

ECONOMETRIC
EXERCISES 7

Bayesian Econometric Methods

Gary Koop
Dale J. Poirier
Justin L. Tobias

CAMBRIDGE

CAMBRIDGE

www.cambridge.org/9780521855716

This page intentionally left blank

Bayesian Econometric Methods

This book is a volume in the Econometric Exercises series. It teaches principles of Bayesian econometrics by posing a series of theoretical and applied questions, and providing detailed solutions to those questions. This text is primarily suitable for graduate study in econometrics, though it can be used for advanced undergraduate courses, and should generate interest from students in related fields, including finance, marketing, agricultural economics, business economics, and other disciplines that employ statistical methods. The book provides a detailed treatment of a wide array of models commonly employed by economists and statisticians, including linear regression-based models, hierarchical models, latent variable models, mixture models, and time series models. Basics of random variable generation and simulation via Markov Chain Monte Carlo (MCMC) methods are also provided. Finally, posterior simulators for each type of model are rigorously derived, and Matlab computer programs for fitting these models (using both actual and generated data sets) are provided on the Web site accompanying the text.

Gary Koop is Professor of Economics at the University of Strathclyde. He has published numerous articles in Bayesian econometrics and statistics in journals such as the *Journal of Econometrics*, *Journal of the American Statistical Association*, and the *Journal of Business and Economic Statistics*. He is an associate editor for several journals, including the *Journal of Econometrics* and *Journal of Applied Econometrics*. He is the author of the books *Bayesian Econometrics*, *Analysis of Economic Data*, and *Analysis of Financial Data*.

Dale J. Poirier is Professor of Economics at the University of California, Irvine. He is a Fellow of the Econometric Society, the American Statistical Association, and the *Journal of Econometrics*. He has been on the Editorial Boards of the *Journal of Econometrics* and *Econometric Theory* and was the founding editor of *Econometric Reviews*. His professional activities have been numerous, and he has held elected positions in the American Statistical Association and the International Society for Bayesian Analysis. His previous books include *Intermediate Statistics and Econometrics: A Comparative Approach* and *The Econometrics of Structural Change*.

Justin L. Tobias is Associate Professor of Economics, Iowa State University, and has also served as an Assistant and Associate Professor of Economics at the University of California, Irvine. Professor Tobias has authored numerous articles in leading journals, including the *International Economic Review*, *Journal of Applied Econometrics*, *Journal of Business and Economic Statistics*, *Journal of Econometrics*, and the *Review of Economics and Statistics*.

Econometric Exercises

Editors:

Karim M. Abadir, *Tanaka Business School,
Imperial College London, UK*

Jan R. Magnus, *CentER and Department of Econometrics and Operations Research,
Tilburg University, The Netherlands*

Peter C. B. Phillips, *Cowles Foundation for Research in Economics,
Yale University, USA*

Titles in the Series (* = planned):

- 1 Matrix Algebra (K. M. Abadir and J. R. Magnus)
- 2 Statistics (K. M. Abadir, R. D. H. Heijmans, and J. R. Magnus)
- 3 Econometric Theory, I (P. Paruolo)
- 4 Empirical Applications, I (A. van Soest and M. Verbeek)
- 5 Econometric Theory, II
- 6 Empirical Applications, II
- 7 Bayesian Econometric Methods (G. Koop, D. J. Poirier, and J. L. Tobias)
- * Time Series Econometrics, I
- * Time Series Econometrics, II
- * Microeconometrics
- * Panel Data
- * Nonlinear Models
- * Nonparametrics and Semiparametrics
- * Simulation-Based Econometrics
- * Computational Methods
- * Financial Econometrics
- * Robustness
- * Econometric Methodology

Bayesian Econometric Methods

Gary Koop

University of Strathclyde, Scotland

Dale J. Poirier

University of California, Irvine, USA

Justin L. Tobias

Iowa State University, USA



CAMBRIDGE
UNIVERSITY PRESS

CAMBRIDGE UNIVERSITY PRESS

Cambridge, New York, Melbourne, Madrid, Cape Town, Singapore, São Paulo

Cambridge University Press

The Edinburgh Building, Cambridge CB2 8RU, UK

Published in the United States of America by Cambridge University Press, New York

www.cambridge.org

Information on this title: www.cambridge.org/9780521855716

© Gary Koop, Dale J. Poirier, and Justin L. Tobias 2007

This publication is in copyright. Subject to statutory exception and to the provision of relevant collective licensing agreements, no reproduction of any part may take place without the written permission of Cambridge University Press.

First published in print format 2007

ISBN-13 978-0-511-29482-2 eBook (EBL)

ISBN-10 0-511-29482-4 eBook (EBL)

ISBN-13 978-0-521-85571-6 hardback

ISBN-10 0-521-85571-3 hardback

ISBN-13 978-0-521-67173-6 paperback

ISBN-10 0-521-67173-6 paperback

Cambridge University Press has no responsibility for the persistence or accuracy of urls for external or third-party internet websites referred to in this publication, and does not guarantee that any content on such websites is, or will remain, accurate or appropriate.

