural way to achieve this rate is by *hard* or *soft thresholding* (Problem 12.3.11). Here

$$\delta_c^h(y) \equiv y 1(|y| \geq c)$$

and

$$\delta_c^S(y) = (y - c \operatorname{sgn} y)1(|y| \geq c)$$

are the *hard* and *soft thresholded* estimates, respectively.

Essentially for hard thresholding, we test $H : \theta_j = 0$ for each j and use a critical value of the form $K(n^{-1}/\log n)^{\frac{1}{2}}$. If we accept, we take $\hat{\theta}_j = 0$. Else, we leave

$$\hat{\theta}_j = Y_j .$$

On theoretical grounds, there is no reason to prefer one type of thresholding over the other.

Summary. In Section 12.3.1 we go more fully into asymptotic theory for classification and regression, introducing optimal rates of convergence in various models, in particular, regular parametric ones. In Section 12.3.2 we introduce asymptotic minimax lower bounds and develop ways of obtaining these explicitly, in particular via testing, Theorem 12.3.2, and Fano’s inequality, and apply these to obtain minimax rates for nonparametric regression. We show the asymptotic rate optimality of the Nadaraya-Watson estimate under a first order Lipschitz condition. In Section 12.3.3 we introduce the Gaussian white noise model (GWN), which enables us to put nonparametric problems in a canonical context just as we did for parametric ones in Chapter 6. Section 12.3.4 discusses minimax rate bounds on IMSE for subsets of the GWN model and Section 12.3.5 establishes such a bound for a sparse submodel of the GWN model.

12.4 Oracle Inequalities

As we mentioned in Section I.1 oracle inequalities are another approach to measuring the performance of prediction methods. In the minimax approach we consider, for a model \mathcal{P} , all possible prediction rules from the point of view of their worst behaviour on \mathcal{P} . The oracle inequality point of view is instead to consider a limited class of rules \mathcal{D} . We compare the statistician’s performance to what an “oracle” who knows P but is restricted to using $\delta \in \mathcal{D}$ could achieve, *at each* $P \in \mathcal{P}$. Abstractly, suppose $\mathcal{D} = \{\delta_t\}, t \in \mathcal{T}$, and we define,

$$R_t(P) = R(P, \delta_t) = E_P \ell(P, \delta_t) ,$$

where ℓ is the loss we consider. An oracle then picks $t^*(P)$ such that,

$$R^*(P) \equiv R_{t^*(P)}(P) = \min_t R_t(P) .$$

Suppose the statistician has a data determined rule for choosing t , call it \hat{t} . We would like

$$R(P, \delta_{\hat{t}}) \approx R^*(P) . \tag{12.4.1}$$

An oracle inequality makes (12.4.1) precise through a statement such as

$$R(P, \delta_{\hat{t}}) \leq R^*(P) + \Delta(P) , \tag{12.4.2}$$

where $\Delta(P)$ can be explicitly bounded independent of P .

Example 12.4.1. As in Section I.4, we consider $\mathbf{X} = (\mathbf{Z}, Y)$ where \mathbf{Z} is a vector of predictors and Y the response variable to be predicted. Let $\lambda(y, d)$ denote the loss when d is the predicted value of y . Given a predictor $\delta : Z \rightarrow Y$ we define the prediction loss of δ as the expected loss under P ,

$$\ell(P, \delta) = \int \lambda(y, \delta(\mathbf{z})) dP(\mathbf{x}). \quad (12.4.3)$$

With this formulation we can also think of the oracle as choosing

$$t^* = \arg \min_t \ell(P, \delta_t).$$

Suppose we have a training sample $\mathbf{X} = X_1, \dots, X_n = (\mathbf{Z}_1, Y_1), \dots, (\mathbf{Z}_n, Y_n)$ from P and use the \hat{t} that minimizes a criterion for how closely $\delta_t(Z_i)$ predicts Y_i , $1 \leq i \leq n$. In this case we want an inequality of the form,

$$P[\ell(P, \hat{\delta}_{\hat{t}}) \leq \ell(P, \delta_{t^*}) + \Delta(P, \hat{P}, \varepsilon)] \geq 1 - \varepsilon \quad (12.4.4)$$

where \hat{P} is the empirical probability and $\Delta(P, \hat{P}, \varepsilon)$ can be bounded independent of P . An inequality of the type (12.4.4) implies one of type (12.4.2) under integrability conditions if $\Delta(P, \hat{P}, \varepsilon)$ is boundable independent of ε . \square

Oracle inequalities have three important features.

- (i) They provide information at each P
- (ii) They are “nonasymptotic” since the bounds $\Delta(\cdot)$ are more or less explicit.
- (iii) They exhibit explicit important features of the problem such as the role of dimension on the performance of δ_t .

The first two features are limited, however, since the first requires one to know how well the best of class \mathcal{D} can do against a particular P (the oracle’s performance) and the second can be thought of as merely making asymptotic calculations carefully since the bounds are only of interest if the bound on Δ tends to 0 as sample size n tends to infinity. Also, the bound when made independent of P corresponds to worst case behaviour, and is very conservative in practice.

How do we choose \hat{t} ? At first sight we might try

$$\hat{t} = \arg \min_t \ell(\hat{P}, \delta_t), \quad (12.4.5)$$

Unfortunately, as we have seen in Section I.7, this will not work unless \mathcal{P} is regular parametric. We are typically considering $\mathcal{P} = \cup_t \mathcal{P}_t$ where \mathcal{P}_t is a sequence of nested parametric models with δ_t based on a minimum contrast estimate of the parameters in \mathcal{P}_t . Then, (12.4.5) inevitably leads to choosing the largest model in \mathcal{P} , i.e. “overfitting”.

As we have noted in Sections I.7, 9.1 and Chapter 11, we need to regularize, typically using a penalty $\text{pen}(t)$ which can also depend on the data. That is we choose \hat{t} as,

$$\hat{t} = \arg \min_t \ell(\hat{P}, \delta_t) + \text{pen}(t) ,$$

and $\Delta(P)$ as defined in (12.4.2) reflects the penalty.

We will follow Tsybakov (2008) in first clarifying these notions in the context of the GWN model.

Example 12.4.2. *Nested linear models.* Consider data $(\mathbf{X}_1, Y_1)^T, \dots, (\mathbf{X}_n, Y_n)^T$ i.i.d. as (\mathbf{X}, Y) where \mathbf{X} is an N vector and Y is scalar, and consider the model

$$Y_i = \mathbf{X}_i \boldsymbol{\beta}_i^{(j)} + \varepsilon_i \text{ for some } 1 \leq j \leq N, \quad i = 1, \dots, n ,$$

where $\{\varepsilon_i\}$ are i.i.d. $\mathcal{N}(0, \sigma_0^2)$, and $\boldsymbol{\beta}_i^{(j)} = (\beta_1, \dots, \beta_j, 0, \dots, 0)^T_{N \times 1}$. This is a sequence of nested regression models, \mathcal{P}_j . As we did in Section 6.1.1 we can by an orthogonal transformation $A(\mathbf{X}_1, \dots, \mathbf{X}_n)$ transform the model to one involving its sufficient statistics only given by

$$\hat{\eta}_k = \eta_k + \frac{\varepsilon'_k}{\sqrt{n}}, \quad 1 \leq k \leq N ,$$

where the model \mathcal{P}_j now corresponds to $\eta_k = 0$, $k > j$, and $\{\varepsilon'_k\}$ are i.i.d. $\mathcal{N}(0, \sigma_0^2)$, $1 \leq k \leq N$. This is just a GWN model with coordinates equal to zero after the N th entry.

12.4.1 Stein's Unbiased Risk Estimate

A key tool for studying and constructing estimates in the GWN model is Stein's unbiased risk estimate (Stein (1981)). We follow our notation in Section 12.3.3 identifying $\hat{\eta}_i$ with Y_i and η_i with μ_i , where $Y_i = \mu_i + \varepsilon_i$, with $\varepsilon_1, \dots, \varepsilon_n$ independent $\mathcal{N}(0, \sigma_0^2)$. Let $\delta(Y_1, \dots, Y_n)$ be an estimate of $\boldsymbol{\mu} \equiv (\mu_1, \dots, \mu_N)^T$. Using (8.3.33) we find that if δ satisfies the conditions of Theorem 8.3 and

$$\boldsymbol{\Delta}(\mathbf{y}) \equiv \delta(\mathbf{y}) - \mathbf{y} ,$$

then,

$$E_{\boldsymbol{\mu}} |\delta(\mathbf{Y}) - \boldsymbol{\mu}|^2 = \frac{N\sigma_0^2}{n} + 2\frac{\sigma_0^2}{n} E_{\boldsymbol{\mu}} \left(\sum_{j=1}^N \frac{\partial \Delta_j}{\partial y_j}(\mathbf{Y}) \right) + E_{\boldsymbol{\mu}} |\boldsymbol{\Delta}(\mathbf{Y})|^2 . \quad (12.4.6)$$

If we define

$$\sigma_n \equiv \sigma_0 / \sqrt{n}$$

we arrive at *Stein's unbiased risk estimate* of $R(\boldsymbol{\mu}, \delta) - N\sigma_0^2/n$,

$$\hat{R}_U(\delta) = 2\sigma_n^2 \sum_{j=1}^N \frac{\partial \Delta_j}{\partial \mu_j}(\mathbf{Y}) + |\boldsymbol{\Delta}(\mathbf{Y})|^2 . \quad (12.4.7)$$

Following our general prescription, given a family $\{\delta_t\}$, $t \in T$, of estimates of $\boldsymbol{\eta}$ with $\delta_t \equiv (\delta_{t1}, \dots, \delta_{tN}, \dots)^T$, such that $\delta_{tj} = 0$, $j > N$, it seems reasonable to pick \hat{t} by minimizing $\widehat{R}_U(\delta_t)$. Consider the classes of estimates

$$\begin{aligned}\mathcal{D}_0 &\equiv \{\delta_\lambda(\hat{\boldsymbol{\eta}}) = (\lambda\hat{\eta}_1, \dots, \lambda\hat{\eta}_N)^T, 0 \leq \lambda \leq 1\} \\ \mathcal{D}_1 &\equiv \{\delta_m(\hat{\boldsymbol{\eta}}) = (\hat{\eta}_1, \dots, \hat{\eta}_m, 0, \dots, 0, \dots)^T, 1 \leq m \leq N\}\end{aligned}$$

indexed by $t = \lambda$ and $t = m$, respectively. The members of \mathcal{D}_0 are *Stein-type shrinkage estimators*. They “shrink” $\hat{\eta}_j = Y_j$ towards zero when $0 \leq \lambda < 1$, uniformly for all coordinates. The members of \mathcal{D}_1 are truncated estimators.

12.4.2 Oracle Inequality for Shrinkage Estimators

We first compute the MSE of $\delta \in \mathcal{D}_0$ for $\boldsymbol{\eta} \in \mathcal{P}_N$ where \mathcal{P}_N is the GWN model with mean vector $(\eta_1, \dots, \eta_N, 0, \dots)$. This MSE is evidently

$$r_{N,n}(\lambda) \equiv \sum_{j=1}^N E(\lambda\hat{\eta}_j - \eta_j)^2 = N\lambda^2 \frac{\sigma_0^2}{n} + (1-\lambda)^2 \sum_{j=1}^N \eta_j^2. \quad (12.4.8)$$

Using calculus, for a given $\boldsymbol{\eta}$ the oracle risk is obtained for

$$\lambda(\boldsymbol{\eta}) = \frac{|\boldsymbol{\eta}|^2}{N\sigma_n^2 + |\boldsymbol{\eta}|^2},$$

where $\sigma_n^2 \equiv \sigma_0^2/n$. By substitution in (12.4.8), the oracle risk is given by

$$R_O(N, n, \boldsymbol{\eta}) = \frac{N\sigma_n^2 |\boldsymbol{\eta}|^2}{N\sigma_n^2 + |\boldsymbol{\eta}|^2}. \quad (12.4.9)$$

To consider this problem in a different form we note that every $\delta \in \mathcal{D}_0$ can for some γ be identified with the penalty estimator

$$\tilde{\delta}_\gamma \equiv \arg \min_{\boldsymbol{\eta}} \left\{ \sum_{i=1}^N (Y_i - \eta_i)^2 + \gamma \sum_{i=1}^N \eta_i^2 \right\},$$

which we recognize as the posterior mode if the η_i are considered as i.i.d. $\mathcal{N}(0, 1/\gamma)$, and also as ridge regression (Section 12.2.3) for the canonical Gaussian regression model of Section 6.1.1.

For $\delta \in \mathcal{D}_0$ and $\boldsymbol{\eta} \in \mathcal{P}_N$, Stein’s unbiased estimate of risk is (Problem 12.4.1)

$$\widehat{R}_O = 2N\sigma_n^2(1-\lambda) + (1-\lambda)^2 \sum_{j=1}^N \hat{\eta}_j^2, \quad (12.4.10)$$

which when minimized on $\lambda \in [0, 1]$ gives Stein's positive part type estimate,

$$\delta_{\hat{\lambda}}(\hat{\eta}) = \left(1 - \frac{N\sigma_n^2}{|\hat{\eta}|^2}\right)_+ \hat{\eta}. \quad (12.4.11)$$

What can we say about the risk of this estimate?

Theorem 12.4.1. For $N \geq 4$ and $\eta \in \mathcal{P}_N$, the risk of $\delta_{\hat{\lambda}}(\hat{\eta})$ for squared error loss satisfies

$$R(\eta, \delta_{\hat{\lambda}}) = E_{\eta} |\delta_{\hat{\lambda}}(\hat{\eta}) - \eta|^2 \leq R_O(N, n, \eta) + 4\sigma_n^2.$$

Proof. (After Tsybakov (2008)). By (8.3.39) if

$$\delta^*(\hat{\eta}) = \left(1 - \frac{N\sigma_n^2}{|\hat{\eta}|^2}\right) \hat{\eta}$$

is the Stein estimator, then

$$R(\eta, \delta_{\hat{\lambda}}) \leq R(\eta, \delta^*).$$

Now, by (8.3.31)

$$R(\eta, \delta^*) = N\sigma_n^2 - (N^2 - 4N)\sigma_n^4 E_{\eta} |\hat{\eta}|^{-2}.$$

But, by Jensen's inequality,

$$E|\hat{\eta}|^{-2} \geq (E|\hat{\eta}|^2)^{-1} = (|\eta|^2 + \sigma_n^2 N)^{-1}.$$

Substituting in, after some arithmetic (Problem 12.4.1),

$$R(\eta, \delta_{\hat{\lambda}}) \leq \frac{N\sigma_n^2 |\eta|^2}{|\eta|^2 + N\sigma_n^2} + \frac{4\sigma_n^4 N}{|\eta|^2 + \sigma_n^2 N} \leq R_O(N, n, \eta) + 4\sigma_n^2.$$

□

Note. The oracle inequality penalty term for not knowing η does not depend on η .

Remark 12.4.1.

- (a) If we know β and can choose N we can achieve the minimax rate of $n^{-\frac{2\beta}{2\beta+1}}$ over Θ_{β} . The oracle risk is achieved by taking $N = n^{\frac{1}{2\beta+1}}$, since $|\eta|$ is uniformly bounded over Θ_{β} (Problem 12.4.1).
- (b) In the absence of the possibility of varying N , an asymptotically minimax rule is not possible using \mathcal{D}_0 .
- (c) On the other hand, as we shall see in the next subsection, using \mathcal{D}_1 , the estimate selected using Stein's unbiased risk estimate is what we shall define as adaptively minimax over all $\beta > 0$.

12.4.3 Oracle Inequality and Adaptive Minimax Rate for Truncated Estimates

We first establish that the oracle minimax rate for $\boldsymbol{\eta} \in \Theta_\beta$ and \mathcal{D}_1 agrees with the general minimax rate of Section 12.3.4, and then show we can obtain this rate adaptively. Here, for $\delta_m \in \mathcal{D}_1$ and $\sigma_n^2 = \sigma_0^2/n$,

$$R(\boldsymbol{\eta}, \delta_m) = m\sigma_n^2 + \sum_{i=m+1}^{\infty} \eta_i^2. \quad (12.4.12)$$

Let

$$m_0(\boldsymbol{\eta}) = \arg \min_m R(\boldsymbol{\eta}, \delta_m).$$

Then, if $\boldsymbol{\eta} \in \Theta_\beta = \{\boldsymbol{\eta} : \sum_{j=1}^{\infty} (j^\beta \eta_j)^2 \leq M\}$, some $M > 0$,

$$R(\boldsymbol{\eta}, \delta_{m_0}(\boldsymbol{\eta})) \leq N\sigma_n^2 + N^{-2\beta}M,$$

since $\sum_{j=N+1}^{\infty} \eta_j^2 \leq N^{-2\beta} \sum_{j=N+1}^{\infty} (j^\beta \eta_j)^2$. Minimizing over N we obtain

$$N = \{2\beta M n / \sigma_0^2\}^{\frac{1}{2\beta+1}}$$

and the oracle minimax rate

$$R(\boldsymbol{\eta}, \delta_{m_0}(\beta)) \asymp n^{-\frac{2\beta}{2\beta+1}}. \quad (12.4.13)$$

We proceed to show using an oracle inequality that for procedures in \mathcal{D}_1 , if we select m using Stein's unbiased estimator of risk to obtain \hat{m} , then $\delta_{\hat{m}}$ has the correct minimax rate for all β . This is an example of *adaptive (asymptotic) minimaxity* which we next define in the context of i.i.d. sampling. Generally, for a statistical procedure $\hat{\delta}_n$ with risk R , $\hat{\delta}_n(X_1, \dots, X_n)$ is *asymptotically minimax* over a model \mathcal{P} iff

$$\sup \{R(P, \hat{\delta}_n) : P \in \mathcal{P}\} \asymp \inf_{\delta} \sup \{R(P, \delta) : P \in \mathcal{P}\}$$

as $n \rightarrow 0$ where δ ranges over all decision procedures. Now suppose $\mathcal{P}_1, \mathcal{P}_2, \dots$ is a sieve of models with $\cup_{m=1}^{\infty} \mathcal{P}_m = \mathcal{M}$ (with closure in law). Then $\hat{\delta}_n$ is *asymptotically adaptive minimax* over \mathcal{M} iff $\hat{\delta}_n$ is asymptotically minimax for each m . Note that $\hat{\delta}_n$ has no knowledge of \mathcal{M} .

Here is our procedure based on Stein's unbiased risk estimator. For rules in \mathcal{D}_1 $\boldsymbol{\Delta} \equiv (\Delta_1, \dots, \Delta_n)^T$ is given by $\Delta_j = 0$, $j \leq m$; $\Delta_j = -y_j$, $m+1 \leq j \leq n$. It follows that Stein's unbiased estimate of risk is

$$R(m) = -2\sigma_n^2(n-m) + \sum_{j=m+1}^n \hat{\eta}_j^2 = -(2\rho^2 n + \sum_{j=1}^n \hat{\eta}_j^2) + (2\rho^2 m - \sum_{j=1}^m \hat{\eta}_j^2). \quad (12.4.14)$$

The second term is just “Mallows’ C_p ” and since this is the only term that depends on m , minimizing Stein’s unbiased risk estimate leads to Mallows’ selector

$$\hat{m} = \arg \min \left\{ 2\sigma_n^2 m - \sum_{j=1}^m \hat{\eta}_j^2 : 1 \leq m \leq n \right\}.$$

Theorem 12.4.2. For all $\boldsymbol{\eta} \in \Theta_\beta$, all M , $0 < \beta < \frac{1}{2}$

$$\sup \left\{ E_{\boldsymbol{\eta}} |\hat{\boldsymbol{\eta}}_{\hat{m}} - \boldsymbol{\eta}|^2 : \boldsymbol{\eta} \in \Theta_\beta \right\} = O(n^{-\frac{2\beta}{2\beta+1}}). \quad (12.4.15)$$

The technical proof of this result is sketched in Problems 12.4.2–12.4.6.

Theorem 12.4.2 does establish rate optimality as $n \rightarrow \infty$, defined by

$$\overline{\lim}_n \sup_{\mathcal{P}} \frac{E_P \ell(P, \delta_t)}{\ell(P, \delta_{t^*(P)})} \leq K_0. \quad (12.4.16)$$

Thus, as $n \rightarrow \infty$, the statistician’s method, in this case Mallows’, achieves the same rate of convergence of the risk to 0 as does the oracle, uniformly on \mathcal{P} .

More difficult to achieve but evidently even more satisfactory is *asymptotic equivalence* defined by

$$\lim_{n \rightarrow \infty} \sup_{\mathcal{P}} E \frac{\ell(P, \delta_t)}{\ell(P, \delta_{t^*})} = 1, \quad (12.4.17)$$

uniformly on \mathcal{P} . More sophisticated methods are able to achieve (12.4.17) for a class of rules. See Tsybakov (2008), Theorem 3.6.

Remark 12.4.2.

1) **Other Models:** Although, as we have seen, the methods discussed apply to regression with Gaussian errors, the Gaussian white noise model can not be appealed to for density estimation or regression with non-Gaussian errors, not to speak of nonlinear semiparametric models such as logistic regression. However, deep results of Nussbaum (1996) and Brown and collaborators (1997, 2004) show that it is possible to generalize these conclusions to models which can be approximated in the strong sense of Le Cam’s distance between experiments and include the models cited.

2) **Bayesian Selection:** If we assume that $P \in \mathcal{P}_{m_0}$ for some m_0 and if m_0 is known we can estimate $\boldsymbol{\eta}$ efficiently using δ_{m_0} and obtain the optimal minimax rate of n^{-1} . This is possible without knowing m_0 if one uses Schwarz’s Bayes criterion (SBC) (Schwarz(1978)), also called Bayes information criterion (BIC). (See also Rissanen (1983) for the equivalent (in this case), Minimum Descriptive Length criterion.) For SBC, for some constant $c > 0$, we select (see Section 12.6.1)

$$\hat{m} = \arg \max \left\{ \sum_{j=1}^m \hat{\eta}_j^2 - 2c^2 m \log n : 1 \leq m \leq n \right\}.$$

If, however, η is in the closure of $\cup_{m=1}^{\infty} \mathcal{P}_m$, then the rate of convergence of the risk of the SBC $\widehat{\eta}_{\widehat{m}}$ is worse than that of Mallows' $\widehat{\eta}_{\widehat{m}}$ by a factor of $\log n$ (Problem 12.6.1). Note that the Mallows' rule has the same formulation as SBC with $2 \log n$ replaced by 2.

For the SBC \widehat{m} (Problem 12.4.7),

$$\sigma_n^{-1} |\widehat{\eta}_{\widehat{m}}| \leq \sqrt{2 \log n} .$$

12.4.4 An Oracle Inequality for Classification

The oracle inequality formulation in the forms (12.4.2) and (12.4.4) arose early on in the work of Vapnik (1998) on classification. Classification is a problem with a more complex loss function than quadratic loss and the difficulty of the problem is determined not only by that of estimating the regression,

$$P[Y = 1|Z] ,$$

but also the contours of the regression, a task governed by what is called the margin (Mammen and Tsybakov (1995)). On the other hand, the boundedness of all variables involved makes inequalities come out as simple consequences of the empirical process bounds of Section 7.1 and Appendix D.2. Specifically consider a two-class problem so that the training samples (Z_i, Y_i) are i.i.d. as $(Z, Y) \sim P$ with $Y = \pm 1$ and $Z \in \mathcal{H}$. In this case,

$$\delta_t : \mathcal{H} \rightarrow \{-1, 1\} .$$

We use 0 – 1 loss so that

$$R(P, \delta_t) = P[Y \delta(Z) = -1] .$$

Let $\lambda(y, d)$ denote the loss when Y is classified as $y \in \{-1, 1\}$. An oracle would use $t^*(P)$ that minimizes the classification loss $\ell(P, \delta) = E_P(\lambda(Y), \delta(Z))$,

$$\ell(P, \delta_{t^*}) = \min_t \ell(P, \delta_t) .$$

Since we don't know P but have a training sample $\mathbf{X} = \{(Z_i, Y_i) : 1 \leq i \leq n\}$, we want to construct $\widehat{\delta}_{\widehat{t}}(\cdot)$ such that

$$R(P, \widehat{\delta}_{\widehat{t}}) = E\{P[Y_{n+1} \widehat{\delta}_{\widehat{t}}(Z_{n+1}) = -1 | \mathbf{X}]\}$$

is small. A natural approach in the context of this section is to find an unbiased estimate \widehat{R}_t of $R(P, \delta_t)$ and minimize it. If we know nothing about P a natural approach is to use

$$\widehat{R}_t \equiv \frac{1}{n} \sum_{i=1}^n \mathbb{1}(Y_i \delta_t(Z_i) = -1)$$

and use $\widehat{\delta}_{\widehat{t}}$ where

$$\widehat{t} = \arg \min_t \widehat{R}_t .$$

We expect that $\widehat{R}_{\widehat{t}}$ is an underestimate of $R(P, \delta_{\widehat{t}})$. We use an oracle inequality to relate $R(P, \delta_{\widehat{t}})$ to $R(P, \delta_{t^*})$ although this does not actually give us an idea of what $R(P, \delta_{\widehat{t}})$ is.

What an oracle inequality requires is Δ_P such that

$$E\ell(P, \delta_{\widehat{t}}) \leq E\ell(P, \delta_{t^*}) + \Delta_P .$$

Before going to an oracle inequality we establish the fundamental,

Theorem 12.4.3

$$R(P, \delta_{t^*}) \leq R(P, \delta_{\widehat{t}}) \leq R(P, \delta_{t^*}) + 2E \sup_t |\widehat{R}_t - \ell(P, \delta_t)| . \quad (12.4.18)$$

Proof. By definition,

$$\widehat{R}_{\widehat{t}} \leq \widehat{R}_{t^*} , \quad R(P, \delta_{t^*}) \leq R(P, \delta_{\widehat{t}}) .$$

Hence,

$$\begin{aligned} \ell(P, \delta_{\widehat{t}}) &\leq \widehat{R}_{\widehat{t}} + \sup_t |\widehat{R}_t - \ell(P, \delta_t)| \leq \widehat{R}_{t^*} + \sup_t |\widehat{R}_t - \ell(P, \delta_t)| \leq \ell(P, \delta_{t^*}) \\ &\quad + 2 \sup_t |\widehat{R}_t - \ell(P, \delta_t)| . \end{aligned}$$

The theorem follows. \square

We now go to empirical process theory. Suppose that the bracketing number of $\mathcal{F}_{\mathcal{T}} \equiv \{\ell(P, \delta_t) : t \in \mathcal{T}\}$ satisfies

$$N_{[]}(\delta, \mathcal{F}_{\mathcal{T}}, L_2(P)) \leq c\delta^{-d} , \quad (12.4.19)$$

for all P, δ sufficiently small, c independent of P . Then, we can use a simplified version of Theorem 7.1.3, given below, which is appropriate since $|\ell(P, \delta_t)| \leq 1$, and we do not need to bound the variance of $|\ell(P, \delta_t)|$ separately. Recall the notation

$$\mathcal{E}_n(f_t) = n^{-1} \sum_{i=1}^n [f_t(X_i) - E_P f_t(X_i)] .$$

Theorem 12.4.4. (van der Vaart and Wellner (1996) Theorem 2.14.29). Given $\mathcal{F}_{\mathcal{T}} : \{f_t : t \in \mathcal{T}\}$ with f_t satisfying, $|f_t|_{\infty} \leq 1$, for all $t \in \mathcal{T}$, and $N(\delta, \mathcal{F}_{\mathcal{T}}, L_2(P)) \leq c\delta^{-d}$ for all P, δ , and c as in (12.4.19), then, for a universal constant $D \equiv D(c)$,

$$P[n^{\frac{1}{2}} \sup_{t \in \mathcal{T}} |\mathcal{E}_n(f_t)|_{\infty} > a] \leq \left(\frac{D}{\sqrt{d}} \right)^d e^{-2a^2} . \quad (12.4.20)$$

We can now give an oracle inequality:

Theorem 12.4.5. If $\mathcal{F}_{\mathcal{T}}$ satisfies (12.4.19) then

$$R(P, \delta_{\widehat{t}}) \leq \inf_t R(P, \delta_t) + n^{-\frac{1}{2}} \left(\frac{D}{\sqrt{d}} \right)^d \frac{\sqrt{2\pi}}{4} . \quad (12.4.21)$$

Proof. Use Theorem 12.4.3 and (12.4.20) and the standard (see Example 4.4.7) formula,

$$EU = \int_0^\infty P[U \geq v]dv \text{ if } U \geq 0 .$$

□

Remark 12.4.3. 1. The bound (12.4.21) does not rely on P belonging to any family \mathcal{P} and as such may be poor. For instance, suppose δ_t is the Bayes rule if $P = Q_t$ where $\{Q_t : t \in R^p\}$ is a regular parametric model. Then, although $\hat{\delta}_t$ is not chosen optimally, as it is in Example 12.3.1, it may still be shown that its Bayes regret is $O(n^{-1})$ whenever $P = Q_{t_0}$ for some t_0 , rather than $O(n^{-\frac{1}{2}})$ given by (12.4.21).

2. On the other hand (12.4.21) is indeed correct for ascertaining minmax bounds over large families \mathcal{P} — see Marron (1983) where the first such results were presented.

3. The bounds can give qualitative information on where ordinary asymptotic theory may not hold well such as the behaviour of the best one can do as a function of the bracketing number “dimension” d . For regular parametric families \mathcal{F}_T , this is just the dimension of the parameter whose effect on performance is not apparent in results such as Theorem 6.2.2. This additional type of information is a valuable feature of oracle inequalities.

4. This particular oracle inequality does not capture the phenomenon of “margin.” Roughly speaking, Bayes classification requires that we estimate, not necessarily $P[Y = 1|Z = z]$ as a function of z well but rather its contours $C = \{z : P[Y|Z = z] = \frac{1}{2}\}$. This may be easy even if the function is estimated badly if the set of z ’s near C has small probability. On the other hand the contours of a polynomial in several dimensions may be very hard to estimate. The correct combination of conditions on \mathcal{P} which determine minmax classification risk was determined by Mammen and Tsybakov (1995).

Summary. In this section we introduce the notion of oracle inequalities, the principal tool in the machine learning literature for characterizing the performance of classification rules both for restricted and unrestricted model classes \mathcal{P} . They are of value not only in their own right but also as principal tools in establishing the performance of potentially minmax rules. We consider the first, oracle inequalities, for submodels of the Gaussian white noise model, in which we also study the behaviour of Stein’s unbiased risk estimate in constructing adaptive estimates. We next consider the basic type of such inequalities valid for all P considered first by Vapnik (1998) in the context of classification. The literature and uses of such inequalities is perpetually growing. For many more examples see Györfi et al (2002), Bishop (2006), and van de Geer (2000). A very thorough treatment is in Bühlmann and van der Geer (2010).

12.5 Performance and Tuning via Cross Validation

In the previous section we have discussed some methods of choosing the tuning parameter t that determines the level of regularization. We continue this theme in Section 12.5 for a highly important approach referred to as cross validation. Not only does this approach

yield good prediction but it also provides methods of assessing prediction performance. We have considered prediction performance from the point of view of oracle inequalities. These bounds are worst case lower bounds on performance and as such are much too conservative. On the other hand we have, in the past, considered data-based measures such as confidence intervals which though not giving the guarantees of oracle bounds are asymptotically correct and do well in practice. The analogue in the “machine learning context” is to give reasonable estimates of misclassification probabilities and confidence bounds in regression.

To illustrate the difference between oracle bounds and data-based methods consider the simple problem of estimating the mean μ of a distribution concentrated on $[-1, 1]$ using a sample X_1, \dots, X_n i.i.d. as X . Our usual data-based approach outside this chapter would be to use

$$\bar{X} \pm z_{1-\alpha/2} \hat{\sigma}_n n^{-\frac{1}{2}}$$

as an approximate confidence interval, with $\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2$. On the other hand, we could give an oracle type guaranteed bound on μ . By Hoeffding’s inequality, (7.1.10), since $|X| \leq 1$,

$$P[\sqrt{n}|\bar{X} - \mu| \geq v] \leq 2e^{-\frac{v^2}{2}}.$$

This leads to guaranteed level $(1 - \alpha)$ confidence bounds for μ ,

$$\bar{X} \pm \sqrt{-\log(\alpha/2)} n^{-\frac{1}{2}}.$$

Evidently, if α and $\text{Var}(X)$ are small, this is a grossly conservative bound and Hoeffding’s inequality gives a poor estimate of the performance of \bar{X} . Thus the data-based bound is preferable. Cross validation is a data-based approach.

Cross validation will be used to estimate criteria functions. Let δ_t denote a decision rule with tuning parameter $t \in \mathcal{T}$. Crude estimates of performance such as the empirical misclassification rate

$$\frac{1}{n} \sum_{i=1}^n \mathbb{1}(\delta_t(Z_i) \neq Y_i) \tag{12.5.1}$$

are poor, as oracle inequalities make clear. Since regularization is based on optimizing an estimate of performance over a family of procedures, using a good estimate of performance is of importance for choosing the tuning parameter that corresponds to an appropriate amount of regularization; see Arlot and Celisse (2010). We pursue this aspect of data determined selection of the tuning parameter that yields the optimal estimated risk, returning to measurement of performance at the end.

12.5.1 Cross Validation for Tuning Parameter Choice

A number of methods involving sample splitting or approximations to such methods have come under the heading of cross validation. The one most widely studied in the statistical literature is

Leave one out cross validation and empirical risk minimization

The cross validation idea works for any prediction problem and is closely related to the jackknife — see Section 10.3. Let $\mathcal{S} \equiv \{X_i = (Z_i, Y_i), i = 1, \dots, n\}$ with $Y_1 \in R$, $Z_i \in R^p$, be an i.i.d. sample, let t be a tuning parameter, and $\{\delta_t(\cdot, X_1, \dots, X_n) : t \in \mathcal{T}\}$ be a family of predictors which has been determined, for example, by the fitting of a sieve. Let $l(y, d)$ be the prediction loss corresponding to the response y to be predicted and decision d . Thus, we have the examples

$$\begin{aligned} l(y, d) &= 1(y \neq d) && \text{0-1 loss} \\ l(y, d) &= (y - d)^2 && \text{squared error loss} \end{aligned}$$

The naive estimate of the Bayes risk of $\delta_t(\cdot)$ is, as we have noted, the empirical risk,

$$\widehat{R}(\delta_t) = \frac{1}{n} \sum_{i=1}^n l(Y_i, \delta_t(Z_i)) .$$

Although typically asymptotically correct for any fixed t , this is, as we have noted, overoptimistic, since the performance of the rule is being judged on the sample used to create it. One way to reduce this negative bias is to use *leave one out cross validation* (jackknife) with risk defined as

$$\widehat{R}_J(\delta_t) = \frac{1}{n} \sum_{i=1}^n l(Y_i, \delta_t^{(-i)}(Z_i)) , \quad (12.5.2)$$

where $\delta_t^{(-i)}$ is a rule based on $X^{(-i)} \equiv \{X_j, j \neq i\}$. Implicit in this is the assumption that $\delta_t(\cdot | X^{(-i)})$ is defined. We consider δ_t that can be written

$$\delta_t(\cdot | X'_1, \dots, X'_m) = \delta_t(\cdot | \widehat{P}_m)$$

where \widehat{P}_m is the empirical distribution of a subsample X'_1, \dots, X'_m , of \mathcal{S} and $\delta_t(\cdot | P)$ is assumed to be defined for all P used in the following discussion. An indicator of the promise of leave one out cross validation is that $\widehat{R}_J(\delta)$ is an unbiased estimate of the Bayes risk of δ for sample size $n - 1$, that is,

$$E\widehat{R}_J(\delta) = E \int l(y, \delta(z | \widehat{P}_{n-1}) dP(z, y) . \quad (12.5.3)$$

We define the *leave one out cross validation* choice of t as

$$\widehat{t}_J = \arg \min \widehat{R}_J(\delta_t) .$$

Barron, Birgé and Massart (1999) exhibit oracle inequalities for $\widehat{\delta}_{\widehat{t}_J}$ and show a close relation between \widehat{t}_J and Mallows' C_p .

V fold Cross Validation

A more general, computationally faster and much more broadly applicable method is V fold cross validation. We consider

$$t \in \mathcal{T} \equiv \{1, \dots, M_n\}.$$

The sample \mathcal{S} is split into V disjoint parts and each part has $m = n/V$ observations. Call the V subsamples $\mathcal{S}_1, \dots, \mathcal{S}_V$. $V = 5$ and $V = 10$ are customary choices. If $n = 500$ and $V = 10$, we will have 10 subsamples of \mathcal{S} , each with $m = 50$ observations. The estimate \widehat{R}_V of the Bayes risk of a rule δ_t is defined as follows. Let $\mathcal{S}^{(-j)} = \mathcal{S} - \mathcal{S}_j$ and let $\delta_t^{(j)} = \delta_t(\cdot | P^{(j)})$, where $P^{(j)}$ is the empirical distribution of $\{X_i : X_i \in \mathcal{S}^{(-j)}\}$. Then, for loss function ℓ ,

$$\widehat{R}_V(\delta_t) \equiv \frac{1}{mV} \sum \{\ell(Y_i, \delta_t^{(j)}(Z_i)) : i \in \mathcal{S}_j, j = 1, \dots, V\}. \quad (12.5.4)$$

That is, leave out the first m X 's, use the remaining $m(V - 1)$ X 's to compute $\delta_t^{(j)}$, and average the loss on the m X 's left out. Then do the same for the next m observations, and so on for the V subsamples, then average over the subsamples. In practice, the division into V pieces may be made several times and the resulting \widehat{R}_V 's averaged.

Define the V fold cross validation tuning parameter selector as

$$\widehat{t} \equiv \arg \min [\widehat{R}_V(\delta_t)].$$

We will consider properties of $\widehat{\delta}_t$ for the problem of estimating $\mu(z) = E(Y|\mathbf{Z} = \mathbf{z})$ using the squared error loss function, $\ell(y, d) = (y - d)^2$. In particular, we shall state a set of assumptions under which V fold cross validation for quadratic loss is optimal in a sense we shall define.

We give the argument for *sample splitting cross validation*. Our proof will make it clear that *V fold cross validation* is argued in the same way. We split the sample $(Z_1, Y_1), \dots, (Z_n, Y_n)$ into a *test sample* of cardinality m and a *training sample* of cardinality $n - m$. The predictor is computed for the training set and its loss is evaluated at the test set. Without loss of generality let the test set be $\mathcal{S}_1 = (Z_1, Y_1), \dots, (Z_m, Y_m)$.

Use the notation, for $a \equiv a(Z, Y)$,

$$\begin{aligned} \|a\|_P^2 &\equiv \int a^2(z, y) dP(z, y) \\ \|a\|_m^2 &\equiv \frac{1}{m} \sum_{i=1}^m a^2(Z_i, Y_i). \end{aligned}$$

We introduce the following assumptions and further notation.

A0: Y is bounded: $|Y| \leq C < \infty$.

A1: Write $\delta_t(\cdot)$ for $\delta_t(\cdot | P^{(1)})$, where $P^{(1)}$ is the empirical distribution of the training set $\{X_i : X_i \in \mathcal{S}_1\}$.

Let

$$t^* = \arg \min_t \|Y - \delta_t(Z)\|_P^2 = \arg \min_t \|\mu(Z) - \delta_t(Z)\|_P^2.$$

Assume there exists a best oracle rule defined by

$$\delta_{\mathcal{O}}(z) \equiv \delta_{t^*}(z)$$

which for fixed $P \in \mathcal{P}$, and some sequence r_n , satisfies:

$$\|\delta_{\mathcal{O}}(Z) - \mu(Z)\|_P^2 = r_n^2 (1 + o(1)).$$

A2: There exists a best “CV selected rule” defined by

$$\widehat{\delta} \equiv \delta_{\widehat{t}}$$

where $\widehat{t} = \arg \min_t \|Y - \delta_t(Z)\|_m^2$.

A3: Let $\Delta_t(Z) = |\delta_t(Z) - \mu(Z)|$, and assume that for fixed $P \in \mathcal{P}$,

$$\max_t \left\{ \left| \frac{\|\Delta_t(Z)\|_m^2}{\|\Delta_t(Z)\|_P^2} - 1 \right| : t \in \{1, \dots, M_n\} \right\} = o_P(1).$$

If we assume $|\delta_t(Z)| \leq |Y| \leq C_n < \infty$ for all $t \in \{1, \dots, M_n\}$ and $n^{-1} \log M_n = o(1)$, then A3 holds by Hoeffding’s inequality.

A4: $\frac{\log^{\frac{1}{2}} M_n}{mr_n^2} = o(1)$.

Let

$$\widehat{r}_n = \|\delta_{\widehat{t}} - \mu\|_P.$$

Theorem 12.5.1. Under A0–A4,

$$\widehat{r}_n = r_n (1 + o_P(1)). \quad (12.5.5)$$

Remark 12.5.1. For Theorem 12.5.1,

- 1) A0 can be weakened with the result still holding.
- 2) The conditions can be strengthened to yield an oracle inequality form of the Theorem.
- 3) The difference between this sample splitting result and the result for $V > 2$ is that sample splitting is only carried out once to form the risk estimate. This is inessential because the average of V such risk estimates is more accurate although the summands are highly correlated.