To Lise

To the Reverend but not the Queen

To Melissa, Madeline, and Drew

Contents

List of exercises	ix
Preface to the series	xv
Preface	xix
1 The subjective interpretation of probability	1
2 Bayesian inference	11
3 Point estimation	29
4 Frequentist properties of Bayesian estimators	37
5 Interval estimation	51
6 Hypothesis testing	59
7 Prediction	71
8 Choice of prior	79
9 Asymptotic Bayes	91
10 The linear regression model	107
11 Basics of Bayesian computation	117
11.1 Monte Carlo integration	119
11.2 Importance sampling	124
11.3 Gibbs sampling and the Metropolis–Hastings algorithm	128
11.4 Other (noniterative) methods for generating random variates	157
12 Hierarchical models	169
13 The linear regression model with general covariance matrix	191
14 Latent variable models	203
15 Mixture models	253
15.1 Some scale mixture of normals models	254
15.2 Other continuous and finite-mixture models	260
16 Bayesian model averaging and selection	281
16.1 Bayesian model averaging	282
16.2 Bayesian variable selection and marginal likelihood calculation	287

17	Some stationary time series models	297
18	Some nonstationary time series models	319
Appendix		335
Bibliography		343
Index		353

List of exercises

Chapter 1: The subjective interpretation of probability

1.1	Let's make a deal	6
1.2	Making Dutch book	7
1.3	Independence and exchangeability	7
1.4	Predicted probability of success of a Bernoulli random variable	7
1.5	Independence and conditional independence	8

Chapter 2: Bayesian inference

2.1	Conjugate Bernoulli analysis	12
2.2	Application of Bayes' theorem	13
2.3	Conjugate normal analysis with unknown mean and known variance	14
2.4	Conjugate normal analysis with unknown mean and variance	17
2.5	Properties of a normal-gamma distribution	18
2.6	Fictitious sample interpretation of natural conjugate priors	19
2.7	Conjugate normal analysis with known mean and unknown variance	20
2.8	Marginal distribution of θ_1 when $\theta_1, \theta_2 \sim NG$	21
2.9	Conditional distribution of $\theta_2 \theta_1$ when $\theta_1, \theta_2 \sim NG$	21
2.10	Conjugate exponential analysis	22
2.11	Conjugate uniform analysis	22
	Conjugate poisson analysis	22
*2.13	Conjugate exponential family analysis	23
*2.14	Conjugate multivariate normal analysis	25
2.15	Identification	27

Chapter 3: Point estimation

3.1	Point estimation under quadratic loss	31
3.2	Point estimation under asymmetric linear loss	32
3.3	Point estimation under all-or-nothing loss	32
3.4	Point estimation under quadratic loss for conjugate Bernoulli analysis	33
3.5	Point estimation under asymmetric linear loss for conjugate Bernoulli analysis	33
3.6	Point estimation under all-or-nothing loss for conjugate Bernoulli analysis	33
3.7	Point estimation for conjugate normal analysis	34
3.8	Point estimation under weighted squared error loss	34
3.9	Point estimation under alternative squared error loss	34
3.10	Point estimation under nonconjugate normal analysis	35
3.11	Point estimation under linex loss: Zellner [1986b]	36

Chapter 4: Frequentist properties of Bayesian estimators

*4.1	Sampling properties under quadratic loss in conjugate normal analysis	38
*4.2	Sampling properties under quadratic loss in conjugate exponential analysis	40
4.3	Expected bias of Bayesian point estimator	43
4.4	Balanced loss in conjugate normal analysis: Zellner [1994]	43
4.5	Empirical Bayes estimation	45
4.6	Point estimation under linex loss	46
4.7	Risk, Bayes risk, and admissibility: Berger [1985]	47
4.8	Likelihood principle	48

Chapter 5: Interval estimation

5.1	Graphical determination of an HPD interval	52
5.2	Calculating a posterior interval probability	53
5.3	HPD intervals under normal random sampling with known variance	54
5.4	HPD intervals under normal random sampling with unknown variance	55
5.5	Berger and Wolpert [1988, pp. 5–6]	57
5.6	Conditional frequentist reasoning and ancillarity	57
5.7	HPD intervals and reparameterization	58

Chapter 6: Hypothesis testing

6.1	Two simple hypotheses and general likelihood	62
6.2	Two simple hypotheses and normal likelihoods	63
*6.3	Two one-sided hypotheses and normal likelihoods	63
6.4	Simple null, two-sided alternative, and general likelihood	66
6.5	Simple null hypothesis, two-sided alternative, and normal likelihoods	67
6.6	Special cases of Exercise 6.5	67
6.7	Relationship among posterior probabilities under different prior probabilities	68
6.8	Savage–Dickey density ratio	69
6.9	Generalized Savage–Dickey density ratio	69

Chapter 7: Prediction

7.1	Predictive p.m.f. for conjugate Bernoulli analysis	73
7.2	Predictive p.m.f. for conjugate Poisson analysis	73
7.3	Predictive density for conjugate normal analysis with known variance	74
7.4	Predictive point and interval estimation for conjugate normal analysis with known variance	75
7.5	Predictive density for conjugate multivariate normal analysis with known covariance matrix	75
7.6	Predictive estimation: conjugate multivariate normal analysis with known covariance matrix	75
7.7	Predictive density for conjugate normal analysis with unknown mean and variance	76
7.8	Predictive estimation: conjugate normal analysis with unknown mean and variance	76
7.9	Conjugate multivariate normal analysis with unknown mean and covariance matrix	77
7.10	Predictive density for conjugate exponential analysis	78

Chapter 8: Choice of prior

8.1	Bernoulli sampling distribution	82
8.2	Negative binomial sampling distribution	83
8.3	Kass and Wasserman [1996, Section 4.3]	84
8.4	Poisson sampling distribution	84
8.5	Exponential sampling distribution and reparameterization	85
8.6	Normal sampling distribution with unknown mean and variance	86
8.7	Hypothesis testing with improper priors	87
8.8	Bayes factor with a diffuse proper prior	87
*8.9	Strong inconsistency of improper priors	88

Chapter 9: Asymptotic Bayes

9.1	Posterior means	94
9.2	Laplace approximation of a gamma integrating constant	95
9.3	Laplace approximation of a gamma integrating constant, 2	95
9.4	Laplace approximation of a gamma integrating constant, 3	96

9.5	Comparison of Laplace approximations of a gamma integrating constant	97
9.6	Laplace approximation of the posterior mean of an exponential random variable	97
9.7	Tierney–Kadane approximation: exponential posterior mean: Achcar and Smith [1990]	98
9.8	Tierney–Kadane approximation: exponential posterior mean, 2: Achcar and Smith [1990]	100
9.9	Approximate Bayes factors	101
9.10	Approximate marginal posteriors	103

Chapter 10: The linear regression model

10.1	Posterior distributions under a normal-gamma prior	108
10.2	A partially informative prior	110
10.3	Ellipsoid bound theorem	111
*10.4	Problems with Bayes factors using noninformative priors: Leamer [1978, p. 111]	112
10.5	Multicollinearity	114

Chapter 11: Basics of Bayesian computation

Section 11.1: Monte Carlo integration

11.1	Numerical standard error of Monte Carlo estimate	119
11.2	Drawing from standard distributions	120
11.3	Monte Carlo integration in a regression problem	121

Section 11.2: Importance sampling

11.4	Importance sampling	124
11.5	Importance sampling: prior sensitivity	126

Section 11.3: Gibbs sampling and the Metropolis–Hastings algorithm

11.6	Gibbs sampling	129
11.7	Gibbs sampling from the bivariate normal	130
11.8	Gibbs sampling in a regression problem	134
11.9	Gibbs sampling in a model of groupwise heteroscedasticity	135
11.10	Gibbs sampling in the SUR model	136
11.11	Gibbs sampling in a regression model with an unknown changepoint	138
11.12	Gibbs sampling in a count data model with an unknown changepoint	140
11.13	Error diagnostics with Gibbs sampling	142
11.14	Calculating numerical standard errors with Gibbs sampling	144
11.15	Gibbs sampling in a regression model with inequality constraints: Geweke [1996b]	146
11.16	Checking for errors in a posterior simulator: Geweke [2004]	150
11.17	Metropolis–Hastings algorithm, 1: Gelman, Carlin, Stern, and Louis [1995]	152
11.18	Metropolis–Hastings algorithm, 2	154

Section 11.4: Other (noniterative) methods for generating random variates

11.19	The inverse transform	157
11.20	Applications of the inverse transform: drawing exponential and truncated normal variates	159
11.21	Acceptance/rejection sampling	160
11.22	Using acceptance/rejection sampling, 1	162
11.23	A generalized acceptance/rejection algorithm	162
11.24	Using acceptance/rejection sampling, 2	165
11.25	The weighted bootstrap	167

Chapter 12: Hierarchical models

12.1	General results for normal hierarchical linear models	169
12.2	Hierarchical modeling with longitudinal data	171
12.3	A posterior simulator for a linear hierarchical model: Gelfand et al. [1990]	173
12.4	The linear hierarchical model, 2: estimation and convergence diagnostics	177
12.5	Blocking steps in random effects models	181
*12.6	Hierarchical recentering: Gelfand, Sahu, and Carlin [1995]	184
12.7	A hierarchical/smoothness prior approach to nonparametric regression	187

Chapter 13: The linear regression model with general covariance matrix

13.1	Theoretical results for the general case	192
13.2	Heteroscedasticity of a known form	193
13.3	Heteroscedasticity of an unknown form: student t errors	195
13.4	Autocorrelated errors	198

Chapter 14: Latent variable models

14.1	Posterior simulation in the probit model	204
14.2	Parameter identification in the probit model	206
14.3	Probit model: application using Fair's [1978] data	208
14.4	Posterior simulation in the panel probit model	210
14.5	Posterior simulation in the ordered probit, 1	211
14.6	Posterior simulation in the ordered probit, 2: Nandram and Chen [1996]	214
14.7	Posterior simulation in the multinomial probit	216
14.8	Posterior simulation in the tobit model	219
14.9	A tobit model with two-sided censoring: a female labor supply application	221
14.10	Posterior simulation in a treatment–effect model	223
14.11	Posterior simulation in a generalized tobit: a model of potential outcomes, 1	225
14.12	A model of potential outcomes, 2	229
14.13	Posterior simulation in an ordered-treatment, ordered-outcome model	232
14.14	Posterior simulation in the stochastic frontier model	236
14.15	Posterior simulation in the two-part [hurdle] model	240
14.16	Posterior simulation in the sample-selection model	243
14.17	Missing data, 1	246
14.18	Missing data, 2	247
14.19	Missing data, 3	248

Chapter 15: Mixture models**Section 15.1: Some scale mixture of normals models**

15.1	The student t -distribution as a scale mixture of normals	254
15.2	Student t errors in regression and binary choice models	256
15.3	Posterior simulation in a regression model with double exponential errors	258