4) It is, as we have seen, common to have for $P \in \mathcal{P}$, nonparametric,

$$r_n \sim cn^{-1+\delta}, \quad \delta > 0.$$

If $M_n = n^K$, $K < \infty$, and $m = n/\omega(n)$, where $\omega(n) \uparrow \infty$ slowly enough and A3 and A4 hold, then the theorem holds. The take-home story is that V fold cross validated choice where $V = \omega(n)$, i.e. is almost bounded, yields oracle prediction for quadratic loss.

Proof of Theorem 12.5.1

By definition,

$$\widehat{R}(\delta_{\hat{t}}) \leq \widehat{R}(\delta_{t^*}) \quad (12.5.6)$$

$$\|Y - \delta_{t^*}\|_P^2 \leq \|Y - \delta_{\hat{t}}\|_P^2. \quad (12.5.7)$$

Rewrite (12.5.6) as

$$\frac{2}{m} \sum_{i=1}^m (\mu(Z_i) - \widehat{\delta}(Z_i)) \varepsilon_i + \|\widehat{\delta} - \mu\|_m^2 \leq \frac{2}{m} \sum_{i=1}^m (\mu(Z_i) - \delta_0(Z_i)) \varepsilon_i + \|\delta_0 - \mu\|_m^2 \quad (12.5.8)$$

where $\varepsilon_i \equiv Y_i - \mu(Z_i)$. Note that

$$\left| \frac{1}{m} \sum_{i=1}^m (\mu(Z_i) - \widehat{\delta}(Z_i)) \varepsilon_i \right|^2 \leq \max_{1 \leq s \leq M_n} \left\{ \left| \frac{1}{m} \sum_{i=1}^m \frac{(\mu(Z_i) - \delta_t(Z_i)) \varepsilon_i}{\|\mu - \delta_t\|_m} \right|^2 \right\} \|\mu - \widehat{\delta}\|_m^2. \quad (12.5.9)$$

The first factor of the RHS of (12.5.9) is of the form

$$U_n \equiv \max \left\{ \left| \frac{1}{m} \sum_{i=1}^m a_{it} \varepsilon_i \right|^2 : t \in \{1, \dots, M_n\} \right\}$$

where $\{a_{it}\}$ and ε_i are independent given Z_1, \dots, Z_m , $E(\varepsilon_i | Z_1, \dots, Z_m) = 0$ and $\sum_{i=1}^m a_{it}^2 = 1$.

By an inequality of Pinelis which is an application of Hoeffding's inequality (7.2.9), and the boundedness of the ε_i , $|\varepsilon_i| \leq 2C$, we have

$$EU_n \leq 16C^2 \log M_n / m. \quad (12.5.10)$$

Conditioning on X_{m+1}, \dots, X_n we obtain from (12.5.8)–(12.5.10) that

$$\|\mu - \widehat{\delta}\|_m^2 \leq 4Cm^{-\frac{1}{2}} (\log M_n)^{\frac{1}{2}} \|\mu - \widehat{\delta}\|_m + \|\mu - \delta_0\|_m^2. \quad (12.5.11)$$

Next

$$\left| \frac{\|\Delta_{\widehat{s}}\|_m^2}{\|\Delta_{\widehat{s}}\|_P^2} - 1 \right| \leq \max_t \left| \frac{\|\Delta_s\|_m^2}{\|\Delta_s\|_P^2} - 1 \right| = o_P(1), \quad (12.5.12)$$

by A3, and the same holds for Δ_{s_0} . From (12.5.11) and A4, we obtain

$$\|\mu - \widehat{\delta}\|_m^2 \leq (1 + o_P(1))\|\mu - \delta_0\|_m^2,$$

and applying (12.5.12) the theorem follows. \square

Remark 12.5.2. This V fold criterion is the most widely used one for tuning parameter choice. It can be shown (Barron, Birgé and Massart (1999), Propositions 1, 2 and Theorem 3) that if we consider multiple regression with all $2^p - 1$ submodels possible as a sieve and $p \sim \log n$, then leave one out cross validation does not regularize least squares sufficiently while V fold, for suitable V, n, p , does.

We now turn briefly to assessment of performance.

12.5.2 Cross Validation for Measuring Performance

As we have noted, the naive estimate \widehat{R} is typically an underestimate of misclassification risk. Efron (1983) shows both theoretically, through higher order expansion, and experimentally, by Monte Carlo, that the leave one out jackknife estimate \widehat{R}_J is a better estimate of the Bayes risk than \widehat{R} . He develops an even better procedure using a combination of 2 fold cross validation and \widehat{R} . In this case, 2 fold cross validation itself is not very satisfactory since one is essentially unbiasedly estimating the behaviour of the rule using a training sample of size $n/2$, thus overestimating the Bayes risk for a sample of size n . Since \widehat{R} is an underestimate it is reasonable to combine the two and Efron shows that the weights .632 for the cross validation estimate and .368 for \widehat{R} are best. It seems plausible, in view of Theorem 12.5.1, that using $\log n$ fold cross validation to estimate the Bayes risk of say $\widehat{\delta}_t$ should behave well. Unfortunately, using $\log n$ cross validation for t , itself chosen using $\log n$ cross validation, is apt to yield an underestimate. Arlot has shown that using Efron's .632 estimate to select s has given excellent results; see Arlot and Celisse (2010). But again the goals of performance assessment and optimization clash although we could apply the .632 approach to the optimized rule.

Summary. In this section we first study cross validation, the chief method for selecting tuning parameters in prediction methods, and show the wide applicability of V fold cross validation as opposed to leave one out cross validation with a theorem on the asymptotic optimality of V fold cross validation for squared error loss. Actually estimating performance of a final data determined choice of decision rule is difficult although Efron's .632 rule may be generally helpful.

12.6 Model Selection and Dimension Reduction

The first of these two topics starts with a model point of view more akin to Chapters 1–5 where we focussed on parametric models. We still, as in the previous sections, assume we have a nested sieve of models whose closure is nonparametric, but add the assumption that one of the finite dimensional members of the sieve is generating the data. Specifically, if

we are, say, considering regression models with a fixed number k of real factors Z_1, \dots, Z_k and we can expand $E(Y|\mathbf{Z})$ by means of tensor bases to a function of an infinite number of regression covariates, as in Example 12.3.1, we assume that one of these, say a polynomial of order d in the factors, is, in fact, generating the data.

Our second topic is the more fundamental but the hardest to quantify. The point of departure is a set of data, say a sample from a population, which is very high dimensional and has a very complex representation. The object is to find a low dimensional, simple representation which can give subject matter insights and lead to efficient prediction and classification procedures. There is no commitment to a particular model a priori. We are dealing, for example, with a sample of small size in relation to the dimension of the data so that there is no hope of nonparametric estimation of the density. We seek to find a representation of the data and/or the model which is sufficiently low dimensional to be worked with, but sufficiently accurate to enable us to find the essential features of the situation we are analyzing. In this context we will introduce Principal Component Analysis (PCA), a classical method for dimension reduction. We will discuss some difficulties of this and related methods in modern applications. We will also discuss the relationship between the importance of factors and sparsity of models.

12.6.1 A Bayesian Criterion for Model Selection

We consider the problem of selecting a model \mathcal{M} from a sieve of regular parametric models, $\mathcal{M}_0 \subset \dots \subset \mathcal{M}_p$, where

$$\mathcal{M}_k \equiv \{P_{\boldsymbol{\theta}^{(k)}}(\cdot) : \boldsymbol{\theta}^{(k)} \in \Theta_k \text{ open in } R^{N_k}, N_0 < \dots < N_p\}.$$

For a given P in the sieve let $k_0 = k_0(P)$ be the smallest k such that $P \in \mathcal{M}_k$, and let $\boldsymbol{\theta}_0 = \boldsymbol{\theta}_0^{(k_0)} \in \Theta_{k_0}$ denote the subscript on this P . Then P is also in the models with $k > k_0$. If P is in the sieve and generates the data, $k_0(P)$ and $\boldsymbol{\theta}_0^{(k_0(P))}$ are assumed identifiable.

We will discuss the asymptotic determination of k_0 and estimation of $\boldsymbol{\theta}_0^{(k_0)}$ within Θ_{k_0} . Here we consider the simplest model \mathcal{M}_{k_0} to be the most parsimonious or “best” model and we want to have high probability of selecting \mathcal{M}_{k_0} for large sample sizes.

To focus ideas, first consider a simple sieve with data $\mathbf{X}_i = (\mathbf{Z}_i, Y_i)$ i.i.d. and the familiar regression model,

$$Y_i = \beta_0 + \sum_{j=1}^d \beta_j Z_{ij} + \varepsilon_i, \quad i = 1, \dots, n$$

where ε_i are i.i.d. $\mathcal{N}(0, \sigma^2)$, $\mathbf{Z}_i = (Z_{ii}, \dots, Z_{id})^T$. Suppose we can order the factors, $\mathbf{Z}_1, \dots, \mathbf{Z}_d$, in terms of importance to Y . Then, \mathcal{M}_k is parametrized by $\boldsymbol{\theta}^{(k)} \equiv (\beta_0, \dots, \beta_k, \sigma^2)^T$, $\Theta_k \subset R^{k+2}$, $N_p = d+2$, and $P \in \mathcal{M}_k$ iff $\beta_{k+1} = \dots = \beta_d = 0$. Determining k_0 , the number of factors \mathbf{Z}_j with non-zero beta coefficients, is the subject of much research. We could try the following: “Consecutively test $H_j : \beta_j = 0$ vs $\beta_j \neq 0$,

$j = 1, \dots, d$, stopping as soon as we do not reject H_j , and set $k_0 = j - 1$ ". The main issue with this approach is what critical value to use, that is, how to balance type I and II error probabilities. Here we turn to the simplest approach, which is based on a Bayesian point of view.

Returning to the general case, we place prior mass α_k on \mathcal{M}_k , $\sum_{k=0}^p \alpha_k = 1$, $\alpha_k > 0$. That is, α_k is the probability that $k_0 = k$. Given $k_0 = k$ and $P_{\boldsymbol{\theta}} \in \mathcal{M}_k$, we assume $\boldsymbol{\theta} = \boldsymbol{\theta}^{(k)}$ has a prior density $\pi_k(\cdot)$ on Θ_k where $\pi_k > 0$ and continuous. A natural rule for choosing $k_0(P)$ is "Select \hat{k} which maximizes the posterior probability $\hat{\alpha}_k$ of \mathcal{M}_k ". That is,

$$\hat{k} = \arg \max_k \hat{\alpha}_k \equiv \arg \max_k \{ \text{Prob}(k_0 = k | \mathbf{X}) \}$$

where $\mathbf{X} = (X_1, \dots, X_n)^T$ with X_1, \dots, X_n i.i.d. as $X \in R^q$, $q \geq 1$. We prove, under suitable regularity conditions, a theorem which gives a general asymptotically optimal Bayes rule \hat{k} . Schwarz (1978) gave this result for the special case of exponential families.

In our Bayesian framework, we assume that given k and $\boldsymbol{\theta}^{(k)}$, X_i has density $p(\cdot | \boldsymbol{\theta}^{(k)})$, $\boldsymbol{\theta}^{(k)} \in \Theta_k$. Then the density of \mathbf{X} given k is

$$p(\mathbf{x}|k) = \int_{\Theta_k} \Pi_{i=1}^n p(x_i | \boldsymbol{\theta}^{(k)}) \pi_k(\boldsymbol{\theta}^{(k)}) d\boldsymbol{\theta}^{(k)}.$$

Let $I_k(\boldsymbol{\theta}^{(k)}) = E_{\boldsymbol{\theta}^{(k)}} \nabla \nabla^T \log p(X | \boldsymbol{\theta}^{(k)})$, $\boldsymbol{\theta}^{(k)} \in \Theta_k$, be the Fisher information matrix and let $K(\boldsymbol{\theta}^{(k)}, \boldsymbol{\theta}_0) = -E_{\boldsymbol{\theta}_0} \log [p(X | \boldsymbol{\theta}^{(k)}) / p(X | \boldsymbol{\theta}_0)]$ be the Kullback-Leibler divergence where $\boldsymbol{\theta}_0 \equiv (\theta_{01}, \dots, \theta_{0N_{k_0}})^T \equiv \boldsymbol{\theta}_0^{(k_0)}$ and $k_0 = k_0(P)$ corresponds to the true P .

We assume

- (i) $\{p(\cdot | \boldsymbol{\theta}^{(k)})\}$, $k \geq 0$, satisfy the conditions of Theorem 6.2.3, for each \mathcal{M}_k .
- (ii) $\theta_*^{(k)} \equiv \arg \min \{K(\boldsymbol{\theta}^{(k)}, \boldsymbol{\theta}_0) : \boldsymbol{\theta}^{(k)} \in \Theta_k\} > 0$ for $k < k_0$.
- (iii) If $k < k_0$, and $P_{\boldsymbol{\theta}_0}$ generates the data, then the assumptions of Theorem 6.2.1 apply with $\psi = \log p(x | \boldsymbol{\theta}^{(k)})$ and $\theta(P) = \theta_*^{(k)}$

Conditions that imply (iii) can be obtained from the M -estimate consistency results of Chapters 6 and 9 because $\hat{\boldsymbol{\theta}}^{(k)}$ maximizes $n^{-1} \sum_{i=1}^n \log [p(X_i | \boldsymbol{\theta}^{(k)}) / p(X_i | \boldsymbol{\theta}_0)]$ so that the corresponding population parameter minimizes $K(\boldsymbol{\theta}^{(k)}, \boldsymbol{\theta}_0)$ over $\boldsymbol{\theta}^{(k)} \in \Theta_k$, $k < k_0$.

By Bayes formula, for the appropriate constant $c > 0$, the posterior probability of model \mathcal{M}_k is

$$\hat{\alpha}_k = \alpha_k \int \Pi_{i=1}^n p(x_i | \boldsymbol{\theta}^{(k)}) \Pi_k(\boldsymbol{\theta}^{(k)}) d\boldsymbol{\theta}^{(k)} / c.$$

We next give approximations to the maximizer \hat{k} of $\hat{\alpha}_k$ and establish their consistency in a frequentist context.

Theorem 12.6.1. Under (i), (ii), and (iii), let $L_k = \prod_{i=1}^n p(\mathbf{X}_i | \widehat{\boldsymbol{\theta}}^{(k)})$ where $\widehat{\boldsymbol{\theta}}^{(k)}$ is the MLE of $\boldsymbol{\theta}^{(k)}$ in model \mathcal{M}_k , $0 \leq k \leq p$. Suppose the data is generated by $\mathcal{P} \equiv P_{\boldsymbol{\theta}_0}$ with $\boldsymbol{\theta}_0 = \boldsymbol{\theta}_0^{(k_0)}$. Then, if $|\cdot|$ denotes determinant,

(a) For $k > k_0$, as $n \rightarrow \infty$,

$$\log \frac{\widehat{\alpha}_k}{\widehat{\alpha}_{k_0}} = \log \frac{L_k}{L_{k_0}} - \frac{k - k_0}{2} \log \left(\frac{n}{2\pi} \right) - \frac{1}{2} \log \frac{|I_{k_0}(\boldsymbol{\theta}_0^{(k_0)})|}{|I_k(\boldsymbol{\theta}_0^{(k)})|} + \log \frac{\alpha_k}{\alpha_{k_0}} + o_P(1). \quad (12.6.1)$$

(b) For $k < k_0$,

$$\frac{\widehat{\alpha}_k}{\widehat{\alpha}_{k_0}} \xrightarrow{P} 0 \text{ as } n \rightarrow \infty. \quad (12.6.2)$$

(c) If \widehat{k} is the maximizer of $\widehat{\alpha}_k$ and $\widehat{\widehat{k}}$ maximizes

$$2 \log L_k - k(\log n - \log 2\pi) + \log \alpha_k + \log |I_k(\widehat{\boldsymbol{\theta}}^{(k)})|, \quad (12.6.3)$$

then $\widehat{\widehat{k}}$ is consistent in the sense that

$$P[\widehat{\widehat{k}} = \widehat{k} = k_0(P)] \longrightarrow 1 \text{ as } n \rightarrow \infty. \quad (12.6.4)$$

(d) The universal criterion

$$k^* = \arg \max \left\{ \log L_k - \frac{1}{2} k \log n \right\} \quad (12.6.5)$$

satisfies

$$P(k^* = \widehat{k} = k_0(P)) \rightarrow 1 \text{ as } n \rightarrow \infty. \quad (12.6.6)$$

Proof. We begin with (12.6.1). Letting $k > k_0$, write

$$\begin{aligned} \log \frac{\widehat{\alpha}_k}{\widehat{\alpha}_{k_0}} &= \log \frac{L_k}{L_{k_0}} - \log \int \prod_{i=1}^n \frac{p(X_i | \boldsymbol{\theta}_0^{(k_0)})}{p(X_i | \widehat{\boldsymbol{\theta}}^{(k_0)})} \pi_{k_0}(\boldsymbol{\theta}^{(k_0)}) d\boldsymbol{\theta}^{(k_0)} \\ &\quad + \log \int \prod_{i=1}^n \frac{p(X_i | \boldsymbol{\theta}^{(k)})}{p(X_i | \widehat{\boldsymbol{\theta}}^{(k)})} \pi_k(\boldsymbol{\theta}^{(k)}) d\boldsymbol{\theta}^{(k)} + \log \left(\frac{\alpha_k}{\alpha_{k_0}} \right). \end{aligned}$$

Arguing as in the proofs of Theorems 5.5.2 and 6.2.3, reasoning as for (5.5.12), noting that $P_{\boldsymbol{\theta}_0^{(k)}} = P_{\boldsymbol{\theta}_0^{(k_0)}}$, and both $\boldsymbol{\theta}_0^{(k)}$ and $\boldsymbol{\theta}_0^{(k_0)}$ are interior points of their respective parameter spaces, we find that for $n \rightarrow \infty$ the two integrals are

$$\begin{aligned} A &= (2\pi)^{\frac{k_0+1}{2}} n^{-\frac{(k_0+1)}{2}} \int \exp \left\{ -\frac{1}{2} \mathbf{t}^T I_{k_0}(\boldsymbol{\theta}_0^{(k_0)}) \mathbf{t} \right\} d\mathbf{t} (1 + o_P(1)) \\ B &= (2\pi)^{\frac{(k+1)}{2}} n^{-\frac{(k+1)}{2}} \int \exp \left\{ -\frac{1}{2} \mathbf{t}^T I_k(\boldsymbol{\theta}_0^{(k)}) \mathbf{t} \right\} d\mathbf{t} (1 + o_P(1)) \end{aligned}$$

where the A integral is over $R^{N_{k_0}}$ and B is over R^{N_k} . Evaluating A and B, using the fact that the integral of a density is 1, we obtain (12.6.1). Note that even though the information in \mathcal{M}_k , \mathcal{M}_{k_0} is evaluated at the point P , they are typically different. Now, (12.6.1) follows.

To prove (12.6.2) note that, for $k < k_0$,

$$\begin{aligned} \log(\hat{\alpha}_{k_0}/\hat{\alpha}_k) &= \log \int \prod_{i=1}^n \left[\frac{p(X_i|\boldsymbol{\theta}^{(k_0)})}{p(X_i|\hat{\boldsymbol{\theta}}^{(k_0)})} \right] \pi_{k_0}(\boldsymbol{\theta}_0^{(k_0)}) d\boldsymbol{\theta}^{(k_0)} \\ &\quad - \log \int \prod_{i=1}^n \left[\frac{p(X_i|\boldsymbol{\theta}^{(k)})}{p(X_i|\hat{\boldsymbol{\theta}}^{(k_0)})} \right] \pi_k(\boldsymbol{\theta}^{(k)}) d\boldsymbol{\theta}^{(k)} + \log\left(\frac{\alpha_{k_0}}{\alpha_k}\right) \\ &\equiv C_1 + C_2 + C_3. \end{aligned}$$

The first term is identical to the corresponding $k > k_0$ term and thus converges to $-\log A$, which is negative and of order $\log n$. We establish in (12.6.2) by showing that the second term is positive and of order n . Note that, by Jensen's inequality,

$$\begin{aligned} C_2 &= \log \int \prod_{i=1}^n \left[\frac{p(X_i|\hat{\boldsymbol{\theta}}^{(k_0)})}{p(X_i|\boldsymbol{\theta}^{(k)})} \right] \pi_k(\boldsymbol{\theta}^{(k)}) d\boldsymbol{\theta}^{(k)} \\ &\geq \int \sum_{i=1}^n \left[\log \frac{p(X_i|\hat{\boldsymbol{\theta}}^{(k_0)})}{p(X_i|\boldsymbol{\theta}^{(k)})} \right] \pi_k(\boldsymbol{\theta}^{(k)}) d\boldsymbol{\theta}^{(k)}. \end{aligned} \tag{12.6.7}$$

Because $\hat{\boldsymbol{\theta}}^{(k)}$ is a MLE, $p(x|\hat{\boldsymbol{\theta}}^{(k)}) \geq p(x|\boldsymbol{\theta}^{(k)})$. After replacing $\boldsymbol{\theta}^{(k)}$ with $\hat{\boldsymbol{\theta}}^{(k)}$ only the π_k part of the integrand depends on $\boldsymbol{\theta}^{(k)}$. Because the integral of a density is one, we obtain

$$\begin{aligned} \frac{1}{n} C_2 &\geq \frac{1}{n} \sum_{i=1}^n \left\{ \log \frac{p(X_i|\hat{\boldsymbol{\theta}}^{(k_0)})/p(X_i|\boldsymbol{\theta}_0)}{p(X_i|\hat{\boldsymbol{\theta}}^{(k)})/p(X_i|\boldsymbol{\theta}_0)} \right\} \\ &= \frac{1}{n} \sum_{i=1}^n \log \frac{p(X_i|\boldsymbol{\theta}_0)}{p(X_i|\boldsymbol{\theta}_*^{(k)})} + \frac{1}{n} \sum_{i=1}^n \log \frac{p(X_i|\hat{\boldsymbol{\theta}}^{(k_0)})}{p(X_i|\boldsymbol{\theta}_0)} \\ &\quad - \frac{1}{n} \sum_{i=1}^n \log \frac{p(X_i|\hat{\boldsymbol{\theta}}^{(k)})}{p(X_i|\boldsymbol{\theta}_*^{(k)})} \equiv D_1 + D_2 + D_3. \end{aligned}$$

The first term D_1 converges by the law of large numbers almost surely to $K(\boldsymbol{\theta}_*^{(k)}, \boldsymbol{\theta}_0) > 0$. For the second term, note that $2nD_2$ is the likelihood ratio statistic for testing $H : \boldsymbol{\theta} = \boldsymbol{\theta}_0$, and thus converges in law to a $\chi_{k_0}^2$ random variable by Theorem 6.3.2. It follows that $D_2 = O_p(n^{-1})$. We can also show that $2nD_3$ converges in law to a χ_k^2 random variable by arguing as we did to obtain Theorem 6.3.2 from Theorem 6.2.2. Thus $D_3 = O_p(n^{-1})$ also. We conclude that C_2 is positive and of order n . Thus $\alpha_k/\alpha_{k_0} \rightarrow 0$ as $n \rightarrow \infty$ when $k < k_0$.

To show consistency of \hat{k} , note that maximizing $\hat{\alpha}_h$ is equivalent to maximizing $\hat{\rho}_k \equiv \log[\hat{\alpha}_k/\hat{\alpha}_{k_0}]$. Because $\hat{\rho}_{k_0} = 0$, then $\max \hat{\rho}_k \geq 0$. We next use this to show that $P(\hat{k} \neq k_0) \rightarrow 0$.

- a) If $k < k_0$, then $P(\hat{k} = k) \leq P(\hat{\rho}_h \geq 0) \rightarrow 0$ by (12.6.2).
- b) If $k > k_0$, then, by (12.6.1),

$$P(\hat{k} = k) \leq P(\hat{\rho}_k \geq 0) = P(2 \log[L_k/L_{k_0}] \geq (k - k_0) \log n + o_P(1)).$$

By Theorem 6.3.2, $2 \log[L_k/L_{k_0}]$ converges to a \mathcal{X}^2 variable. Thus, $P(\hat{k} = k) \rightarrow 0$.

Next consider the consistency of \hat{k} . It is of the form $\arg \max\{\hat{\rho}_k + a_k\}$ where $a_k = O_p(1)$, and thus if $k \neq k_0$, then $P(\hat{k} = k) \leq P(\hat{\rho}_k \geq O_p(1)) \rightarrow 0$ as before. Similar arguments apply to \hat{k} and k^* . \square

Remark 12.6.1

(a) As we have noted in Section I.7, in the context of regression models, the asymptotic Bayes criteria k^* in (12.6.5), (called SBC and BIC) has the same structure as Mallows' C_p but with $\log n$ replacing 2. So $\log n$ corresponding to a test with significance level tending to 0 gives the correct threshold according to this criterion. The equivalence up to threshold continues to hold for any sequence of smooth nested models if C_p is replaced by the Akaike AIC criterion.

(b) As we have noted in the Gaussian white noise (GWN) framework, Section 12.4, if the models are not nested, for instance, if we consider all regression models with d or fewer variables, the correct threshold for prediction changes to something close to BC. For instance, this is the case if we modify Section I.7 slightly and suppose we observe

$$X_{ij} = \mu_j + \varepsilon_{ij},$$

$j = 1, \dots, d$, $i = 1, \dots, n$, with ε_{ij} i.i.d. $N(0, 1)$, that is, one way ANOVA. Let any subset $\{\mu_{j_1}, \dots, \mu_{j_k}\}$, $k \leq d$, correspond to the model where $\mu_{j_1} = \dots = \mu_{j_k} = 0$, all other μ 's vary freely. Then, for fixed j , the best test is to reject $H : \mu_j = 0$ if $|\bar{X}_j| \geq c/\sqrt{n}$. If we want to choose the correct model with probability tending to 1 we need to take c of order $\sqrt{2 \log d}$. But if we consider the sparse set

$$\Theta_M \equiv \left\{ \boldsymbol{\mu} : \sum_{j=1}^d 1(\mu_j \neq 0) \leq M \right\}$$

we saw in Section 12.3.3 that, if M is bounded, hard thresholding with c of order $(\log d)^{\frac{1}{2}}$ yields the minimax MSE order $(n^{-1} \log d)^{\frac{1}{2}}$.

(c) Software for the SBC rule can be found under "BIC" in "R" and other software packages.

(d) The prior α_k does not appear in the rule k^* . \square

12.6.2 Inference after Model Selection

Again consider the sieve of Section 12.6.1 and let \hat{k} denote a consistent estimate of k_0 . Given consistency of \hat{k} it follows that acting as if $\mathcal{M}_{\hat{k}}$ is the true model gives asymptotically

correct results for testing and estimation of the parameters $\theta_1, \dots, \theta_{k_0}$. To see this suppose T_n is a standardized estimate or test statistic such that $T_n \xrightarrow{\mathcal{L}} T$ as $n \rightarrow \infty$ for model \mathcal{M}_{k_0} , then $P_0(T_n \leq t) = P_0(T_n \leq t | \hat{k} = k_0)P_0(\hat{k} = k_0) + P_0(T_n \leq t | \hat{k} \neq k_0)P_0(\hat{k} \neq k_0)$ which implies $P_0(T_n \leq t | \hat{k} = k_0) \rightarrow P_0(T \leq t)$ as $n \rightarrow \infty$.

Consider our regression example with d coefficients β_1, \dots, β_d and $p = d + 2$. It seems rather unlikely in most cases that the number of factors $p - k_0$ which have zero beta coefficients bears no relation to the selection of k and asymptotic inference. Put another way, if $p - k_0 \rightarrow \infty$, it is unlikely that there exists a $k_0 = k_0(P)$ that is bounded independent of p . We expect indeterminacy of k_0 to affect inference procedures. In this case, the regression example makes clear what happens. Suppose $Z_j = h_j(Z)$ are a basis of $L_2(Z)$ where Z is a univariate regressor and consider the model

$$Y_i = \sum_{j=1}^d \beta_j h_j(Z_i) + \varepsilon_i, \quad i = 1, \dots, n$$

where $\{\varepsilon_i\}$ are i.i.d. $\mathcal{N}(0, \sigma^2)$ and $Eh_j(Z_i) = 0$. If the h_j are orthogonal, i.e. $\int h_j(z)h_k(z) dP(z) = 0$, $j \neq k$, the MLE $\hat{\beta}_j$ of β_j stabilizes no matter how big p fixed is since the vectors $(h_j(Z_1), \dots, h_j(Z_n))^T$ are asymptotically orthogonal.

Using Example 6.2.1, we can show that

$$\hat{\beta}_j = \sum_{i=1}^n Y_i h_j(Z_i) / \sum_{i=1}^n h_j^2(Z_i) + o_P(1) \quad (12.6.8)$$

and arguing formally

$$E(\hat{\beta}_j) = E E(\hat{\beta}_j | Z) = \beta_j + o(1)$$

because

$$E(\hat{\beta}_j | Z) = \frac{E(Y h_j(Z) | Z)}{Eh_j^2(Z)} + o_P(1) = \beta_j + o_P(1), \quad (12.6.9)$$

and $E(Y | Z) = \sum_{j=1}^d \beta_j h_j(Z)$. Similarly, by (1.4.6),

$$\text{Var}(\hat{\beta}_j) = n^{-1} \frac{E[h^2(Z) \text{Var}(Y | Z)]}{[Eh_j^2(Z)]^2} + o(n^{-1}) \quad (12.6.10)$$

since $E(\hat{\beta}_j | Z)$ is nearly constant. Thus inference does not depend on the number of nonzero coefficients to first order. However if the $h_j(Z)$ are not orthogonal, Example 6.2.1 shows how $\hat{\beta}_j$ depends on $h_1(Z), \dots, h_d(Z)$ in general. For fixed $1 \leq k_0 < d$

$$E(\hat{\beta}_j | Z) = \frac{E(Y h_{jk_0}(Z) | Z)}{Eh_{jk_0}^2(Z)} + o_P(1) \quad (12.6.11)$$

where $h_{jk}(Z) = h_j(Z) - \Pi(h_j(Z)|h_l(Z), 1 \leq l \leq k_0, l \neq j)$ and Π is the projection of Z on the linear span of $h_1(Z), \dots, h_{k_0}(Z)$. This reduces to

$$\begin{aligned} E(\hat{\beta}_j) &= E\left\{\frac{\sum_{l=1}^d h_{k_0}(Z)h_{jk_0}(Z)\beta_l}{E_{h_{k_0}}^2(Z)}\right\} + o(1) \\ &= \beta_j + \frac{\sum_{l=k_0+1}^d E(h_l(Z)h_{jk_0}(Z))}{E_{h_{k_0}}^2(Z)}\beta_l + o(1). \end{aligned} \quad (12.6.12)$$

Thus, $\hat{\beta}_j$ is biased with the bias depending on d . Its true variance is not the same as that obtained by acting as if $\hat{k} = k_0$ is correct. If we let $d \rightarrow \infty$ which makes $\hat{k} \rightarrow \infty$, the bias goes away but at a rate lower than n^{-1} . This is the same phenomenon we have observed with nonparametric regression and kernel density estimates. The minimax root MSE rate is of order smaller than $n^{-\frac{1}{2}}$ and both bias and variance are of that order. Although asymptotic normality still holds, mean and variance depend on the smoothness assumptions made on the nonparametric function being estimated.

12.6.3 Dimension Reduction via Principal Component Analysis

Approximating nonparametric models by finite dimensional parametric models is one way of reducing the dimension of the problem and in Sections 9.1.4, 11.4.2, 12.1.1 and 12.2.3 we have seen how sieves play a role in this approach. Techniques which also play a major role involve approximating high dimensional data by lower dimensional data. A very important method in this regard is principal component analysis (PCA) which is applied to samples $\{\mathbf{X}_i : 1 \leq i \leq n\}$ of p dimensional vectors where p is large and can exceed n . It is a way of reducing the number of covariates by replacing them by weighted sums of covariates thereby reducing the number of covariates. The theory is developed in Rao (1973), Seber (1984) and Anderson (2003) while applications can be found in Mardia et al (1979); see also Examples 8.3.11(continued), 9.1.5, Problem 8.3.24, Johnson and Wichern (2003), and Hastie, Tibshirani and Friedman (2009).

The focus is on the covariance matrix $\Sigma(P)$ of $\mathbf{X} = (X_1, \dots, X_p)^T$, where the X 's have been centered so that $E\mathbf{X} = \mathbf{0}$. By spectral theory (see Appendix B.10),

$$\Sigma(P) = \sum_{j=1}^K \lambda_j(P) \boldsymbol{\xi}_j(P) \boldsymbol{\xi}_j^T(P)$$

where $K \leq p$, the λ_j are (not necessarily distinct) eigenvalues of $\Sigma(P)$, and the p dimensional eigenvectors $\{\boldsymbol{\xi}_j\}$ are orthonormal. The linear combination

$$Z_j = \boldsymbol{\xi}_j^T \mathbf{X}, \quad 1 \leq j \leq K$$

is called the j th *principal component* (PC). A small K corresponds to a sparse subspace of the parameter space of Σ .

Although the eigenvalues up to this point appear as artificial constructs they are quite natural parameters. If $\lambda_1(P) > \dots > \lambda_p(P)$ are distinct, the eigenvector ξ_1 corresponding to $\lambda_1(P)$ solves the problem

$$\xi_1 = \arg \max \{ \text{Var}(\mathbf{a}^T \mathbf{X}) : |\mathbf{a}| = 1 \} .$$

That is, the first principal component Z_1 is the linear combination $\sum a_{1j} X_j$ subject to $\sum a_{1j}^2 = 1$ with the maximum variance. Moreover (Problem 12.6.3)

$$\lambda_1(P) = \text{Var}\left(\sum_{j=1}^p \xi_{1j} X_j\right) = \xi_1 \Sigma \xi_1 . \quad (12.6.13)$$

That is, ξ_1 is the maximizer of $\mathbf{a}^T \Sigma \mathbf{a}$.

The second principal component Z_2 is the linear combination $\sum a_{2j} X_j$ subject to $\sum a_{2j}^2 = 1$, $\mathbf{a}_1^T \mathbf{a}_2 = 0$, with the maximum variance, and so on. Also

$$\lambda_2(P) = \text{Var}\left(\sum_{j=1}^p a_{2j} X_j\right)$$

and $\sum_{l=1}^p \lambda_l(P) = \sum_{j=1}^p \text{Var} X_j$.

Even more, ξ_1 equivalently gives the direction of the line in R^p through the origin closest, on average, to \mathbf{X} . More precisely, let $\mathbf{x} = \alpha \mathbf{a}$, $\alpha \in R$, be the line through the origin in direction \mathbf{a} , where $\mathbf{a}^T \mathbf{a} = 1$. The distance between an arbitrary point $\mathbf{y} \in R^p$ and the line is $\mathbf{y}^T \mathbf{y} - (\mathbf{a}^T \mathbf{y})^2$ (Problem 12.6.4). Next consider an unobserved random vector \mathbf{X} in R^p with $E(\mathbf{X}) = \mathbf{0}$ and $\text{Cov}(\mathbf{X}) = \Sigma$. Finding the line $\alpha \mathbf{a}$ such that the expected squared distance of \mathbf{X} from the line is a minimum means maximizing

$$E(\mathbf{a}^T \mathbf{X})^2 = \mathbf{a}^T E(\mathbf{X} \mathbf{X}^T) \mathbf{a} = \mathbf{a}^T \Sigma \mathbf{a} .$$

Or in other words, ξ_1 solves the minimization problem

$$(\alpha_1, \xi_1) = \arg \min_{\alpha, \xi} \{ E|\mathbf{X} - \alpha \xi|^2 : \alpha \in R, |\xi| = 1 \}$$

We have shown that ξ_1 defines the one dimensional (dependent on $\mathcal{L}(\mathbf{X})$) linear subspace of R^p which is, on average, closest to \mathbf{X} . Similarly ξ_2 defines the line in R^p through the origin perpendicular to $\alpha \xi_1$ closest to \mathbf{X} , etc.

The eigenvectors and eigenvalues can be estimated from the empirical covariance matrix $\widehat{\Sigma}$ using established methods such as singular value decomposition, yielding $\widehat{\lambda}_j$ and $\widehat{\xi}_j = (\widehat{\xi}_{1j}, \dots, \widehat{\xi}_{pj})^T$, $1 \leq j \leq p$. The estimated eigenvalues are arranged in decreasing order and thresholded to 0 when they become too small. The resulting first $d < p$ eigenvectors and corresponding eigenvalues give a dimension reduction by replacing the original $\{\mathbf{X}_i ; 1 \leq i \leq n\}$ by the *sample principal components* $\{\widehat{\mathbf{Z}}_i ; 1 \leq i \leq n\}$, where

$$\widehat{Z}_{ij} = \sum_{k=1}^p \widehat{\xi}_{kj} X_{ik}, \quad 1 \leq j \leq d, \quad 1 \leq i \leq n .$$

These principal components are uncorrelated. It follows that if we regress $\{Y_i\}$ on $\{\widehat{Z}_{ij}\}$ using least squares, the fitted regression is

$$\widehat{\mathbf{Y}} = \bar{Y} \mathbf{1} + \sum_{j=1}^d \widehat{\eta}_j \widehat{\mathbf{Z}}_j$$

where $\widehat{\mathbf{Z}}_j = (\widehat{Z}_{1j}, \dots, \widehat{Z}_{nj})^T$ and $\widehat{\eta}_j = \sum_{i=1}^n (Y_i - \bar{Y})(\widehat{Z}_{ij} - \bar{\widehat{Z}}_j) / \sum_{i=1}^n (\widehat{Z}_{ij} - \bar{\widehat{Z}}_j)^2$. In the context of the regression model where we observe $(\mathbf{X}_1, Y_1), \dots, (\mathbf{X}_n, Y_n)$ i.i.d., $\mathbf{X}_i \in R^p$, $Y \in R$ with p large, sparse modeling with small d makes it possible to do classification and prediction using $(\widehat{\mathbf{Z}}_1, Y_1), \dots, (\widehat{\mathbf{Z}}_n, Y_n)$.

Association studies

In the preceding PCA approach the effects of the individual original variables on the response is not apparent. In genome-wide association studies (GWAS) this is fixed by singling out one variable X_k one at a time and regressing Y_i on $\{X_{ik}, \widehat{Z}_{ij}\}$, $1 \leq i \leq n$, $1 \leq j \leq d$. When the number of X 's p is larger than n , this approach is difficult to implement because \widehat{Z}_{ij} is unstable. However, for testing the hypothesis that X_k and Y are uncorrelated, the software package “eigensoft” give tests that produce approximately correct Bonferroni p -values for balanced case-control studies where the number of cases and the number of controls are nearly the same, and other assumptions are satisfied (see Yang (2013), and Yang, Doksum, and Tsui (2014)).

Remark 12.6.2. Because Σ and $\widehat{\Sigma}$ depend on the scale of the X 's, the corresponding correlation matrices are often used in their place. \square

Remark 12.6.3. One approach to *sparse PCA* is to use only the first 10 principal components which often can be justified by checking that $\widehat{\lambda}_j$, $j > 10$, are nearly zero. When p is large, using $d = 10$ PC's often produce nearly the same results as d chosen by model selection procedures, e.g. cross-validation. Another approach is to replace $\sum_{j=1}^p a_{kj}^2 = 1$ in the optimization PC problem with adjusted versions of the Lasso type condition $\sum_{j=1}^p |a_{kj}| = 1$. Adjustments are made to obtain convex optimization problems. See Hastie, Tibshirani, and Wainwright (2015) and Remark 8.3.2.

We do not discuss this further here but note again that for large p , the empirical eigenvectors and eigenvalues can be very poor estimates of the population quantities of interest. Here too, as in regression, we typically regularize reflecting our beliefs that the data nearly live in a much lower dimensional subspace. For instance, $\lambda_{k+1} = \dots = \lambda_p = 0$ correspond to a sparse subspace where $p - k$ of the predictors are linearly dependent on the others and \mathbf{X} is in a k dimensional subspace of R^p . We will discuss such issues even more briefly in our last Section 12.7 in which we point to enormous areas, barely touched by this book, and current areas of research interest.

Summary

In this section we have briefly focused on the huge body of research on model selection and dimension reduction of data, one of the most important issue in modern statistical practice. The reduction in model dimension centered on an asymptotic Bayes criteria while

dimension reduction of data briefly discussed principal component analysis. We had already considered, in a slightly different context, dimension reduction via Mallows' C_p and cross validation in Section 12.4.

12.7 Topics Briefly Touched and Current Frontiers

Complex data structures

As we have noted in the past, modern statistical data are full of structures. To the old types of temporal data, time series, have now been added spatio-temporal data, matrix time series, images which can be thought of as spatial data, but have a special complexity and structure, networks as in genomics or the World Wide Web, dynamical computer models, hierarchical structures such as text, and so on.

Beyond a few examples in time series such as autoregressive processes we have not discussed modelling such situations parametrically or nonparametrically mainly because models are evolving for these complex situations.

For stationary Gaussian time series, the spectrum defines the most general semiparametric model and its estimation and its analysis has been highly developed (see, for example, Brillinger (2001)). However, just as with linear regression, the tool of choice in the Gaussian case, while the spectral techniques have been of great value even in the absence of Gaussianity, robustness and other considerations have led to time domain modelling.