Section 15.2: Other continuous and finite-mixture models

15.4	The negative binomial as a continuous mixture of the Poisson	260
15.5	Experiments with two-component normal mixtures	262
15.6	Using finite normal mixtures in a regression model	265
15.7	A normal mixture with covariates affecting the mixture assignment: response delays	266
15.8	Are my students randomly guessing?	270
15.9	A skewed regression model	276

Chapter 16: Bayesian model averaging and selection**Section 16.1: Bayesian model averaging**

16.1	Basic g-prior results commonly used with BMA	282
16.2	Markov chain Monte Carlo model composition [MC ³]	283
16.3	BMA: an application in economic growth	284

Section 16.2: Bayesian variable selection and marginal likelihood calculation

16.4	Stochastic search variable selection	287
*16.5	Nonparametric regression via variable selection: Smith and Kohn [1996]	289
16.6	Calculating the marginal likelihood with simulated output: Gelfand–Dey [1994]	293
16.7	Calculating the marginal likelihood with simulated output: Chib [1995]	293
16.8	Calculating marginal likelihoods with generated data	295

Chapter 17: Some stationary time series models

17.1	Using the AR(p) model to understand the properties of a series	298
17.2	The threshold autoregressive (TAR) model	300
17.3	Extensions of the basic TAR model, 1: other threshold triggers	304
17.4	Extensions of the basic TAR model, 2: switches in the error variance	307
17.5	Autoregressive conditional heteroscedasticity	310
17.6	Vector autoregressive models	312

Chapter 18: Some nonstationary time series models

18.1	The AR(1) model with a flat prior	320
18.2	Sims and Uhlig [1991]	321

18.3	Unit roots and Jeffreys' prior	324
18.4	Berger and Yang [1994]	326
18.5	Kadane, Chan, and Wolfson [1996]	327
18.6	Cointegration	330

Preface to the series

The past two decades have seen econometrics grow into a vast discipline. Many different branches of the subject now happily coexist with one another. These branches interweave econometric theory and empirical applications and bring econometric method to bear on a myriad of economic issues. Against this background, a guided treatment of the modern subject of econometrics in volumes of worked econometric exercises seemed a natural and rather challenging idea.

The present series, *Econometric Exercises*, was conceived in 1995 with this challenge in mind. Now, almost a decade later it has become an exciting reality with the publication of the first installment of a series of volumes of worked econometric exercises. How can these volumes work as a tool of learning that adds value to the many existing textbooks of econometrics? What readers do we have in mind as benefiting from this series? What format best suits the objective of helping these readers learn, practice, and teach econometrics? These questions we now address, starting with our overall goals for the series.

Econometric Exercises is published as an organized set of volumes. Each volume in the series provides a coherent sequence of exercises in a specific field or subfield of econometrics. Solved exercises are assembled together in a structured and logical pedagogical framework that seeks to develop the subject matter of the field from its foundations through to its empirical applications and advanced reaches. As the Schaum series has done so successfully for mathematics, the overall goal of *Econometric Exercises* is to develop the subject matter of econometrics through solved exercises, providing a coverage of the subject that begins at an introductory level and moves through to more advanced undergraduate- and graduate-level material.

Problem solving and worked exercises play a major role in every scientific subject. They are particularly important in a subject like econometrics in which there is a rapidly growing literature of statistical and mathematical technique and an ever-expanding core to the discipline. As students, instructors, and researchers, we all benefit by seeing carefully

worked-out solutions to problems that develop the subject and illustrate its methods and workings. Regular exercises and problem sets consolidate learning and reveal applications of textbook material. Clearly laid out solutions, paradigm answers, and alternate routes to solution all develop problem-solving skills. Exercises train students in clear analytical thinking and help them in preparing for tests and exams. Teachers, as well as students, find solved exercises useful in their classroom preparation and in designing problem sets, tests, and examinations. Worked problems and illustrative empirical applications appeal to researchers and professional economists wanting to learn about specific econometric techniques. Our intention for the *Econometric Exercises* series is to appeal to this wide range of potential users.

Each volume of the series follows the same general template. Chapters begin with a short outline that emphasizes the main ideas and overviews the most relevant theorems and results. The introductions are followed by a sequential development of the material by solved examples and applications, and by computer exercises when appropriate. All problems are solved and they are graduated in difficulty with solution techniques evolving in a logical, sequential fashion. Problems are asterisked when they require more creative solutions or reach higher levels of technical difficulty. Each volume is self-contained. There is some commonality in material across volumes to reinforce learning and to make each volume accessible to students and others who are working largely, or even completely, on their own.

Content is structured so that solutions follow immediately after the exercise is posed. This makes the text more readable and avoids repetition of the statement of the exercise when it is being solved. More importantly, posing the right question at the right moment in the development of a subject helps to anticipate and address future learning issues that students face. Furthermore, the methods developed in a solution and the precision and insights of the answers are often more important than the questions being posed. In effect, the inner workings of a good solution frequently provide benefit beyond what is relevant to the specific exercise.

Exercise titles are listed at the start of each volume, following the table of contents, so that readers may see the overall structure of the book and its more detailed contents. This organization reveals the exercise progression, how the exercises relate to one another, and where the material is heading. It should also tantalize readers with the exciting prospect of advanced material and intriguing applications.

The series is intended for a readership that includes undergraduate students of econometrics with an introductory knowledge of statistics, first- and second-year graduate students of econometrics, as well as students and instructors from neighboring disciplines (such as statistics, psychology, or political science) with interests in econometric methods. The volumes generally increase in difficulty as the topics become more specialized.

The early volumes in the series (particularly those covering matrix algebra, statistics, econometric models, and empirical applications) provide a foundation to the study of econometrics. These volumes will be especially useful to students who are following the first-year econometrics course sequence in North American graduate schools and need to prepare for

graduate comprehensive examinations in econometrics and to write an applied econometrics paper. The early volumes will equally be of value to advanced undergraduates studying econometrics in Europe, to advanced undergraduates and honors students in the Australasian system, and to masters and doctoral students in general. Subsequent volumes will be of interest to professional economists, applied workers, and econometricians who are working with techniques in those areas, as well as students who are taking an advanced course sequence in econometrics and statisticians with interests in those topics.

The *Econometric Exercises* series is intended to offer an independent learning-by-doing program in econometrics and it provides a useful reference source for anyone wanting to learn more about econometric methods and applications. The individual volumes can be used in classroom teaching and examining in a variety of ways. For instance, instructors can work through some of the problems in class to demonstrate methods as they are introduced; they can illustrate theoretical material with some of the solved examples; and they can show real data applications of the methods by drawing on some of the empirical examples. For examining purposes, instructors may draw freely from the solved exercises in test preparation. The systematic development of the subject in individual volumes will make the material easily accessible both for students in revision and for instructors in test preparation.

In using the volumes, students and instructors may work through the material sequentially as part of a complete learning program, or they may dip directly into material in which they are experiencing difficulty to learn from solved exercises and illustrations. To promote intensive study, an instructor might announce to a class in advance of a test that some questions in the test will be selected from a certain chapter of one of the volumes. This approach encourages students to work through most of the exercises in a particular chapter by way of test preparation, thereby reinforcing classroom instruction.

Further details and updated information about individual volumes can be obtained from the *Econometric Exercises* Web site:

<http://us.cambridge.org/economics/ee/econometricexercises.htm>

The Web site also contains the basic notation for the series, which can be downloaded along with the \LaTeX style files.

As series editors, we welcome comments, criticisms, suggestions, and, of course, corrections from all our readers on each of the volumes in the series as well as on the series itself. We bid you as much happy reading and problem solving as we have had in writing and preparing this series.

York, Tilburg, New Haven
June 2005

Karim M. Abadir
Jan R. Magnus
Peter C. B. Phillips

Preface

Bayesian econometrics has enjoyed an increasing popularity in many fields. This popularity has been evidenced through the recent publication of several textbooks at the advanced undergraduate and graduate levels, including those by Poirier (1995), Bauwens, Lubrano, and Richard (1999), Koop (2003), Lancaster (2004), and Geweke (2005). The purpose of the present volume is to provide a wide range of exercises and solutions suitable for students interested in Bayesian econometrics at the level of these textbooks.

The Bayesian researcher should know the basic ideas underlying Bayesian methodology (i.e., Bayesian theory) and the computational tools used in modern Bayesian econometrics (i.e., Bayesian computation). The Bayesian should also be able to put the theory and computational tools together in the context of substantive empirical problems. We have written this book with these three activities – theory, computation, and empirical modeling – in mind. We have tried to construct a wide range of exercises on all of these aspects. Loosely speaking, Chapters 1 through 9 focus on Bayesian theory, whereas Chapter 11 focuses primarily on recent developments in Bayesian computation. The remaining chapters focus on particular models (usually regression based). Inevitably, these chapters combine theory and computation in the context of particular models. Although we have tried to be reasonably complete in terms of covering the basic ideas of Bayesian theory and the computational tools most commonly used by the Bayesian, there is no way we can cover all the classes of models used in econometrics. Accordingly, we have selected a few popular classes of models (e.g., regression models with extensions and panel data models) to illustrate how the Bayesian paradigm works in practice. Particularly in Chapters 12 through 18 we have included substantive empirical exercises – some of them based closely on journal articles. We hope that the student who works through these chapters will have a good feeling for how serious Bayesian empirical work is done and will be well placed to write a Ph.D. dissertation or a journal article using Bayesian methods.

For the student with limited time, we highlight that a division in this book occurs between the largely theoretical material of Chapters 1 through 9 and the largely regression-based material in Chapters 10 through 18. A student taking a course on Bayesian statistical theory could focus on Chapters 1 through 9, whereas a student taking a Bayesian econometrics course (or interested solely in empirical work) could focus more on Chapters 10 through 18 (skimming through the more methodologically oriented material in the early chapters).