Here semiparametric models such as general autoregressive and more general Markov and hidden Markov models have been used and analyzed. Asymptotics as the time series lengthens have been the major theoretical tools. Bootstrap and other Monte Carlo techniques have been developed for some of these models (Politis, Romano and Wolf (1999), Bühlmann and Künsch (1995), Künsch (1989), and Bickel, Götze and van Zwet (1997)).

Hierarchical, latent variable and Bayesian graphical models

There has been an enormous revival of Bayesian modelling. It has taken the form of hierarchical models, or, more generally, Bayesian graphical models of the type also familiar in the frequentist world, such as factor analysis, random effects, and mixed models, i.e. models where the observed variable structure is postulated to be due to unobservable or unknown hidden variables. A mixed model, for example, might be modelling genetic factors in Type II diabetes. To what extent do genetic factors account for Type II diabetes? This development has been driven by the possibility of incorporating partial scientific knowledge into the statistical model in this way and also because, with the advent of the MCMC techniques discussed in Section 10.4, computation of posterior distributions becomes, in principle, feasible in complex situations. Accounts may be found in Berger (1985) and Gelman, Carlin, Stern, Dunson, Vehtari and Rubin (2014). These approaches have proved successful in the context of prediction and machine learning and are favored by many Bayesian, frequentists, or the most common type of statistician, these days, pragmatists. See Wainwright and Jordan (2008). The reason for this, as we indicated earlier, is that no matter how a prediction algorithm is constructed its value is judged nonparametrically through checking predictions from cross validation based on training samples on known cases.

From a scientific point of view there is an interest in establishing causal relationships, importance of variables, etc. Equivalently, we may start with a graph of variables (nodes) and edges (associative relations). We want to end with a directed graph where the directions represent causation. This can be problematic. We typically observe association between variables which can be causally explained in many ways. For instance, if we observe $(X, Y) \sim N_2(0, 0, 1, 1, \rho)$ we can't tell whether Y was generated from $Y = \rho X + \varepsilon_1$, $\varepsilon_1 \perp X$, $\varepsilon_1 \sim N(0, 1 - \rho^2)$, or from $X = \rho Y + \varepsilon_2$ with $\varepsilon_2 \perp Y$, $\varepsilon_2 \sim \varepsilon_1$. However, some causative relations can be ruled out as being inconsistent with others. For instance, a loop in the graph, $A \rightarrow B \rightarrow C \rightarrow A$ is problematic since either A causes B or B causes A . Important applications of these notions have appeared in statistics and biostatistics — see Pearl (2009), Rubin (1990), and van der Laan and Robins (2003). Not infrequently, the hierarchical Bayesian approach appears dangerous. Distributional and causal assumptions are thrown in at various points of the hierarchy only for convenience of computation and if final posterior probabilities are taken seriously, it is difficult to disentangle what is coming from the data, what is coming from reasonable scientific assumptions, and what is coming from convenience assumptions. This is a problem shared by frequentists building hierarchical models as much as Bayesians. It seems plausible that what matters in a complex hierarchical model are the parametric assumptions made along the way rather than whether one puts a prior on at the root of a hierarchical model step or uses maximum likelihood. This last remark is supported by the Bernstein-von Mises theorem (Theorems 5.5.2 and 6.2.3).

Modelling Deterministic Situations

Computer models of great complexity for traffic, wildfires, and atmospheric science are now common; see Sacks et al (1989). Although, in principle, these situations are deterministic and can, again in principle, be modelled only by subjective Bayesians, in fact probabilistic models are very useful and important tools in their analysis. For instance, numerical PDE computations running on super computers for climate prediction are “assimilated” with data (with stochastic error structure) to produce far more accurate predictions than could be done using data alone.

Computational questions

An issue that high dimensionality and possibly sample size have made paramount is computation time. Despite the ever increasing speed and capabilities of computers, running all possible regressions on a set of n observations with p covariates requires on the order of $2^p n$ operations, an impossibly large number if p is even in the hundreds. Things are, of course, much worse for optimization algorithms for functions on R^p arising routinely in maximum likelihood or for that matter for Markov Chains Monte Carlo simulations in high dimensional spaces. It has been clear that to get good statistical performance, we need to add computational performance, to get so called Polynomial (P) rather than Non-Polynomial (NP) algorithms although even that may be excessive. That is, if p characterizes the number of parameters, numbers, objects that the algorithm needs to deal with, then, for some α , a (P) algorithm takes at most on the order of p^α operations to yield an answer whatever be the values of the parameters. An (NP) algorithm either will need more than p^α for any α for some situations or have an unknown worst case performance. The issues are,

in fact, rather subtle since as for decision procedures, this is based on a worst case analysis and an algorithm may perform well in practice even though it is known to be NP. A famous example is the simplex method for solving linear frequency problems. These issues are only now starting to be treated systematically; see Chandrasekaran and Jordan (2013) for instance.

Behaviour of classical procedures when p, n are comparable or $p \gg n$

It has been known for some time in the statiscal physics and probability communities that if $p/n \rightarrow c > 0$, then the behaviours of various statistical functions of the empirical covariance matrix are very different than if $p/n \rightarrow 0$. A striking example studied by Wachter (1978), Geman(1984), and more recently by El Karoui (2005) following classical work in statistical physics is that of the empirical covariance matrix of n i.i.d vectors which are multivariate normal with mean 0 and variance covariance matrix the $p \times p$ identity. The distribution of the eigenvalues of this matrix does not tend to point mass at 1, as they should and do for fixed p . Moreover, for slight perturbations to the identity covariance, the empirical eigenvectors have strange properties. Thus, principal component analysis (PCA) becomes suspect. Similar difficulties arise for the classical Gaussian linear discriminant function (Bickel and Levina (2008a,b)). These can be studied by using models with different forms of sparsity and regularization such as those we have analysed earlier in this chapter for regression. See Remark 8.3.2 for references to these studies of sparse PCA.

It is worth considering situations where the vector of parameters is unstructured a priori such as the $p \times p$ population covariance matrix when sampling from a p -dimensional population and situations where some of the population parameters are structured, for instance, the Fourier coefficients of a smooth density function appearing in a semiparametric model. Here, smoothness corresponds to a sparsity assumption i.e. most of the coefficients vanish. Examples of this type of large p situation have been dealt with in Sections 12.3, 12.4, and 12.6. Another example of situations where many parameters appear but sparsity is assumed is the following:

Studying many parameters simultaneously from testing and related points of view

In genomics and other fields, it has become routine to test many null hypotheses at the same time, in the expectation that only some small fraction of these if any will not be null. The classical approach here has been to use the Bonferroni inequality (see Section 4.4). That is, if there are N hypotheses, to ensure that the probability of type I error, in the sense of making any mistake whatsoever, is no more than α , one tests at level α/N . The probability of at least one Type I error when testing N hypotheses is called the *family-wise error rate (FWER)*. That is, if V denotes the number of true null hypotheses rejected, then $FWER = P(V \geq 1)$. If the model distribution P assumes that all N null hypotheses are true, then $FWER \leq \alpha$ is called *weak control* of FWER, while if P is not required to satisfy this assumption, $FWER \leq \alpha$ is called *strong control* of FWER. The popularity of the *Bonferroni multiple testing* method is in part due to:

Proposition 12.7.1. *Suppose we have available tests $\varphi_1, \dots, \varphi_N$ for testing N null hypotheses H_1, \dots, H_N . Let \hat{p}_j denote the p-value of φ_j for testing H_j , $j = 1, \dots, N$. Then the multiple testing procedure that rejects the H_j with $\hat{p}_j \leq \alpha/N$ strongly controls FWER.*

Proof. Let $J_0 = \{j : H_j \text{ is true}\}$ and $N_0 = \text{cardinality of } J_0$. Then

$$FWER = P(V \geq 1) = P\left(\bigcup_{j \in J_0} \{\hat{p}_j \leq \alpha/N\}\right) \leq \sum_{j \in J_0} P(\hat{p}_j \leq \alpha/N) \leq \frac{N_0}{N} \alpha \leq \alpha.$$

□

Holm (1979) constructed a simultaneous testing procedure that strictly improves on the Bonferroni method: Let $p_{(1)} \leq \dots \leq p_{(N)}$ denote the ordered p -values for N tests. Reject H_j if

$$p_{(j)} \leq \frac{\alpha}{N - j + 1}.$$

A proof of strong control of FWER for the Holm method is outlined in Problem 12.7.1.

When the number of hypotheses is large and $\alpha \leq 0.1$, the Bonferroni and Holm methods rule out all but the very strongest signals and then makes discovery of strong but less extreme signals impossible. To remedy this failing, Benjamini and Hochberg (1995) proposed replacing the Bonferroni criteria by bounding the *False Discovery Rate* (FDR), the expected fraction of falsely rejected hypotheses to the total number of rejections, and derived a simple rule for doing so with desirable properties. This rule rejects H_j if $p_{(j)} \leq p_{(\hat{j})}$, where

$$\hat{j} = \arg \max \{j : p_{(j)} \leq \alpha/N\}.$$

If the p -values are independent for j with H_j true, the rule controls the FDR at level α for such j . An overview may be found in papers of Benjamini and Hochberg (1995), Lehmann, Romano and Shaffer (2003), Dudoit, Shaffer and Boldrick (2003), and others.

The monograph of Efron (2010) gives an important overview of issues in modern large data contexts from an empirical Bayesian point of view. This is an area of continuing research — see, for example, the problems and a related solution to a different problem in Q. Li et al (2011) as well as the study of other features of the distribution of p values which underlies the work of Benjamini and Hochberg.

Dimension reduction, clustering, and related concepts

Sufficiency is not usually a useful concept in the high dimensional world we are now dealing with. But dimension reduction and the related concepts, clustering, or, as the information theorists call it, “lossy compression,” are essential for analysis and understanding. Such methodologies interact intimately with regularization methods we have mentioned. If we do not regularize or, more specifically, set some parameters to 0, the analysis we then give may end up giving us an incorrect qualitative picture of the mechanics behind the data.

We end with an augmentation of a famous quotation from Box (1979):

“Models, of course, are never true but fortunately it is only necessary that they be useful” to which we would like to add

“Useful models generate conclusions which can point to new verifiable questions.”

Randomized methods as part of statistics

A number of procedures have arisen in recent years in addition to the bootstrap, MCMC and similar methods, which provide new methodology for high dimensional data. An intriguing example now called “sketching” is a class of methods for dimension reduction.

Here random projections of the data points into lower dimensional subspaces are analyzed, yielding both computational and possibly inferential advantages at a possible cost of information loss. We see this as an emerging frontier on the borders of statistics and computation. See Baraniuk, Cevher, and Wakin (2010) as a recent nonstatistical reference.

A method using randomization intrinsically that we discuss more fully is Random Forests (RF). This method for classification and regression, or supervised learning, and clustering or unsupervised learning, was introduced by Breiman (2001), (see also Amit and Geman (1997)), as a follow up to CART. In this approach, ensembles of random classification or regression trees are grown on randomly chosen training sets. Classification is determined by majority voting, and regression by averaging empirical predictors produced the trees. The randomness is introduced in two ways.

- 1) Each tree is grown on a bootstrap sample of the same sample size as the training sample. Each bootstrap sample, because of sampling with replacement, leaves out about 30% of the data vectors. This leaves, as a complement of the bootstrap training sample, an independent subsample for assessing importance of predictors (also called features). But the trees are typically dependent since the training samples overlap.
- 2) Even more significantly, when the vector of features/predictors is large, each tree uses only a small random subset of features on which to determine a split. That is, the CART splits are each made by optimizing over their own small subset of variables, of the order of 5, with sets varying randomly from split to split. For classification, although this is not necessary, the trees are grown to purity. Each leaf contains only cases of one data type.

Various advantages accrue. The methods are much more stable than CART since they smooth by averaging, even though the trees are not independent. More significantly, the use of different variables at each split and the use of subsamples enables efficient exploration of high dimensional spaces. The details of RF make analysis of the method difficult. However, its empirical performance is superb, and recently research and understanding have advanced. See Hastie, Tibshirani and Friedman (2009) for recent references, and Scornet, Biau, and Vert (2015).

Summary

In this final section we have listed a number of topics which are of active interest at least to us. Many equally workable directions have not been mentioned. But this is a book of “Selected Topics.”

12.8 Problems and Complements

Problems for Section 12.1

1. With the notation of Section 12.1, suppose, initially, $\mathbf{Z} = (1, Z_1)^T$ where $Z_1 \in [0, 1]$. Show that if we enlarge \mathbf{Z} to $\mathbf{Z}^* = (1, Z_1, Z_1^2, \dots, Z_1^d)^T$, then for any continuous $\mu(\mathbf{Z}_1)$, and $\varepsilon > 0$, there exists d , $\mathbf{a} \equiv (a_0, \dots, a_d)$ such that,

$$E(\mu(Z_1) - \mathbf{a}^T \mathbf{Z}^*)^2 \leq \varepsilon^2 .$$

Hint: Use the Weierstrass approximation theorem.

- 2.** Suppose $\mathbf{x}_1, \dots, \mathbf{x}_N$ are N points in general position in R^d and (y_1, \dots, y_N) are associated signs $y \in \{-1, 1\}$. Show that there exists a 1–1 map $\Phi : R^d \rightarrow R^M$ for some $M \geq d$ and a vector $\mathbf{v} \in R^M$ such that

$$\operatorname{sgn}(\mathbf{v}^T \mathbf{x}_i) = y_i, \quad i = 1, \dots, N.$$

Hint: Let $\mathbf{e}_1, \dots, \mathbf{e}_N$ be the coordinate axes in R^N , map \mathbf{x}_i onto \mathbf{e}_i , and let $\mathbf{v} = (y_1, \dots, y_N)^T$.

Problems for Section 12.2

- 1.** Let $\hat{f}_h(\mathbf{x})$ denote the multivariate convolution kernel estimate. Assume that K has convex compact support and is of order 1; and assume that $D^2 f(\mathbf{x}) \leq M$ for all $\mathbf{x} \in S(f)$ and all f , some $M > 0$. Show that for \mathbf{x} in the interior S^0 of $S(f)$

$$E_F \hat{f}_h(\mathbf{x}) = f(\mathbf{x}) + O(|h|^2)$$

$$\operatorname{Var}_F f_h(\mathbf{x}) = (n|h|^d)^{-1} f(\mathbf{x}) \int K^2(\mathbf{u}) d\mathbf{u} (1 + o(1)).$$

Hint: Follow the steps in the proofs of Propositions 11.2.1 and 11.2.2.

- 2.** Assume that $0 < \int |D^2 f(\mathbf{x})|^2 d\mathbf{x} < \infty$. Let $\hat{f}_h(\mathbf{x})$ be as defined by (12.2.1).

- (a) Show that AIMSE expression (12.2.5) for $\hat{f}_h(x)$ is minimized by

$$h = a_f n^{-\frac{1}{d+4}}$$

for some $a_f > 0$. Give an expression for a_f .

- (b) Show that the minimum of the AMISE expression (12.2.5) is

$$b_f n^{-\frac{4}{d+4}}$$

for some $b_f > 0$. Give an expression for b_f .

- (c) Evaluate the constant $a_f \equiv a(\Sigma)$ in part (a) of this problem when \mathbf{X} is $\mathcal{N}_d(\mathbf{0}, \Sigma)$ and $K_h(\mathbf{u})$ is a product kernel with $K_j(u) = 1[|u| \leq 1]/2$. Let $\widehat{\Sigma}$ be the sample covariance matrix. Then

$$\widehat{h} = a(\widehat{\Sigma}) n^{-\frac{1}{d+4}}$$

is an estimate of the h that yields the optimal asymptotic IMSE if the true density is $\mathcal{N}(\boldsymbol{\mu}, \Sigma)$. Here $\mathcal{N}_d(\mathbf{0}, \Sigma)$ is called a *reference distribution*. See Section 11.2.3.

- 3. Boundary kernel density estimates.** Consider $\mathbf{h} = (h, \dots, h)$. Suppose the support of f is a rectangle $\prod_{j=1}^d [a_j, b_j]$ and that $D^2 f(\mathbf{x}) \leq M$ for all $\mathbf{x} \in S(f)$ and some $M > 0$. Also assume that $K(\mathbf{u}) = \prod_{j=1}^d K_0(u_j)$, where K_0 is symmetric and has support $[-1, 1]$. Let $B_\lambda = \prod_{j=1}^d [a_j, a_j + \lambda h]$, $0 < \lambda < 1$, denote the left boundary region, and let $\mathbf{x}_h = (a_1 + \lambda h, \dots, a_d + \lambda h)$ denote a point in B_λ .

(a) Show that the kernel estimate is asymptotically biased in B_λ . That is, as $h \rightarrow 0$

$$|\widehat{f}_h(x_h) - f_h(x_h)| = O_P(1).$$

(b) Let $W(\mathbf{x}) = \int_{S(f)} K_h(\mathbf{z} - \mathbf{x}) d\mathbf{z}$ and define

$$\tilde{f}_h(\mathbf{x}_h) = \frac{\sum_{i=1}^n K_h(\mathbf{X}_i - \mathbf{x})}{nW(\mathbf{x})} \mathbf{1}(\mathbf{x} \in S(f)).$$

Show that $\tilde{f}_h(\mathbf{x}_h)$ is asymptotically unbiased in B_λ , That is, as $h \rightarrow 0$,

$$|\tilde{f}_h(\mathbf{x}_h) - f_h(x_h)| = o_P(1).$$

Hint: See Problem 11.3.1. Jiang and Doksum (2003) give further asymptotic properties of $\tilde{f}_h(\mathbf{x})$.

(c) Show that $\tilde{f}_h(\mathbf{x}) = \beta_0(\mathbf{x}, \widehat{F})$, where

$$\beta_0(\mathbf{x}, F) = \arg \min E\{|[f(\mathbf{Z}) - \beta_0]^2 | \mathbf{X} = \mathbf{x}\}$$

and $(\mathbf{Z} | \mathbf{X} = \mathbf{x})$ has density

$$q_h(\mathbf{z} | \mathbf{x}) = \frac{K_h(\mathbf{z} - \mathbf{x}) \prod_{j=1}^d \mathbf{1}(a_j \leq z_j \leq b_j)}{W(\mathbf{x})}.$$

Hint: See Section 11.3.

4. Let D^k be the differential operator defined in Section B.8 and let $\| \cdot \|_\infty$ denote sup norm over \mathcal{S} , and let m_j and ν_j be as defined in Section 11.6.2. Suppose that $\inf\{f(\mathbf{x}) : \mathbf{x} \in \mathcal{S}\} > 0$, $\|Df\|_\infty < \infty$, $\|D^{J+2}\mu\|_\infty < \infty$, $\|D\sigma^2\|_\infty < \infty$, $m_{J+2}(K) < \infty$, $\nu_5(K) < \infty$, $m_j(K) = 0$, $j = 1, \dots, J$, and J is even. Show that

(a) as $n \rightarrow \infty$, $h \rightarrow 0$, $nh^3 \rightarrow \infty$, uniformly for $\mathbf{x} \in \mathcal{S}$,

$$\text{MSE}(\widehat{\boldsymbol{\mu}}_{\text{NW}}(\mathbf{x}) | \mathbf{X}^{(n)}) \equiv E\{|\widehat{\boldsymbol{\mu}}_{\text{NW}}(\mathbf{x}) - \boldsymbol{\mu}(\mathbf{x})|^2 | \mathbf{X}^{(n)}\} = O_P(h^{J+2}) + O_P(n^{-1}h^{-d}).$$

(b) The minimizer of the asymptotic version of $\text{MSE}(\widehat{\boldsymbol{\mu}}_{\text{NW}}(\mathbf{x}) | \mathbf{X}^{(n)})$ is of the form

$$h = c \left[\frac{1}{(J+2)n} \right]^{\frac{1}{J+2+d}}$$

which leads to $\text{MSE}(\widehat{\boldsymbol{\mu}}_{\text{NW}}(\mathbf{x}) | \mathbf{X}^{(n)})$ of order

$$n^{-\frac{J+2}{J+2+d}}.$$

Hint: See Section 11.6.2.

5. In Example 12.2.2, show that if $\hat{\Pi}_j = C^{-1}$, $1 \leq j \leq C$, and if $\hat{p}_{jk}(\cdot)$ denotes the k th nearest neighbour estimate, then the classifier based on replacing $p_j(\cdot)$ with $\hat{p}_{jk}(\cdot)$ in the Bayes rule classifies I_{n+1} as $I_{\hat{j}_k}$ where \hat{j}_k is the subscript on the k th nearest neighbour to X_{n+1} .

6. Assume that $d = 1$, that the support of f is $[a, b]$ with a and b finite, that $\sup\{|f''(x)| : x \in [a, b]\} \leq M$, and that $k = k_n \rightarrow \infty$, $n^{-1}k_n \rightarrow 0$ as $n \rightarrow \infty$. Then the k th nearest neighbour classifier is consistent.

Hint: See Section 11.2 and Problem 11.4.3.1

7. Assume that f satisfies the conditions of Problem 6 above. Show that the bandwidth \hat{h} that corresponds to the nearest neighbour classifier satisfies $\hat{h} = O_P(\frac{1}{n})$.

8. Suppose (\mathbf{X}_i, Y_i) are i.i.d. where $Y_i = 1$ with probability π_0 , $Y_i = -1$ with probability $1 - \pi_0$, and $\mathbf{X}_i \sim \mathcal{N}(\boldsymbol{\mu}_0 Y_i, \Sigma_0)$ given Y_i . Assume 0-1 loss and that $\boldsymbol{\mu}_0$, Σ_0 and π_0 are known.

(a) Show the Bayes classification rule for Y_{n+1} given \mathbf{X}_{n+1} is given by

$$\begin{aligned}\hat{Y}_{n+1} &= 1 \text{ if } (\mathbf{X}_{n+1} - \boldsymbol{\mu}_0)^T \Sigma_0^{-1} (\mathbf{X}_{n+1} - \boldsymbol{\mu}_0) \\ &\quad - (\mathbf{X}_{n+1} + \boldsymbol{\mu}_0)^T \Sigma_0^{-1} (\mathbf{X}_{n+1} + \boldsymbol{\mu}_0) > 2 \log \frac{\pi_0}{1 - \pi_0} \\ &= -1 \text{ otherwise}\end{aligned}$$

or equivalently

$$\hat{Y}_{n+1} = 1 \text{ iff } \boldsymbol{\mu}_0^T \Sigma_0^{-1} \mathbf{X}_{n+1} > \log \frac{1 - \pi_0}{\pi_0} .$$

(b) Show that the Bayes risk if $\pi_0 = \frac{1}{2}$ is $R_\theta = 1 - \Phi(\boldsymbol{\mu}_0^T \Sigma_0^{-1} \boldsymbol{\mu}_0)$.

9. Assume (\mathbf{X}_i, Y_i) , $i = 1, \dots, n$ are i.i.d. as (\mathbf{X}, Y) with $Y = 1$ or -1 and

$$\text{logit}(P(Y = 1 | \mathbf{X})) = \sum_{j=1}^d \alpha_j X_j + \alpha_0$$

where $\text{logit}(u) = \log(u/(1-u))$, $0 < u < 1$. This is the logistic regression model.

(a) Show that the Bayes classification rule given $\mathbf{X}_{n+1} = (X_{1,n+1}, \dots, X_{d,n+1})^T$ is

$$\hat{Y}_{n+1} = 1 \text{ iff } \sum_{j=1}^d \alpha_j X_{i,n+1} > -\alpha_0 .$$

(b) Suppose that α is unknown. Show that as $n \rightarrow \infty$ the rule

$$\widehat{Y}_{n+1} = 1 \text{ iff } \sum_{j=1}^d \widehat{\alpha}_j X_{j,n+1} + \widehat{\alpha}_0 > 0$$

where $\widehat{\alpha} = \arg \max \left\{ \sum_{i=1}^n \left\{ \sum_{j=1}^d (\alpha_j X_{ji} + \alpha_0) - \log(1 + e^{\sum_{j=1}^d \alpha_j X_{ji} + \alpha_0}) \right\} \right\}$ converges to the Bayes rule.

- (c) Show that if the density of \mathbf{X} is $(1-\pi_0)\mathcal{N}(\boldsymbol{\mu}_0, \Sigma_0) + \pi_0\mathcal{N}(-\boldsymbol{\mu}_0, \Sigma_0)$, then the Bayes rule of (a) coincides with that of Problem 12.2.8.
- (d) Assume 0-1 classification loss. If we replace $\Sigma_0, \boldsymbol{\mu}_0, \pi_0$ by their MLEs under the Gaussian model mentioned in Section 12.2.3, show that the resulting classifier (LDA) is, as $n \rightarrow \infty$, no better than that of the logistic regression model classifier and, in general, is worse.

10.(a) Show that the minimizer of

$$L(x, \mu) \equiv (x - \mu)^2 + \lambda|\mu|$$

in μ is given by *soft thresholding*

$$\begin{aligned}\widehat{\mu} &= x, \quad |x| \leq c. \\ &= x - c(sgn x), \quad |x| > c\end{aligned}$$

(b) Suppose Y is $n \times 1$ and \mathbf{X} is $n \times p$. Deduce that if $\lambda > 0$ and $n > d$ the *Lasso*:

$$\text{Minimize } |Y - \mathbf{X}\beta|^2 + \lambda|\beta|_1$$

has a unique sparse solution $\widehat{\beta} = (\widehat{\beta}_1, \dots, \widehat{\beta}_d)^T$ where for some $S(\lambda) \subset \{1, \dots, d\}$,

$$\widehat{\beta}_j = 0, \quad j \in S(\lambda).$$

(c) Show that if we replace $= \sum_{j=1}^d |\beta_j|$ by $|\beta|_c = \sum_{j=1}^d |\beta_j|^c$, $c > 1$, then if Y has a density $\widehat{\beta}$ is never sparse with probability 1.

Hint. Use Lagrange multipliers or equivalently the Karush-Kuhn-Tucker conditions (KKT).

Remark. The adaptive Lasso (Zou (2006)) replaces $\lambda|\beta|_1$ with $\lambda \sum w_j |\beta_j|$, where w_j are weights of the form $1/|\beta_j^*|$ with β_j^* consistent estimates of β_j , $1 \leq j \leq d$. It has desirable consistency and oracle properties. Software is available online.

11. Let $M(H)$ denote the margin of the separating hyperplane H . Show that the hyperplane that maximizes $M(H)$ can be found by solving (12.2.19).

Hint. $H = \{\mathbf{x} : \beta^T(\mathbf{x} - \mathbf{x}_0) = 0\}$ where $\beta^T \mathbf{x}_0 = \beta_0$ and $\pi(\mathbf{x}|\beta) = (\beta^T \mathbf{x})\beta$.

12. Verify (12.2.29).

13. Argue geometrically that the support vector machine optimization problem, may be stated as in (12.2.19).

14. Assume model (12.2.30). For M fixed show that the first Adaboost algorithm converges for fixed $n > M$ to the classifier specified by the sample version of (12.2.30) i.e.

$$\delta(\mathbf{Z}) = 1 \text{ iff } \sum_{j=1}^M \hat{\alpha}_j h_j(\mathbf{Z}) > 0$$

where $\hat{\boldsymbol{\alpha}} = \arg \min Q(\boldsymbol{\alpha}, \hat{P}_n)$ is assumed to exist.

Hint. Apply the methods used to prove Theorem 12.2.2.

15. Assume model (12.2.30). Show that as $n \rightarrow \infty$,

- (a) the minimizer of $Q(\boldsymbol{\alpha}, \hat{P}_n)$ converges to the minimizer of $Q(\boldsymbol{\alpha}, P)$, and
- (b) if the true distribution $P(\mathbf{z}, y)$ is such that \mathbf{Z} has finite support, then the first Adaboost algorithm is Bayes consistent in the sense of (12.2.10).

16. Suppose $\text{logit}(P(Y = 1|\mathbf{Z})) = \sum_{j=1}^{\infty} \alpha_j h_j(\mathbf{Z})$, $\alpha_j \neq 0$ for all j and Adaboost 1 is iterated forever using the function h_1, h_2, h_3, \dots . Show that Adaboost 1 doesn't converge.

17. (a) Show that minimizing $R(\boldsymbol{\alpha}, P)$ in (12.2.31) with $\lambda = (2\gamma)^{-1}$ is equivalent to solving (12.2.24).

(b) Show that if $\lambda = 0$ in (12.2.31) and the Bayes classifier for model (12.2.30) is of the form $\delta_B = \text{sgn}(\sum \alpha_j^* h_j(\mathbf{Z}))$, then δ_B minimizes both $R(\boldsymbol{\alpha}, P)$ with $\lambda = 0$ and the minimum Bayes risk $P[Y \sum_{j=1}^M \alpha_j^* h_j(\mathbf{Z}) < 0]$.

(c) Show that under the assumptions of (b) support vector machines can be put in the framework of boosting.

18. (Breiman et al. (1984)). Define a *measure of impurity* of a finite probability distribution, $\{p_1, \dots, p_K\} \in \mathcal{S}_K$, the K simplex, as a function $\phi : \mathcal{S}_K \rightarrow R$ such that

- (i) ϕ is maximized uniquely for $p_j = \frac{1}{K}$, $1 \leq j \leq K$
- (ii) ϕ is minimized for any $\mathbf{e}_j \equiv \{\delta_{ij} : i = 1, \dots, K\}$ where $\delta_{ij} = 1(i = j)$.
- (iii) ϕ is symmetric.

- (a) Show that with $0 \log 0 = 0$ the following are both impurity measures with minimums zero:

$$\phi_1 = \sum_{j=1}^K p_j^2 ; \quad \phi_2 = - \sum_{j=1}^K p_j \log p_j .$$

- (b) Show that both ϕ_1 and ϕ_2 are strictly concave functions of \mathbf{p} , that is,

$$\phi(\alpha \mathbf{p}_1 + (1 - \alpha) \mathbf{p}_2) \geq \alpha \phi(\mathbf{p}_1) + (1 - \alpha) \phi(\mathbf{p}_2)$$

for all $\alpha \in [0, 1]$ with equality iff $\mathbf{p}_1 = \mathbf{p}_2$.

Hint. Because the sum is concave, it is enough to check concavity for $k = 1$.

19. Suppose $Y \in \{1, \dots, K\}$ so that there are K classes. Let t be a node with class fractions (p_1, \dots, p_K) . Split the node into K pieces with corresponding compositions $\pi_j(p_{1j}, \dots, p_{Kj})$, $j = 1, \dots, K$, $\sum_{i=1}^K p_{ij} = 1$ where π_j is the fraction assigned to descendant node j .

- (a) Show that for ϕ_1, ϕ_2 as in Problem 18, for any such split S , if we define the total *purity gain* of the descendant nodes, by

$$\Delta(S) \equiv \sum_{j=1}^K \pi_j \phi(p_{1j}, \dots, p_{Kj}) - \phi(p_1, \dots, p_K),$$

then, $\Delta(S) \geq 0$.

Hint. Use concavity.

- (b) Show that for $K = 2$, $\phi = \phi_2$, the splitting rule described in the text as the T algorithm is

$$\widehat{S} \equiv \arg \max_S \Delta(S).$$

- (c) Show that $\Delta(S) = 0$ iff $p_{ij} = p_i$ for all i, j .

20. (a) Show that if f, g are densities of P and Q with respect to μ ,

$$\sup_A |P(A) - Q(A)| = \int |f - g| d\mu.$$

(b) Deduce that if $\mathcal{X} = \{x_1, \dots, x_m\}$ and $(p_1, \dots, p_n), (q_1, \dots, q_m)$ are corresponding vectors of probabilities,

$$\max_A \left| \sum_{j \in A} \{p_j : j \in A\} - \sum_{j \in A} \{q_j : j \in A\} \right| = \sum |p_j - q_j|$$

Hint. $|P(A) - Q(A)| = |\int_A (f - g) d\mu|$. Take $A = \{x : \frac{q}{f}(x) \geq 1\}$.

21. Consider the following splitting scheme, called twicing, for a node t with fractions (p_1, \dots, p_K) . Let $\mathcal{C} = \{j_1, \dots, j_m\}$, $\mathcal{C} \subset \{1, \dots, K\}$, $\bar{\mathcal{C}} = \{1, \dots, K\} - \mathcal{C}$. Split the node t into t_L, t_R such that the node t_L corresponds to (p_{1L}, \dots, p_{KL}) and similarly for t_R . If π_L, π_R are the fractions of the two nodes, let $p(1|t) = \sum \{p_j : j \in \mathcal{C}\}$. Define $p(1|t_L), p(1|t_R)$ similarly. Then

$$p(1|t) = \pi_L p(1|t_L) + \pi_R p(1|t_R).$$

Show that the purity gain from a split into two classes \mathcal{C} and $\bar{\mathcal{C}}$ is

$$\Delta(S, \mathcal{C}) = -\{2p(1|t)(1-p(1|t)) - \pi_L p(1|t_L)(1-p(1|t_L)) - \pi_R p(1|t_R)(1-p(1|t_R))\}.$$

22. Suppose $K = 2$ and $\phi = \phi_2$ in the previous problem. Given a split S , let t, t_L, t_R be the parent left and right nodes with compositions $(p(1|t), 1 - p(1|t))$ and

$$\pi_L(p(1|t_L), 1 - p(1|t_L)), \pi_R(p(1|t_R), 1 - p(1|t_R))$$

where $p(1|t)$, $p(1|t_L)$, $p(1|t_R)$ are the (conditional) probability masses for class 1 and $\pi_L = 1 - \pi_R$ is the marginal probability of being assigned to t_L . Show that

$$\Delta(S, \mathcal{C}) = 2\pi_L\pi_R(p(1|t_L) - p(1|t_R))^2.$$

Hint. Write $p(1|t_L) = \frac{1}{2} + \Delta_1$, $p(1|t_R) = \frac{1}{2} + \Delta_2$, $p(1|t) = \frac{1}{2} + \pi_L\Delta_1 + \pi_R\Delta_2$. Note that $p^2(1|t) + (1 - p(1|t))^2 = 1 - 2p(1|t)(1 - p(1|t))$.

$$\Delta(S, \mathcal{C}) = -2\left\{\pi_L\left(\frac{1}{4} - \Delta_1^2\right) + \pi_R\left(\frac{1}{4} - \Delta_2^2\right) - \left(\frac{1}{4} - (\pi_L\Delta_1 + \pi_R\Delta_2)^2\right)\right\}.$$

23. Let

$$p(\mathcal{C}|t) = \sum \{p_j : j \in \mathcal{C}\}, p(\mathcal{C}|t_L) = \sum \{p_{jL} : j \in \mathcal{C}\}, p(\mathcal{C}|t_R) = \sum \{p_{jR} : j \in \mathcal{C}\}$$

and let $\Delta(S, \mathcal{C})$ be as in Problem 21. Show that the split $S(\mathcal{C}^*)$ which maximizes $\Delta(S, \mathcal{C})$ over all S, \mathcal{C} may be obtained by maximizing

$$\pi_L\pi_R \left(\sum_j |p_{jL} - p_{jR}| \right)^2$$

over all splits, yielding s^* , and the optimal

$$\mathcal{C}^* = \{j : p(j|t_L^*) \geq p(j|t_R^*)\}$$

where t_L^* , t_R^* are the results of s^* .

24. Show that in the T algorithm for classification trees, if $p_i > 0$, $i = 0, 1, \dots, C_{j_m}$, $j = 1, \dots, 2^m$ is the partition of R^m defined by level m of the tree and $C_m(z)$ is the member of the partition containing z , then $C_m(z) \supset C_{m+1}(z)$ for all m and $\cap_m C_m(z) = \{z\}$.

Problems for Section 12.3

1. Justify the calculations in (12.3.7) using the conditions of Theorem 6.2.2.

2. Suppose (\mathbf{Z}, I) has joint density p defined by the logistic regression model

$$\log \left(\frac{p(\mathbf{z}, \boldsymbol{\theta})}{1 - p(\mathbf{z}, \boldsymbol{\theta})} \right) = \boldsymbol{\theta}^T \mathbf{z}$$

where \mathbf{z} is $d \times 1$, $\boldsymbol{\theta} \in R^d$ and

$$p(\mathbf{z}, \boldsymbol{\theta}) \equiv P[I = 1 | \mathbf{Z} = \mathbf{z}] = 1 - P[I = 0 | \mathbf{Z} = \mathbf{z}].$$

- (a)** Show that if $\mathbf{X}_i = (\mathbf{Z}_i, I_i)$, $1 \leq i \leq n$, are i.i.d., the joint conditional density of the I_i given $\mathbf{Z}_i \equiv (Z_{i1}, \dots, Z_{id})^T$ follows a canonical experimental family with statistics

$$T_j(\mathbf{X}) = \sum_{i=1}^n Z_{ij} I_i .$$

- (b)** Using classical MLE theory show that the conditional MLE $\hat{\theta}$ is asymptotically Gaussian and give its limiting mean and variance covariance matrix.

- 3.** Let (\mathbf{Z}, I) be as in Problem 12.3.2. Suppose the conditional density of \mathbf{Z} given $I = j$ is $\mathcal{N}(\boldsymbol{\mu}_j, \Sigma)$, $j = 0, 1$, and that $P[I = j] = \pi_j$, $j = 0, 1$.

- (a)** Show that the joint distribution of (\mathbf{Z}, I) can be written, for suitable $\boldsymbol{\eta} = (\boldsymbol{\mu}, \Sigma)$, g ,

$$g(\mathbf{Z}, \boldsymbol{\theta}, \boldsymbol{\eta}) \exp(I\boldsymbol{\theta}^T \mathbf{Z})(1 + e^{\boldsymbol{\theta}^T \mathbf{Z}})^{-1} .$$

- (b)** Assume 0-1 classification loss. Deduce that if $\mathbf{Z} \sim g(\cdot, \boldsymbol{\theta}, \boldsymbol{\eta})$ then the linear discriminant analysis (LDA) classification rule is better than the logistic regression classification rule in terms of regret, but that the asymptotic Bayes risk of LDA is the same as that of logistic regression.

- (c)** Show that if g is not of the form in **(a)**, LDA can be arbitrarily worse than logistic regression and even not asymptotically Bayes.

Hint. See Problem 12.2.9.

- 4.** Show that in the nonparametric regression example, (12.3.27) and (12.3.28) define a prior distribution with

$$|\mu(\mathbf{z}_1, \varepsilon) - \mu(\mathbf{z}_2, \varepsilon)| \leq d^{-\frac{1}{2}} |\mathbf{z}_1 - \mathbf{z}_2| .$$

- 5.** Establish the Fourier identities (12.3.47).

- 6.** For the framework of Section 12.3.4 show that for any prior π and quadratic loss there is a prior with independent components which has the same Bayes risk.

Hint. Given π consider the prior with independent components and the same marginals as π .

- 7.** Show that in the Gaussian contiguous case, $c = \frac{1}{2}$ is the best asymptotic choice to put in the Bayes test φ_B and then the Bayes risk (12.3.16) converges to $1 - \Phi(\frac{\sigma}{2})$ as $n \rightarrow \infty$.

Hint. $\Lambda_n \implies \mathcal{N}(\frac{\sigma^2}{2}, \sigma^2)$ under $P^{(n)} \in \Omega_n$ and $\frac{1}{2}(\bar{\Phi}(\frac{t+\frac{\sigma^2}{2}}{\sigma}) + \Phi(\frac{t-\frac{\sigma^2}{2}}{\sigma}))$ is maximized for $t = 0$. See Example 3.3.2.

- 8.** Consider the notation in the proof of Fano's inequality. Establish the existence and uniqueness of the decision rule $\hat{\delta}(x) = j$ iff $\rho(\hat{\theta}, P_j) < r_m/2$.

- 9.** For \mathcal{P} as in Remark 12.3.2(b), show that the minmax risk of the Nadaraya-Watson estimate is no smaller in order than $n^{-\frac{2s}{2s+d}}$.

Hint: Apply Fano's inequality for suitable P_1, \dots, P_m .

10. Exhibit an estimate achieving the rate of Problem 12.3.8.

Hint: Use Nadaraya-Watson with kernel having vanishing moments.

11. In the framework of Proposition 12.3.2, show that by using the hard thresholding rule $\delta_c^h(y)$ with suitable $c = c_n$ one can construct estimates minmax on $\Theta_{0,n}$.

12. Establish (12.3.49) and (12.3.50).

13. Consider the Gaussian shift model $x_i = \theta_i + \varepsilon_i$, $i = 1, 2, \dots$, where x_i 's are the observations and θ_i 's are the parameters to be estimated. Assume that the parameter vector $\theta = (\theta_1, \theta_2, \dots)$ lies in an ellipsoid $\sum_{i=1}^{\infty} i^{2m} \theta_i^2 \leq M$, where M is a fixed positive constant and m is a known positive integer. The ε_i 's are independent normal random variables with mean 0 and variance n^{-1} .

In the following problems, you may use this fact: there exists a positive constant C_1 such that $\sum_{i=1}^{\infty} (1 + \lambda \rho_i)^{-2} \leq C_1 \lambda^{-1/2m}$ for all $\lambda > 0$, where $\rho_i = i^{2m} - 1$. You may also assume the existence of the solutions to the problems below.