Although there have been some attempts to create specifically Bayesian software (e.g., BUGS, which is available at <http://www.mrc-bsu.cam.ac.uk/bugs>, or BACC, which is available at <http://www2.cirano.qc.ca/~bacc>), in our estimation, most Bayesians still prefer to create their own programs using software such as Matlab, OX, or GAUSS. We have used Matlab to create answers to the empirical problems in this book. Our Matlab code is provided on the Web site associated with this book:

http://www.econ.iastate.edu/faculty/tobias/Bayesian_exercises.html

A few notational conventions are applied throughout the book, and it is worthwhile to review some of these prior to diving into the exercises. In regression-based problems, which constitute a majority of the exercises in the later chapters, lowercase letters such as y and x_i are reserved to denote scalar or vector quantities whereas capitals such as \mathbf{Y} or \mathbf{X}_j are used to denote matrices. In cases in which the distinction between vectors and scalars is critical, this will be made clear within the exercise. In the regression-based problems, \mathbf{y} is assumed to denote the $n \times 1$ vector of stacked responses for the dependent variable, y_i the i th element of that vector, \mathbf{x}_i a vector of covariate data, and \mathbf{X} the $n \times k$ matrix obtained from stacking the \mathbf{x}_i over i . Latent variables, which are often utilized in the computational chapters of the book, are typically designated with a “*” superscript, such as y_i^* . In Chapters 1 through 9, many exercises are presented that are not directly related to linear regression models or models that can be viewed as linear on suitably defined latent data. In these exercises, the distinction between random variables and realizations of those variables is sometimes important. In such cases, we strive to use capital letters to denote random variables, which are unknown *ex ante*, and lowercase letters to denote their realizations, which are known *ex post*. So, in the context of discussing a posterior distribution (which conditions on the data), we will use \bar{y} , but if we are interested in discussing the sampling properties of the sample mean, \bar{y} would be the appropriate notation. Finally, “ \times ” is used to denote multiplication in multiline derivations, and specific parameterizations of various densities are provided in the Appendix associated with this book.

On the issue of parameterization, the reader who is somewhat familiar with the Bayesian literature may realize that researchers often employ different parameterizations for the same model, with no particular choice being “correct” or “ideal.” A leading example is the linear regression model, in which the researcher can choose to parameterize this model in terms of the error variance or the error precision (the reciprocal of the variance). In this book, we try and remain consistent in terms of parameterization within individual chapters, though some departures from this trend do exist, particularly in Chapters 11 and 16. These differences arise from our own individual tastes and styles toward approaching these models, and they

are superficial rather than substantive. In our view it is quite valuable to expose the student to the use of different parameterizations, since this is the reality that he or she will face when exploring the Bayesian literature in more detail. In all cases, the parameterization employed is clearly delineated within each exercise.

We would like to thank the editors of the *Econometrics Exercises* series – Karim Abadir, Jan Magnus, and Peter Phillips – for their helpful comments and support during the planning and writing of this book. Hoa Jia, Babatunde Abidoye, and Jingtao Wu deserve special recognition for reviewing numerous exercises and helping to reduce the number of typographical errors. The list of other colleagues and students who have helped us – through designing, solving, and pointing out errors in our problems or solutions – is too long to enumerate here. We would, however, like to thank our students at the University of California, Irvine; Leicester University; University of Toronto; and the Institute for Advanced Studies and CIDE (Italy) for their participation, wise insights, and enthusiasm.

The subjective interpretation of probability

Reverend Thomas Bayes (born circa 1702; died 1761) was the oldest son of Reverend Joshua Bayes, who was one of the first ordained nonconformist ministers in England. Relatively little is known about the personal life of Thomas Bayes. Although he was elected a Fellow of the Royal Society in 1742, his only known mathematical works are two articles published posthumously by his friend Richard Price in 1763. The first dealt with the divergence of the Stirling series, and the second, “An Essay Toward Solving a Problem in the Doctrine of Chances,” is the basis of the paradigm of statistics named for him. His ideas appear to have been independently developed by James Bernoulli in 1713, also published posthumously, and later popularized independently by Pierre Laplace in 1774. In their comprehensive treatise, Bernardo and Smith (1994, p. 4) offer the following summarization of *Bayesian statistics*:

Bayesian Statistics offers a rationalist theory of personalistic beliefs in contexts of uncertainty, with the central aim of characterizing how an individual should act in order to avoid certain kinds of undesirable behavioral inconsistencies. The theory establishes that expected utility maximization provides the basis for rational decision making and that Bayes’ Theorem provides the key to the ways in which beliefs should fit together in the light of changing evidence. The goal, in effect, is to establish rules and procedures for individuals concerned with disciplined uncertainty accounting. The theory is not descriptive, in the sense of claiming to model actual behavior. Rather, it is prescriptive, in the sense of saying “if you wish to avoid the possibility of these undesirable consequences you must act in the following way.”

Bayesian econometrics consists of the tools of Bayesian statistics applicable to the models and phenomena of interest to economists. There have been numerous axiomatic formulations leading to the central unifying Bayesian prescription of maximizing subjective

utility as the guiding principle of Bayesian statistical analysis. Bernardo and Smith (1994, Chapter 2) is a valuable segue into this vast literature. Deep issues are involved regarding meaningful separation of probability and utility assessments, and we do not address these here.

Non-Bayesians, who we hereafter refer to as *frequentists*, argue that situations not admitting repetition under essentially identical conditions are not within the realm of statistical enquiry, and hence “probability” should not be used in such situations. Frequentists define the probability of an event as its long-run relative frequency. This frequentist interpretation cannot be applied to (i) unique, once-and-for-all type of phenomenon, (ii) hypotheses, or (iii) uncertain past events. Furthermore, this definition is nonoperational since only a finite number of trials can ever be conducted. In contrast, the desire to expand the set of relevant events over which the concept of probability can be applied, and the willingness to entertain formal introduction of “nonobjective” information into the analysis, led to the subjective interpretation of probability.

Definition 1.1 (Subjective interpretation of probability) Let \mathcal{K} denote the body of knowledge, experience, or information that an individual has accumulated about the situation of concern, and let E denote an uncertain event (not necessarily repetitive). The probability of E afforded by \mathcal{K} is the “degree of belief” in E held by an individual in the face of \mathcal{K} .

Since at least the time of Ramsey (1926), such degrees of belief have been operationalized in terms of agreed upon reference lotteries. Suppose you seek your degree of belief, denoted $p = P(E | \mathcal{K})$, that an event E occurs. Consider the following two options.

1. Receiving a small reward $\$1$ if E occurs, and receiving $\$0$ if E does not occur.
2. Engaging in a lottery in which you win $\$1$ with probability p , and receiving $\$0$ with probability $1 - p$.

If you are indifferent between these two choices, then your degree of belief in E occurring is p . Requiring the reward to be “small” is to avoid the problem of introducing utility into the analysis; that is, implicitly assuming utility is linear in money for small gambles.

Bruno de Finetti considered the interesting situation in which an individual is asked to quote betting odds (ratios of probabilities) on a set of uncertain events and accept any wagers others may decide to make about these events. According to de Finetti’s *coherence principle* the individual should never assign “probabilities” so that someone else can select stakes that guarantee a sure loss (*Dutch book*) for the individual whatever the eventual outcome. A sure loss amounts to the “undesirable consequences” contained in the earlier quote of Bernardo and Smith. This simple principle implies the axioms of probability discussed in Abadir, Heijmans, and Magnus (2006, Chapter 1) except that the additivity of probability of intersections for disjoint events is required to hold only for *finite* intersections. Nonetheless, for purposes of convenience, we consider only countably additive probability in this volume.

De Finetti's Dutch book arguments also lead to the standard rule for conditional probability. Consider two events A and B. By using the *factorization rule for conditional probability* [Abadir et al. (2006, p. 5)],

$$(A \text{ and } B) = (A) (B | A) = (B) (A | B),$$

the simplest form of Bayes' theorem follows immediately:

$$(B | A) = \frac{(A) (B | A)}{(A)}$$

In words, we are interested in the event A to which we assign the *prior* probability (A) for its occurrence. We observe the occurrence of the event B . The probability of B occurring given that A has occurred is the *posterior* probability $(B | A)$. More generally, we have the following result.

Theorem 1.1 (Bayes' theorem for events) Consider a probability space $[\Omega, \mathcal{F}, (\cdot)]$ and a collection $E_n \in \mathcal{F} \ (n = 1, 2, \dots)$ of mutually disjoint events such that $(E_n) \geq 0 \ (n = 1, 2, \dots)$ and $E_1 \cup E_2 \cup \dots \cup E_N = \Omega$. Then

$$(E_n | A) = \frac{(A | E_n) (E_n)}{\sum_{j=1}^N (A | E_j) (E_j)} \quad (n = 1, 2, \dots, N) \quad (1.1)$$

for every $A \in \mathcal{F}$ such that $(A) > 0$.

Proof: The proof follows directly upon noting that the denominator in (1.1) is (A) .

An important philosophical topic is whether the conditionalization in Bayes theorem warrants an unquestioned position as the model of learning in the face of knowledge of the event A . Conditional probability $(B | A)$ refers to *ex ante* beliefs on events not yet decided. *Ex post* experience of an event can sometimes have a striking influence on the probability assessor (e.g., experiencing unemployment, stock market crashes, etc.), and the experience can bring with it more information than originally anticipated in the event. Nonetheless, we adopt such conditionalization as a basic principle.

The subjective interpretation reflects an individual's personal assessment of the situation. According to the subjective interpretation, probability is a property of an individual's perception of reality, whereas according to classical and frequency interpretations, probability is a property of reality itself. For the subjectivist there are no "true unknown probabilities" in the world out there to be discovered. Instead, "probability" is in the eye of the beholder.

Bruno de Finetti assigned a fundamental role in Bayesian analysis to the concept of *exchangeability*, defined as follows.

Definition 1.2 A finite sequence $E_t \ (t = 1, 2, \dots, T)$ of events (or random variables) is *exchangeable* iff the joint probability of the sequence, or any subsequence, is invariant under permutations of the subscripts, that is,

$$(E_{\pi(1)}, E_{\pi(2)}, \dots, E_{\pi(T)}) = (E_1, E_2, \dots, E_T), \quad (1.2)$$

where $(\cdot)(\cdot) = 1, 2, \dots, \cdot$ is a permutation of the elements in $\{1, 2, \dots, \cdot\}$. An infinite sequence is exchangeable iff any finite subsequence is exchangeable.

Exchangeability provides an operational meaning to the weakest possible notion of a sequence of “similar” random quantities. It is “operational” because it only requires probability assignments of *observable* quantities, although admittedly this becomes problematic in the case of infinite exchangeability. For example, a sequence of Bernoulli trials is *exchangeable* iff the probability assigned to particular sequences does not depend on the order of “successes” (\cdot) and “failures” (\cdot) . If the trials are exchangeable, then the sequences FSS, SFS, and SSF are assigned the same probability.