(a) The penalized likelihood estimator $\hat{\theta}_i$ is the solution to the optimization problem:

$$\min_{\theta} \left\{ \sum_{i=1}^{\infty} (x_i - \theta_i)^2 + \lambda \sum_{i=1}^{\infty} \rho_i \theta_i^2 \right\}.$$

Show that if λ is of order $O(n^{-2m/(2m+1)})$, then there exists $C_2 > 0$ such that $\sum_{i=1}^{\infty} E(\hat{\theta}_i - \theta_i)^2 \leq C_2 n^{-2m/(2m+1)}$.

(b) The projection estimator $\tilde{\theta}_i$ is defined by

$$\tilde{\theta}_i = \begin{cases} x_i, & i \leq n^{1/(2m+1)} \\ 0, & \text{otherwise.} \end{cases}$$

Show that there exists $C_3 > 0$ such that $\sum_{i=1}^{\infty} E(\tilde{\theta}_i - \theta_i)^2 \leq C_3 n^{-2m/(2m+1)}$.

(c) Show that the estimator resulting from L_1 penalty estimation:

$$\min_{\theta} \left\{ \sum_{i=1}^{\infty} (x_i - \theta_i)^2 + 2\lambda \sum_{i=1}^{\infty} |\theta_i| \right\}$$

has the form $\text{sgn}(x_i)[|x_i| - \lambda]_+$.

(d) Show that the estimator resulting from L_0 penalty estimation:

$$\min_{\theta} \left\{ \sum_{i=1}^{\infty} (x_i - \theta_i)^2 + \lambda^2 \{ \# \text{ of } \theta_i \text{ that is not 0} \} \right\}$$

has the form $x_i 1_{\{|x_i| > \lambda\}}$.

- (e) Consider now a rectangular subset of the original ellipsoid set of parameters: $\theta_i \leq \xi_i$, where ξ_i 's are given positive numbers satisfying $\sum_{i=1}^{\infty} (1 + \rho_i) \xi_i^2 \leq M$. Find the minimax linear estimator of the form $\theta_i = \alpha_i x_i$ for all i , in terms of mean squared error.

Problems for Section 12.4

1. (a) Verify (12.4.10) and (12.4.11).
(b) Complete the proof of Theorem 12.4.1 by establishing the last inequality in the proof.
(c) Derive the results of Remark 12.4.1 (a).
2. **χ^2 inequality.** Suppose $X \sim \chi_m^2$, $X = \sum_{i=1}^m Z_i^2$, $Z_i \sim N(0, 1)$ independent. Show that

$$P[|X - m| > t\sqrt{m}] \leq 2 \exp \left[-t^2 \left(1 + \frac{t}{\sqrt{m}} \right)^{-1} \right] \leq 4 \left(e^{-\frac{t^2}{2}} + e^{-\frac{\sqrt{m}t}{2}} \right).$$

Hint: $P[X \geq m - \frac{t}{\sqrt{m}}] \leq \inf_s e^{-s(\sqrt{m}+t)} \left(1 - \frac{2s}{\sqrt{m}} \right)^{-1}$ for $s < \frac{\sqrt{m}}{2}$. Use,

$$\arg \min \left\{ -s\sqrt{m} - st - \frac{m}{2} \log \left(1 - \frac{2s}{\sqrt{m}} \right) \right\} = \frac{t}{2\sqrt{m}} \left(1 + \frac{t}{\sqrt{m}} \right)^{-1} < \frac{1}{2},$$

and $-\log(1 - x) \leq x + x^2$, $0 \leq x < \frac{1}{2}$. Similarly, $P[X < m + \frac{t}{\sqrt{m}}] = P[m - X > -t\sqrt{m}] \leq \inf_s e^{s(\sqrt{m}+t)} \left(1 + \frac{2s}{\sqrt{m}} \right)^{-\frac{m}{2}}$.

3. Deduce, if $\mathbf{X}_m \sim \mathcal{X}_m^2$, that for some universal c, D

$$E \sup \left\{ |X_m - m| - c\sqrt{m} \log n \right\}_+ : 1 \leq m \leq n \leq D.$$

Hint: If $\sup_m \Delta_{m,n}$ is the random variable above

$$E \sup \Delta_m \leq K + \int_K^\infty P[\sup \Delta_{m,n} > t] dt \leq K + \sum_{m=1}^n \int_K^\infty P[\Delta_{m,n} > t] dt. \quad (12.8.1)$$

By Problem 12.4.2,

$$\begin{aligned} & \int_K^\infty P \left[\frac{|X_m - m|}{\sqrt{m}} > c \log n + \frac{t}{\sqrt{m}} \right] dt \\ & \leq 4 \left(\int_K^\infty e^{-(c \log n + \frac{t}{\sqrt{m}})^2/2} dt + \int_K^\infty e^{-(t + c\sqrt{m} \log n)/2} dt \right) \\ & \leq 4\sqrt{m} \left(\int_{c \log n}^\infty e^{-\frac{v^2}{2}} dv \right) + 4n^{-c\sqrt{m}}. \end{aligned}$$

The claimed bound follows for $c > 1$ by substituting this bound in (12.8.1).

4. Show that if $\widehat{\eta}_i$ are as in the GWN model

$$E \sup \left\{ \left| \sum_{i=m+1}^n \eta_i (\widehat{\eta}_i - \eta_i) \right| - c\rho \left(\sum_{i=m+1}^n \eta_i^2 \right)^{\frac{1}{2}} \right\} \leq D\sqrt{\log n}.$$

Hint: $\rho^2 \sum_{i=m+1}^n \eta_i^2 = \text{Var} \sum_{i=m+1}^n \eta_i (\hat{\eta}_i - \eta_i)$.

5. If $\tilde{R}(m)$ is defined by

$$\tilde{R}(m) = 2\sigma_n^2 m - \sum_{j=1}^m \hat{\eta}_j^2 ,$$

show that

(a) $E\tilde{R}(m) = \sigma_n^2 m - \sum_{j=1}^m \eta_j^2$

(b) $\text{Var}\tilde{R}(m) = \sum_{j=1}^m \text{Var} \hat{\eta}_j^2 = 2\sigma_n^4 m + \sigma_n^2 \sum_{j=1}^m \eta_j^2$.

(c) $\max_{\Theta_\beta} \min_{1 \leq m \leq n} E\tilde{R}(m) = O(n^{-\frac{2\beta}{2\beta+1}})$.

Hint: Plug in $m = n^{\frac{1}{2\beta+1}}$.

6. Show that if \hat{m} is defined as in Theorem 12.4.2

$$\max \left\{ E \left[\sum_{i=1}^{\hat{m}} (\hat{\eta}_i - \eta_i)^2 + \sum_{i=\hat{m}+1}^n \eta_i^2 \right] : \Theta_\beta \right\} \leq K n^{-\frac{2\beta}{2\beta+1}} \log n \text{ for } K = K(\beta) .$$

Hint: By Problem 12.4.6 (using C, D, etc. generically)

$$\left| E \sum_{i=1}^{\hat{m}} (\hat{\eta}_i - \eta_i)^2 - E\sigma_n^2 \hat{m} \right| \leq (D + C \log n E\sqrt{\hat{m}}) \sigma_n^2 \equiv \Delta .$$

Therefore,

$$\begin{aligned} E \left(2\sigma_n^2 \hat{m} - \sum_{i=1}^{\hat{m}} \hat{\eta}_i^2 \right) &= E \left(\sigma_n^2 \hat{m} + E \sum_{i=1}^{\hat{m}} (\hat{\eta}_i - \eta_i)^2 - \sum_{i=1}^{\hat{m}} \hat{\eta}_i^2 \right) + \Delta \\ &= E \left(\sigma_n^2 \hat{m} - 2 \sum_{i=1}^{\hat{m}} \hat{\eta}_i \eta_i + \sum_{i=1}^{\hat{m}} \eta_i^2 \right) + \Delta \\ &= E \left(\sigma_n^2 \hat{m} - 2 \sum_{i=1}^{\hat{m}} (\hat{\eta}_i - \eta_i) \eta_i - \sum_{i=1}^{\hat{m}} \eta_i^2 \right) + \Delta \\ &= E \left(\sum_{i=1}^{\hat{m}} (\hat{\eta}_i - \eta_i)^2 + \sum_{i=\hat{m}+1}^n \eta_i^2 \right) - 2E \left(\sum_{i=1}^{\hat{m}} (\hat{\eta}_i - \eta_i) \eta_i - \sum_{i=1}^n \eta_i^2 \right) + 2\Delta \\ &= E \left[\sum_{i=1}^{\hat{m}} (\hat{\eta}_i - \eta_i)^2 + \sum_{i=\hat{m}+1}^n \eta_i^2 \right] \\ &\quad + \left[2 \sum_{i=\hat{m}+1}^n (\hat{\eta}_i - \eta_i) \eta_i - 2 \left(\sum_{i=1}^n (\hat{\eta}_i - \eta_i) \eta_i - \sum_{i=1}^n \eta_i^2 \right) \right] + 2\Delta . \end{aligned}$$

Using Problem 12.4.6,

$$\left| E \sum_{i=\hat{m}+1}^n (\hat{\eta}_i - \eta_i) \eta_i \right| \leq C \sigma_n E \left(\sum_{i=\hat{m}+1}^n \eta_i^2 \right)^{\frac{1}{2}} \sqrt{\log n} .$$

Finally,

$$\begin{aligned} & E \left(\sum_{i=1}^{\hat{m}} (\hat{\eta}_i - \eta_i)^2 + \sum_{i=\hat{m}+1}^n \eta_i^2 \right) \\ & \leq 2\Delta + \left(\sigma_n E \left(\sum_{i=\hat{m}+1}^n \eta_i^2 \right)^{\frac{1}{2}} \sqrt{\log n} \right) + E \left(\sum_{i=1}^{m_0} (\hat{\eta}_i - \eta_i)^2 + \sum_{i=m_0+1}^n \eta_i^2 \right) \end{aligned}$$

since $E(2\sigma_n^2 \hat{m} - \sum_{i=1}^{\hat{m}} \hat{\eta}_i^2) \leq E(2\sigma_n^2 m_0 - \sum_{i=1}^{m_0} \eta_i^2)$. But

$$\begin{aligned} E\sigma_n^2 \sqrt{\hat{m}} & \leq \sigma_n^2 (E\hat{m})^{\frac{1}{2}} \\ \sigma_n E \left(\sum_{i=\hat{m}+1}^n \eta_i^2 \right)^{\frac{1}{2}} & \leq \frac{1}{2} \left(\sigma_n^2 + E \sum_{i=\hat{m}+1}^n \eta_i^2 \right) \end{aligned}$$

and the results follows.

7. Show that in Remark 12.4.1, if \hat{m} is selected by SBC (BIC), then

$$|\hat{\eta}_{\hat{m}}| \leq \sigma_0 n^{-\frac{1}{2}} \sqrt{2 \log n} .$$

Problems for Section 12.5

1. Establish the sufficient conditions for A3 leading to Theorem 12.5.1.

Hint. Apply the law of large numbers using (1) to reduce to the case $\|\Delta_S^{(Z)}\|_P^2 = 1$. Apply Hoeffding's or a similar inequality.

2. (The Efron–Stein Inequality). Let $T(x_1, \dots, x_{n-1})$ be a symmetric function of $n-1$ variables and let X_1, \dots, X_n be i.i.d. as $X \in R$. Set $T_i = T(X^{(-i)})$ and $T_+ = n^{-1} \sum_{i=1}^n T_i$, where

$$\mathbf{X}^{(-i)} = (X_1, \dots, X_{i-1}, X_{i+1}, \dots, X_n)^T .$$

(a) Show that,

$$\text{Var } T_n(X_1, \dots, X_{n-1}) \leq E \sum_{i=1}^n (T_i - T_+)^2 .$$

Hint. Use $\text{Var } U = \frac{1}{2} E(U - U')^2$, where U, U' are i.i.d., as U . Also, condition on $\mathbf{X}^{(-i)}$.

(b) Use (a) to show that the leave one out cross validation estimate of variance tends to overestimate the variance.

3. Why crossvalidate? Consistent estimation of prediction error (PE). Let $(\mathbf{X}_1, Y_1), \dots, (\mathbf{X}_n, Y_n)$, where $\mathbf{X}_i = (X_{i1}, \dots, X_{ip})^T$, be i.i.d. as $\mathbf{X} \in R^p$, $Y \in R$. Assume the model

$$Y_i = \mathbf{X}_i^T \boldsymbol{\beta} + \varepsilon_i, \quad \varepsilon_i \sim i.i.d. \mathcal{N}(0, \sigma^2)$$

with $E|\mathbf{X}|^4 < \infty$, $\text{Var}(\mathbf{X}) = \Sigma_0$, where Σ_0 is nonsingular and known. Let $\mathbf{X}_D \equiv (X_{ij})_{n \times p}$ be the design matrix and let $\hat{\boldsymbol{\beta}} = (X_D^T X_D)^{-1} \mathbf{Y}$ be the least squares (and maximum likelihood) estimate of $\boldsymbol{\beta}$ (see (6.1.14)). Show that if $\boldsymbol{\beta} = \boldsymbol{\beta}_0$ is known,

(a) The average prediction error when we use $\mathbf{X}_i^T \boldsymbol{\beta}_0$ to predict Y_i , $1 \leq i \leq n$, is

$$PE = E\left(\frac{1}{n} \sum (Y_i - \mathbf{X}_i^T \boldsymbol{\beta}_0)^2\right) = \sigma^2.$$

(b) The estimate of PE obtained by replacing $\mathbf{X}_i \boldsymbol{\beta}_0$ by $\mathbf{X}_i \hat{\boldsymbol{\beta}}$ in the formula in **(a)** is biased, in fact,

$$E\left(n^{-1} \sum_{i=1}^n (Y_i - \mathbf{X}_i^T \hat{\boldsymbol{\beta}})^2\right) = \sigma^2 \left(1 - \frac{p}{n}\right).$$

(c) Show that if p is fixed,

$$\hat{\boldsymbol{\beta}} = \boldsymbol{\beta}_0 + n^{-1} \Sigma_0^{-1} \sum_{i=1}^n \mathbf{X}_i \varepsilon_i + O_P(n^{-1}).$$

(d) Show that if we use the least squares estimate $\hat{\boldsymbol{\beta}}_m$ based on (\mathbf{X}_i, Y_i) , $i = 1, \dots, m$ to estimate $\boldsymbol{\beta}$, and we use (\mathbf{X}_i, Y_i) ; $i = m+1, \dots, n$ to estimate prediction error, then

$$\widehat{PE} \equiv \frac{1}{n-m} \sum_{i=m+1}^n (Y_i - \mathbf{X}_i^T \hat{\boldsymbol{\beta}}_m)^2 = \sigma^2 + O_P\left(\frac{1}{n}\right)$$

provided $(m/n) \rightarrow \lambda$ for some $\lambda \in (0, 1)$.

Hint. **(b)**, condition on \mathbf{X}_D . For **(c)** and **(d)** use the delta method and $E|\mathbf{X}_i \Sigma_0^{-1} \mathbf{X}_i^T| = E|\mathbf{X}_i \Sigma_0^{-\frac{1}{2}}|^2$.

Problems for Section 12.6

1. Show that if we apply the Bayesian criterion (12.6.5) to choose m in the sieve method for estimating θ in Θ_β of the GWN model as in Theorem 12.4.2, and if $\boldsymbol{\eta}$ is in the closure of $\cup_{n=1}^\infty \mathcal{P}_m$, then we obtain the minimax risk rate $\log n n^{-(\frac{2\beta}{2\beta+1})}$ rather than $n^{-(\frac{2\beta}{2\beta+1})}$ obtained by using Stein's unbiased risk estimate (Mallows' C_p), as shown in Problems 12.4.2–12.4.6.

2. Consider the one-sample model

$$Y_i = \mu + \varepsilon_i, \quad i = 1, \dots, n$$

with $\{\varepsilon_i\}$ i.i.d. $\mathcal{N}(0, 1)$. Let \mathcal{M}_0 be the model with $\mu = 0$ and let \mathcal{M}_1 be the model with $\mu \in R$ arbitrary.

- (a) Find the model selection rule based on the Bayesian criterion (12.6.5).
- (b) Show that the rule in (a) selects the correct model with probability tending to one as $n \rightarrow \infty$.
- (c) The AIC criterion with known $\sigma^2 = 1$ takes the form

$$AIC(\mathcal{M}_k) = 2 \sum_{i=1}^n \log p(y_i | \hat{\theta}_k) - 2k.$$

Find the model selection rule based on AIC.

- (d) Show that the rule in (c) satisfies $P_{\mu=0}$ (Decide model \mathcal{M}_1) = $P(\chi_1^2 > 2)$. That is, it provides a test wth level $P(\chi_1^2 > 2)$. It does not select the correct model with probability tending to one as $n \rightarrow \infty$.

3. Show that the maximal eigenvalue of the covariance matrix Σ is given by

$$\lambda_1(P) = \text{Var}\left(\sum_{j=1}^p a_j^* X_j\right)$$

where $|\mathbf{a}^*| = 1$ and $\mathbf{a}^* = \arg \max \left\{ \text{Var}\left(\sum_{j=1}^p a_j X_j\right) : |\mathbf{a}| = 1 \right\}$.

4. In Section 12.6.3, show that the distance between a point $\mathbf{y} \in R^p$ and the line $\mathbf{x} = \alpha \mathbf{a}$ is $\mathbf{y}^T \mathbf{y} - (\mathbf{a}^T \mathbf{y})^2$.

Hint. Consider the point $\mathbf{t} \in R^p$ where the perpendicular from \mathbf{y} to the line intersects the line. Then $\mathbf{t} = \alpha \mathbf{a}$ where α must satisfy $\mathbf{a}^T(\mathbf{y} - \alpha \mathbf{a}) = 0$. Because $\mathbf{a}^T \mathbf{a} = 1$, this gives $\alpha = \mathbf{a}^T \mathbf{y}$. Thus

$$|\mathbf{y} - \mathbf{t}|^2 = |\mathbf{y} - (\mathbf{a}^T \mathbf{y}) \mathbf{a}|^2.$$

Next expand the right hand side.

Problems for Section 12.7

1. Let $p^{(1)} \leq \dots \leq p^{(N)}$ be the ordered p -values and let $H^{(j)}$ denote the null hypothesis corresponding to $p^{(j)}$, $1 \leq j \leq N$. Set $p^{(N+1)} = 1$ and define

$$j^* = \min_{1 \leq j \leq N+1} \{j : p^{(j)} > \alpha/(N+1-j)\}.$$

The *Holm* (1979) *multiple testing procedure* rejects $H^{(1)}, \dots, H^{(j^*-1)}$. Show that this method strongly controls FWER at level α .

Hint. Let $J_0 = \{j : H^{(j)} \text{ is true}\}$, $N_0 = \text{cardinality of } J_0$, and $j_0 = N - N_0 + 1$. Let A be the event “ $p_{(j^*)} < \alpha/N_0$ ” and B = “ $p_{(j)} > \alpha/N_0$ for all $j \in J_0$.” Show that $B \implies “j^* < j_0” \implies A$, and that $P(B) \geq 1 - \alpha$.

2. *Optimal multiple testing.*

- (a) (Spjótvoll, 1972). Let g_{01}, \dots, g_{0N} and g_1, \dots, g_N be integrable functions of a data vector \mathbf{X} and let $S(\gamma)$ be all multiple tests $\psi = \{\psi_1, \dots, \psi_N\}$ of the null hypotheses H_1, \dots, H_N that satisfy

$$\sum_{j=1}^N \int \psi_j(\mathbf{x}) g_{0j}(\mathbf{x}) d\nu(\mathbf{x}) = \gamma \quad (12.8.2)$$

for a given probability ν . Show that for an appropriate c , the multiple test

$$\psi = (\psi_1, \dots, \psi_N) = \{1[g_j(\mathbf{x}) > cg_{0j}] : 1 \leq j \leq N\}, \quad c > 0 \quad (12.8.3)$$

maximizes

$$\sum_{j=1}^N \int \psi_j(\mathbf{x}) g_j(\mathbf{x}) d\nu(\mathbf{x}) \quad (12.8.4)$$

among all tests $\psi \in S(\gamma)$.

Hint. Extend the proof of the Neyman-Pearson Lemma.

Remark. The functions g_j and g_{0j} need not be densities. They are general integrable functions.

- (b) Let $R = \sum_{j=1}^N 1[\psi_j(\mathbf{x}) = 1]$ be the number of “discoveries,” $f_{0j}(\mathbf{x})$ and $f_j(\mathbf{x})$ densities of \mathbf{X} under the hypothesis H_j and alternative A_j , respectively, $g_{0j}\nu = [f_{0j}/R]1(R > 0)$ if H_j is false, 0 otherwise, and $g_j\nu = [f_j/R]1(R > 0)$ if H_j is false, 0 otherwise. With this notation, by definition, (12.8.2) is the False Discovery Rate (FDR) and (12.8.4) is the Correct Discovery Rate (CDR). Give the rule that maximizes the CDR subject to FDR = γ .

- (c) (Spjótvoll (1972), Storey (2007), Sun and Cai (2007)). Give the multiple test that maximizes the expected number of true positives for a fixed level γ of the expected number of false positives.
- (d) Suppose we express ψ_j in a multiple test rule as $\psi_j = \psi(\hat{p}_j) = 1(\hat{p}_j \leq a)$ where \hat{p}_j is the p -value for the hypothesis H_j . Express the solutions to (b) and (c) in terms of p -values.

Hint. When H_j is true, $\hat{p}_j \sim Unif[0, 1]$.

- (e) **Bayes.** Let π_j denote the prior probability that H_j is true; and let f_0 and f_j denote the densities of \mathbf{X} when H_j is true and false, respectively. The posterior probability of H_j is

$$P(H_j|\mathbf{x}) = \frac{\pi_j f_0(\mathbf{x})}{\pi_j f_0(\mathbf{x}) + (1 - \pi_j) f_j(\mathbf{x})}.$$

Consider the Bayes rule $\psi_B = (\psi'_B, \dots, \psi_B^N)$ that rejects the H_j with $P(H_j|\mathbf{x}) \leq q$ for some preassigned $q \in (0, 1)$. Then

$$\psi^B = \{1(f_j(\mathbf{x}) > (1 - q)\pi_j f_0(\mathbf{x})) / (1 - \pi_j) : 1 \leq j \leq N\}.$$

Thus, if $\pi_j = \pi$, ψ^B is optimal in the sense of problem (a).

- (f) Empirical Bayes (Efron 2008, 2010). Suppose we write ψ_j in the multiple test rule as $\psi_j(\mathbf{Z}) = \mathbb{1}(Z_j \leq b_j)$ where $Z_j = \Phi^{-1}(\hat{p}_j)$ and $\Phi = N(0, 1)$ df. Then $Z_j \sim N(0, 1)$ when H_j is true. Let f_0 denote the $N(0, 1)$ density, let π_j denote the prior probability that H_j is true, and let f_j denote the density of Z_j when H_j is false. Then the posterior probability of the null hypothesis H_j given $Z_j = z_j$ is

$$P(H_j|z_j) = \frac{\pi_j f_0(z_j)}{\pi_j f_0(z_j) + (1 - \pi_j) f_j(z_j)}.$$

Now reject H_j if $P(H_j|z_j) \leq q$, $q \in (0, 1)$.

Efron (2010) operationalizes this rule by constructing estimates of Π_j and f_j , thereby obtaining an empirical Bayes rule which can be found in the software *R* under “loc FDR” (local False Discovery Rate).

- 3. False Discovery Distributions.** Let T_1, \dots, T_N be test statistics for testing $H_j : \mu_j = 0$, $1 \leq j \leq N$. Each is based on a sample of size n . Assume that the joint distribution of T_1, \dots, T_N is defined by the equation

$$T_j = \sqrt{n}\mu_j + \rho Z + (1 - \rho^2)^{\frac{1}{2}}\varepsilon_j$$

where $Z, \varepsilon_1, \dots, \varepsilon_N$ are i.i.d. $\mathcal{N}(0, 1)$. Let p_j be the *p*-value of the test that rejects H_j for $|T_j|$ large, let

$$R(u) = \sum_{j=1}^N \mathbb{1}(p_j \leq u) = \sum_{j=1}^N \mathbb{1}(|T_j| > -z_{u/2})$$

be the *number of discoveries*, let $J_0 = \{j : \mu_j = 0\}$, and let the number of *false discoveries* be

$$V(u) = \sum_{j \in J_0} \mathbb{1}(p_j \leq u) = \sum_{j \in J_0} \mathbb{1}(|T_j| > -z_{u/2}).$$

- (a) Let $N_0 = \{\#j : \mu_j = 0\}$. Show that the conditional distribution of $V(u)$ given Z can be approximated by the distribution of

$$W(\mu) \equiv N_0 \left\{ \Phi \left(\frac{z_{u/2} + \rho Z}{\sqrt{1 - \rho^2}} \right) + \Phi \left(\frac{z_{u/2} - \rho Z}{\sqrt{1 - \rho^2}} \right) \right\}, \quad Z \sim \mathcal{N}(0, 1).$$

- (b) Show that (i) $W(u) \xrightarrow{P} N_0 u$ as $\rho \rightarrow 0$; and (ii) $W(u) \xrightarrow{P} N_0$ as $\rho \rightarrow 1$. Find the limit as $\rho \rightarrow -1$.

Hint for (ii). Condition on $I = \mathbb{1}(Z > 0)$.

- (c) Give the range of values of the probability $P(V(u) \geq 1)$ of at least one false rejection as ρ ranges from -1 to 0 and from 0 to 1 .

- (d) Let $J_1 = \{j : \mu_j \neq 0\}$ and $N_1 = \{\#j : \mu_j \neq 0\}$.
- (i) Find a variable $S(u)$ whose distribution approximates the conditional distribution of $R(u)$ given Z when $\mu_i = \mu/\sqrt{n}$ and
- (ii) $N_1/N_0 \rightarrow 0$, (ii) $N_1/N_0 \rightarrow \lambda \in (0, 1)$.
- (e) In (d) above, find the probability limit of $S(u)$ as $\rho \rightarrow 0$ and as $\rho \rightarrow \pm 1$ for the cases (i) and (ii).

Fan, Han and Gu (2012), among others, have given approaches to FDR for dependent p -values.

Appendix D

SOME AUXILIARY RESULTS

D.1 Probability Results

A collection $\{Y_{ni}; 1 \leq i \leq n\}$ of random variables is called a *double array* if the probability distribution P_{ni} of Y_{ni} depends on both i and n . An example is the regression model

$$Y_{ni} = \alpha + \beta_n x_{ni} + \varepsilon_i$$

where $\beta_n = c/\sqrt{n}$, $n^{-1} \sum (x_{ni} - \bar{x}_n)^2 \rightarrow \tau \in (0, 1)$ and $\varepsilon_1, \dots, \varepsilon_n$ are i.i.d. Such models are used when investigating asymptotic power. See Section 6.3.2. Another example is the “bootstrap” asymptotic theory of Section 10.3.4. A key result for double arrays is

1. The Lindeberg-Feller Central Limit Theorem.

Theorem D.1 (Lindeberg-Feller) Suppose that for each fixed n , Y_{n1}, \dots, Y_{nn} are independent with $0 < \sigma_{ni}^2 \equiv \text{Var}(Y_{ni}) < \infty$. Set $\sigma_n^2 = \sum_{i=1}^n \sigma_{ni}^2$. Then, for $n \rightarrow \infty$,

$$\max_{1 \leq i \leq n} \frac{\sigma_{ni}^2}{\sigma_n^2} \rightarrow 0 \quad (\text{D.1.1})$$

and

$$\frac{1}{\sigma_n} \sum_{i=1}^n [Y_{ni} - EY_{ni}] \xrightarrow{\mathcal{L}} \mathcal{N}(0, 1) \quad (\text{D.1.2})$$

iff for each $\varepsilon > 0$,

$$\lim_{n \rightarrow \infty} \frac{1}{\sigma_n^2} \sum_{i=1}^n E [(Y_{ni} - EY_{ni})^2 1(|Y_{ni} - EY_{ni}| > \varepsilon \sigma_n)] = 0. \quad (\text{D.1.3})$$

Corollary D.1 Suppose Y_1, \dots, Y_n are i.i.d. as $Y \sim P$, where P does not depend on n . Let $\mu = EY$, $\sigma^2 = \text{Var}Y$, and $T_n = \sum_{i=1}^n c_i Y_i$, then for $n \rightarrow \infty$ the following are equivalent

$$\frac{T_n - \mu \sum_{i=1}^n c_i}{(\sigma^2 \sum_{i=1}^n c_i^2)^{\frac{1}{2}}} \xrightarrow{\mathcal{L}} \mathcal{N}(0, 1), \quad (\text{D.1.4})$$

$$v_n^2 \equiv \frac{(\sum_{i=1}^n c_i^2)}{\max_{1 \leq i \leq n} c_i^2} \rightarrow \infty. \quad (\text{D.1.5})$$

Proof. See Problem D.1.2.

Theorem D.2 The condition (D.1.3) holds for all $\varepsilon > 0$ if for some $\delta > 0$

$$\sum_{i=1}^{k_n} E|Y_{ni} - EY_{ni}|^{2+\delta} = o(\sigma_n^{2+\delta}). \quad (\text{D.1.6})$$

Remark D.1 (D.1.3) is called *Lindeberg's Condition* and (D.1.6) is called *Liapunov's Condition*.

Remark D.2 Suppose $n^{-1}\sigma_n^2 \rightarrow c$ for some $c > 0$. Then (D.1.3) becomes

$$\frac{1}{n} \sum E[(Y_{ni} - EY_{ni})^2 1(|Y_{ni} - EY_{ni}| > \varepsilon\sigma_n)] \rightarrow 0. \quad (\text{D.1.7})$$

A multivariate version is

Theorem D.3. (*Multivariate Lindeberg-Feller*) Suppose for each n , Y_{n1}, \dots, Y_{nk_n} are independent random vectors in R^d with $\text{Cov}(Y_{ni})$, $1 \leq i \leq k_n$, finite. Let $|\cdot|$ denote Euclidean norm. If for $\varepsilon > 0$,

$$\lim_{n \rightarrow \infty} \sum_{i=1}^{k_n} E\{|Y_{ni}|^2 1(|Y_{ni}| > \varepsilon)\} = 0$$

and

$$\sum_{i=1}^{k_n} \text{Cov}(Y_{ni}) \rightarrow \Sigma$$

for Σ positive definite, then

$$\sum_{i=1}^{k_n} (Y_{ni} - EY_{ni}) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \Sigma).$$

Remark D.3. In Theorem D.3, the Lindeberg condition is for uncentered Y_{ni} . This turns out to be convenient for the asymptotics of bootstrap methods. Clearly Theorem D.3 also holds if Lindeberg's condition is for centered Y_{ni} .

2. Almost Sure Convergence and the Borel–Cantelli Lemma

Almost sure convergence of a sequence of random variables and Kolmogorov's Strong Law of Large Numbers (SLLN) appear in Section B.7 of Volume I. Here are some useful additional facts about a.s. convergences.

D.4 *Convergence in probability and a.s. convergence.* $Z_n \xrightarrow{P} Z$ as $n \rightarrow \infty$ iff every subsequence $\{Z_{n_k}\}$, $k \geq 1$, has a subsubsequence $\{Z_{n'_k}\}$, $\{n'_k\} \subset \{n_k\}$, $k \geq 1$ such that $Z_{n'_k} \xrightarrow{a.s.} Z$ as $n'_k \rightarrow \infty$.

D.5 If $Z_n \xrightarrow{a.s.} Z$, then $Z_n \xrightarrow{P} Z$.

D.6 If $\mathbf{Z}_n \xrightarrow{a.s.} \mathbf{Z}$ iff $\lim_{n \rightarrow \infty} P(\sup_{m \geq n} |\mathbf{Z}_n - \mathbf{Z}| > \varepsilon) = 0$.

D.7 Uniform SLLN (Chung (1951)). Let $\mathbf{Z}_1, \mathbf{Z}_2, \dots$ be i.i.d. as $\mathbf{Z} \in R^d$, $\mathbf{Z} \sim P \in \mathcal{P}$. Let $|\cdot|$ be Euclidean norm and let $\bar{\mathbf{Z}}_n = n^{-1} \sum_{i=1}^n \mathbf{Z}_i$. If $E|\mathbf{Z}| < \infty$, and

$$\lim_{M \rightarrow \infty} \sup_{P \in \mathcal{P}} E_P(|\mathbf{Z}| \mathbf{1}(|\mathbf{Z}| \geq M)) = 0,$$

then the SLLN holds uniformly, that is, for $\varepsilon > 0$,

$$\lim_{n \rightarrow \infty} \sup_{P \in \mathcal{P}} P\left(\sup_{m \geq n} |\bar{\mathbf{Z}}_m - E_P(\mathbf{Z})| \geq \varepsilon\right) = 0.$$

D.8 The Borel–Cantelli Lemma. Suppose

$$\sum_{n=1}^{\infty} P[|\mathbf{Z}_n - \mathbf{Z}| \geq \varepsilon] < \infty$$

for all $\varepsilon > 0$. Then, $\mathbf{Z}_n \xrightarrow{a.s.} \mathbf{Z}$ as $n \rightarrow \infty$.

3. Kolmogorov's Inequality

The following result is useful for establishing uniform convergence in probability of nonparametric regression estimates. See Problems 11.6.8 and 11.6.9.

Theorem D.4. Let X_1, X_2, \dots be independent random variables with $E(X_j) = 0$, $E(X_j^2) < \infty$, and let $S_j = \sum_{i=1}^j X_i$. Then, for every $\varepsilon > 0$,

$$P\left(\max_{1 \leq j \leq n} |S_j| > \varepsilon\right) \leq \text{Var}(S_n)/\varepsilon^2.$$

D.2 Weak Convergence of Functions. Supplement to Section 7.1

Proof of Theorem 7.1.1. The full proof of this theorem may be found in van der Vaart and Wellner (1996), for instance. However, it is easier and less technical to see that if q is Lipschitz continuous, then (i) and (ii) imply that $\mathcal{L}(q(Z_n(\cdot))) \rightarrow \mathcal{L}(q(Z(\cdot)))$. Here's the argument for q Lipschitz. Let t_{mj} , $1 \leq j \leq k_m$, $m \geq 1$, be points such that $t_{mj} \in T_{mj}$. Let $Z_n^{(m)}(t) = \sum_{j=1}^{k_m} \mathbf{1}(t \in T_{mj}) Z_n(t_{mj})$ and define $Z^{(m)}$ similarly. Comparing the characteristic functions of $q(Z_n)$ and $q(Z)$, we compute

$$\begin{aligned} & |E \exp\{isq(Z_n)\} - E \exp\{isq(Z)\}| \\ & \leq |E \exp\{isq(Z_n)\} - E \exp\{isq(Z_n^{(m)})\}| \\ & + |E \exp\{isq(Z_n^{(m)})\} - E \exp\{isq(Z^{(m)})\}| \\ & + |E \exp\{isq(Z^{(m)})\} - E \exp\{isq(Z)\}|. \end{aligned} \tag{D.2.1}$$

Let $\Delta_m(Z_n) = \max\{\sup[|Z_n(s) - Z_n(t)| : s, t \in T_{mj}] : 1 \leq j \leq k_m\}$ and let $A_{mn} = \{|\Delta_m(Z_n)| \leq \varepsilon_m\}$. Note that, $|Z_n - Z_n^{(m)}|_\infty \leq \Delta_m$. We claim that

$$\begin{aligned} & |E \exp\{isq(Z_n)\} - E \exp\{isq(Z_n^{(m)})\}| \\ & \leq |s| |E[q(Z_n) - q(Z_n^{(m)})]| \mathbf{1}(A_{mn}) + 2P[A_{mn}^c]. \end{aligned} \quad (\text{D.2.2})$$

This follows from $|e^{ia} - e^{ib}| \leq |a - b|$ and $|e^{ia}| \leq 1$. But by the Lipschitz nature of q ,

$$E|q(Z_n) - q(Z_n^{(m)})| \mathbf{1}(A_{mn}) \leq M\varepsilon_m$$

and by (7.1.5), $P[A_{mn}^c] \leq \delta_m$. By the separability of Z and the FIDI convergence of $Z_n(\cdot)$, we have $P[\Delta_m(Z) \geq \varepsilon_m] \leq \delta_m$ also (Problem 7.1.1), and the third term in (D.2.1) can be bounded just as the first term has been. Combining these bounds, we have from (D.2.1),

$$\begin{aligned} & \limsup_n |E \exp\{isq(Z_n)\} - E \exp\{isq(Z)\}| \\ & \leq 2(M|s|\varepsilon_m + \delta_m) + \limsup_n |E \exp\{isq(Z_n^{(m)})\} - E \exp\{isq(Z^{(m)})\}|. \end{aligned} \quad (\text{D.2.3})$$

The second term in (D.2.3) is 0 by the FIDI convergence of $Z_n^{(m)}(\cdot)$ to $Z^{(m)}(\cdot)$. Now, let $m \rightarrow \infty$ to complete the proof. Note that $Z^{(m)}$ as defined yield tightness of $Z(\cdot)$. \square

The general proof requires showing that (D.2.3) implies that, for every $\epsilon > 0$, there exists a compact set K_ϵ such that $P[Z_n(\cdot) \in K_\epsilon] \geq 1 - \epsilon$ and $P[Z(\cdot) \in K_\epsilon] \geq 1 - \epsilon$. By a classical result, functions which are continuous on a compact set are uniformly continuous.

Proof of Theorem 7.1.2. Consider the bracket set collection, \mathcal{T}_m , corresponding to $\delta = m^{-\alpha}$, $\alpha > 1$ to be chosen later, $m = 1, 2, \dots$. Without loss of generality, we can suppose that the bracket sets T_{mj} form a partition of T in the sense that if $T_{mj} = (\underline{f}_{mj}, \bar{f}_{mj}) \in \mathcal{T}_m$, $j = 1, \dots, J_m$, then $T_{mj} \cap T_{mk} = \emptyset$ unless $j = k$. This can be achieved in such a way (Problem D.2.2) that (ii) will still hold for some constant c . Further, let g_{mj} be a representative member of T_{mj} for each j, m . It follows that we can identify each $f \in T$ by a sequence $(j_1(f), j_2(f), \dots)$ where $f \in T_{m,j_m(f)}$, $m = 1, 2, \dots$. That such $j_m(f)$ exist follows from the partition property.

Let $g_m = g_{m,j_m(f)}$, and define $\bar{f}_m, \underline{f}_m$ similarly. That $(j_1(f), j_2(f), \dots)$ characterize $W_P^0(f)$ (and f) (up to sets of probability 0) follows from

$$W_P^0(f) = W_P^0(g_1) + \sum_{m=1}^{\infty} (W_P^0(g_{m+1}) - W_P^0(g_m)) \quad (\text{D.2.4})$$

in the sense that

$$W_P^0(g_1) + \sum_{j=1}^m (W_P^0(g_{j+1}) - W_P^0(g_j)) = W_P^0(g_{m+1}) \quad (\text{D.2.5})$$

and

$$E_P(f - g_{m+1})^2(X) \leq E_P(\bar{f}_{m+1} - \underline{f}_{m+1})^2(X) \leq m^{-\alpha} \rightarrow 0 \quad (\text{D.2.6})$$

as $m \rightarrow \infty$. Let b, w_1, w_2, \dots be non-negative weights which will in fact depend on λ with $b + \sum_{m=1}^{\infty} w_m = 1$. Pick m_0 which will also depend on λ and write

$$W_P^0(f) = W_P^0(g_{m_0}) + \sum_{m=m_0}^{\infty} (W_P^0(g_{m+1}) - W_P^0(g_m)).$$

Note that

$$[|W_P^0(f)| \geq \lambda] \subset [|W_P^0(g_{m_0})| \geq b\lambda] \cup \bigcup_{m=m_0}^{\infty} [|W_P^0(g_{m+1}) - W_P^0(g_m)| \geq w_m \lambda].$$

Hence, using (D.2.5),

$$\begin{aligned} & P[\sup\{|W_P^0(f)| \geq \lambda : f \in T\}] \\ & \leq P[\max\{|W_P^0(g_{m_0,j})| : 1 \leq j \leq J_{m_0}\} \geq b\lambda] \\ & \quad + \sum_{m=m_0+1}^{\infty} P[\max\{|W_P^0(g_{m+1,j}) - W_P^0(g_{mj'})| : 1 \leq j \leq J_{m+1}, \\ & \quad 1 \leq j' \leq J_m\} \geq w_m \lambda] \mathbf{1}(T_{m+1,j} \cap T_{mj'} \neq \emptyset). \end{aligned} \tag{D.2.7}$$

The bound (D.2.7) is fundamental. It comes from a generalization of the classical $P[|X + Y| \geq \epsilon] \leq P[|X| \geq \frac{\epsilon}{2}] + P[|Y| \geq \frac{\epsilon}{2}]$, bounding $|W_P^0(g_m)|$ by $\max\{|W_P^0(g_{mj})| : 1 \leq j \leq J_m\}$ and using a similar bound for the differences. Go further and replace the terms on the right hand side of (D.2.7) by

$$\begin{aligned} & \sum_{j=1}^{J_{m_0}} P[|W_P^0(g_{mj})| \geq b\lambda] \\ & + \sum_{m=m_0+1}^{\infty} \sum_{j=1}^{J_{m+1}} \sum_{j'=1}^{J_m} P[|W_P^0(g_{m+1,j}) - W_P^0(g_{mj'})| \geq w_m \lambda] \mathbf{1}(T_{m+1,j} \cap T_{mj'} \neq \emptyset). \end{aligned} \tag{D.2.8}$$

Although $g_{m+1,j}$ may not belong to $T_{m,j_m(f)}$, because $f \in T_{m,j_m(f)} \cap T_{m+1,j_{m+1}(f)}$, it is still true that

$$E(g_{m+1,j} - g_{mj})^2 \leq 2\{E(g_{m+1,j} - f)^2 + E(g_{mj} - f)^2\} \leq 4m^{-2\alpha}.$$

Therefore, we can use the bounds (7.2.8) and (D.2.6) to bound (D.2.8) by

$$2J_{m_0} \exp\left\{-\frac{\lambda^2 b^2 \gamma^{-2}}{2}\right\} + 2 \sum_{m=m_0+1}^{\infty} J_m J_{m+1} \exp\left\{-\frac{\lambda^2}{8}(w_m^2 m^{2\alpha})\right\}. \tag{D.2.9}$$

Finally, use (ii) to bound (D.2.9) by

$$C_1 m_0^{\alpha d} \exp\left\{-\frac{\lambda^2 b^2 \gamma^{-2}}{2}\right\} + C_2 \sum_{m=m_0+1}^{\infty} m^{2\alpha d} \exp\left\{-\frac{\lambda^2}{2}(w_m^2 m^{2\alpha})\right\}, \tag{D.2.10}$$

where C_1 and C_2 are generic constants. By taking $w_m = cm^{-\alpha'}$, $\alpha' > 1$, it is not hard to show (Problem D.2.4) that we can bound (D.2.10) as specified in the statement of the theorem. \square

Discussion: This argument uses the idea of “chaining” introduced by Kolmogorov. It exploits the representation of $W_P^0(f)$ given by (D.2.4) which in turn reflects the “decimal” representation of f . The key observation is that the increments in (D.2.4) have variances which go down and, as a consequence of the bound (7.2.10), even if the deviation λ from the mean 0 is weighted by a small w_m , the probability of that deviation drops much more rapidly than w_m does. Note that this argument shows that the order of magnitude of the probability of deviation by λ or more of the maximum is essentially of the same order

$$\exp\{-\lambda^2/2\gamma^2\}$$

as that of a single term.

D.3 Functional Derivatives. Supplement to Section 7.2

Recall that if $f : O \rightarrow R$ where O is open $\subset R^d$ then, at a point $\mathbf{x} \in R^d$, f is partially differentiable with *partial derivatives* $\frac{\partial f}{\partial x_j}(\mathbf{x})$ iff the limits

$$\frac{\partial f}{\partial x_j}(\mathbf{x}) = \lim_{h \rightarrow 0} \frac{f(\mathbf{x} + h\mathbf{e}_j) - f(\mathbf{x})}{h}$$

exist, where e_j is the j th basis vector. The function f has a *total differential* iff a linear approximation to f holds. That is, suppose

$$Df(\mathbf{x})(\mathbf{t}) = \lim_{h \rightarrow 0} \frac{f(\mathbf{x} + h\mathbf{t}) - f(\mathbf{x})}{h} \quad (\text{D.3.1})$$

holds uniformly for $|\mathbf{t}|$ bounded. Then $Df(\mathbf{x})$ is the linear map $\mathbf{t} \rightarrow \sum_{j=1}^p \frac{\partial f}{\partial x_j}(\mathbf{x})t_j$, $\mathbf{t} = (t_1, \dots, t_p)^T$ given by (D.3.1). As is well known, functions with partial derivatives but no total differential are easy to exhibit.

If $f : B \rightarrow R$ where B is a linear space and

$$D_G f(x)(t) = \lim_{h \rightarrow 0} \frac{f(x + ht) - f(x)}{h}, \quad x, t \in B \quad (\text{D.3.2})$$

exists and is well defined, D_G is called the *Gâteaux derivative* of f at x in the direction t . The Gâteaux derivative is the analogue of the partial differential. The more useful and stronger notions correspond to uniformity in t statements about the limit and B , a normal linear space (Banach). That is, we want D such that

$$\sup \{|f(x + ht) - f(x) - hDf(x)(t)| : t \in \mathcal{T}\} = o(h). \quad (\text{D.3.3})$$

If \mathcal{T} is any bounded set in B it follows that $Df(x)(\cdot)$ is a bounded linear functional on B and $Df(x)(\cdot)$ is called the *Fréchet derivative* while if \mathcal{T} is any compact set, $Df(x_{\mathcal{T}})(\cdot)$

is called the *Hadamard* derivative and $Df(x)$ is a linear functional. If $B = R^d$ the two notions produce the same total differential.

Our interest typically focusses on $B = \{\text{All finite signed } \mu \text{ measures on } R^d \text{ with total variation norm, } \|\mu\| = \sup \{|\mu(A)| : A \in \mathcal{A}\} \text{ where } \mathcal{A} \text{ is the class of all sets in } R^d \text{ for which we can assign probability. Here "all signed measures } \mu \text{ on } R^d\text{" is the class of all } \mu \text{ on } \mathcal{A} \text{ of the form } \{aP - bQ : a, b \in R, P \text{ and } Q \text{ are probabilities on } R^p\}\}$. For this B , a natural subspace of linear functionals can be identified with the space of all bounded continuous functions q on R^d , with corresponding linear functional, $L_q(\mu) = \int q(v)d\mu(v)$ where $d\mu(v) = adF(v) - bdG(v)$ with F and G equal to the df's corresponding to P and Q , $v \in R^d$. If $Df(x)$ is of the form L_q and Df is Hadamard then if t is a signed measure,

$$Df(x)(t) = \int q(v)dt(v) \quad (\text{D.3.4})$$

and $q(x) = Df(x)(\delta_x)$ where δ_x is point mass at x .

Identifying these quantities with objects in R^d we see that (D.3.4) corresponds to the usual formula for the derivative in a particular direction and $q(x)$ corresponds to the partial derivatives in the directions of the coordinate axes. See Example 7.2.1, for these analogies. If B is not the space of all signed measures but just the convex cone of probabilities, we can still carry out limit (D.3.2) in a direction $t = G - F$ at $x = F$ where F and G are df's since, then, $F + h(G - F) = (1 - h)F + hG$, a member of the cone if $0 \leq h \leq 1$. The q we obtain is then not unique since $\int (q(v) + c)d(G - F_0)(v) = \int q(v)d(G - F_0)(v)$, but if we require $\int q(v)dF_0(v) = 0$ we obtain the derivative as a unique influence function.

The Hadamard sense is what is needed to carry out the delta method while its computation is via Gateaux. See van der Vaart (1998) for a careful treatment using this notion of derivative and the preceding treatment of Huber (1974) and Reeds (1976).

D.4 Asymptotics for the Cox Estimate. Supplement to Section 9.2.2

The Influence Function of the Cox Partial Likelihood Estimate

1. *Derivation of the Gâteaux derivative of $\dot{\Gamma}_\tau(\beta, P)$.* To find the Gâteaux derivative set

$$P_\varepsilon = (1 - \varepsilon)P + \varepsilon Q, \quad P_{j\varepsilon} = (1 - \varepsilon)P_j + \varepsilon Q_j; \quad p_{j\varepsilon} = dP_{j\varepsilon}, \quad j = 1, 2.$$

Then we seek

$$\frac{\partial}{\partial \varepsilon} \dot{\Gamma}_\tau(\beta, P_\varepsilon) = \frac{\partial}{\partial \varepsilon} \left[\int_0^\tau -\frac{\dot{S}_0}{S_0}(t, \beta, P_\varepsilon) dP_{2\varepsilon}(t) + \int 1(t \leq \tau) \mathbf{z} dP_\varepsilon(\mathbf{z}, t) \right] \Big|_{\varepsilon=0} \equiv I + II.$$

By taking Q_1 to be pointmass at \mathbf{Z} , we find

$$II = Z1(T \leq \tau) - E[Z1(T \leq \tau)].$$

Write $S_0(\varepsilon)$, $\dot{S}_0(\varepsilon)$ for $S_0(t, \beta, P_\varepsilon)$, $\dot{S}(t, \beta, P_\varepsilon)$ and S_0, S_0 for $S_0(0)$, $\dot{S}_0(0)$. For I , we need

$$\frac{\partial}{\partial \varepsilon} \frac{\dot{S}_0}{S_0}(\varepsilon) = \frac{\partial}{\partial \varepsilon} \dot{S}_0(\varepsilon) S_0^{-1} - \dot{S}_0 S_0^{-2} \left(\frac{\partial}{\partial \varepsilon} S_0(\varepsilon) \right),$$

where

$$\frac{\partial}{\partial \varepsilon} S_0(\varepsilon) = \frac{\partial}{\partial \varepsilon} E_{P_\varepsilon}(e^{\boldsymbol{\beta}^T \mathbf{z}} \mathbf{1}(T \geq t)) = e^{\boldsymbol{\beta}^T \mathbf{z}} \mathbf{1}(T \geq t) - S_0$$

is obtained by letting Q be pointmass at $X = (\mathbf{Z}, T)$. Similarly,

$$\frac{\partial}{\partial \varepsilon} \dot{S}_0(\varepsilon) = \mathbf{Z} e^{\boldsymbol{\beta}^T \mathbf{z}} \mathbf{1}(T \geq t) - \dot{S}_0 .$$

Thus

$$\begin{aligned} \frac{\partial}{\partial \varepsilon} \frac{\dot{S}_0}{S_0}(\varepsilon) &= S_0^{-1} [\mathbf{Z} e^{\boldsymbol{\beta}^T \mathbf{z}} \mathbf{1}(T \geq t) - \dot{S}_0] - S_0^{-2} \dot{S}_0 [e^{\boldsymbol{\beta}^T \mathbf{z}} \mathbf{1}(T \geq t) - S_0] \\ &= S_0^{-1} e^{\boldsymbol{\beta}^T \mathbf{z}} \mathbf{1}(T \geq t) [\mathbf{Z} - (\dot{S}_0 / S_0)] . \end{aligned}$$

Because $p_{2\varepsilon}|_{\varepsilon=0} = p_2$ and $\partial p_{2\varepsilon}(t)/\partial \varepsilon = q_2 - p_2$ for $T \leq \tau$, this gives

$$I = -e^{\boldsymbol{\beta}^T \mathbf{z}} \int_0^T S_0^{-1} [\mathbf{Z} - \frac{\dot{S}_0}{S_0}(t, \beta, P)] dP_2(t) - \frac{\dot{S}_0}{S_0}(T, \beta, P) + \int_0^\tau \frac{\dot{S}_0}{S_0}(t, \beta, P) dP_2(t) .$$

Now $I + II = \gamma(X, P)$.

2. Proof of Theorem 9.2.2. Suppose we can show that $\gamma(X, P)$ is an influence function for $\ddot{\Gamma}_\tau^{-1}(\beta, \widehat{P})$ in the sense of (7.2.2). Then

$$\sqrt{n}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}) = \ddot{\Gamma}_\tau^{-1}(\beta^*, \widehat{P}) \sqrt{n} \left[n^{-1} \sum_{i=1}^n \gamma(X_i, P) + o_P(1) \right] .$$

By Slutsky's theorem (Corollary 7.1.2) for processes, we can replace $\ddot{\Gamma}_\tau^{-1}(\beta^*, \widehat{P})$ with $\ddot{\Gamma}_\tau^{-1}(\beta, P)$ and the results (ii) and (iii) follow. We leave details to the reader in Problem D.4.1. \square

D.5 Markov Dependent Data. Supplement to Section 10.4

A. Markov Sequences

Let X_1, X_2, \dots be a sequence of \mathcal{X} valued random variables. The i.i.d. sequences we have considered so far have two critical properties that we will generalize in this section.

(i) Independence

$$\mathcal{L}(X_m | X_{m-1}, \dots, X_1) = \mathcal{L}(X_m), \quad m \geq 2 .$$

(ii) Identical distribution

$$\mathcal{L}(X_m) = \mathcal{L}(X_1), \quad m \geq 2 .$$

The natural generalization of the first property is the *Markov property* defined by

$$\mathcal{L}(X_m | X_{m-1}, \dots, X_1) = \mathcal{L}(X_m | X_{m-1}), \quad m \geq 2 . \tag{D.5.1}$$

The most important generalization of the second property is *stationarity* defined by

$$\mathcal{L}(X_1, \dots, X_m) = \mathcal{L}(X_{1+k}, \dots, X_{m+k}), \quad m \geq 1, \quad k \geq 0. \quad (\text{D.5.2})$$

An intermediate property (Problem D.5.1) implied by (D.5.1) and (D.5.2), but not by either one singly, is called (e.g. Grimmett and Stirzaker, Section 6.1) *homogeneous Markov*. It is defined by (D.5.1) and “ $\mathcal{L}(X_m|X_{m-1})$ does not depend on m ,” that is,

$$\mathcal{L}(X_m|X_{m-1}) = \mathcal{L}(X_2|X_1), \quad m \geq 2.$$

A sequence is *stationary* and *Markov* iff it is homogeneous Markov and

$$\mathcal{L}(X_m) = \mathcal{L}(X_1), \quad m \geq 2. \quad (\text{D.5.3})$$

Thus, an i.i.d. sequence is stationary Markov.

To study Markov sequences $\{X_n, n \geq 1\}$ it is best to think of n as time and X_n as the state of a dynamical system initiated at X_1 . Thus, we refer to \mathcal{X} as the *state space* and its members as *states*. We think of the evolution mechanism, $\mathcal{L}(X_2|X_1)$, as being given by a *Markov kernel* $K : \mathcal{X} \times \mathcal{X} \rightarrow R$ defined by

$$K(x, y) = P(X_2 = y|X_1 = x), \quad x, y \in \mathcal{X}$$

if \mathcal{X} is discrete, countable. If \mathcal{X} is Euclidean and $X_2|X_1$ has a continuous case density $P_{X_2|X_1}(x_2|x_1)$, we define the Markov kernel by

$$K(x, y) = P_{X_2|X_1}(y|x), \quad x, y \in \mathcal{X}.$$

We call $\mathcal{L}(X_1)$ the *initial distribution* and if the Markov sequence is stationary, $\mathcal{L}(X_1)$ is a *stationary distribution* because in this case $\mathcal{L}(X_n) = \mathcal{L}(X_1)$ for all $n \geq 1$.

In what follows, we will define stationary distributions more generally and describe how they, under certain conditions, approximate the distribution of X_n for n large.

B. Homogeneous Finite Markov Chains

We sketch the theory for $\mathcal{X} = \{1, \dots, N\}$, that is for finite Markov chains. In fact, for Markov Chain Monte Carlo (MCMC), and many other applications, we need Markov sequences for \mathcal{X} countable or \mathcal{X} general Euclidean. Such generalizations hold only under additional conditions which we shall give appropriate references to, as needed.

Homogenous chains satisfy $P(X_{m+n} = y|X_m = x) = P(X_{n+1} = y|X_1 = x)$, so we can call

$$p_n(x, y) \equiv P(X_{n+1} = y|X_1 = x)$$

n-step transition probabilities. The *n-step transition matrix* \mathbf{P}_n is defined by

$$\mathbf{P}_n = (p_n(i, j))_{N \times N}.$$

A key property is

Lemma D.5.1. *If X_1, X_2, \dots is a finite homogeneous Markov chain, then*

$$\mathbf{P}_n = \mathbf{K}^n, \quad n = 1, 2, \dots$$

where \mathbf{K}^n is the n th power of the transition matrix $\mathbf{K} = (K(i,j))_{N \times N}$. In particular

$$P(X_{n+1} = j | X_1 = i) = (\mathbf{K}^n)_{i,j},$$

the (i,j) entry of \mathbf{K}^n .

Proof. Argue by induction: $\mathbf{P}_1 = \mathbf{K}^1$ by definition. By A.4.4, the Markov property, homogeneity, and induction,

$$\begin{aligned} P(X_n = j | X_1 = i) &= \sum_{k=1}^N P(X_n = j | X_{n-1} = k, X_1 = i) P(X_{n-1} = k | X_1 = i) \\ &= \sum_{k=1}^N P(X_n = j | X_{n-1} = k) P(X_{n-1} = k | X_1 = i) \\ &= (\mathbf{K}^{n-1} \mathbf{K})_{i,j} = (\mathbf{K}^n)_{i,j}. \end{aligned}$$

□

Next, consider the marginal probabilities

$$\mathbf{q}^{(n)} = (p_1^{(n)}, \dots, p_N^{(n)})^T, \quad p_i^{(n)} = P(X_n = i), \quad i \in \{1, \dots, N\}.$$

Lemma D.5.2. Under the conditions of Lemma D.5.1, $(\mathbf{q}^{(n)})^T = (\mathbf{q}^{(1)})^T \mathbf{K}^n$.

Proof. By A.4.4 and Lemma D.5.1, $p_j^{(n)} = \sum_i P(X_n = j | X_1 = i) P(X_1 = i) = \sum_i p_j^{(1)} p_n(i, j) = (\mathbf{q}^{(1)} \mathbf{P}_n)_j = ((\mathbf{q}^{(1)})^T \mathbf{K}^n)_j$. □

Thus the behavior of the chain is determined by the initial transition matrix \mathbf{K} and the *initial distribution* $\mathbf{q}^{(1)}$. Moreover, we can determine the asymptotic distribution of X_n by examining the limit of \mathbf{K}^n .

C. Some Definitions

Important features of homogeneous Markov chains include

- (i) **Irreducibility:** All points in \mathcal{X} are accessible from any other point. That is, $P(X_m = j | X_1 = i) > 0$ for some m , all i, j . For finite chains this is equivalent to “ \mathbf{K}^m has all elements positive for some m on.”
- (ii) **Recurrence:** A chain is *recurrent* if for all $x \in \mathcal{X}$,

$$P[X_k = x \text{ for some } k | X_1 = x] = 1$$

- (iii) **Positive recurrent:** This refers to recurrent chains with $E\tau_x < \infty$ for all $x \in \mathcal{X}$, where $\tau_x = \min\{n \geq 1 : X_n = x\}$ is the first visit time to x .
- (iv) **Aperiodicity:** The *period* k of a state x is the largest integer k for which returns to state x occurs in multiples of k steps. Formally, $k = \gcd\{n : P(X_n = x | X_1 = x) > 0\}$, where “*gcd*” denotes greatest common divisor. A Markov chain is *aperiodic* if $k = 1$ for all $x \in \mathcal{X}$.

- (v) A **stationary probability distribution** on \mathcal{X} is a vector $\boldsymbol{\pi} = (\pi(1), \dots, \pi(N))^T$ such that $\pi(i) \geq 0$, $\sum \pi(i) = 1$, $\sum_i \pi(i) K(i, j) = \pi(j)$, that is,

$$\boldsymbol{\pi}^T \mathbf{K} = \boldsymbol{\pi}^T.$$

Equivalently, if $\boldsymbol{\pi}$ is the distribution of X_n and $\boldsymbol{\pi}^T \mathbf{K} = \boldsymbol{\pi}^T$, then this sequence is stationary Markov. A stationary distribution may not exist.

- (vi) A stationary Markov chain with kernel K and $p_i = P(X_n = x_i)$ satisfies *detailed balance* if

$$p_i K(i, j) = p_j K(j, i).$$

Summing over j we obtain

$$p_i = \sum_{j=1}^N p_j K(j, i)$$

so that $(p_1, \dots, p_N)^T = \boldsymbol{\pi}$ is a stationary distribution. Such chains are also called *reversible* because detailed balance is equivalent to $(X_1, X_2) \sim (X_2, X_1)$. Algebraically, if $D = \text{diag}(p_1, \dots, p_N)$, then $\mathbf{K}D = D\mathbf{K}^T$.

- (vii) If $\boldsymbol{\pi} = (\pi(1), \dots, \pi(N))^T$ is a stationary distribution, the *stationary transition matrix* is

$$\boldsymbol{\Pi} = \begin{pmatrix} \pi(1), & \dots, & \pi(N) \\ \vdots & & \vdots \\ \pi(1), & \dots, & \pi(N) \end{pmatrix}.$$

If Y_1, Y_2, \dots is a homogeneous Markov chain with transition matrix $\boldsymbol{\Pi}$, then Y_1, Y_2, \dots satisfy detailed balance and Y_1, Y_2, \dots are i.i.d. as $\boldsymbol{\pi}$.

- (viii) The *operator norm* on the class of $N \times N$ matrices is defined by

$$\|M\| = \sup\{|M\mathbf{x}| : |\mathbf{x}| \leq 1\}$$

where \mathbf{x} is $N \times 1$ and $|\cdot|$ denotes Euclidean norm. The *Frobenius norm* for $M = (m_{ij})_{N \times N}$ is $(\sum_{i,j} m_{ij}^2)^{\frac{1}{2}}$. It is equivalent to $\|\cdot\|$ for finite matrices.

D. Fundamental Theorem of Finite Markov Chains. Let X_1, X_2, \dots be an irreducible, aperiodic, homogeneous finite Markov chain. Then there exists a unique stationary distribution $\boldsymbol{\pi}$, and if \mathbf{P}_n is the n -step transition matrix, then

$$\|\mathbf{P}_n - \boldsymbol{\Pi}\| \rightarrow 0. \tag{D.5.4}$$

If $\boldsymbol{\pi}$ satisfies detailed balance, then more is true: there exists $c < \infty$ and $\rho < 1$ such that for all n ,

$$\|\mathbf{P}_n - \boldsymbol{\Pi}\| \leq c\rho^n. \tag{D.5.5}$$

Note that (D.5.4) implies that any initial distribution $\mathbf{q}^{(1)}$,

$$|(\mathbf{q}^{(1)})^T \mathbf{P}_n - \boldsymbol{\pi}^T| \rightarrow 0. \quad (\text{D.5.6})$$

Moreover, (D.5.5) is equivalent to (Problem D.5.4)

$$\sum_{i=1}^N |P(X_n = i) - \pi(i)| \leq M\rho^n, \quad (\text{D.5.7})$$

for some $M < \infty$. That is we can compute the stationary distribution arbitrarily closely by starting at any distribution and running the Markov chain. This is sometimes called the *forgetting property*.

The proof of the theorem and bounds on c are in Diaconis and Strook (1991). See also Grimmett and Stirzaker (2001, page 295).

Remark D.5.1. As we noted, such results are needed for countable state spaces. What modifications are needed for this case? See Grimmett and Stirzaker (2001, Chapter 6, page 232) for details. The conditions needed are irreducibility, aperiodicity, and positive recurrence. Unfortunately only (D.5.4) can be established under these conditions for countable \mathcal{X} .

The potential slowness of convergence if $|\mathcal{X}| = \infty$ is not only a weakness of the proof. In fact, (see Gilks et al (1995)) arbitrarily slowly converging Markov chains can be constructed. The situation when $\mathcal{X} = \mathbb{R}^d$ is even worse. Conditions for (D.5.4) with $\|\cdot\|$ the variational norm are more difficult to formulate and convergence in high dimensional situations is known empirically and theoretically to typically be very slow. See Meyn and Tweedie (2009).

D.6 Asymptotics for Locally Polynomial Estimates. Supplement to Section 11.6

1. Proof of Theorem 11.6.1. We will use (11.6.11). Write

$$S_{nj} = h^{j-1} \sum_{i=1}^n U_i^j K(U_i), \quad j = 0, 1, \dots \quad (\text{D.6.1})$$

where $U_i = (X_i - x)/h$ with density $hf(x + hu)$. By Chebychev's inequality (A.15.2) and Taylor expansion,

$$\begin{aligned} S_{nj} &= ES_{nj} + O_P(\sqrt{\text{Var } S_{nj}}) \\ &= nh^j \int u^j K(u) f(x + hu) du + O_P(h^{j-1} \sqrt{nE[U^{2j} K^2(U)]}) \quad (\text{D.6.2}) \\ &= nh^j \{f(x)m_j + hf'(x)m_{j+1} + O_P(c_n)\} \end{aligned}$$

where $c_n = (nh)^{-\frac{1}{2}}(1 + h^2)$. The last equality follows because if K is symmetric, so is K^2 , $\nu_5(K) = 0$, and

$$\begin{aligned} E[U^{2j}K^2(U)] &= h \int u^{2j}K^2(u)f(x+hu)du \\ &= h \int u^{2j}K^2(u)\{f(x) + huf'(x) + \frac{1}{2}f''(x+hu^*)h^2\}du, |u^*| \leq |u| \\ &= h[f(x)\nu_{2j}(K) + O(h^2)]. \end{aligned}$$

Next, for $\epsilon = \epsilon_n = O_P(c_n)$ and any constant c , we have

$$(c + \epsilon)^{-1} = c^{-1} - c^{-2}\epsilon + O_P(\epsilon^2) \quad \text{as } \epsilon \rightarrow 0$$

which gives

$$S_{no}^{-1} = n^{-1}\{[f(x)]^{-1} + O_P(c_n)\}. \quad (\text{D.6.3})$$

By combining (11.6.11), (D.6.1), (D.6.2), and (D.6.3), the result follows. \square

2. Proof of Theorem 11.6.2. For X_i close to x consider the approximation

$$\sigma^2(X_i) = \sigma^2(x) + O_P(X_i - x).$$

Using this and the arguments leading to (D.6.2) with $j = 0$, we have

$$\sum_{i=1}^n \sigma^2(X_i) K_h^2(X_i - x) = h^{-1} n f(x) \sigma^2(x) \nu(K) \{1 + o_P(1)\}.$$

Similarly, $S_{no}^{-2} = n^{-2}\{f^{-2}(x) + o_P(1)\}$, and the result follows from (11.6.12). \square

3. The Bias of Local Polynomial Estimates. Let $\mathbf{S}_n = \mathbf{Z}^T \mathbf{W} \mathbf{Z} = (S_{n,j+\ell})_{0 \leq j \leq p, 1 \leq \ell \leq p}$, $\mathbf{T}_n = (S_{n,p+1}, \dots, S_{n,2p+1})$, $\mathbf{S} = (m_{j+\ell})_{0 \leq j \leq p, 0 \leq \ell \leq p}$, $\mathbf{S}^{-1} = (S^{j\ell})_{0 \leq j \leq p, 0 \leq \ell \leq p}$, $\mathbf{m}_p = (m_{p+1}, \dots, m_{2p+1})$, and $\mathbf{H} = \text{diag}(1, h, \dots, h^p)$.

Theorem D.6.1 Under the conditions of Theorem 11.6.3,

$$(i) \quad \text{Bias} [\widehat{\boldsymbol{\beta}}(x) | \mathbf{X}] = \mathbf{H}^{-1} \mathbf{S}^{-1} \mathbf{m}_p \boldsymbol{\beta}_{p+1} h^{p+1} \{1 + o_P(1)\}. \quad (\text{D.6.4})$$

Proof. Using (11.6.17), (11.6.18), and (D.6.2), the conditional bias is

$$\begin{aligned} \mathbf{S}_n^{-1} \mathbf{Z}^T \mathbf{W} [\beta_{p+1}(X_i - x)^{p+1} + o_P(X_i - x)^{p+1}]_{1 \leq i \leq n} \\ = \mathbf{S}_n^{-1} \beta_{p+1} \mathbf{T}_n + o_P(nh^{p+1}) \\ = \mathbf{S}_n^{-1} \beta_{p+1} [nh^{p+1} f(x) \mathbf{H} \mathbf{m}_p + o(h)] + o_P(nh^{p+1}). \end{aligned} \quad (\text{D.6.5})$$

Next note that (D.6.2) also gives

$$\mathbf{S}_n = n f(x) \mathbf{H} \mathbf{S} \mathbf{H} \{1 + o_P(1)\} \quad (\text{D.6.6})$$

which together with (D.6.5) yields Theorem D.6.1. Theorem 11.6.3 follows by using $\mathbf{H}^{-1} = \text{diag}(1, h^{-1}, \dots, h^{-p})$ and a little algebra (Problem D.6.2). \square

Let $\eta_{p+1} \equiv \eta_{p+1}(K)$ be the first entry of the vector $\mathbf{S}^{-1}\mathbf{m}_p$. A useful result is (see Problem D.6.1):

Lemma D.6.1. *Under the conditions of Theorem 11.6.3, if p is even, then $\eta_{p+1}(K) = 0$.*

Proof. Recall that $\mathbf{S} = (m_{j+l})_{0 \leq j \leq p, 0 \leq l \leq p}$ and $\mathbf{m}_p = (m_{p+1}, \dots, m_{2p+1})$. When K is symmetric, then $m_j = 0$ for j odd. It follows that the vector $\mathbf{S}^{-1}\mathbf{m}_p$ has zero as its first entry, that is $\eta_{p+1}(K) = 0$. \square

The local linear case

Lemma D.6.2. *Assume the conditions of Theorem 11.6.3. If $p = 1$, then $\eta_2(K) = m_2(K)$.*

Proof. The symmetry of K implies that $m_j = m_j(K) = 0$ for j odd. Thus, in the local linear case where $p = 1$, we have $\mathbf{S}^{-1} = \text{diag}(1, m_2^{-1})$ and $\mathbf{m}_1 = (m_2, 0)$, which implies $\eta_2(K) = m_2(K)$. \square

We have shown:

Theorem D.6.2. *Under the conditions of Theorem 11.6.3, if $p = 1$, then*

$$\text{Bias}[\hat{\mu}(x)|\mathbf{X}] = \frac{1}{2}m_2(K)\mu''(x)h^2 + o_P(h^2).$$

4. The Variance of Local Polynomial Estimates. Let $\sigma^2(x) = \text{Var}(Y|X = x)$, $\Sigma = \text{diag}[K_h^2(X_i - x)\sigma^2(X_i)]_{1 \leq i \leq n}$, then from (11.6.16),

$$\text{Var}(\hat{\beta}(x)|\mathbf{X}) = (\mathbf{Z}^T \mathbf{W} \mathbf{Z})^{-1} (\mathbf{Z}^T \Sigma \mathbf{Z}) (\mathbf{Z}^T \mathbf{W} \mathbf{Z})^{-1}. \quad (\text{D.6.7})$$

Let $\mathbf{S}_n^* = \mathbf{Z}^T \Sigma \mathbf{Z} = (S_{n,j+\ell}^*)_{0 \leq j \leq p, 0 \leq \ell \leq p}$, where

$$S_{n,j}^* = \sum_{i=1}^n (X_i - x)^j K_h^2(X_i - x) \sigma^2(X_i),$$

and let $\mathbf{S}^* = (\nu_{j+\ell})_{0 \leq j \leq p, 0 \leq \ell \leq p}$, where $\nu_p(K) = \int [K(u)]^2 du$.

Theorem D.6.3. Under the conditions of Theorem 11.6.4,

$$\text{Var}(\hat{\beta}(x)|\mathbf{X}) = \frac{\sigma^2(x)}{f(x)nh} H^{-1} S^{-1} S^* S^{-1} H^{-1} \{1 + o_P(1)\}.$$

Proof: By using the expansion leading to (D.6.2), we find

$$S_n^* = nh^{-1} f(x) \sigma^2(x) H S^* H \{1 + o_P(1)\}. \quad (\text{D.6.8})$$

Theorems D.6.3 and 11.6.4 follow from this and (D.6.7) (Problem D.6.3).

D.7 Supplement to Section 12.2.2

Proof of Theorem 12.2.1.

We temporarily modify the estimate $\hat{\mu}_{\text{NW}}(\mathbf{x})$ slightly for proving that the IMSE converges to zero at an appropriate rate. Consider the modified estimate,

$$\hat{\mu}_{\text{NW}}(\mathbf{x}, a_n) \equiv \hat{\mu}_{\text{NW}}(\mathbf{x}) \mathbf{1}[\hat{f}(\mathbf{x}) \geq a_n]$$

where $a_n \downarrow 0$ will be chosen in the proof.

We begin by establishing (i) and (ii) for $\hat{\mu}_{\text{NW}}(\mathbf{x}, a_n)$. Write $K_i^* = K_h(\mathbf{X}_i - \mathbf{x})$, $J = \mathbf{1}(\hat{f}(\mathbf{x}) \geq a_n)$, and $(0/0) = 0$. We want MSE $(\hat{\mu}_{\text{NW}}(\mathbf{X}, a_n)) = E(D^2)$ where

$$D = (n\hat{f})^{-1} J \sum_{i=1}^n K_i^* Y_i - \mu(\mathbf{x}) .$$

Next write $\mu(\mathbf{x}) = J\mu(\mathbf{x}) + (1 - J)\mu(\mathbf{x})$ and note that

$$J\mu(\mathbf{x}) = J(n\hat{f})^{-1} \sum_{i=1}^n \mu(\mathbf{x}) K_i^* .$$

It follows that

$$D = (n\hat{f})^{-1} J \sum_{i=1}^n K_i^* [Y_i - \mu(\mathbf{x})] - (1 - J)\mu(\mathbf{x}) . \quad (\text{D.7.1})$$

Using $\hat{f}^{-1} = f^{-1} + (\hat{f}^{-1} - f^{-1})$, we have

$$\begin{aligned} D &= n^{-1} f^{-1} J \sum_{i=1}^n K_i^* [Y_i - \mu(\mathbf{x})] - (1 - J)\mu(\mathbf{x}) \\ &\quad + n^{-1} (\hat{f}^{-1} - f^{-1}) J \sum_{i=1}^n K_i^* [Y_i - \mu(\mathbf{x})] \\ &\equiv D_1 + R_1 + R_2 . \end{aligned} \quad (\text{D.7.2})$$

We will show that D_1 is the main term with $E(D_1^2)$ of order $O(h^2) + O((nh^d)^{-1})$ and that R_1 and R_2 are remainder terms with $E(R_1^2)$ and $E(R_2^2)$ of smaller order than $E(D_1^2)$. We will use the inequality

$$E(D^2) \leq 3[E(D_1^2) + E(R_1^2) + E(R_2^2)] .$$

Note that because $0 \leq J \leq 1$, $E(D_1^2) \leq E(D_0^2)$ where

$$D_0 = n^{-1} f^{-1} \sum K_i^* [Y_i - \mu(\mathbf{x})] .$$

Next we use $E(D_0^2) = \text{Var}(D_0) + E^2(D_0)$ and condition on $\mathbf{X}^{(n)} = (X_1, \dots, X_n)$. Then, since $E(Y_i|\mathbf{X}^{(n)}) = \mu(\mathbf{X}_i)$,

$$\begin{aligned} E(D_0) &= E[E\{D_0|\mathbf{X}^{(n)}\}] \\ &= n^{-1}f^{-1}E\left[\sum_{i=1}^n K_i^*E\{Y_i - \mu(\mathbf{x})\}|\mathbf{X}^{(n)}\}\right] \\ &= n^{-1}f^{-1}E[nK_h(\mathbf{X} - \mathbf{x})[\mu(\mathbf{X}) - \mu(\mathbf{x})]] . \end{aligned}$$

Make the change of variable $\mathbf{Z} = (\mathbf{X} - \mathbf{x})h^{-1}$; then \mathbf{Z} has density $h^d f(\mathbf{x} + h\mathbf{z})$ and $K_h(\mathbf{X} - \mathbf{x}) = h^{-d}K(\mathbf{Z})$. Thus,

$$\begin{aligned} |E(D_0)| &\leq \varepsilon^{-1} \int K(\mathbf{z})|\mu(\mathbf{x} + h\mathbf{z}) - \mu(\mathbf{x})|f(\mathbf{x} + h\mathbf{z})d\mathbf{z} \\ &\leq M\varepsilon^{-1}h \int K(\mathbf{z})|\mathbf{z}|f(\mathbf{x} + h\mathbf{z})d\mathbf{z} = O(h) , \end{aligned} \quad (\text{D.7.3})$$

uniformly for $\mathbf{x} \in \mathcal{S}$ and $G \in \mathcal{G}$.

To bound $\text{Var}(D_0)$, we use $f^{-1} \leq \varepsilon^{-1}$, and write

$$\text{Var}(D_0) \leq n^{-2}\varepsilon^{-2}\text{Var}\left\{\sum_{i=1}^n K_i^*[Y_i - \mu(\mathbf{x})]\right\} = n^{-1}\varepsilon^{-2}\text{Var}\{K[Y - \mu(\mathbf{x})]\} ,$$

where $K^* = K_h(\mathbf{X} - \mathbf{x})$ and, by (1.4.6),

$$\begin{aligned} \text{Var}\{K^*[Y - \mu(\mathbf{x})]\} &= E\{\text{Var}[K^*[Y - \mu(\mathbf{x})]|\mathbf{X}]\} + \text{Var}\{K^*E[Y - \mu(\mathbf{x})|\mathbf{X}]\} \\ &= E[(K^*)^2\sigma^2(\mathbf{X})] + \text{Var}\{K^*[\mu(\mathbf{X}) - \mu(\mathbf{x})]\} \\ &\leq ME[(K^*)^2] + E\{(K^*)^2[\mu(\mathbf{X}) - \mu(\mathbf{x})]^2\} . \end{aligned}$$

Next note that

$$E[(K^*)^2] = \int h^{-d}K^2(\mathbf{z})f(\mathbf{x} + h\mathbf{z})d\mathbf{z} = O(h^{-d}) . \quad (\text{D.7.4})$$

Using $\mathbf{X} = \mathbf{x} + h\mathbf{Z}$ and $|\mu(\mathbf{x} + h\mathbf{Z}) - \mu(\mathbf{x})| \leq h|\mathbf{Z}|$, we have

$$E\{(K^*)^2[\mu(\mathbf{X}) - \mu(\mathbf{x})]^2\} \leq h^{2-d} \int \mathbf{z}^2 K^2(\mathbf{z})f(\mathbf{x} + h\mathbf{z})d\mathbf{z} = O(h^{2-d}) . \quad (\text{D.7.5})$$

It follows from (D.7.3), (D.7.4), and (D.7.5) that the leading term in the MSE satisfies

$$E(D_1^2) = O(h^2) + O((nh^d)^{-1})$$

uniformly for $\mathbf{x} \in \mathcal{S}$, $G \in \mathcal{G}$. Thus $E(D_1^2)$ equals $A(h) \equiv C_1h^2 + C_2(nh^d)^{-1}$ plus smaller order terms for some constants C_1 and C_2 depending only on the constants in A_1, \dots, A_5 . The expression $A(h)$ is minimized by taking

$$h = bn^{-1/(2+d)}$$

for some $b > 0$ depending on C_1 and C_2 . This gives

$$\inf_{h>0} E(D_1^2) = O(n^{-2/(2+d)}) ,$$

which is the result (i) in Theorem 12.2.1.

We next show that $E(R_1^2)$ is of smaller order than $E(D_1^2)$. Note that

$$E(R_1^2) = \mu^2(\mathbf{x}) E(1 - J)^2 = \mu^2(\mathbf{x}) E|1 - J| = \mu^2(\mathbf{x}) P(\hat{f}(\mathbf{x}) < a_n) .$$

Next set $f_h(\mathbf{x}) = EK_h(\mathbf{X} - \mathbf{x})$, and write

$$P(\hat{f}(\mathbf{x}) < a_n) = P\left(n^{-1} \sum [K_h(\mathbf{X}_i - \mathbf{x}) - f_h(\mathbf{x})] < a_n - f_h(x)\right)$$

where $f_h(\mathbf{x}) = \int K(\mathbf{z}) f(\mathbf{x} + h\mathbf{z}) d\mathbf{z}$. By the multivariate version of (11.2.5),

$$f_h(x) = f(x) + o(1)$$

as $h \rightarrow 0$ uniformly for $\mathbf{x} \in \mathcal{S}$. Because $f(\mathbf{x}) \geq \varepsilon$ for all $\mathbf{x} \in \mathcal{S}$, we can select $\delta > 0$ so that $f_h(x) \geq \frac{1}{2}\varepsilon \equiv \varepsilon_0$ for all $\mathbf{x} \in \mathcal{S}$ and all $h \leq \delta$. It follows that with $Y'_i = f_h(\mathbf{x}) - K_h(\mathbf{X}_i - \mathbf{x})$, if we select N large enough so that $a_n < \varepsilon_0$ for $n \geq N$, then

$$P(\hat{f}(\mathbf{x}) < a_n) \leq P(n^{-1}|\Sigma Y'_i| > \varepsilon_0 - a_n) .$$

We will apply Hoeffding's inequality (7.1.10). Thus we need

$$\sigma_n^2 = \text{Var}(\Sigma Y'_i) = n \text{Var}(K_h(\mathbf{X} - \mathbf{x})) \leq nE[K_h^2(\mathbf{X} - \mathbf{x})] = O(nh^{-d}) \quad (\text{D.7.6})$$

by (D.7.4). It follows from (7.1.10) that for some constants c_1, c_2 , and c_3 ,

$$P(\hat{f}(\mathbf{x}) > a_n) \leq c_1 \exp\left\{-c_2 n^2 (\varepsilon_0 - a_n)^2 \left[1 + \frac{c_3(\varepsilon_0 - a_n)}{nh^{-d}}\right]^{-1}\right\} .$$

When $n^{-1}h^d \rightarrow 0$, the right hand side is of the same order as

$$\exp\{-\delta n^2 n^{-1} h^d\} = \exp\{-\delta nh^d\}, \text{ some } \delta > 0 .$$

It follows that for $a_n \downarrow 0$, uniformly for $\mathbf{x} \in \mathcal{S}, G \in \mathcal{G}$,

$$E_G(R_1^2) = o((n^{-1}h^d)^{-1}) .$$

We turn to $E(R_2^2)$ and note that because $0 \leq J \leq 1$,

$$E(R_2^2) \leq n^{-2} E\left[(\hat{f}^{-1} - f^{-1})^2 \{\Sigma K_i [Y_i - \mu(\mathbf{x})]\}^2\right] .$$

Using $\hat{f}^{-1} - f^{-1} = (\hat{f}f)^{-1}(f - \hat{f})$, $\hat{f}^{-1} \leq a_n$, and $(\hat{f} - f)^2 \leq 2[(f_h - f)^2 + (\hat{f} - f_h)^2]$, we have

$$E(R_2^2) \leq 2(na_n)^{-2} E\left\{[(f_h - f)^2 + (\hat{f} - f_h)^2] \{\Sigma K_i [Y_i - \mu(\mathbf{x})]\}^2\right\} \equiv R_{21} + R_{22} .$$

To bound R_{21} , note that

$$\begin{aligned} |f_h(\mathbf{x}) - f(\mathbf{x})| &\leq \int K(\mathbf{z}) |f(\mathbf{x} + h\mathbf{z}) - f(\mathbf{z})| f(\mathbf{x} + h\mathbf{z}) d\mathbf{z} \\ &\leq M \int K(\mathbf{z}) h |\mathbf{z}| d|\mathbf{z}| = O(h) . \end{aligned} \quad (\text{D.7.7})$$

Moreover, writing $Y_i - \mu(\mathbf{x}) = [Y_i - \mu(\mathbf{X}_i)] + [\mu(\mathbf{X}_i) - \mu(\mathbf{x})]$ and $\mu_i^* = \mu(\mathbf{X}_i) - \mu(\mathbf{x})$,

$$\begin{aligned} E[\Sigma K_i^*[Y_i - \mu(\mathbf{x})]]^2 &\leq 2E[E(\{\Sigma K_i^*[Y_i - \mu(\mathbf{X}_i)]\}^2 | \mathbf{X}^{(n)})] + 2E[\{\Sigma K_i^*[\mu(\mathbf{X}_i) - \mu(\mathbf{x})]\}^2] \\ &= 2nE[(K^*)^2 \sigma^2(\mathbf{X})] + 2E\left(\sum_{i=1}^n K_i^* \mu_i^*\right)^2 \equiv 2(T_1 + T_2) . \end{aligned}$$

The first term T_1 is of order $O(nE(K^2))$ because $\sigma^2(\mathbf{X})$ is bounded, and $E[(K^*)^2] = O(h^{-d})$ by (D.7.4). Next note that

$$T_2 = n\text{Var}(K^* \mu^*) + [nE(K^* \mu^*)]^2 ,$$

where $\mu^* = \mu(\mathbf{X}) - \mu(\mathbf{x})$, $E(K^* \mu^*) = O(h)$ by (D.7.3), and $\text{Var}(K^* \mu^*) = O(h^{2-d})$ by (D.7.5). Thus $T_1 + T_2 = O(nh^{-d}) + O(n^2 h^2)$ and by (D.7.7), $R_{21} = O(a_n^{-2} h^2 (nh^d)^{-1}) + O(a_n^{-2} h^4)$ which leads us to require $h^2 a_n^{-2} \rightarrow 0$ as $n \rightarrow \infty$. For the optimal $h = bn^{-1/(2+d)}$ this means $a_n n^{1/(2+d)} \rightarrow \infty$. If a_n is of the form $a_n = cn^{-a}$ with $c > 0$ we need $a < 1/(2+d)$. With these choices R_{21} is dominated by D_1 , that is,

$$R_{21} = o(h^2) + o((nh^d)^{-1}) .$$

To bound R_{22} , follow the steps in (??) to get

$$\begin{aligned} E\left[(\hat{f} - f_h)^2 \left\{ \sum_{i=1}^n K_i [Y_i - \mu(\mathbf{x})] \right\}^2\right] &= E\left[(\hat{f} - f_h)^2 E\left[\left\{ \sum_{i=1}^n K_i^* [Y_i - \mu(\mathbf{x})] \right\}^2 | \mathbf{X}^{(n)}\right]\right] \\ &\leq 2 \left[E(\hat{f} - f_h)^2 \left[\sum_{i=1}^n (K_i^*)^2 \sigma^2(\mathbf{X}_i) + \left(\sum_{i=1}^n K_i^* \mu_i^* \right)^2 \right] \right] \\ &\leq C \left[E(\hat{f} - f_h)^2 \left[\sum_{i=1}^n (K_i^*)^2 + \left(\sum_{i=1}^n K_i^* |\mathbf{Z}_i| \right)^2 h^2 \right] \right] \\ &\equiv C[S_1 + S_2] \end{aligned}$$

for some generic C because $\sigma^2(\cdot)$ is bounded and $|\mu_i^*| \leq h|\mathbf{Z}_i|$. To bound S_1 , write

$\hat{f} - f_h = n^{-1} \sum_{i=1}^n [K_i^* - E(K_i^*)]$ and compute

$$\begin{aligned} S_1 &= n^{-2} E \left[\sum_{i,j,k} (K_i^*)^2 [K_j^* - E(K_j^*)] [K_k^* - E(K_k^*)] \right] \\ &\leq n^{-2} \left[nE\{(K^*)[K^* - E(K^*)]^2\} + n^2 E((K^*)^2) \text{Var}(K^*) \right] \\ &= n^{-1} h^{-3d} \int K^2(\mathbf{z}) [K(\mathbf{z}) - f_h(\mathbf{z})]^2 f(\mathbf{x} + h\mathbf{z}) dz + O(h^{-2d}) \\ &= O(n^{-1} h^{-3d}) + O(h^{-2d}). \end{aligned}$$

Next, S_2 is bounded as follows:

$$\begin{aligned} S_2 &\leq n^{-2} h^2 E \left\{ \sum_{i,j,k,l} K_i^* |\mathbf{Z}_i| K_j^* |\mathbf{Z}_j| [K_k^* - E(K_k^*)] [K_l^* - E(K_l^*)] \right\} \\ &\leq n^{-2} h^2 \left\{ nE[(K^*)^2 |\mathbf{Z}|^2 [K^* - E(K^*)]^2] + n^2 E((K^*)^2 |\mathbf{Z}|^2) \text{Var}(K^*) \right. \\ &\quad \left. + n^3 E^2(K^* |\mathbf{Z}|) \text{Var}(K^*) \right\} \\ &\leq h^2 \left\{ n^{-1} h^{-3d} \int K^2(\mathbf{z}) [K(\mathbf{z}) - f_h(\mathbf{z})]^2 |\mathbf{z}| f(\mathbf{x} + h\mathbf{z}) dz + O(h^{-2d}) \right. \\ &\quad \left. + nO(h^{-d}) \right\} = O(n^{-1} h^{2-3d}) + O(h^{2-2d}) + O(nh^{2-d}). \end{aligned}$$

Finally, by singling out the dominating terms in R_{22} , we find

$$R_{22} = O(a_n^{-2} h^2 (nd)^{-1}) + O(a_n^2 (nh^d)^{-1} (nh^d)^{-1}) + o(h^2) + o((nh^d)^{-1})$$

provided $h^2 \rightarrow 0$, $(nh^d)^{-1} \rightarrow 0$, $a_n^{-2} h^2 \rightarrow 0$, and $a_n^{-2} (nh^d)^{-1} \rightarrow 0$. It follows that $R_1 + R_2$ is of smaller order than the main term D_1 under these conditions and (i) follows for $\hat{\mu}_{\text{NW}}(\cdot, a_n)$.

Part (ii) follows from part (i) save that we have to deal with neighborhoods $[\delta\mathcal{S}]_\varepsilon$ of the boundary $\delta\mathcal{S}$ of \mathcal{S} . That is, we must bound

$$\int_{[\delta\mathcal{S}]_\varepsilon} E(\hat{\mu}_{\text{NW}}(\mathbf{x}, a_n) - \mu(\mathbf{x}))^2 d\mathbf{x} \tag{D.7.8}$$

for M sufficiently large so that $\mathcal{S} \subset S(\mathbf{0}, M)$, where $S(\mathbf{0}, M)$ is a M sphere around $\mathbf{0}$, and

$$[\delta\mathcal{S}]_\varepsilon \equiv \{\mathbf{x} \in \mathcal{S} : |\mathbf{x} - \mathbf{v}| \leq \varepsilon \text{ for some } \mathbf{v} \in \delta\mathcal{S}\}.$$

In this range of \mathbf{x} 's the variance bounds continue to hold, but we have a different bound for the squared bias. In fact, if $d(\mathbf{x}, \delta\mathcal{S})$ is the distance of \mathbf{x} from $\delta\mathcal{S}$,

$$(E[\hat{\mu}_{\text{NW}}(\mathbf{x}, a_n)] - \mu(\mathbf{x}))^2 \leq C(h^2 + (Mh - d(\mathbf{x}, \delta\mathcal{S}))^2),$$

since

$$\int_{A(h, \mathbf{x})} \mu(\mathbf{x} + h\mathbf{z}) K(\mathbf{z}) d\mathbf{z} = 0$$

for $A(h, \mathbf{x}) = \{\mathbf{z} : \mathbf{x} + h\mathbf{z} \notin \mathcal{S}\}$. For simplicity take $\mathcal{S} = \{|\mathbf{x}| : |\mathbf{x}| \leq 1\}$. Then,

$$(D.7.8) \leq Ch^2 V([\delta\mathcal{S}]_{Mh}) + \int \left(Mh - \left|\mathbf{x} - \frac{\mathbf{x}}{|\mathbf{x}|}\right|\right)^2 \cdot 1(1 - Mh \leq |\mathbf{x}| \leq 1) d\mathbf{x} \quad (D.7.9)$$

where V is volume so that the first term in (D.7.9) is bounded by Ch^{d+2} .

The second term in (D.7.9) can be bounded by switching to polar coordinates with radius r , changing variable from r to $1-r$, and using

$$\int_{1-Mh}^1 (1-r)^{d+1} dr = Ch^{d+2}$$

where C is a constant. The calculation we have just made applies to any compact \mathcal{S} with a smooth boundary but we do not pursue this.

Thus (ii) is proved for $\widehat{\mu}_{\text{NW}}(\mathbf{x}, a_n)$, since all bounds we give are either bounded by (D.7.9) or are uniform on \mathcal{S} .

It remains to prove that (i) and (ii) hold for $\widehat{\mu}_{\text{NW}}$ itself. It suffices to show that if $h = Cn^{-\frac{1}{2+d}}$, then

$$E(\widehat{\mu}_{\text{NW}}(\mathbf{x}, a_n) - \widehat{\mu}_{\text{NW}}(\mathbf{x}))^2 = E(\widehat{\mu}_{\text{NW}}^2(\mathbf{x}) 1(\widehat{f}(\mathbf{x}) \leq a_n)) \rightarrow 0$$

uniformly in $\mathbf{x} \in \mathcal{S}$.

Note that if

$$Y_j = \mu(\mathbf{X}_j) + \varepsilon_j, \quad 1 \leq j \leq n,$$

then, by Assumption A1,

$$\widehat{\mu}_{\text{NW}}(\mathbf{x}) \leq M_0 + \max\{|\varepsilon_1|, \dots, |\varepsilon_n|\}. \quad (D.7.10)$$

Since $|\varepsilon_j|$ given $\mathbf{X} = \mathbf{x}$ has a moment generating function uniformly bounded for $|t| \leq \delta$ and $\mathbf{x} \in \mathcal{S}, \delta > 0$

$$P[\max\{|\varepsilon_1|, \dots, |\varepsilon_n|\} \geq s] = P[\cup_{i=1}^n \{|\varepsilon_i| \geq s\}] \leq \sum_{i=1}^n P(|\varepsilon_i| \geq s) \leq nM'e^{-s\delta} \quad (D.7.11)$$

by Markov's inequality. Thus,

$$E(\widehat{\mu}_{\text{NW}}^2(\mathbf{x}) 1(|\widehat{f}(\mathbf{x})| \leq a_n)) \leq 2E[(M_0^2 + \max_{1 \leq j \leq n} \varepsilon_j^2) 1(\widehat{f}(\mathbf{x}) \leq a_n)] \quad (D.7.12)$$

$$\leq 2M_0^2 P[\widehat{f}(\mathbf{x}) \leq a_n] + 2(E \max_{1 \leq j \leq n} \varepsilon_j^4)^{\frac{1}{2}} P^{\frac{1}{2}} [\widehat{f}(\mathbf{x}) \leq a_n] \quad (D.7.13)$$

where the second inequality is by the Cauchy-Schwarz inequality. We also have

$$P^{\frac{1}{2}} [\widehat{f}(\mathbf{x}) \leq a_n] = o(n^{-e}) \quad (D.7.14)$$

for all $c > 0$, uniformly on \mathcal{G} for all $\mathbf{x} \in \mathcal{S}$. Moreover, by (D.7.11)

$$E \max_{1 \leq j \leq n} \varepsilon_j^4 \leq nM' \int_0^\infty e^{-s^{\frac{1}{4}}\delta} ds = O(n). \quad (\text{D.7.15})$$

Combining (D.7.12)–(D.7.15) we obtain

$$E \widehat{\mu}_{\text{nw}}^2(\mathbf{x}) \mathbf{1}(\widehat{f}(\mathbf{x}) \leq a_n) = o(n^{-\frac{2}{d+2}}),$$

uniformly as needed, and the theorem follows.

D.8 Problems and Complements

Problems for Appendix D.1

- 1.** Suppose X_{n1}, \dots, X_{nn} are independent with $X_{ni} \sim \text{Bernoulli}(p_{ni})$. Show that if $\sigma_n^2 \equiv \sum_{i=1}^n \text{Var}X_{ni} \rightarrow \infty$ as $n \rightarrow \infty$, then

$$\frac{1}{\sigma_n} \sum_{i=1}^n (X_{ni} - p_{ni}) \xrightarrow{\mathcal{L}} \mathcal{N}(0, 1).$$

Hint: Use (D.1.6) with $\delta = 1$. Show that $E|X_{ni} - p_{ni}|^3 \leq 2p_i(1 - p_i)$.

- 2.** Establish Corollary D.1.1.

Hint: Let F denote the *df* of $(Y_i - \mu)$, then

$$\begin{aligned} & E(Y_i - EY_i)^2 \mathbf{1}(|Y_i - EY_i| > \varepsilon\sigma_n) \\ &= c_i^2 \int y^2 \mathbf{1}(|c_i y| > \varepsilon\sigma_n) dF(y) \\ &\leq c_i^2 \int y^2 \mathbf{1}(|y| > \varepsilon\sigma_n v_n) dF(y). \end{aligned}$$

- 3.** Suppose X_{n1}, \dots, X_{nn} are independent with $X_{ni} \sim \text{uniform}(-a_{ni}, a_{ni})$. Let $\sigma_n^2 = \sum_{i=1}^n \text{Var}X_{ni}$. Show that if $\max \sigma_{ni} \leq M$ for some bound M and if $\sigma_n^2 \rightarrow \infty$, then $\sigma_n^{-1} \sum_{i=1}^n X_{ni} \xrightarrow{d} \mathcal{N}(0, 1)$ as $n \rightarrow \infty$.

- 4.** Let X_1, \dots, X_n be i.i.d. as $X \sim P$. Suppose $\psi : R \rightarrow R$ is nondecreasing, $\psi(-\infty) < 0 < \psi(\infty)$, and $|\psi(x)| \leq M < \infty$ for all $x \in R$. Let $\lambda(\theta) = E_P \psi(X - \theta)$ and $\tau^2(\theta) = \text{Var}_P \psi(X - \theta)$. Show that if $\lambda'(\theta)$ exists and $\lambda'(\theta) < 0$, then

$$\frac{1}{\sqrt{n}\tau(\theta)} \sum [\psi(X_i - \theta_n) - \lambda(\theta_n)] \xrightarrow{\mathcal{L}} \mathcal{N}(0, 1)$$

for every sequence $\{\theta_n\}$ with $\theta_n = \theta + \frac{c}{\sqrt{n}}$, $c \in R$.

Remark: This result establishes the assumption of Problem 5.4.1(c) in Volume I.

- 5.** Consider the regression model $\mathbf{Y} = \mathbf{X}\beta + \varepsilon$ where ε has i.i.d. components with mean zero and variance σ^2 , and \mathbf{X} is a $n \times d$ matrix of constants of full rank. Let $\widehat{\beta} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Y}$ be the least squares estimate. Show that

- (a) $(\mathbf{X}^T \mathbf{X})^{\frac{1}{2}}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}) = (\mathbf{X}^T \mathbf{X})^{-\frac{1}{2}} \mathbf{X}^T \boldsymbol{\varepsilon} = \sum_{i=1}^n \mathbf{c}_{ni} \varepsilon_i$
where \mathbf{c}_{ni} is the i th column of the $d \times n$ matrix $\mathbf{c} \equiv (\mathbf{X}^T \mathbf{X})^{-\frac{1}{2}} \mathbf{X}^T$.

- (b) If $\max\{|\mathbf{c}_{ni}| : 1 \leq i \leq n\} \rightarrow 0$, then $(\mathbf{X}^T \mathbf{X})^{\frac{1}{2}}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \xrightarrow{\mathcal{L}} \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I})$.

Hint. Use Lindeberg-Feller. Note that $\sum_{i=1}^n |\mathbf{c}_{ni}|^2 = \text{trace}(\mathbf{C}\mathbf{C}^T) = d$.

- (c) In the simple linear regression model $Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$,

$$(i) \quad (\mathbf{X}^T \mathbf{X})^{-\frac{1}{2}} \mathbf{X}^T = n^{-\frac{1}{2}} \left(\frac{1}{x} \frac{\bar{x}}{x^2} \right)^{\frac{1}{2}} \begin{pmatrix} 1 & 1 & \dots & 1 \\ x_1 & x_2 & \dots & x_n \end{pmatrix};$$

- (ii) if $n^{-1} \sum x_i^2$ is bounded, and $n^{-\frac{1}{2}} \max\{|x_i| : 1 \leq i \leq n\} \rightarrow 0$, then

$$(\mathbf{X}^T \mathbf{X})^{\frac{1}{2}}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \sigma^2).$$

Problems for Appendix D.2

1. Show that if $Z_n(\cdot)$ obeys the conditions of Theorem 7.2.1 and (7.1.5) holds, then $Z(\cdot)$ must also have $P[\Delta_m(Z) > \varepsilon_m] \leq \delta_m$ where

$$\Delta_m(Z) = \max \sup \{ |Z(s) - Z(t)| : s, t \in T_{mj}; 1 \leq j \leq k_m \}.$$

Hint. $P[\Delta_m(Z) > \varepsilon_m] = P[\sup\{ |Z(s_i) - Z(s_j)| : s_i, s_j \in T_{jm} \cap C, 1 \leq j \leq k_m \} > \varepsilon_m]$

by separability. By FIDI convergence

$$\limsup_n P[\max_{1 \leq i, j \leq N} |Z_m(s_i) - Z_m(s_j)| > \varepsilon_m] \geq P[\max_{1 \leq i, j \leq N} |Z(s_i) - Z(s_j)| > \varepsilon].$$

2. Let $\mathcal{T}_\delta \equiv \{(\underline{f}_{1\delta}, \bar{f}_{1\delta}), \dots, (\underline{f}_{N(\delta)\delta}, \bar{f}_{N(\delta)\delta})\}$, $\delta > 0$ be the minimal set of brackets for a given T , P and δ . Suppose that condition (ii) of Theorem 7.1.2 holds. Show that it is possible to determine $\mathcal{T}_\delta^* \{S_{1\delta}, \dots, S_{N'(\delta)\delta}\}, (\underline{g}_{1\delta}, \bar{g}_{1\delta}), \dots, (\underline{g}_{N'(\delta)\delta}, \bar{g}_{N'(\delta)\delta})$ such that

$$(i) \quad S_{j,\delta} \subset (\underline{g}_{j,\delta}, \bar{g}_{j,\delta}),$$

$$(ii) \quad \cup_{j=1}^{N'(\delta)} S_{j,\delta} = T,$$

$$(iii) \quad S_{j,\delta} \cap S_{i,\delta} = \emptyset \text{ if } i \neq j,$$

$$(iv) \quad E(\bar{g}_{j,\delta} - \underline{g}_{j,\delta})^2(X) \leq \delta^2,$$

$$(v) \quad \text{For positive constants } c \text{ and } d, N'(\delta) \leq c\delta^{-d} \text{ for all } \delta \in (0, c).$$

These are the properties we need for Theorem 7.1.2.

Hint. Given \mathcal{T}_δ , let

$$\begin{aligned} \widetilde{S}_{1\delta} &= (\underline{f}_{1\delta}, \bar{f}_{1\delta}) \\ \widetilde{S}_{2\delta} &= (\underline{f}_{2\delta}, \bar{f}_{2\delta}) \cap (\underline{f}_{1\delta}, \bar{f}_{1\delta})^C, \end{aligned}$$

etc. This achieves (i), (ii), (iii) while maintaining $\underline{f}_{j\delta} = \bar{g}_{j\delta}$.

3. Show using separability and the Borel–Cantelli Lemma (D.1.6) that for a suitable (Ω, A, P) , the Wiener process $W(\cdot)$ and the Brownian bridge $W^0(\cdot)$ on $[0, 1]$ are continuous with probability one as defined by (7.1.24).

4. Complete the proof of Theorem 7.1.2.

Hint. Choose $m_0(\lambda)$ to make the first term in (D.2.10) of the right order.

5. (a) Prove Theorem 7.1.5 by paralleling the argument for Donsker’s theorem.

(b) Show how to use the theorem to prove Donsker’s theorem for X_1, \dots, X_n i.i.d. F_0 .

6. Prove the Glivenko–Cantelli theorem (7.1.15) by establishing that (7.1.14) holds a.s.

Hint. Use the Borel–Cantelli Lemma and (7.1.10).

7. (a) For $d = 1$, show that replacing $2d(1 + \epsilon)$ in the statement of Theorem 7.1.2 by $d - 1$ is best possible.

Hint. Consider

$$T = \left\{ f : f = \cos \theta (1(0, \frac{1}{2}] - 1(\frac{1}{2}, 1]) + \sqrt{2} \sin \theta (1(0, \frac{1}{4}] - 1(\frac{1}{4}, \frac{1}{2}]), 0 \leq \theta < 2r \right\}.$$

Note that if Z_1, Z_2 are independent $\mathcal{N}(0, 1)$,

$$\sup_{0 \leq \theta < 2\pi} (\cos \theta Z_1 + \sin \theta Z_2)^2 = Z_1^2 + Z_2^2,$$

and $Z_1^2 + Z_2^2 \sim \mathcal{X}_2^2 = \mathcal{E}(\frac{1}{2})$.

(b) Establish this result for general d .

Hint. $P[\mathcal{X}_{d+1}^2 \geq \lambda^2] = c \int_{\lambda^2}^{\infty} t^d e^{-\frac{t}{2}} dt \asymp \lambda^{2d} e^{-\frac{\lambda^2}{2}}$.

8. Show that Theorem 7.1.2 can be generalized as follows.

Let $f \rightarrow Z(f)$, $(f, g) \rightarrow Z(g) - Z(f)$ be separable Gaussian processes on $T \subset L_2(P)$. Suppose $EZ(f) \equiv 0$, $E(Z(f) - Z(g))^2 \leq (E_P(f - g))^2(X)$, and T obeys (ii) in Theorem 7.1.2.

(a) Show that the conclusion of Theorem 7.1.2 holds with W^0 replaced by Z .

(b) Conclude that W , the Wiener process, has almost surely continuous sample functions on T in the sense that if $E_P(f_n - f)^2(X) \rightarrow 0$, then $W(f_n) \rightarrow W(f)$ with probability 1. For better and more general results, see Talagrand (1994).

9. Show that in Theorem 7.1.2 (ii) implies (i).

10. Show that if Z is normal, $\mathcal{N}(0, 1)$, and $\lambda \geq 1$, then

$$P(|Z| \geq \lambda) \leq 2\varphi(\lambda)/\lambda.$$

Hint. $P(Z \leq -\lambda) = \int_{-\infty}^{-\lambda} \varphi(t) dt$. For $t \leq -\lambda$, $\varphi(t) \leq (-t/\lambda)\varphi(t)$, $-\int_{-\infty}^{-\lambda} t\varphi(t) dt = \varphi(\lambda)$.

11. Establish (7.1.24).

Problems for Appendix D.4.

1. Show that $\gamma(x, P)$ is an influence function for $-\dot{\Gamma}_\alpha(\hat{\beta}, \hat{P})$ in the sense of (7.2.2) provided that the assumptions of Theorem 9.2.2 hold.

Problems for Appendix D.5.

1. Show that the sequence X_1, X_2, \dots is homogeneous Markov if it satisfies (D.5.1) and (D.5.2).

2. Let $\pi(i) > 0$ satisfy $\sum_{i=1}^N \pi(i) = 1$. If Y_1, Y_2, \dots is a homogeneous finite Markov chain with transition kernel $K(i, j) = \pi(j)$, then Y_1, Y_2, \dots satisfy detailed balance and Y_1, Y_2, \dots are i.i.d. with $P(Y_i = j) = \pi(j)$, $1 \leq j \leq N$.

3. Show that a stationary Markov Chain satisfies detailed balance iff it is reversible.

4. (a) Show that under the assumptions given, (D.5.5) implies that there is $M > 0$ such that for all $i \in \{1, \dots, N\}$,

$$|P(X_n = i) - \pi(i)| \leq M\rho^n.$$

(b) Establish (D.5.7).

Problems for Appendix D.6.

1. Suppose K is symmetric and $p = 2$. Assume the regularity conditions of Section 11.6. Find the asymptotic expressions for (a) Bias $(\hat{\mu}(x)|\mathbf{X})$, (b) Var $(\hat{\mu}(x)|\mathbf{X})$.

Hint: See Lemma D.6.1.

2. Show that Theorem 11.6.3 follows from Theorem D.6.1.

Hint: See the proof of Lemma D.6.1.

3. Show that Theorem 11.6.4 follows from Theorem D.6.2.

4. Establish (D.6.8).

5. Show that at boundary points, the bias of the locally linear kernel estimate of $\mu(x)$ is of order h^2 , while the bias of locally constant estimate $\hat{\mu}_{\text{NW}}(x)$ is of order h .

Appendix E

SOLUTIONS FOR VOLUME II

Solutions to I.7.3

(a)

$$\begin{aligned}
 R(\boldsymbol{\mu}, \hat{\boldsymbol{\mu}}_0) &= \frac{1}{n} \sum_{i=1}^n E[\bar{Y} - (\beta_0 + \beta_1 z_i)]^2 \\
 &= \frac{1}{n} \sum_{i=1}^n E[(\bar{Y} - \beta_0) - \beta_1 z_i]^2 = \frac{1}{n} \left\{ n \left(\frac{\sigma^2}{n} \right) + \beta_1^2 \sum z_i^2 - 2.0 \right\} \\
 &= \frac{\sigma^2}{n} + \beta_1^2 \left(\frac{\sum z_i^2}{n} \right) \text{ (or use (6.1.22))}. \\
 R(\boldsymbol{\mu}, \hat{\boldsymbol{\mu}}_1) &= \frac{1}{n} \sum_{i=1}^n E[\hat{\beta}_0 + \hat{\beta}_1 z_i - (\beta_0 + \beta_1 z_i)]^2 \\
 &= \frac{1}{n} \sum_{i=1}^n \sigma^2 \left(\frac{1}{n} + \frac{z_i^2}{\sum z_i^2} \right) = \frac{1}{n} \{ \sigma^2 + \sigma^2 \} = \frac{2\sigma^2}{n}.
 \end{aligned}$$

(b) $E(\hat{\beta}_1^2) = [E(\hat{\beta}_1)]^2 + \text{Var}(\hat{\beta}_1) = \beta_1^2 + \frac{\sigma^2}{\sum z_i^2}$. The first results follows. We know (Section 6.1) that $\frac{RSS_1}{(n-2)}$ is unbiased for σ^2 , which shows the second result.

(c) $\frac{2s^2}{n} < \hat{\beta}_1^2 \left(\frac{\sum z_i^2}{n} \right) \Leftrightarrow t^2 < 2$.

(d) If the errors are normal we know that the t -distribution converges to the normal distribution (see Example 5.3.7). If the ε_i are not normal we need to use the Lindeberg-Feller Central Limit theorem to conclude that $\frac{(\sum z_i Y_i)}{(\sigma/\sum z_i^2/n)} \xrightarrow{\mathcal{L}} \mathcal{N}(0, 1)$ under H . Now replace σ with s and use Slutsky's theorem.

(e) $R(\boldsymbol{\mu}, \hat{\boldsymbol{\mu}}_0) = \frac{\sigma^2}{n} + \frac{\beta_1^2 \sum_{i=1}^n z_i^2}{n}$, $R(\boldsymbol{\mu}, \hat{\boldsymbol{\mu}}_1) = \frac{2\sigma^2}{n}$ $R(\boldsymbol{\mu}, \hat{\boldsymbol{\mu}}_1) < R(\boldsymbol{\mu}, \hat{\boldsymbol{\mu}}_0)$ iff $\frac{\beta_1^2}{(\sigma^2/ns_z^2)} > 1$

(f) t has a noncentral t -distribution, that is $t \stackrel{d}{=} \frac{(Z+\mathcal{T})}{\sqrt{V|K}}$ where $Z \sim \mathcal{N}(0, 1)$, $V \sim \mathcal{X}_k^2$, $k = n - 2$. Thus

$$E(t^2) = \{[E(Z + \mathcal{T})]^2 + \text{Var}(Z)\} k E(V^{-1}) = \frac{(\mathcal{T}^2 + 1)(n - 2)}{(n - 4)}.$$

Solve $\frac{(\mathcal{T}^2 + 1)(n - 2)}{(n - 4)} = t^2$ for \mathcal{T}^2 ; get $\widehat{\mathcal{T}}^2 = \frac{(n-4)}{(n-2)}t^2 - 1$.

(g)

$$\begin{aligned} t^2 &= \frac{n\widehat{\beta}_1^2\widehat{\sigma}_z^2}{[RSS_1/(n-2)]}. \\ \Sigma[Y_i - \bar{Y}]^2 &= RSS_1 + \widehat{\beta}_1^2\Sigma z_i^2. \quad RSS_1 = n(\widehat{\sigma}_4^2 - \widehat{\beta}_1^2\widehat{\sigma}_z^2). \\ t^2 &= \frac{n(n-2)\widehat{\beta}_1^2\widehat{\sigma}_z^2}{n(\widehat{\sigma}_Y^2 - \widehat{\beta}_1\widehat{\sigma}_z^2)} = \frac{(n-2)}{\left(\frac{\widehat{\sigma}_r^2}{\widehat{\beta}_1^2\widehat{\sigma}_z^2} - 1\right)} = \frac{(n-2)}{\left(\frac{1}{r^2} - 1\right)} = \frac{(n-2)r^2}{(1-r^2)}. \\ r^2 &= \frac{t^2}{t^2 + (n-2)} \text{ is increasing in } t^2 \text{ for } n \geq 3. \\ t^2 \geq 2 &\Leftrightarrow r^2 \geq \frac{2}{n}, \quad n \geq 3. \end{aligned}$$

Solutions to I.7.4

(a) $E(\widehat{\mu}_i - \mu_i)^2 = (\mu_i - \mu_{0i})^2 + \text{Var}\widehat{\mu}_{0i}$. By (6.1.15), $\text{Var}(\widehat{\mu}_0) = \text{Var}(H_1\mathbf{Y}) = \sigma^2 H_1$. It follows that $\sum_{i=1}^n \text{Var}\widehat{\mu}_{0i} = \sigma^2 \text{trace } H_1 = \sigma^2 q$, thus $n^{-1}E|\widehat{\mu} - \mu|^2 = \frac{\sigma^2 q}{n} + n^{-1}|\widehat{\mu} - \widehat{\mu}_0|^2$.

(b) (i) $E|\widehat{\mu} - \widehat{\mu}_0|^2 = E|\mathbf{Y} - \widehat{\mu}_0|^2 - E|\mathbf{Y} - \widehat{\mu}|^2$ by (6.1.22). We can write $\mathbf{Y} - \widehat{\mu}_0 = (I - H_1)\mathbf{Y} = (I - H_1)(\mu + \varepsilon)$, thus

$$|\mathbf{Y} - \widehat{\mu}_0|^2 = (\mathbf{Y} - \widehat{\mu})^T(\mathbf{Y} - \mu_0) = \mu^T(I - H_1)\mu + 2\mu^T(I - H_1)\varepsilon + \varepsilon^T(I - H_1)\varepsilon.$$

Here $\mu^T(I - H_1)\mu = \mu^T(\mu - H\mu) = \mu^T(\mu - \mu_0) = |\mu - \mu_0|^2$. Also $E[2\mu^T(I - H_1)\varepsilon] = 0$ and $E[\varepsilon^T(I - H_1)\varepsilon] = \sigma^2 \text{trace } (I - H_1) = \sigma^2(n - q)$. It follows that $E|\widehat{\mu} - \widehat{\mu}_0|^2 = n\sigma^2 + |\mu - \mu_{0i}|^2 - q\sigma^2$. Thus unbiased estimates are

$$\begin{aligned} \widehat{R}_q &= n^{-1}|\widehat{\mu} - \widehat{\mu}_0|^2 + \frac{2qs^2}{n} - s^2, \quad \widehat{R}_p = \frac{2ps^2}{n} - s^2; \\ \widehat{R}_p - \widehat{R}_q &= \frac{2(p-q)s^2}{n} - n^{-1}|\widehat{\mu} - \widehat{\mu}_0|^2. \end{aligned}$$

(c) (i) $R_p - R_q = \frac{(p-q)\sigma^2}{n} - \frac{|\mu - \mu_0|^2}{n} < 0$ iff $\sigma^{-2}|\mu - \mu_0|^2 > p - q$.

(ii) $F = \frac{\text{(noncentral } \chi^2_{p-q})}{\text{(central } \chi^2_{n-p})}$ with noncentrality parameter θ^2 by Proposition 6.1.2. Thus

$E(F) = cE\{(Z_1 + \theta)^2 + \sum_{i=2}^{p-q} Z_i^2\}E(V^{-1})$ where $c = \frac{(n-p)}{(p-q)}$. We find $E\{(Z_1 + \theta)^2 + \sum_{i=2}^{p-q} Z_i^2\} = (p-q) + \theta^2$. By Problem B.2.4, if $k = n - p$,

$$E(V^{-1}) = \frac{\frac{1}{2}\Gamma(\frac{1}{2}k - 1)}{\Gamma(\frac{1}{2}k)} = \frac{\frac{1}{2}}{(\frac{1}{2}k - 1)} = \frac{1}{(k-2)}.$$

Thus, $E(F) = \frac{(n-p)(p-q+\theta^2)}{(n-p-2)(p-q)}$. Solve $E(F) = F$ for θ^2 ; get $\widehat{\theta}^2 = \frac{(n-p-2)(p-q)}{n-p}F - (p-q)$. $\widehat{\theta}^2 > 1 \Leftrightarrow F > [1 + (p-q)]\frac{n-p}{(p-q)(n-p-2)}$.

Solution to Problem 7.1.3. For any $\epsilon > 0$ and any set of random variables $\{W_1, \dots, W_{k_m}\}$, let A_j be the event “ $|W_j| > \epsilon$.” Then

$$P[\max_j |W_j| > \epsilon] = P\left(\bigcup_{j=1}^{k_m} A_j\right) \leq \sum_{j=1}^{k_m} P(A_j) \leq k_m \max_j P(A_j).$$

Solution to Problem 7.1.13(a).

$$|Z_n(t) - Z(t)| = \left| \frac{1}{n} \sum_{i=1}^{r_n} 1(W_i \leq t) \right| < \frac{r_n}{n} \rightarrow 0 \text{ as } n \rightarrow \infty$$

uniformly in $t \in \mathcal{R}$. Therefore $Z_n - Z \xrightarrow{\text{FIDI}} 0$, and using Slutsky’s theorem we have $Z_n \xrightarrow{\text{FIDI}} Z$.

(b) To establish weak convergence we need to show that in addition to the above, for $T_{mj} = [\frac{j-1}{k_m}, \frac{j}{k_m}]$ there exist sequences ε_m and δ_m such that

$$\limsup_{n \rightarrow \infty} P\left[\max\left\{\sup\{|Z_n(s) - Z_n(t)| : s, t \in T_{mj}\} : 1 \leq j \leq k_m\right\} > \varepsilon_m\right] \leq \delta_m.$$

We have

$$\begin{aligned} \sup |Z_n(s) - Z_n(t)| &= \sup \left| (s-t)V + \frac{1}{n} \sum_{i=1}^{r_n} \{1(W_i \leq s) - 1(W_i \leq t)\} \right| \\ &\leq \sup |s-t|V + \frac{r_n}{n} = \frac{1}{k_m} + \frac{r_n}{n}, \end{aligned}$$

and, because $|V| \leq 1$ and $s, t \in T_{mj}$,

$$\limsup_{n \rightarrow \infty} P\left[\frac{1}{k_m} + \frac{r_n}{n} > \varepsilon_m\right] = 0 \leq \delta_m$$

for $\varepsilon_m = \frac{1}{k_m} + \frac{1}{m}$ and $\delta_m = \frac{1}{m}$.

Solution to Problem 7.2.4(a). In Example 7.2.2, set $g_1(x, y) = x$, $g_2(x, y) = y$, $g_3(x, y) = xy$, and $h(\mu_1, \mu_2, \mu_3) = \mu_3 - \mu_1\mu_2$. Then $(\partial/\partial\mu_1)h = -\mu_2$, $(\partial/\partial\mu_2)h = -\mu_1$, and $(\partial/\partial\mu_3)h = 1$. Here $\mu_1 = E(X)$ and $\mu_2 = E(Y)$. Now (7.2.10) gives

$$\psi(x, y, P) = -\mu_2(x - \mu_1) - \mu_1(y - \mu_2) + xy - E(XY).$$

Solution to Problem 7.2.25(a).

$$\begin{aligned}
\text{CORR}(\mu(X), Y) &= \frac{\text{Cov}(\mu(X), Y)}{\sqrt{\text{Var}(Y)} \sqrt{\text{Var}(\mu(X))}} = \frac{\text{Cov}(\mu(X), \mu(X) + \varepsilon)}{\sqrt{\text{Var}(Y)} \sqrt{\text{Var}(\mu(X))}} \\
&= \frac{\text{Var}(\mu(X))}{\sqrt{\text{Var}(Y)} \sqrt{\text{Var}(\mu(X))}} = \sqrt{\frac{\text{Var}(\mu(X))}{\text{Var}(Y)}} \\
&= \sqrt{\frac{E(\mu^2(X)) - [E(\mu(X))]^2}{\text{Var}(Y)}} = \sqrt{\frac{E(\mu^2(X)) - \mu_Y^2}{\text{Var}(Y)}}.
\end{aligned}$$

Therefore

$$\eta^2 = \frac{E(\mu^2(X)) - \mu_Y^2}{\text{Var}(Y)}.$$

Solution to Problem 8.2.24 When $\rho = 0$, \mathbf{T} is complete and sufficient by exponential family theory (Theorem 8.2.2). Moreover, because $\hat{\rho}$ is location and scale invariant its distribution does not depend on $(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2)$ ($\hat{\rho}(\mathbf{X}, \mathbf{Y}) = \hat{\rho}(\mathbf{X}', \mathbf{Y}')$ where $X'_i = (X_i - \mu_1)/\sigma_1$, and $Y'_i = (Y_i - \mu^2/\sigma^2)$). Thus \mathbf{T} and $\hat{\rho}$ are independent by Basu's theorem.

Solution to Problem 8.2.25

(a) The joint density is

$$\frac{(\prod_{i=1}^n y_i^2) \exp\{-\sum_{i=1}^n y_i/\tau - \sum_{i=1}^n x_i y_i/\sigma\tau\}}{(\tau^3 \sigma)^n}$$

so $(\sum_{i=1}^n Y_i, \sum_{i=1}^n X_i Y_i)$ is minimal sufficient using exponential family theory.

(b) For a single observation, the moment generating function is

$$M_{Y,XY}(s, t) = E(e^{sY+tXY}) = \int_0^\infty \int_0^\infty e^{sy+txy} y^2 e^{-y/\tau-xy/\tau\sigma} dy dx / \tau^3 \sigma.$$

Substituting $u = y$, $v = xy$ we get

$$= \int_0^\infty \int_0^\infty u e^{-u(1/\tau-s)} du e^{-v(1/\tau\sigma-t)} dv = \frac{1}{(1-s\tau)^2(1-tr\sigma)},$$

and

$$M_{\sum Y_i, \sum X_i Y_i}(s, t) = (M_{Y,XY}(s, t))^n = (1-s\tau)^{-2n} (1-tr\sigma)^{-n}$$

so that $\sum_{i=1}^n Y_i$ has a Gamma($2n, \tau$) density, $\sum_{i=1}^n X_i Y_i$ has a Gamma($n, \tau\sigma$) density, and they are statistically independent.

(c) Solving the derivative equations for the log likelihood,

$$\begin{aligned}\mathcal{L} &= \ln \left(\prod y_i^2 \right) - \sum y_i / \tau - \sum x_i y_i / \tau \sigma - n \ln(\tau^3 \sigma) \\ \frac{\partial \mathcal{L}}{\partial \tau} &= \frac{\sum y_i}{\tau^2} + \frac{\sum x_i y_i}{\tau^2 \sigma} - \frac{3n\tau^2}{\tau^3} = 0 \\ \frac{\partial \mathcal{L}}{\partial \sigma} &= \frac{\sum x_i y_i}{\tau \sigma^2} - \frac{n}{\sigma} = 0\end{aligned}$$

from which the maximum likelihood estimators are

$$\hat{\tau} = \frac{\sum_{i=1}^n Y_i}{2n}, \quad \hat{\sigma} = \frac{2 \sum_{i=1}^n X_i Y_i}{\sum_{i=1}^n Y_i}.$$

(d)

$$E(\hat{\tau}) = \frac{2n\tau}{2n} = \tau, \quad E(\hat{\sigma}) = 2n\tau\sigma E(1/\sum Y_i) = 2n\tau\sigma \frac{1}{\tau(2n-1)} = \sigma \frac{2n}{(2n-1)},$$

and the UMVU estimators for τ and σ , respectively, are

$$\hat{\tau} = \frac{1}{2n} \sum_{i=1}^n Y_i$$

and

$$\hat{\sigma} = \sqrt{\frac{(2n-1)}{2n} \frac{\sum_{i=1}^n X_i Y_i}{\sum_{i=1}^n Y_i} \frac{(2n-1)}{2n}}.$$

Solution to Problem 8.2.26.

(a) The joint density of X_1, \dots, X_n is

$$f(x|\theta) = f(x_1, \dots, x_n|\theta) = \frac{1}{\theta^n} 1(1 \leq x_{(n)} < \theta). \quad (\text{E.1})$$

By the factorization theorem, $T = X_{(n)}$ is a sufficient statistic for θ . We now show that it is also complete. Suppose that $U(T)$ is an unbiased estimate of 0, then

$$E(U(T)) = U(1) \frac{1}{\theta^n} + \int_1^\theta U(t) n \frac{t^{n-1}}{\theta^n} dt = 0,$$

or

$$U(1) + \int_1^\theta U(t) n t^{n-1} dt = 0,$$

which implies that

$$U(1) = 0, \quad U(t) \equiv 0.$$

Therefore, $T = X_{(n)}$ is complete and sufficient for θ . By the Lehmann-Scheffé Theorem, we need only find an unbiased estimate $W(T)$ of θ . Since

$$W(1) + \int_1^\theta W(t)nt^{n-1}dt = \theta^{n+1},$$

$$E(W(t)) = W(1)\frac{1}{\theta^n} + \int_1^\theta W(t)n\frac{t^{n-1}}{\theta^n}dt = \theta,$$

or, equivalently,

$$W(1) + \int_1^\theta W(t)nt^{n-1}dt = \theta^{n+1},$$

we have

$$W(1) = 1, \quad W(t) = \frac{n+1}{n}t.$$

Then

$$W = 1(X_{(n)} = 1) + \frac{n+1}{n}X_{(n)}1(X_{(n)} > 1)$$

is the UMVUE of θ .

(b) By (E.1) in part (a), the likelihood function of θ is

$$L(\theta) = \frac{1}{\theta^n}1(\theta > x_{(n)}),$$

which clearly attains its maximum at $x_{(n)}$, hence $\hat{\theta} = X_{(n)}$. Since $\eta = P(Y_1 > 1) = \frac{\theta-1}{\theta}$, by the invariance property of MLE, $\hat{\eta} = \frac{\hat{\theta}-1}{\hat{\theta}}$.

(c) It is easy to check that the joint pdf $f(\mathbf{x}|\theta)$ is MLR in $X(n)$; therefore a UMP test based on $X_{(n)}$ is

$$\delta = 1(X_{(n)} > c).$$

Since $\theta_0 > (1 - \alpha)^{-1/n}$, we have $c > 1$, and c is determined by

$$E_{\theta_0}(\delta) = P(X_{(n)} > c) = 1 - \left(\frac{c}{\theta_0}\right)^n = \alpha.$$

Then $c = \theta_0(1 - \alpha)^{1/n}$.

Solution to 9.1.9. Given an i.i.d. sample (Y_i, Z_i, U_i) consider the conditional density at $Y = y$ given the covariates $(Z, U) = (z, u)$,

$$p_{\beta, \eta}(x) = \{v[\beta' z + \eta(u)]\}^y \{1 - v[\beta' z + \eta(u)]\}^{1-y}.$$

Assume that u is continuously distributed on $(0, 1)$, then the u_i will be distinct with probability 1. Consider a function $\hat{\eta}_m$ for which $\hat{\eta}_m(u_i) = m$ if $y_i = 1$ and $\hat{\eta}_m(u_i) = -m$ if $y_i = 0$. For a fixed sample of size n , since the u_i are distinct, it is possible to construct

a polynomial of degree $n - 1$ on $(0, 1)$ that takes the required values. (For example the relevant coefficients could be obtained by a linear projection of the required function values onto the collection of vectors, $\{1, u^1, \dots, u^{n-1}\}$.) Since $\widehat{\eta}_m$ is a polynomial it trivially satisfies the smoothness requirements. We can show directly that the supremum of the likelihood computed over this sequence of functions is equal to unity by computing directly the following pointwise limits.

Case 1: $y_i = 1, \widehat{\eta}_m(u) = m$

$$\lim_{m \rightarrow \infty} p_{\beta, \eta}(x) = \lim_{m \rightarrow \infty} \frac{1}{1 + e^{-\beta' z - m}} = 1.$$

Case 2: $y_i = 0, \widehat{\eta}_m(u) = -m$

$$\lim_{m \rightarrow \infty} p_{\beta, \eta}(x) = \lim_{m \rightarrow \infty} \frac{1}{1 + e^{-\beta' z + m}} = \lim_{m \rightarrow \infty} \frac{e^{-\beta' z + m}}{1 + e^{-\beta' z + m}} = 1.$$

Solution to Problem 9.1.4(a). For $z \geq 1$,

$$P(V = z | V \geq 1) = P(v = z) / P(V \geq 1) = p(z) / [1 - P(V = 0)].$$

(b)

$$E(Z) = \frac{1}{1 - e^{-\mu}} \sum_{z=1}^{\infty} \frac{ze^{-\mu} \mu^z}{z!}$$

where the sum is the mean μ of a Poisson (μ) variable.

(c) $\frac{\partial}{\partial \mu} \sum_{i=1}^n \log p_F(x_i) = 0$ has the given solution which is easily seen to provide the maximum. Or use $\mathcal{L}(X) = \mathcal{L}(V + 1)$ to argue the conclusion.

Solution to Problem 9.1.22 If T is PQH then $\lambda(t_\alpha | \mathbf{z}) = \theta \lambda(s_\alpha)$ for some $\lambda = f/[1 - F]$ where $t_\alpha = F^{-1}(\alpha | \mathbf{z}), s_\alpha = F^{-1}(\alpha)$.

Note that

$$\begin{aligned} PQH &\iff \frac{f(t_\alpha | \mathbf{z})}{1 - F(t_\alpha | \mathbf{z})} = \theta \frac{f(s_\alpha)}{1 - F(s_\alpha)} \\ &\iff \frac{f(t_\alpha | \mathbf{z})}{1 - \alpha} = \theta \frac{f(s_\alpha)}{1 - \alpha} \\ &\iff f(t_\alpha | \mathbf{z}) = \theta f(s_\alpha) \iff AFT. \end{aligned}$$

Solution to Problem 9.2.15

$$\begin{aligned} (\text{a}) \quad E[\psi(Z, T; b)] &= E\{E[\psi(Z, T; b)|Z]\} \\ E[\partial\psi(Z, T; b)/\partial b] &= E\{\dot{r}(Z, b)/r(Z, b) - \dot{r}(Z, b)E(T|Z)\}, \end{aligned}$$

where $E(T|z) = [r(z; \beta_0)]^{\frac{1}{\alpha}} \Gamma(\alpha^{-1} + 1)$. Thus, $E[\psi(Z, T; \beta_0)] = 0$ iff $\alpha = 1$.

(b) $\log r = -\beta z$, $\dot{r}(Z; \beta)/r(Z; \beta) = -z$; $\alpha = 1$. Set $\Delta(b) = E[\psi(Z, T; b)]$, then $\Delta(b) = -E(Z) + (Z \exp\{-bZ\}/\exp\{-\beta_0 Z\})$. $\Delta'(b) = -E(Z^2 \exp\{-Z(b-\beta_0)\}) < 0$. Thus, $b = \beta_0$ is the unique solution to $\Delta(b) = 0$.

$$(c) \quad D(b) = -n^{-1} \sum_{i=1}^n Z_i + n^{-1} \sum_{i=1}^n Z_i \exp\{-bZ_i\} T_i$$

$$D'(b) = -n^{-1} \sum_{i=1}^n Z_i^2 \exp\{-bZ_i\} T_i < 0.$$

So, the solution to $D(b) = 0$ is unique.

(d) Since $D(b)$ is decreasing (see the solution to part(c)), then

$$P(\hat{\beta} \leq b) = P(D(b) \leq 0).$$

Thus $P(\sqrt{n}(\hat{\beta} - \beta_0) \leq s) = P(D(b) \leq 0)$ where $b_n = \beta_0 + s/\sqrt{n}$. Let $\Delta(b) = E[\psi(Z, T; b)]$ and $\sigma^2(b) = \text{Var}(\psi(Z, T; b))$, then

$$P(D(b_n) \leq 0) = P\left(\frac{D(b_n) - \Delta(b_n)}{\sigma(b_n)/\sqrt{n}} \leq \frac{-\sqrt{n}\Delta(b_n)}{\sigma(b_n)}\right) \rightarrow \Phi(l) + o(1),$$

by Lindeberg-Feller and Polya, where $l = \lim \frac{-\sqrt{n}\Delta(b_n)}{\sigma(b_n)}$. Note that because $\Delta(\beta_0) = 0$,

$$\sqrt{n}\Delta(b_n) = \sqrt{n}[\Delta(\beta_0 + s/\sqrt{n}) - \Delta(\beta_0)] \rightarrow \Delta'(\beta_0)s = E(Z^2)s.$$

Similarly,

$$E[\psi^2(Z, T; b_n)] = E[\psi^2(Z, T; \beta_0 + s/\sqrt{n})] \rightarrow E[\psi^2(Z, T; \beta_0)] = \sigma^2(\beta_0).$$

Thus,

$$P(\sqrt{n}(\hat{\beta} - \beta_0) \leq s) \rightarrow \Phi\left(\frac{E(Z^2)s}{\sigma(\beta_0)}\right).$$

That is, $\sqrt{n}(\hat{\beta} - \beta_0) \rightarrow_D \mathcal{N}(0, \sigma^2(\beta_0)/E^2(Z^2))$.

Solution to Problem 9.3.7. The influence function of $\hat{\theta}$ is the first coordinate of

$$I^{-1}(\theta_0, \boldsymbol{\eta}_0)(\dot{\ell}_1, \dot{\ell}_{21}, \dots, \dot{\ell}_{2d})^T = \Psi(x, \theta_0, \boldsymbol{\eta}_0),$$

where

$$I(\theta, \boldsymbol{\eta}) = E_{(\theta, \boldsymbol{\eta})}[(\dot{\ell}_1, \dot{\ell}_{21}, \dots, \dot{\ell}_{2d})^T (\dot{\ell}_1, \dot{\ell}_{21}, \dots, \dot{\ell}_{2d})].$$

Let $E_0(\cdot) = E_{(\theta_0, \boldsymbol{\eta}_0)}$, then $I(\theta_0, \boldsymbol{\eta}_0)$ can be written as a block matrix,

$$\begin{bmatrix} E_0 \dot{\ell}_1^2 & \sum_{\dot{\ell}_1, \dot{\ell}_2} \\ \sum_{\dot{\ell}_2, \dot{\ell}_1} & \sum_{\dot{\ell}_2, \dot{\ell}_2} \end{bmatrix}_{(d+1) \times (d+1)}$$

where $\sum_{\dot{\ell}_1, \dot{\ell}_2} = (E_0(\dot{\ell}_1 \dot{\ell}_{21}), \dots, E_0(\dot{\ell}_1 \dot{\ell}_{2d}))$, $\sum_{\dot{\ell}_2, \dot{\ell}_1} = (E_0(\dot{\ell}_{21} \dot{\ell}_1), \dots, E_0(\dot{\ell}_{2d} \dot{\ell}_1))^T = \sum_{\dot{\ell}_1, \dot{\ell}_2}^T$, $\sum_{\dot{\ell}_2, \dot{\ell}_2} = E_0[\dot{\ell}_2 \dot{\ell}_2^T]$, $\dot{\ell}_2 = (\dot{\ell}_{21}, \dots, \dot{\ell}_{2d})^T$.

Because we are interested in the first coordinate of $\Psi(x, \theta_0, \eta_0)$, we just need the first row of $I^{-1}(\theta_0, \eta_0)$, which for this type of block matrix is

$$\left(\frac{1}{E_0 \dot{\ell}_1^2 - \sum_{\dot{\ell}_1 \dot{\ell}_2} (\sum_{\dot{\ell}_2 \dot{\ell}_2})^{-1} \sum_{\dot{\ell}_2 \dot{\ell}_1}}, - \frac{\sum_{\dot{\ell}_1 \dot{\ell}_2} (\sum_{\dot{\ell}_2 \dot{\ell}_2})^{-1}}{E_0 \dot{\ell}_1^2 - \sum_{\dot{\ell}_1 \dot{\ell}_2} (\sum_{\dot{\ell}_2 \dot{\ell}_2})^{-1} \sum_{\dot{\ell}_2 \dot{\ell}_1}} \right)_{1 \times (d+1)}.$$

The first coordinate of $\Psi(X, \theta_0, \eta_0)$, the influence function of $\hat{\theta}$, is

$$\psi_1(X, \theta_0, \eta_0) = \frac{\dot{\ell}_1 - \sum_{\dot{\ell}_1 \dot{\ell}_2} (\sum_{\dot{\ell}_2 \dot{\ell}_2})^{-1} \dot{\ell}_2}{E_0 \dot{\ell}_1^2 - \sum_{\dot{\ell}_1 \dot{\ell}_2} (\sum_{\dot{\ell}_2 \dot{\ell}_2})^{-1} \sum_{\dot{\ell}_2 \dot{\ell}_1}}.$$

On the other hand, according to B.10.20 and the fact that $E_0(\dot{\ell}_1) = 0$, $E_0(\dot{\ell}_2) = \mathbf{0}$,

$$\dot{\ell}_1 - \pi(\dot{\ell}_1 | [\dot{\ell}_{21}, \dots, \dot{\ell}_{2d}]) = \dot{\ell}_1 - \sum_{\dot{\ell}_1 \dot{\ell}_2} (\sum_{\dot{\ell}_2 \dot{\ell}_2})^{-1} \dot{\ell}_2.$$

Also

$$\begin{aligned} & \| \dot{\ell}_1 - \pi(\dot{\ell}_1 | [\dot{\ell}_{21}, \dots, \dot{\ell}_{2d}]) \|^2 \\ &= E_0 \left(\dot{\ell}_1^2 - 2 \sum_{\dot{\ell}_1 \dot{\ell}_2} (\sum_{\dot{\ell}_2 \dot{\ell}_2})^{-1} \dot{\ell}_2 \dot{\ell}_1 \right. \\ &\quad \left. + \sum_{\dot{\ell}_1 \dot{\ell}_2} (\sum_{\dot{\ell}_2 \dot{\ell}_2})^{-1} \dot{\ell}_2 \dot{\ell}_2^T (\sum_{\dot{\ell}_2 \dot{\ell}_1})^{-1} \sum_{\dot{\ell}_2 \dot{\ell}_1} \right) \\ &= E_0 \dot{\ell}_1^2 - \sum_{\dot{\ell}_1 \dot{\ell}_2} (\sum_{\dot{\ell}_2 \dot{\ell}_2})^{-1} \sum_{\dot{\ell}_2 \dot{\ell}_1} a; \end{aligned}$$

therefore, the influence function of $\hat{\theta}$ is given by

$$\psi_1(X, \theta_0, \eta_0) = \frac{\dot{\ell}_1 - \pi(\dot{\ell}_1 | [\dot{\ell}_{21}, \dots, \dot{\ell}_{2d}])}{\| \dot{\ell}_1 - \pi(\dot{\ell}_1 | [\dot{\ell}_{21}, \dots, \dot{\ell}_{2d}]) \|^2}.$$

Finally according to Remark 9.3.4 the efficient influence function for a regular estimate of $\hat{\theta}$ is given by

$$\tilde{\ell}(x, P_0; \theta, \mathcal{P}) = \frac{\dot{\ell}_1 - \pi(\dot{\ell}_1 | \tilde{\mathcal{P}}_2(P_0))}{\| \dot{\ell}_1 - \pi(\dot{\ell}_1 | \tilde{\mathcal{P}}_2(P_0)) \|^2}$$

and since in this framework, $\tilde{\mathcal{P}}_2(P_0) = [\dot{\ell}_{21}, \dots, \dot{\ell}_{2d}]$, the influence function of the MLE of θ is the efficient influence function; therefore $\hat{\theta}$ is efficient in the sense of Section 9.3.

Solution to Problem 9.3.13.

$$p(x, y; f, g, \mu) = g(y - \mu(x)) f(x) = g(\varepsilon) f(x).$$

P_0 will be f_0 , g_0 , μ_0 , so let $\log f_t(x) = \log f_0(x) + ta(x)$,

$$\log g_t(\varepsilon) = \log g_0(\varepsilon) + tb(\varepsilon),$$

$$\mu_t(x) = \mu_0(x) + tc(x),$$

$\log p_t(x, y) = \log g_t(\varepsilon_t) + \log f_t(x)$, where $\varepsilon_t = y - \mu_t(x) = y - \mu_0(x) - tc(x)$.

So $\log p_t(x, y) = \log g_0(\varepsilon_t) + tb(\varepsilon_t) + \log f_0(x) + ta(x)$

$$\frac{\partial \log p_t(x, y)}{\partial t} \Big|_{t=0} = -\frac{g'_0(\varepsilon_0)c(x)}{g_0(\varepsilon_0)} - b(\varepsilon_0)c(x) + a(x) \text{ where } \varepsilon_0 = y - \mu_0(x).$$

$$\text{Thus, } \dot{\mathcal{P}}(P_0) = \left\{ a(x) + b(\varepsilon_0)c(x) + d(x)\frac{g'_0(\varepsilon_0)}{g_0(\varepsilon_0)} \right\}$$

for some a, b, c, d arbitrary functions in $L_2(P_0)$. With a different parametrization

$$f_t(x) = f_0(x) + ta(x), \quad g_t(\varepsilon) = g_0(\varepsilon) + tb(\varepsilon), \quad \mu t(x) = \mu_0(x) + tc(x)$$

$$\frac{\partial \log p_t(x)}{\partial t} \Big|_{t=0} = -\frac{g'_0(\varepsilon_0)c(x)}{g_0(\varepsilon_0)} - \frac{b(\varepsilon_0)c(x)}{g_0(\varepsilon_0)} + \frac{a(x)}{f_0(x)}.$$

This time $\dot{\mathcal{P}}(P_0) = \left\{ \frac{a(x)}{f_0(x)} + \frac{b(\varepsilon_0)c(x)}{g_0(\varepsilon_0)} + d(x)\frac{g'_0(\varepsilon_0)}{g_0(\varepsilon_0)} \right\}$ which is equivalent to the first one in the sense that it is of the form “function of x ” plus “a function of x times a function of ε_0 .”

Solution to Problem 9.3.14. Using Remark 9.3.4 we can find the tangent space one parameter at a time: Let $\dot{\mathcal{P}}_1(P_0)$, $\dot{\mathcal{P}}_2(P_0)$, and $\dot{\mathcal{P}}_3(P_0)$ be the tangent spaces of P_{ρ, f_0, g_0} , P_{ρ_0, f, g_0} , and $P_{\rho_0, f_0, g}$ respectively.

(a) By Problem 9.3.7, the P_{ρ, f_0, g_0} case is obtained by differentiating the log likelihood of $\mathcal{N}(0, 0, 1, 1, \rho)$ with respect to ρ .

(b) Case P_{ρ_0, f, g_0} . The density of the bivariate normal copula is

$$p(x, y; \rho, f, g) = \frac{\varphi_\rho(z, w)}{\varphi(z)\varphi(w)} f(x)g(y) \quad (\text{A})$$

where φ_ρ is the density of $N(0, 0, 1, 1, \rho)$.

Let $f_t(x) = f_0(x) + tb(x)$; $Z_t = \Phi^{-1}(F_t(x))$; $F_t(x) = \int_{-\infty}^x f_t(u) du = F_0(x) + t \int_{-\infty}^x b(u) du$. From (A) evaluated at ρ_0, f_t, g_0 , we have

$$\log p_{zt} = -\log \sqrt{1 - \rho_0^2} - \frac{(z_t^2 - 2\rho_0 z_t w_0 + w_0^2)}{2(1 - \rho_0^2)} + \log f_t(x) + \log g_0(y) + \frac{z_t^2}{2} + \frac{w_0^2}{2}$$

$$\frac{2 \log p_{zt}}{2t} = -\frac{1}{2(1 - \rho_0^2)} (2z_t z'_t - 2\rho_0 w_0 z'_t) + \frac{b(x)}{f_t(x)} + z_t z'_t$$

where

$$Z'_t = \frac{d\Phi^{-1}(F_t(x))f_t(x)b(x)}{dx} = \frac{f_t(x)}{\varphi(F_t(x))} b(x) = \frac{f_t(x)}{\varphi(z_t)} b(x).$$

Thus

$$\dot{\mathcal{P}}_2(P_0) = \frac{\partial \log p_{zt}}{\partial t} \Big|_{t=0} = \frac{f_0(x)b(x)}{\varphi(z_0)} \left(\frac{\rho_0 w_0 - z_0}{1 - \rho_0^2} + z_0 \right) + \frac{b(x)}{f_0(x)}$$

for some function $b(x)$ in $L_2(P_0)$.

The case $P_{\rho_0, f_0, g}$ is the same.

Solution to Problem 10.2.6.

(a) $F_L(x) = \{1 + \exp(-x)\}^{-1} = u \Leftrightarrow 1 + \exp(-x) = u^{-1} \Leftrightarrow \exp(-x) = u^{-1} - 1 \Leftrightarrow x = -\log(u^{-1} - 1)$; therefore $F_L^{-1}(u) = -\log(u^{-1} - 1)$. If $U \sim \text{Unif}[0, 1]$, then $F_L^{-1}(u) \sim f_L(x)$, then a simple Monte Carlo technique will be

1. Simulate U_1, \dots, U_n from $\text{Unif}(0, 1)$.
2. Obtain X_1, \dots, X_n by computing $X_i = F_L^{-1}(U_i) = -\log(U_i^{-1} - 1)$.

(b) $\sup_x \frac{f(x)}{f_L(x)} = \sup_x \frac{a}{1+e^{-|x|}} = a$ which is finite because f is a density, then use the following rejective sampling algorithm:

1. Draw a sample from $f_L(x) : x_1, x_2, \dots$ We can use part (a) for this.
2. Draw Bernoulli variables I_1, I_2, \dots with

$$P(I_j = 1) = \frac{1}{a} \frac{f(x_j)}{f_L(x_j)} = \frac{1}{1 + e^{-|x_j|}}.$$

3. Set $\tau = \min\{j : I_j = 1\}$ and reject $x_1, \dots, x_{\tau-1}$. Keep $x_\tau = x_1^*$. Repeat until we get x_1^*, \dots, x_n^* ; this sample will have density $f(x)$.

(c) Matlab command: $\text{int}\left(\exp(-x)/((1+\exp(-x))^2*(1-\exp(-\text{abs}(x))))\right), -\inf, \inf$) result $a = 4/3$. and from Theorem 10.2.2, $E_0(\tau) = \sup_x \frac{f(x)}{f_L(x)} = a = 4/3$

(d) 1000 simulations took less than 1 second.

Solution to Problem 10.5.2. Assuming f_0 has bounded derivatives of order up to 3, by Taylor expansion of $f_0^2(y - \boldsymbol{\theta}^T \mathbf{Z}_i)$ for $\boldsymbol{\theta}$ around $\mathbf{0}$ and (10.5.11)

$$\begin{aligned} \text{Var}_0 \left(\prod_{i=1}^n \frac{f_{\boldsymbol{\theta}}(Y_i | \mathbf{Z}_i)}{f_0(Y_i)} \right) &= \prod_{i=1}^n \left(\int \frac{f_0^2(y) + 2f_0(y)f'_0(y)\mathbf{Z}_i^T \boldsymbol{\theta}}{f_0(y)} dy + O(|\boldsymbol{\theta}|^2) \right) - 1 \\ &= \prod_{i=1}^n \left(1 + 2 \sum_{j=1}^d \frac{c_j z_{ij}}{\sqrt{n}} \int f'_0(y) dy + O(|\boldsymbol{\theta}|^2) \right) - 1. \end{aligned}$$

Using $\int f_0^1(y) dy = \frac{d}{dy} \int f_0(y) dy = \frac{d}{dy} 1 = 0$,

$$\prod_{i=1}^n \left(1 + 0\left(\frac{1}{n}\right) \right) - 1 = O(1).$$

To justify the $O(|\boldsymbol{\theta}|^2)$ term, note that the next term in the expansion is

$$2 \int \{f_0''(y) + [f'_0(y)]^2/f_0(y)\} dy$$

which is finite when $f_0 = \varphi$.

Solution to Problem 11.6.9. From the definition of $\widehat{\theta}(x)$, we see

$$\widehat{\theta}(x) - E(\widehat{\theta}(x)) = \frac{1}{k} \sum_{j=i+1}^{i+k} (Y_j - E(Y_j)), \quad x \in I_i, \quad i = 0, 1, 2, \dots, n-k.$$

Define

$$S_i = \sum_{j=1}^i \frac{1}{k} (Y_j - E(Y_j)), \quad Z_j = \frac{1}{k} (Y_j - E(Y_j)),$$

then

$$S_i = \sum_{j=1}^i Z_j, \quad \widehat{\theta}(x) - E(\widehat{\theta}(x)) = \sum_{j=i+1}^{i+k} Z_j = S_{i+k} - S_i, \quad i = 1, 2, \dots, n,$$

where $S_0 = 0$. Also the $\{Z_j\}$ are independent with $E(Z_j) = 0$,

$$\begin{aligned} E(Z_j^2) &= \text{Var}(Z_j) = \sigma^2(Z_j) = \frac{1}{k^2} \text{Var}(Y_j) \\ &= \frac{1}{k^2} \theta(x_i)(1 - \theta(x_i)) \leq \frac{1}{4k^2} < \infty, \quad j = 1, 2, \dots, n, \end{aligned}$$

since $0 \leq \theta(x_i) \leq 1$. Now, consider

$$P\left(\sup_{-\infty < X < \infty} |\widehat{\theta}(x) - E(\widehat{\theta}(x))| \leq \varepsilon\right).$$

Note that

$$\sup_{-\infty < X < \infty} |\widehat{\theta}(x) - E(\widehat{\theta}(x))| = \max_{0 \leq i \leq n-k} \left| \sum_{j=i+1}^{i+k} Z_j \right|.$$

Thus

$$\begin{aligned} P\left(\sup_{-\infty < X < \infty} |\widehat{\theta}(x) - E(\widehat{\theta}(x))| \leq \varepsilon\right) \\ = P\left(\max_{0 \leq i \leq n-k} \left| \sum_{j=i+1}^{i+k} Z_j \right| \leq \varepsilon\right) = P\left(\max_{0 \leq i \leq n-k} |S_{i+k} - S_i| \leq ve\right). \end{aligned}$$

Since

$$\max_{0 \leq i \leq n-k} |S_{i+k} - S_i| \leq \max_{0 \leq i \leq n-k} (|S_{i+k}| + |S_i|)$$

then

$$\left\{ w : \max_{0 \leq i \leq n-k} |S_{i+k} - S_i| \leq \varepsilon \right\} \supset \left\{ w : \max_{0 \leq i \leq n} |S_i| \leq \frac{\varepsilon}{2} \right\} = \left\{ w : \max_{1 \leq i \leq n} |S_i| \leq \frac{\varepsilon}{2} \right\}.$$

Therefore

$$\begin{aligned} P\left(\sup_{-\infty < X < \infty} |\hat{\theta}(x) - E(\hat{\theta}(x))| \leq \varepsilon\right) \\ \geq P\left(\max_{1 \leq i \leq n} |S_i| \leq \frac{\varepsilon}{2}\right) = 1 - P\left(\max_{1 \leq i \leq n} |S_i| > \frac{\varepsilon}{2}\right). \end{aligned}$$

Apply the Kolmogorov Inequality in Appendix D.1 to get

$$P\left(\max_{1 \leq i \leq n} |S_i| > \frac{\varepsilon}{2}\right) \leq \frac{\sigma^2(S_n)}{\left(\frac{\varepsilon}{2}\right)} = \frac{4\sigma^2(S_n)}{\varepsilon^2}. \quad (\text{E.2})$$

Now,

$$\begin{aligned} \sigma^2(S_n) &= \text{Var}(S_n) = \sum_{i=1}^n \frac{1}{k^2} \theta(x_i)(1 - \theta(x_i)) \\ &\leq \frac{1}{k^2} \sum_{i=1}^n \frac{1}{4} = \frac{n}{4k^2}. \end{aligned} \quad (\text{E.3})$$

We get

$$P\left(\sup_{-\infty < X < \infty} |\hat{\theta}(x) - E(\hat{\theta}(x))| \leq \varepsilon\right) \geq 1 - \frac{4\sigma^2(S_n)}{\varepsilon^2} \geq 1 - \frac{n}{k^2\varepsilon^2}.$$

REFERENCES

- AÏT-SAHLIA, Y., BICKEL, P.J. and STOKER, T.M., Goodness-of-fit tests for kernel regression with an application to option implied volatilities, *Journal of Econometrics* 105, 363–412, 2001.
- AALEN, O., Nonparametric inference for a family of counting processes, *Ann. Statist.* 6, 701–726, 1978.
- ABRAMOVICH, F., BENJAMINI, Y., DONOHO, D.L. and JOHNSTONE, I.M., Adapting to unknown sparsity by controlling the false discovery rate, *Ann. Statist.* 34, 584–653, 2006.
- AKAIKE, H., Fitting autoregressive models for prediction. *Ann. Inst. Statist. Math.* 21, 243–247, 1969.
- AMINI A.A. and WAINWRIGHT M.J., High-dimensional analysis of semidefinite relaxations for sparse principal components, *Ann. Statist.* 37, 2877–2921, 2009.
- AMIT, Y. and GEMAN, D., Shape quantization and recognition with randomized trees. *Neural computation* 9, 1545–1588, 1997.
- ANDERSEN, P.K., BORGAN, O., GILL, R.D. and KEIDING, N., Censoring, truncation, and filtering in statistical models based on counting processes, *Contemporary Mathematics* 80, 19–60, Providence, American Mathematical Society, 1988.
- ANDERSEN, P.K., BORGAN, O., GILL, R.D. and KEIDING, N., *Statistical Models Based on Counting Processes*. New York, Springer, 1993.
- ANDERSEN, P.K. and GILL, R.D., Cox’s regression model for counting processes: A large sample study. *Ann. Statist.* 10, 1100–1120, 1982.
- ANDERSON, T.W., *Introduction to Multivariate Statistical Analysis*, 3rd ed. J. Wiley, 2003.
- ANDREWS, D.F., BICKEL, P.J., HAMPEL, F.R., HUBER, P.J., ROGERS, W.H. and TUKEY, J. W., *Robust estimates of Location: Survey and Advances*. Princeton, NJ, Princeton University Press, 1972.
- APOSTOL, T.M., *Mathematical Analysis*, 1st ed., Addison Wesley, 1957.
- ARLOT, S. and CELISSE, A., A survey of cross validation procedures for model selection, *Statistics Surveys* 4, 40–79, 2010.

- ASSOUAD, P., Deux remarques sur l'estimation, *L.R. Acad. Sci. Paris, ser I*, 296, 1021–1024, 1983.
- BABU, G.J. and FEIGELSON, E.D., *Astrostatistics*, London, Chapman and Hall, 1996.
- BARANIUK, R.G., CEVHER, V., and WAKIN, M.B., Low-dimensional models for dimensionality reduction and signal recovery: A geometric perspective, *Proceedings of the IEEE* 98, 959–971, 2010.
- BARLOW, R.E. and PROSCHAN, F., Inequalities for linear combinations of order statistics from restricted families, *Ann. Math. Statist.* 37, 1574–1591, 1966.
- BARRON, A., BIRGÉ, L. and MASSART, P., Risk bounds for model selection via penalization, *Probab. Theory Related Fields* 113(3), 301–413, 1999.
- BELL, C.B. and DOKSUM, K.A., “Optimal” one-sample distribution-free tests and their two-sample extensions, *Ann. Math. Statist.* 37, 120–132, 1966.
- BELL, C.B., BLACKWELL, D. and BREIMAN, L., On the completeness of order statistics, *Ann. Math. Statist.* 31, 794–796, 1960.
- BENJAMINI, Y. and HOCHBERG, Y., Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. Royal Statist. Soc. Ser. B* 57, 289–300, 1995.
- BERGER, J.O., *Statistical Decision Theory and Bayesian Analysis*, 2nd ed., New York, Springer, 1985.
- BERNARDO, J.M. and SMITH, A.F.M., *Bayesian Theory*, New York, Wiley, 1994.
- BICKEL, P.J., Some contributions to the theory of order statistics, *Proceedings of Vth Berkeley Symposium on Probability and Statistics*, 575–592, 1967.
- BICKEL, P.J., Tests for monotone failure rate II, *Ann. Math. Statist.* 40, 1250–1260, 1970.
- BICKEL, P.J., Minimax estimation of the mean of a normal distribution when the parameter space is restricted”, *Ann. Statist.* 9, 1301–1309, 1981.
- BICKEL, P.J. and DOKSUM, K.A., Tests for monotone failure rate based on normalized spacings, *Ann. Math. Statist.* 40, 1216–1235, 1969.
- BICKEL, P. J. and DOKSUM, K. A., *Mathematical Statistics. Basic Ideas and Selected Topics*, 1st ed., Oakland, CA, Holden-Day, 1977.
- BICKEL, P.J. and FREEDMAN, D.A. Some asymptotic theory for the bootstrap, *Ann. Statist.* 9, 1196–1217, 1981.
- BICKEL, P.J., GÖTZE, F. and VAN ZWET, W.R., Resampling fewer than n observations: Gains, losses, and remedies for losses, *Statistica Sinica* 7, 1–31, 1997.
- BICKEL, P.J., KLAASSEN, C.A.J., RITOV, Y. and WELLNER, J.A., *Efficient and Adaptive Estimation for Semiparametric Models*, John Hopkins Univ. Press, Baltimore, 1993. Reissued by Springer, New York, 1998.
- BKRW, 1993, 1998). Short for previous entry.
- BICKEL, P.J. and KRIEGER, Confidence bands for a distribution function using the bootstrap, *J. Amer. Statist. Assoc.* 84, no.405, 95–100, 1989.

- BICKEL, P. J. and LEVINA, E., Regularized estimation of large covariance matrices, *Ann. Statist.* 36, 199–227, 2008a.
- BICKEL, P. J. and LEVINA, E., Covariance regularization by thresholding, *Ann. Statist.* 36(6), 2577–2604, 2008b.
- BICKEL, P. J. and LI, B., Regularization in statistics, Discussants: A.B. Tsybakov, S.A. van de Geer, B. Yu, T. Valdés, C. Rivero, J. Fan, and A. van der Vaart, *Test*, 271–344, 2006.
- BICKEL, P. J. and RITOY, Y., Estimating integrated squared density derivatives: Sharp best order of convergence estimates, *Sankhya*, A381-393, 1988.
- BICKEL, P.J. and RITOY, Y., Local asymptotic normality of ranks and covariates in transformation models, *Festschrift for Lucien Le Cam*, D. Pollard and G.L. Yang, eds., New York, Springer, 1997.
- BICKEL, P.J., RITOY, Y. and STOKER, T., Tailor-made tests for goodness of fit to semiparametric hypotheses, *Ann. Statist.* 34, 721–741, 2006.
- BICKEL, P. J. and ROSENBLATT, M., On some global measures of the deviation of density function estimates, *Ann. Statist.* 1, 1071–1095, 1973.
- BICKEL, P. J. and SAKOV, A., On the choice of m in the m out of n bootstrap and confidence bounds for extrema, *Statistica Sinica* 18, 967-985, 2008.
- BILLINGSLEY, P., *Convergence of Probability Measures*, New York, Wiley, 1968.
- BIRKHOFF, G. and MACLANE, S., *A Survey of Modern Algebra*, AKP Classics, 1998.
- BIRNBAUM, A., JOHNSTONE, I.M., NADLER, B. and PAUL, D., Minimax bounds for sparse PCA with noisy high-dimensional data, *Ann. Statist.* 41, 1055–1084, 2013.
- BISHOP, C., *Pattern Recognition and Machine Learning*, New York, Springer, 2006.
- BJERVE, S., DOKSUM, K.A. and YANDELL, B.S., Uniform confidence bounds for regression based on a simple moving average, *Scand. J. Statist.* 12, 159–169, 1995.
- BLACKWELL, D. and GIRSHICK, M.A. *Theory of Games and Statistical Decisions*. Wiley, New York, 1954. Reissued by Dover, New York, 1979.
- BLYTH, C.R., On minimax statistical procedures and their admissibility, *Ann. Math. Statist.* 22, 22-42, 1951.
- BOX, G.E.P and COX, D.R., An analysis of transformations, *J. Royal Statist. Soc. Ser. B* 26, 211–252, 1964.
- BOX G.E.P., HUNTER, W.G. and HUNTER, J.S. *Statistics for Experimenters*, New York, Wiley, 1978.
- BOYD, S. and VANDENBERGHE, L., *Convex Optimization*, Cambridge University Press, Cambridge, 2004.
- BREIMAN, L., *Probability*, Reading, MA, Addison-Wesley 1968. Reprinted in SIAM Classics in Applied Mathematics, 2000.

- BREIMAN, L., Random forests, *Machine learning* 45, 5–32, 2001.
- BREIMAN, L., FRIEDMAN, J.H., OLSHEN, R.A. and STONE, J., *Classification and Regression Trees*, Belmont, Wadsworth, 1984.
- BRILLINGER, D., *Time Series. Data Analysis and Theory*. Reprint of the 1981 edition. *Classics in Applied Mathematics* 36, Philadelphia, PA, Society for Industrial and Applied Mathematics (SIAM), 2001.
- BROOKS, S., GELMAN, A., JONES, G. and MENG, X.L., eds., *Handbook of Markov Chain Monte Carlo*. CRC press, 2011.
- BROWN, L., CARTER, A., LOW, M., and ZHANG, C., Equivalence theory for density estimation, Poisson processes and Gaussian white noise with drift, *Ann. Statist.* 32, 2399–2430, 2004.
- BROWN, L. and LOW, M., Asymptotic equivalence of nonparametric regression and white noise, *Ann. Statist.* 24, 2384–2398, 1996.
- BUHLMANN, P. and KÜNSCH, H.R., The blockwise bootstrap for general parameters of a stationary time series, *Scand. J. Statist.* 22, 35–54, 1995.
- BUHLMANN, P. and VAN DE GEER, S., *Statistics for High-Dimensional Data: Methods, Theory and Applications*, Heidelberg, Springer, 2011.
- CASELLA, G. and STRAWDERMAN, W. E., Estimating a bounded normal mean, *Ann. Statist.* 9, 870–878, 1981.
- CHANDRASEKARAN, V. and JORDAN, M.I., Computational and statistical tradeoffs via convex relaxation. *PNAS* 110, 1181–1190, 2013.
- CHEN, H., Convergence rates for parametric components in a partly linear model, *Ann. Statist.* 16, 136–146, 1988.
- CHEN, A. and BICKEL, P.J., *Efficient Independent Component Analysis*, Technical Report, UC Berkeley, 2003.
- CHERNOFF, H., GASTWIRTH, J.L. and JOHNS, M.V. Asymptotic distribution of linear combinations of functions of order statistics with applications to estimation. *Ann. Math. Statist.* 38, 52–72, 1967.
- CHUNG, K.L., The strong law of large numbers, *Proc. 2nd Berkeley Symp. Math. Statist. Prob.*, 341–352, Berkeley, Univ. California Press, 1951.
- CLEVELAND, W., Robust locally weighted regression and smoothing scatterplots, *J. Amer. Statist. Assoc.* 74, 829–836, 1979.
- CLEVELAND, W. and DEVLIN, S., Locally weighted regression: An approach to regression analysis by local fitting, *J. Amer. Statist. Assoc.* 83, 596–610, 1988.
- COMON, P., Independent component analysis, a new concept?, *Signal Processing* 36, 287–314, 1994.
- COVER, T.M. and HART, P.E., Nearest neighbor pattern classification, *IEEE Transactions on Information Theory* 13, 21–27, 1967.
- COVER, T.M. and THOMAS, J.A., *Elements of Information Theory*, New York, Wiley, 1991.

- Cox, D.R., Regression models and life-tables, *J. Royal Statist. Soc. Ser. B* 34, No. 2, 187–220, 1972.
- Cox, D.R., Partial likelihood, *Biometrika* 62, 269–276, 1975.
- COX, D.R. and OAKES, D., *Analysis of Survival Data*, New York, Chapman and Hall, 1984.
- DARLING, D.A. The Cramér–Smirnov test in the parametric case, *Ann. Inst. Statist. Math.* 26, 1–20, 1955.
- DAUBECHIES, T., *Ten Lectures on Wavelets*, SIAM, Philadelphia, 1992.
- DE BOOR, C., *A Practical Guide to Splines*. New York, Springer, 1978.
- DE HAAN, L. and FERREIRA, A., *Extreme Value Theory. An Introduction*, New York, Springer, 2006.
- DE MONTRICHER, G.F., TAPIA, R.A. and THOMPSON, J.R., Nonparametric maximum likelihood estimation of probability densities by penalty function methods, *Ann. Statist.* 3, 1329–1348, 1975.
- DEVROYE, L., GYÖRFI, L. and LUGOSI, G., A Probabilistic Theory of Pattern Recognition, *Applications of Mathematics* 31, New York, Springer-Verlag, 1996.
- DIACONIS, P., The Markov chain Monte Carlo revolution. *Bull. American Math. Soc.* 46, 179–205, 2009.
- DIACONIS, P. and FREEDMAN, D. On the consistency of Bayes estimates (with a discussion and a rejoinder by the authors), *Ann. Statist.* 14, 1–67, 1986.
- DIACONIS, P. and STROOCK, D., Geometric bounds for eigenvalues of Markov chains, *The Annals of Applied Probability* 1(1), 36–61, 1991.
- DIACONIS, P. and STRUMFELS, S., Algebraic algorithms for sampling from conditional distributions, *The Annals of Statist.* 26, 363–397, 1998.
- DOKSUM, K.A., An extension of partial likelihood methods for proportional hazard models to general transformation models, *Ann. Statist.* 15, 325–345, 1987.
- DOKSUM, K.A. and OZEKI, A., Semiparametric models and likelihood - the power of ranks, *Optimality. The Third Erich L. Lehmann Symposium*, Javier Rojo, ed. IMS Lecture Notes-Monograph Series, 67–92, 2009.
- DOKSUM, K.A. and NABEYA, S., Estimation in proportional hazard and log-linear models, *J. Statistical Planning and Inference*, 297–303, 1984.
- DOKSUM, K.A. and SAMAROV, A., Nonparametric estimation of global functionals and a measure of explanatory power of covariates in regression, *Ann. Statist.* 23, 1443–1473, 1995.
- DONOHO, D. and JOHNSTONE, I.M., Minimax risk over ℓ_p -balls for ℓ_q -error, *Probab. Theory Related Fields* 99, 277–303, 1994.
- DONOHO, D. and JOHNSTONE, I., Adapting to unknown smoothness via wavelet shrinkage, *J. Amer. Statist. Assoc.* 90, 1200–1224, 1995.
- DONOHO, D. and LU, R.C., Geometrizing Rates of Convergence II, III , *Ann. Statist.* 19, 633–667, 1991.

- DONSKER, M.D., Justification and extension of Doob's heuristic approach to the Kolmogorov-Smirnov theorems. *Ann. Math. Statist.* 23, 277–281, 1952.
- DOOB, J.L., Heuristic approach to the Kolmogorov-Smirnov theorems. *Ann. Math. Statist.* 20, 393–403, 1949.
- DOOB, J.L. *Stochastic Processes*, Wiley, New York, 1953.
- DOUKHAN, R. *Mixing: Properties and Examples*. Springer Lecture Notes in Statistics, New York, Springer-Verlag, 1995.
- DUDOIT, S., SHAFFER, J.P., and BOLDRICK, J.C., Multiple hypothesis testing in microarray experiments, *Statist. Sci.* 18(1), 71–103, 2003.
- DURBIN, J. Distribution theory for tests based on the sample distribution function. *Regional Conference Series in Applied Mathematics* 9, SIAM Philadelphia, PA, 1973.
- EFRON, B., Bootstrap methods: Another look at the jackknife, *Ann. Statist.* 7(1), 1–26, 1979.
- EFRON, B., Estimation the error rate of a prediction rule: Improvement on cross-validation, *J. Amer. Statist. Assoc.* 78(382), 316–331, 1983.
- EFRON, B., Microarrays, empirical Bayes and the two-groups model, *Statistical Science* 2, 197–223, 2008.
- EFRON, B., *Large Scale Inference. Empirical Bayes Methods for Estimation, Testing and Prediction*, Cambridge University Press, Cambridge, 2010.
- EFRON, B. and MORRIS, C.N., Stein's estimation rule and its competitors – an empirical Bayes approach, *J. Amer. Statist. Assoc.* 68, 117–130, 1973.
- EFRON, B. and TIBSHIRANI, R.J., *An Introduction to the Bootstrap*, Monographs on Statistics and Applied Probability 57, Chapman and Hall, New York, 1993.
- EL KAROUI, N., Recent results about the largest eigenvalue of random covariance matrices and statistical application, *Acta Physica Polonica B* 36, 2005.
- EL KAROUI, N., Operator norm consistent estimation of large dimensional sparse covariance matrices, *Ann. Stat.* 36, 2717–2756, 2008.
- ENGLE, R.F., GRANGER, C.W.J., RICE, J. and WEISS, A., Semiparametric estimates of the relation between weather and electricity sales, *J. Amer. Statist. Assoc.* 81, 310–320, 1986.
- FAN, J. and GIJBELS, I., *Local Polynomial Modelling and Its Applications*. London, Chapman and Hall, 1996.
- FAN, J., HÄRDLE, W. and MAMMEN, E., Direct estimation of low dimensional components in additive models, *Ann. Statist.* 26, 943–971, 1998.
- FAN, J. and LI, R., New estimation and model selection procedures for semiparametric modeling in longitudinal data analysis, *J. Amer. Statist. Assoc.* 99, 710–723, 2004.
- FAN, J., HAN, X. and GU, W., Estimating false discovery proportion under arbitrary covariance dependence, *J. Amer. Statist. Assoc.* 107, 1019–1035, 2012.

- FANO, R. M., *Class Notes for Transmission of Information*, Course 6.574, MIT, Cambridge MA, 1952.
- FELLER, W., *An Introduction to Probability Theory and Its Applications*, Vol. I, 3rd ed., New York, Wiley, 1968.
- FITHIAN W., SUN, D. and TAYLOR, J., Optimal inference after model selection, arXiv preprint arXiv:1410.2597s, 2014.
- FIX, E. and HODGES, J., Discriminatory analysis–nonparametric discrimination: Consistency properties. Technical Report 21-49-004 4, Randolph Field, Texas, U.S. Air Force, School of Aviation Medicine, 1951.
- FREUND, Y. and SCHAPIRE, R., A decision-theoretic generalization of online learning and application to boosting, *Journal of Computer and System Sciences* 55, 119–139, 1997.
- FRIEDMAN, J., HASTIE, T. and TIBSHIRANI, R., Additive logistic regression: A statistical view of boosting (with discussion), *Ann. Statist.* 28, 337–407, 2000.
- GELMAN, A., CARLIN, J.B., STERN, H.S. and RUBIN, D.B., *Bayesian Data Analysis*. 2nd Ed., Boca Raton, FL, Chapman & Hall/CRC, 2004.
- GELMAN, A., CARLIN, J.B., STERN, H.S., DUNSON, D.B., VEHTARI, A. and RUBIN, D.B., *Bayesian Data Analysis*, Vol. 2, Boca Raton, FL, CRC Press, 2014.
- GEMAN, S. and GEMAN, D. Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images, *IEEE Transactions on Pattern Analysis and Machine Intelligence* 6, 721–741, 1984.
- GEYER, C.J., *Optimization of Functions, Markov Chain Monte Carlo in Practice.*, p.241, CRC Press, 1995.
- GHOSH, J.K. and RAMAMOORTHI, R.V., *Bayesian nonparametrics*. New York, Springer 2003.
- GILKS, W.R. (ed.), *Markov Chain Monte Carlo in Practice*. CRC Press, 1995.
- GILL, R.D., VARDI, Y. and WELLNER, J.A., Large sample theory of empirical distributions in biased sampling models, *Ann. Statist.* 16, 1069–1112, 1988.
- GINÉ, E. and ZINN, J., Bootstrapping general empirical measures. *Ann. Probab.* 18, 851–869, 1990.
- GOOD, I. J. and R. A. GASKINS, Nonparametric roughness penalties for probability densities, *Biometrika* 58, 255–277, 1971.
- GÖTZE, F., Abstract. *Bulletin of Institute of Mathematical Statistics*, 1993.
- GRENNANDER, U., *Abstract Inference*, Wiley, New York, 1981.
- GRIMMETT, G. and STIRZAKER, D., *Probability and Random Processes*, Oxford Univ. Press, Oxford, 2001.
- GYÖRFI L., KOHLER, M., KRZYZAK, A. and WALK, H., *A Distribution-Free Theory of Nonparametric Regression*, New York, Springer, 2002.
- HÁJEK, J., A characterization of limiting distributions of regular estimates, *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, 14, 323–330, 1970.

- HÁJEK, J., Local asymptotic minimax and admissibility in estimation. *Proceedings of the Sixth Berkeley Symposium on Math. Statist. and Prob.*, 1, 175–194, 1972.
- HÁJEK, J. and SIDÁK, Z., *Theory of Rank Tests*, Academic Press, New York, 1967.
- HALL, P. *The Bootstrap and Edgeworth Expansion*. Springer, New York, 1997.
- HALL, P. and HYDE, C.C., *Martingale Theory and its Application*, Academic Press, 1980.
- HALL, P., On the number of bootstrap simulations required to construct a confidence interval, *Ann. Statist.* 14, 1453–1462, 1986.
- HAMMERSLEY, J.M. and HANDSCOMB, D.C., *Monte Carlo Methods*. Chapman and Hall, London, 1965.
- HÄRDLE, W., LIANG, H. and GAO, J., *Partially Linear Models*, Springer, New York, 2000.
- HASTIE, T. and TIBSHIRANI, R., *Generalised Additive Models*, Chapman and Hall, London, 1990.
- HASTIE, T., TIBSHIRANI, R. and FRIEDMAN, J., *The Elements of Statistical Learning*, 2nd ed., New York, Springer, 2001, 2009.
- HASTIE, T., TIBSHIRANI, R. and WAINWRIGHT, M., *Statistical Learning with Sparsity: The Lasso and Generalizations*, CRC Press, 2015.
- HODGES, J.L., Jr. and LEHMANN, E.L., Some applications of the Cramér Rao inequality, *Proc. Second Berkeley Symp. on Math. Statist. and Prob.*, 1, Univ. of California Press, 1951, 13–22.
- HOEFFDING, W., A class of statistics with asymptotically normal distributions. *Ann. Math. Statist.* 19, 293–325, 1948.
- HOEFFDING, W., Optimum nonparametric tests, *Proc. of 2nd Berkeley Symposium on Mathematical Statistics and Probability*, Univ. California Press, 1951, 83–92.
- HOEFFDING, W., Probability inequalities for sums of bounded random variables.” *J. Amer. Statist. Assoc.* 58, 13–30, 1963.
- HOERL, A.E. and KENNARD, R.W., Ridge regression: Biased estimation for nonorthogonal problems, *Technometrics*, 12, 55–67, 1970.
- HOLM, S., A simple sequentially rejective multiple test procedure, *Scand. J. Statist.* 6, 65–70, 1979.
- HUBER, P., The behaviour of maximum likelihood estimates under non standard conditions. *Proc. of 5th Berkeley Symposium on Math. Statist. and Prob.*, 221–234, 1967.
- HUBER, P., *Robust Statistics*, New York, Wiley, 1981.
- HYVÄRINEN, A., KARHUNEN, J. and OJA, E., *Independent Component Analysis*, New York, John Wiley and Sons, 2001.
- IBRAGIMOV, I. and HASMINSKII, R.Z., *Statistical Estimation: Asymptotic Theory*, New York, Springer, 1981.
- IBRAGIMOV, I. and LINNIK, Y., *Independent and Stationary Sequences of Random Variables*, WoltersNordhoff Publishing, Groningen, 1971.

- JAMES, W. and STEIN, C., Estimation with quadratic loss, *Proc. Fourth Berkeley Symposium on Math. Statist. and Prob.*, I, Univ. of California Press, 1961, 311–319.
- JIANG, J. and DOKSUM, K.A., Empirical plug-in curve and surface estimates”, *Mathematical and Statistical Methods in Reliability*, B.H. Lindquist and K.A. Doksum, eds., New Jersey, World Scientific, 2003.
- JOHNSON, R.A. and WICHERN, D. W., *Applied Multivariate Statistical Analysis*, Upper Saddle River, New Jersey, Prentice-Hall, 2003.
- JOHNSTONE, I.M. and LU, A.Y., On consistency and sparsity for principal components analysis in high dimensions, *J. Amer. Statist. Assoc.*, 104, 682–693, 2009.
- JOHNSTONE, I.M. and SILVERMAN, B.W., Empirical Bayes selection of wavelet thresholds, *Ann. Statist.* 33, 1700–1752, 2005.
- JUNG, S. and MARRON, J.S., PCA consistency in high dimension, low sample size context, *Ann. Statist.* 37, 4104–4130, 2009.
- KAGAN, A.M., LINNIK, Y.V. and RAO, C.R., *Characterization Problems of Mathematical Statistics*, New York, Wiley, 1973.
- KALBFLEISCH, J.D. and PRENTICE, R.L., Marginal likelihoods based on Cox’s regression and life model, *Biometrika* 60, 267–278, 1973.
- KALBFLEISCH, J.D. and PRENTICE, R.L., *The Statistical Analysis of Failure Time Data*, 2nd ed., Wiley, New York, 2002.
- KALBFLEISCH, J. and SPROTT, D.A., Application of likelihood methods involving a large number of parameters (with discussion), *J. Royal Statist. Soc. Ser. B* 32, 175–208, 1970.
- KERKYACHARIAN, G. and PICKARD, D., Wavelet shrinkage: Asymptopia? (with discussion), *J. Royal Statist. Soc. Ser. B* 57, 201–337, 1995.
- KIEFER, J. and WOLFOWITZ, J. Consistency of the maximum likelihood estimator in the presence of infinitely many nuisance parameters, *Ann. Math. Statist.* 27, 887–906, 1956.
- KLAASSEN, C.A.J. and WELLNER, J. A., Efficient estimation in the bivariate normal copula model: Normal margins are least favourable, *Bernoulli* 3, 55–77, 1997.
- KOSOROK, M.R., *Introduction to Empirical Processes and Semiparametric Inference*, New York, Springer, 2008.
- KÜNSCH, H.R., The jackknife and the bootstrap for general stationary observations, *Ann. Statist.* 17, 1217–1241, 1989.
- LAWLESS, J.E., *Statistical Models and Methods for Lifetime Data*, New York, Wiley, 1982.
- LE CAM, L., On the asymptotic theory of estimation and testing hypotheses, *Proc. Third Berkeley Symp. Math. Statist. Prob.* 1, Univ. California Press, Berkeley, 1956, 129–156.
- LE CAM, L., Locally asymptotically normal families of distributions, *Univ. California Publ. Statist.* 3, 37–98, 1960.
- LE CAM, L., Sufficiency and approximate sufficiency, *Ann. Math. Statist.* 35, 1419–1455, 1964.

- LE CAM, L., Limits of experiments, *Proc. Sixth Berkeley Symp. Math. Statist. Prob. 1*, Univ. California Press, Berkeley, 1972, 245–261.
- LE CAM, L., *Asymptotic Methods in Statistical Decision Theory*, New York, Springer-Verlag, 1986.
- LE CAM, L. and YANG, G.L., *Asymptotics in Statistics; Some Basic Concepts*, New York, Springer, 1990.
- LEHMANN, E.L., Consistency and unbiasedness of certain nonparametric tests, *Ann. Math. Statist.* 22, No. 2, 165–179, 1951.
- LEHMANN, E.L., The power of rank tests, *Ann. Math. Statist.* 24, 23–43, 1953.
- LEHMANN, E.L., Ordered families of distributions, *Ann. Math. Statist.* 26, 399–419, 1955.
- LEHMANN, E.L., *Nonparametrics. Statistical Methods Based on Ranks*, New York, Springer, 2006.
- LEHMANN, E.L. and CASELLA, G., *Theory of Point Estimation*, 2nd ed., New York, Springer, 1998.
- LEHMANN, E.L. and ROMANO, J., *Testing Statistical Hypotheses*, 3rd ed., New York, Springer, 2005.
- LEHMANN, E.L. and SCHEFFÉ, H., Completeness, similar regions, and unbiased estimation, *Part 1, Sankhyā* 10, 305–340, *Part 2, Sankhyā* 15, 219–236, 1950, 1955.
- LEHMANN, E.L., ROMANO, J.P. and SHAFFER, J.P., On optimality of stepdown and stepup multiple test procedures, *Ann. Statist.* 33(3), 1084–1108, 2005.
- LEVIT, B.Y., Infinite dimensional informational lower bounds, *Theory Prob. Applic.* 23, 388–394, 1978.
- LIN, D.Y. and ZENG, D., Correcting for population stratification in genomewide association studies, *J. Amer. Statist. Assoc.* 106, 997–1008, 2011.
- LITTLE, R.J.A. and RUBIN, D.B., *Statistical Analysis with Missing Data*, New York, Wiley, 1986. Republished by New York, Wiley, 2014.
- LIU, J.S., *Monte Carlo Strategies in Scientific Computing*. New York, Springer, 2001.
- LOADER, C., *Local Regression and Likelihood*, Springer, New York, 1999.
- MALLOWS, C.L., A note on asymptotic joint normality, *Ann. Math. Statist.*, 43, 508–515, 1972.
- MALLOWS, C.L., Some comments on C_p , *Technometrics* 15, 661–675, 1973.
- MAMMEN, E. and TSYBAKOV, A.B., Asymptotic minimax recovery of sets with smooth boundaries, *Ann. Statist.* 23, 502–524, 1995.
- MAMMEN, E. and van de GEER, S.A., Penalized quasi-likelihood estimation in partial linear models, *Ann. Statist.* 25, 1014–1035, 1997.
- MARCINKIEWICZ, J. and ZYGMUND, A., Quelques inégalités pour les opérations linéaires, *Fundamenta Mathematica* 32, 113–121, 1939.
- MARDIA, K.V., KENT, J.T. and BIBBY, J.M., *Multivariate Analysis*, Academic Press, 1979.

- MARRON, J.S., Optimal rates of convergence to Bayes risk in nonparametric discrimination, *Ann. Statist.*, 11, 1142–1155, 1983.
- MECKLIN, C.J. and MUNDFROM, D.J., An appraisal and bibliography of tests for multivariate normality, *International Stat. Review* 72, 123–138, 2004.
- MEIR, R and RÄTSCH, G., An introduction to boosting and leveraging, In *Advanced Lectures on Machine Learning*, S. Mendelson and A. Smola, eds., Lecture Notes in Computer Science, Springer, 2003, 119–184.
- METROPOLIS, N., ROSENBLUTH, A. W., ROSENBLUTH, M. N., TELLER, A. H. and TELLER, E., Equations of state calculations by fast computing machines, *J. Chemical Physics* 21, 1087–1091, 1953.
- MEYN, S.P. and TWEEDIE, R.L., *Markov Chains and Stochastic Stability*. New York, Springer, 1993.
- MORGAN, J.N. and SONQUIST, J.A., Problems in the analysis of survey data, and a proposal, *J. Amer. Statist. Assoc.* 58, 415, 1963.
- MURPHY, S.A., Asymptotic theory for the frailty model, *Ann. Statist.* 23, No. 1, 182–198, 1995.
- MURPHY, S.A., Consistency in a proportional hazards model incorporating a random effect, *Ann. Statist.* 22, 712–731, 1994.
- MURPHY, S.A. and van der VAART, A.W., On profile likelihood (with discussion), *J. Amer. Statist. Assoc.* 95, 449–485, 2000.
- MURPHY, S.A., ROSSINI, T.J. and van der VAART, A.W., MLE in the proportional odds model, *J. Amer. Statist. Assoc.* 92, 968–976, 1997.
- NACHBIN, L., *The Haar Integral*, Von Nostrand, New York, 1965.
- NADARAYA, E.A., *Nonparametric Estimation of Probability Densities and Regression Curves*, Boston, Kluwer Academic Publishers, 1989.
- NUSSBAUM, M., Asymptotic equivalence of density estimation and Gaussian white noise, *Annals of Statistics*, 24, 2399–2430, 1996.
- OSTLAND, M. *A Monte Carlo algorithm applied to travel time estimation and vehicle matching*, UC Berkeley Thesis, 1999.
- OWEN, A., Empirical likelihood ratio confidence intervals for a single functional, *Biometrika* 75, 237–249, 1988.
- OWEN, A., *Empirical Likelihood*, Chapman and Hall, 2001.
- PARNER, E., Asymptotic theory for the correlated gamma-frailty model, *Ann. Statist.* 26, 183–214, 1998.
- PASULA, H., RUSSELL, S., OSTLAND, M. and RITOY, Y., Tracking many objects with many sensors, Proc. IJCAI-99, 1999.
- PATTERSON, N., PRICE, A.L. and REICH, D., Population structure and eigenanalysis, *PLOS Genetics* 12, 2074–2093, 2006.

- PAUL, D., Asymptotics of sample eigenstructure for a large dimensional spiked covariance model, *Statistica Sinica* 17, 1617–1642, 2007.
- PEARL, J., *Causality, Models, Reasoning and Inference*, 2nd ed., Cambridge University Press, Cambridge, 2009.
- PETTY, K., BICKEL, P., JIANG, J., OSTLAND, M., RICE, J., RITOY, Y. and SCHOENBERG, F., Accurate estimation of travel times from single-loop detectors. *Transportation Research Part A: Policy and Practice*, 32(1), 1–17, 1998.
- PETTY, K., OSTLAND, M., KWON, J., RICE, J. and BICKEL, P., A new methodology for evaluating incident detection algorithms, *Transportation Research Part C: Emerging Technologies* 10, 189–204, 2002.
- PINSKER, M.S., Optimal filtering of square integrable signals in Gaussian white noise, *Problems of Information Transmission*, 16, 120–133, 1980.
- POLITIS, D.N. and ROMANO, J.P., Large sample confidence regions based on subsamples under minimal assumptions (in resampling), *Ann. Statist.* 22, No. 4, 2031–2050, 1994.
- POLITIS, D.N., ROMANO, J.P. and WOLF, M., *Subsampling*. New York, Springer, 1999.
- POLLARD, D. *Convergence of Stochastic Processes*, New York, Springer-Verlag, 1984.
- POLLARD, D., Empirical Processes: Theory and Applications, NSF-CBMS Regional Conference Series in Probability and Statistics, Volume 2, Hayward, CA., IMS, 1990.
- PRICE, A., PATTERSON, N., PLENGE, R., WEINBLATT, M., SHADICK, N. and REICH, D., Principal components analysis corrects for stratification in genome-wide association studies, *Nature Genetics* 38, 904–909, 2006.
- PROSCHAN, F. and PYKE, R., Tests for monotone failure rate, *Proc. Fifth Berkeley Symp. Math. Statist. Prob.* 3, Univ. of California Press, 1967.
- QUENOUILLE, M.H., Approximate Tests of Correlation in Time-Series, *J. Royal Statist. Soc. Ser. B* 11, 68–84, 1949.
- QUINLAN, J.R., Simplifying decision trees, *International Journal of Man-Machine Studies* 27(3), 221–234, 1987.
- QUINLAN, J.R., *C4.5: Programs for Machine Learning*, Morgan Kaufmann Series in Machine Learning, San Francisco, Morgan Kaufmann Publishers, 1993.
- RAO, C.R., *Linear Statistical Inference and Its Applications*, 2nd ed., Wiley, New York, 1973.
- RAO, C.R. and SHINOZAKI, N., Precision of individual estimators in simultaneous estimation of parameters, *Biometrika* 65, 23–30, 1978.
- RAO, P., *Nonparametric Functional Estimation*, Orlando, Academic Press, 1983.
- REID, N., A conversation with Sir David Cox, *Statistical Science* 9, 439–455, 1994.
- RIPLEY, B.D., *Stochastic Simulation*, New York, Wiley, 1987.
- RIPLEY, B.D., *Pattern Recognition and Neural Networks*, Cambridge University Press, Cambridge, 1996.

- RISSANEN, J., Universal Prior for Integers and Estimation by Minimum Description Length, *Ann. Statist.* 11(2), 416–431, 1983.
- ROBBINS, H. An empirical Bayes approach to statistics, *Proc. Third Berkeley Symp. Math. Statist. and Prob 1*, Univ. of Calif. Press, Berkeley CA, 1956, 157–164.
- ROBBINS, H., The empirical Bayes approach to statistical decision problems, *Ann. Math. Statist.* 35, 1–20, 1964.
- ROCKAFELLAR, R.T., *Convex Analysis*, Princeton, NJ, Princeton University Press, 1969.
- ROSENTHAL, J., Asymptotic variance and convergence rates of nearly periodic Markov chain Monte Carlo algorithms, *J. Amer. Statist. Assoc.* 98, 169–177, 2003.
- RUBIN, D.B., Comments on J. Neyman and causal inference in experiments and observational studies. On the application of probability theory to agriculture experiments. *Statist. Sci.* 4, 472–480, 1990.
- RUDEMO, M., Empirical choice of histograms and kernel density estimators *Scand. J. Statist.* 9, 65–78, 1982.
- RUDIN, C., DAUBECHIES, I. and SCHAPIRE, R.E., The dynamics of AdaBoost: Cyclic behavior and convergence of margins, *J. Mach. Learn. Res.* 5, 1557–1595, 2003.
- RUPPERT, D. and WAND, M.P., Multivariate weighted least squares regression, *Ann. Statist.* 22, 1346–1370, 1994.
- RUPPERT, D., WAND, M.P., HOLST, U. and HÖSSJER, O., Local polynomial variance-function estimation. *Technometrics* 39, 262–273, 1997.
- SACKS, J., WELCH, W.J., MITCHELL, T.J. and WYNN, H.P., Design and analysis of computer experiments, *Statist. Sci.* 4, 1989.
- SAVAGE, I.R., Contributions to the theory of rank order statistics, the two-sample case, *Ann. Math Statist.* 27, 590–615, 1956.
- SAVAGE, I.R., Contributions to the theory of rank order statistics – the “trend” case, *Ann. Math Statist.* 28, 968–977, 1957.
- SAVAGE, I.R., Lehmann alternatives, Proceedings of Conference on Nonparametric Statistical Inference, Budapest, Hungary, 1980, 795–821.
- SCHAPIRE, R.E., The strength of weak learnability, *Machine Learning* 5, 197–227, 1990.
- SCHERVISH, M., *Theory of Statistics*. New York, Springer, 1995.
- SCHWARZ, G., Estimating the dimension of a model, *Ann. Statist.* 6(2), 461–464, 1978.
- SCORNET, E., BIAU, G., and VERT, J., Consistency of random forests, *Ann. Statist.* 43, 1716–1741, 2015.
- SCOTT, D. W., On optimal and data-based histograms, *Biometrika* 66, 605–610, 1979.
- SCOTT, D.W., *Multivariate Density Estimation. Theory, Practice, and Visualization*. Wiley, New York, 1992.
- SEBER, G.A.F., *Multivariate Observations*, New York, Wiley, 1984.

- SERFLING, R.J., *Approximation Theorems of Mathematical Statistics*, New York, Wiley, 1980.
- SHAO, J. and TU, D., *The Jackknife and Bootstrap*, New York, Springer, 1995.
- SHEN, D., SHEN, H., and MARRON, J. S., Consistency of sparse PCA in high dimension, low sample size contexts, Technical report, 2011. Available at <http://arxiv.org/pdf/1104.4289v1.pdf>
- SHIBATA, R., An optimal selection of regression variables. *Biometrika* 68, 45–54, 1981.
- SHORACK, G.R., Weak convergence of the general quantile process in $\|\cdot/q\|$ -metrics. *Bull. Inst. Math. Statist.* 11, 60–71, 1982.
- SHORACK, G.R. and WELLNER, J.A., *Empirical Processes with Applications to Statistics*, New York, Wiley, 1986.
- SIBUYA, M., Generating doubly exponential random numbers, *Ann. Inst. Statist. Math. Tokyo Suppl.* VI–7, 1968.
- SILVERMAN, B., *Density Estimation for Statistics and Data Analysis*, London, Chapman and Hall, 1986.
- SKLAR, A., Fonctions de répartition à n dimensions et leurs marges, *L'Institut de Statistique de L'Université de Paris*, 8, 1959, 229–231.
- SPIÓTVOLL, E., On the optimality of some multiple comparison procedures”, *Ann. Math. Statist.* 43, 398–411, 1972.
- STANLEY, R.P., *Enumerative Combinations*, Wadsworth, Monterey, 1986.
- STEIN, C.M., Efficient nonparametric testing and estimation, *Proc. Third Berkeley Symp. Math. Statist. Prob. 1*, Univ. California Press, Berkeley, 1956a, 187–195.
- STEIN, C.M., Inadmissibility of the usual estimator for the mean of a multivariate distribution *Proc. Third Berkeley Symp. Math. Statist. and Prob 1*, Univ. of Calif. Press, Berkeley CA, 1956b, 197–206.
- STEIN, C.M., Estimation of the mean of a multivariate normal distribution, *Ann. Statist.*, 9, 1135–1151, 1981.
- STEIN, C.M., Approximate computation of expectations, *Lecture Notes and Monograph Series V.7*, Hayward, CA, Institute of Mathematical Statistics, 1986.
- STIGLER, S., Linear functions of order statistics.” *Ann. Math. Statist.* 40, 770–788, 1969.
- STIGLER, S.M., Completeness and unbiased estimation, *The American Statistician* 26, 28–29, 1972.
- STONE, C.J., Optimal global rates of convergence for nonparametric regression, *Ann. Statist.* 10, 1040–1053, 1982.
- STONE, C.J., HANSEN, M., KOOPERBERG, C. and TRUONG, Y., Polynomial splines and their tensor products (with discussion and rejoinder by authors and Jianhua Z. Huang), *Ann. Statist.* 25, 1341–1470, 1997.
- STONE, C. J. and KOO, J. Y., Logspline density estimation, *Contemporary Mathematics* 59, 1–15, 1986.

- STOREY, J.D., The optimal discovery procedure: A new approach to simultaneous significance testing, *J. Roy. Statist. Soc. Ser. B* 69, 347–368, 2007.
- SUN, W. and CAI, T.T., Oracle and adaptive compound decision rules for false discovery rate control, *J. Amer. Statist. Assoc.* 102, 901–912, 2007.
- TALAGRAND, M., Sharper bounds for Gaussian and empirical processes. *Ann. Prob.* 22, 28–76, 1994.
- TAPIA, R. and THOMPSON, J., *Nonparametric Probability Density Estimation*, Baltimore, John Hopkins University Press, 1978.
- TIBSHIRANI, R., Regression shrinkage and selection via the lasso, *J. Roy. Statist. Soc. Ser. B* 58, 267–288, 1996.
- TIERNEY, L., Introduction to general state space Markov chain Theory, *Markov Chain Monte Carlo in Practice.*, CRC Press, 1995, p. 59.
- TIKHONOV, A., Solution of incorrectly formulated problems and the regularization method, *Soviet Math. Dokl.* 4, 1035–1038, 1963.
- TSIATIS, A.A., A large sample study of Cox's regression model, *Ann. Statist.* 9, 93–108, 1981.
- TSYBAKOV, A., *Introduction to Nonparametric Estimation*, New York, Springer, 2008.
- TUKEY, J.W., Bias and confidence in not-quite large samples, *Ann. Math. Statist.* 29, 614, 1958.
- VAPNIK, V.N., *The Nature of Statistical Learning Theory*, 2nd Ed., 1998, New York, Springer, 1996.
- VAPNIK, V.N., *Statistical Learning Theory*. New York, Wiley, 1998.
- VAN DE GEER, S.A., *Applications of Empirical Processes Theory*, Vol. 6 of Cambridge Ser. in Statist. and Probabilistic Mathematics, 2000(a).
- VAN DE GEER, S.A., *Empirical Processes in M-Estimation*, Cambridge Univ. Press, Cambridge, 2000(b).
- VAN DER LAAN, M.J. and ROBINS, J.M. *Unified Methods for Censored Longitudinal Data and Causality*, New York, Springer, 2003.
- VAN DER VAART, A.W., *Asymptotic Statistics*, Cambridge, Cambridge Univ. Press, 1998.
- VAN DER VAART, A.W. and WELLNER, J., *Weak Convergence and Empirical Processes: With Applications to Statistics*. New York, Springer, 1996.
- VAN ZWET, W.R., *Convex Transformation of Random Variables*, Math. Centrum, Amsterdam, 1964.
- VAN ZWET, W.R., A Berry-Esseen bound for symmetric statistics. *Zeitschrift fur Wahrscheinlichkeitstheorie und Verw Gebiete* 66, 425–440, 1984.
- VAUPEL, J.W., MANTON, K.G. and STALLARD, E., The impact of heterogeneity in individual frailty on the dynamics of mortality, *Demography* 16, 439–454, 1979.

- VENTER, J. and DE WET, T., Asymptotic distributions of certain test statistics, *South African Statist. J.* 6, 135–149, 1972.
- VIOLLAZ, A.J., Nonparametric estimation of probability density functions based on orthogonal expansions, *Rev. Mat. Uni. Complut. Madrid*, 41–84, 1989.
- VON MISES, R., On the asymptotic distribution of differentiable statistical functions. *Ann. Math. Statist.* 18, 309–348, 1947.
- VON NEUMANN, J., Various techniques used in connection with random digits, *National Bureau of Standards Applied Mathematics Series* 12, 36–38, 1951.
- WACHTER, K.W., The strong limits of random matrix spectra for sample matrices of independent elements”, *The Annals of Probability* 6, 1–18, 1978.
- WAHBA, G., On the numerical solution of Fredholm integral equations of the first kind, Tech. Report, Math. Research Center, Univ. of Wisconsin, Madison, 1969.
- WAHBA, G., *Spline Models for Observational Data*, SIAM, Philadelphia, 1990.
- WAINWRIGHT, M.J. and JORDAN, M.I., Graphical models, exponential families, and variational inference, *Foundations and Trends® in Machine Learning* 1, 2008, 1–305.
- WALD, A., Test of statistical hypothesis concerning several parameters when the number of observations is large, *Transactions of the American Math. Soc.* 54, 426–482, 1943.
- WALD, A., *Statistical Decision Functions*, Oxford, Wiley, 1950.
- WAND, M. P. and JONES, M. C., *Kernel Smoothing*, London, Chapman and Hall, 1995.
- WANG, Y., A likelihood ratio test against stochastic ordering in several populations, *J. Amer. Statist. Assoc.* 91, 1676–1683, 1996.
- WIDDER, D.V., *The Laplace Transform*, Princeton NJ, Princeton University Press, 1941.
- WIJSMAN, R., Invariant measures on groups and their use in statistics, Hayward, CA, IMS Lecture Notes, 1990.
- WITTEN, D. M., TIBSHIRANI, R. and HASTIE, T., A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis, *Biostatistics* 10, 515–534, 2009.
- YANG, F., On high dimensional data analysis and biomedical genomics, Ph.D. Thesis, Dept. of Statistics, Univ. of Wisconsin, 2013.
- YANG, F., DOKSUM, K. and TSUI, K.W., Principal component analysis (PCA) for high dimensional data. PCA is dead. Long live PCA. In: *Proceedings for Workshop on Perspectives on High Dimensional Data Analysis II*, Montreal, S.E. Ahmed, ed., Comtemporary Mathematics, American Mathematical Society, Providence, RI, 2014.
- YUAN, M. and LIN, Y., Model selection and estimation in regression with grouped variables. *J. Royal Statist. Soc. Ser. B* 68(1), 49–67, 2007.
- ZENG, D. and LIN, D.Y., Efficient estimation for the accelerated failure time model. *J. Amer. Statist. Assoc.* 102, 1387–1396, 2007.
- ZENG, D. and LIN, D.Y., Maximum likelihood estimation in semiparametric regression models with censored data, *J.R. Statist. Soc. B* 69, 507–564, 2007a.

ZHAO, P., ROCHA, G. and YU, B., The composite absolute penalties family for grouped and hierarchical variable selection, *Ann. Statist.* 37, 3468-3497, 2009.

ZOU, H., The adaptive Lasso and its oracle properties, *J. Amer. Statist. Assoc.* 101, 1418-1429, 2006.

Accessing the E-book edition

Using the VitalSource® ebook

Access to the VitalBook™ ebook accompanying this book is via VitalSource® Bookshelf – an ebook reader which allows you to make and share notes and highlights on your ebooks and search across all of the ebooks that you hold on your VitalSource Bookshelf. You can access the ebook online or offline on your smartphone, tablet or PC/Mac and your notes and highlights will automatically stay in sync no matter where you make them.

1. **Create a VitalSource Bookshelf account at <https://online.vitalsource.com/user/new> or log into your existing account if you already have one.**
2. **Redeem the code provided in the panel below to get online access to the ebook.**
Log in to Bookshelf and select **Redeem** at the top right of the screen. Enter the redemption code shown on the scratch-off panel below in the **Redeem Code** pop-up and press **Redeem**. Once the code has been redeemed your ebook will download and appear in your library.



No returns if this code has been revealed.

DOWNLOAD AND READ OFFLINE

To use your ebook offline, download Bookshelf to your PC, Mac, iOS device, Android device or Kindle Fire, and log in to your Bookshelf account to access your ebook:

On your PC/Mac

Go to <https://support.vitalsource.com/hc/en-us> and follow the instructions to download the free **VitalSource Bookshelf** app to your PC or Mac and log into your Bookshelf account.

On your iPhone/iPod Touch/iPad

Download the free **VitalSource Bookshelf** App available via the iTunes App Store and log into your Bookshelf account. You can find more information at <https://support.vitalsource.com/hc/en-us/categories/200134217-Bookshelf-for-iOS>

On your Android™ smartphone or tablet

Download the free **VitalSource Bookshelf** App available via Google Play and log into your Bookshelf account. You can find more information at <https://support.vitalsource.com/hc/en-us/categories/200139976-Bookshelf-for-Android-and-Kindle-Fire>

On your Kindle Fire

Download the free **VitalSource Bookshelf** App available from Amazon and log into your Bookshelf account. You can find more information at <https://support.vitalsource.com/hc/en-us/categories/200139976-Bookshelf-for-Android-and-Kindle-Fire>

N.B. The code in the scratch-off panel can only be used once. When you have created a Bookshelf account and redeemed the code you will be able to access the ebook online or offline on your smartphone, tablet or PC/Mac.

SUPPORT

If you have any questions about downloading Bookshelf, creating your account, or accessing and using your ebook edition, please visit <http://support.vitalsource.com/>

Mathematical Statistics: Basic Ideas and Selected Topics, Volume II presents important statistical concepts, methods, and tools not covered in the authors' previous volume. This second volume focuses on inference in non- and semiparametric models. It not only reexamines the procedures introduced in the first volume from a more sophisticated point of view but also addresses new problems originating from the analysis of estimation of functions and other complex decision procedures and large-scale data analysis.

The book covers asymptotic efficiency in semiparametric models from the Le Cam and Fisherian points of view as well as some finite sample size optimality criteria based on Lehmann-Scheffé theory. It develops the theory of semiparametric maximum likelihood estimation with applications to areas such as survival analysis. It also discusses methods of inference based on sieve models and asymptotic testing theory. The remainder of the book is devoted to model and variable selection, Monte Carlo methods, nonparametric curve estimation, and prediction, classification, and machine learning topics.

Features

- Develops basic asymptotic tools, including weak convergence for random processes, empirical process theory, and the functional delta method
- Discusses the classical theory of statistical optimality in a decision-theoretic context
- Presents inference procedures and their properties in a variety of applications and models, such as Cox's regression model, models for censored data, and partial linear models
- Describes properties of Monte Carlo/simulation-based methods, including the bootstrap and Markov chain Monte Carlo (MCMC)
- Examines the nonparametric estimation of functions of one or more variables
- Covers many topics related to statistical learning, including support vector machines and classification and regression trees (CART)

Using the tools and methods developed in this textbook, you will be ready for advanced research in modern statistics. Numerous examples and problems illustrate statistical modeling and inference concepts. Measure theory is not required for understanding.

WITH VITALSOURCE®
EBOOK



- Access online or download to your smartphone, tablet or PC/Mac
- Search the full text of this and other titles you own
- Make and share notes and highlights
- Copy and paste text and figures for use in your own documents
- Customize your view by changing font size and layout



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

K25643

ISBN: 978-1-4987-2268-1

90000



WWW.CRCPRESS.COM