Exchangeability involves recognizing symmetry in beliefs concerning only observables, and presumably this is something about which a researcher may have intuition. Ironically, subjectivists emphasize observables (data) and objectivists focus on unobservables (parameters). Fortunately, Bruno de Finetti provided a subjectivist solution to this perplexing state of affairs. De Finetti’s representation theorem and its generalizations are interesting because they provide conditions under which exchangeability gives rise to an isomorphic world in which we have iid observations conditional on a mathematical construct, namely, a parameter. These theorems provide an interpretation of parameters that differs substantively from the interpretation of an objectivist.

As in the case of iid sequences, the individual elements in an exchangeable sequence are identically distributed, but they are *not* necessarily independent, and this has important predictive implications for learning from experience. The importance of the concept of exchangeability is illustrated in the following theorem.

Theorem 1.2 (de Finetti’s representation theorem) Let $t(\cdot) = 1, 2, \dots$ be an infinite sequence of Bernoulli random variables indicating the occurrence (1) or nonoccurrence (0) of some event of interest. For any finite sequence $t(\cdot) = 1, 2, \dots, \cdot$, define the average number of occurrences

$$\bar{t}_T = \frac{1}{T} \sum_{t=1}^T t \quad (1.3)$$

Let $(\cdot_1, \cdot_2, \dots, \cdot_T) = \Pr(\cdot_1 = 1, \cdot_2 = 2, \dots, \cdot_T = T)$ denote a probability mass function (p.m.f.) reflecting *exchangeable* beliefs for an arbitrarily long finite sequence $t(\cdot) = 1, 2, \dots, \cdot$, and let $(\cdot) = \Pr(\cdot \leq \cdot)$ denote its associated cumulative distribution function (c.d.f.). Then (\cdot) has the representation

$$(\cdot_1, \cdot_2, \dots, \cdot_T) = \int_0^1 (\cdot) (\cdot), \quad (1.4)$$

where

$$(\cdot) = \prod_{t=1}^T y_t (1 - y_t)^{(1-y_t)}, \quad (1.5)$$

$$(\cdot) = \lim_{T \rightarrow \infty} H(\bar{t}_T \leq \cdot), \quad (1.6)$$

and $H(\cdot)$ denotes probability with respect to the c.d.f. (\cdot) corresponding to p.m.f. (1.4).

Proof: See de Finetti (1937) or the simpler exposition of Heath and Sudderth (1976).

Theorem 1.1 implies that it is *as if*, given $\{p_t\}$ ($t = 1, 2, \dots$), are iid Bernoulli trials where the probability of a success is p_t , and the “parameter” p_t is assigned a probability distribution with c.d.f. $F(p)$ that can be interpreted as belief about the long-run relative frequency of $\bar{p}_T \leq p$ as $T \rightarrow \infty$. From de Finetti’s standpoint, both the quantity \bar{p}_T and the notion of independence are “mathematical fictions” implicit in the researcher’s subjective assessment of arbitrarily long *observable* sequences of successes and failures. The parameter p_t is of interest primarily because it constitutes a limiting form of predictive inference about the observable \bar{p}_T via (1.6). The mathematical construct p_t may nonetheless be useful. However, Theorem 1.2 implies that the subjective probability distribution need not apply to the “fictitious” p_t but only to the *observable* exchangeable sequence of successes and failures. When the c.d.f. is absolutely continuous, so that $F(p) = \int_0^p f(x) dx$ exists, then (1.4) becomes

$$P(p_1, p_2, \dots, p_T) = \int_0^1 \prod_{t=1}^T p_t^{y_t} (1-p_t)^{(1-y_t)} f(p) dp \quad (1.7)$$

It is clear from (1.4) and (1.7) that exchangeable beliefs assign probabilities acting as if the p_t ’s are iid Bernoulli random variables given $f(p)$, and then average over values of p using the weight $f(p)$ to obtain a marginal density for the p_t ’s. Let $T = \bar{p}_T$ be the number of successes in T trials. Since there are $\binom{T}{r}$ ways in which to obtain $T = r$ successes in T trials, it follows immediately from (1.4) and (1.5) that

$$\Pr(T = r) = \binom{T}{r} \int_0^1 p^r (1-p)^{T-r} f(p) dp \quad (r = 0, 1, \dots, T), \quad (1.8)$$

where

$$f(p) = \lim_{T \rightarrow \infty} \Pr(\bar{p}_T^{-1} T \leq p) \quad (1.9)$$

Thus, given $f(p)$, it follows from (1.8) that exchangeable beliefs assign probabilities acting as if \bar{p}_T has a binomial distribution given $f(p)$, and then average over values of p using the weight $f(p) = \binom{T}{r} \Pr(\bar{p}_T = r)$. Bayes and Laplace suggest choosing the “mixing” distribution $f(p)$ for p to be uniform over $[0, 1]$, in which case (1.8) reduces to

$$\Pr(T = r) = \binom{T}{r} \int_0^1 p^r (1-p)^{T-r} dp = \frac{1}{T+1} \binom{T}{r}, \quad r = 0, 1, \dots, T, \quad (1.10)$$

In words, (1.10) describes beliefs that in T trials, any number r of successes are equally likely. In the degenerate case in which the distribution of p assigns probability one to some value p_0 , then de Finetti’s theorem implies that \bar{p}_T follows the standard binomial distribution

$$\Pr(T = r) = \binom{T}{r} p_0^r (1-p_0)^{T-r}, \quad (1.11)$$

and (1.9) implies

$$\lim_{T \rightarrow \infty} \bar{p}_T = p_0 \quad (1.12)$$

with “probability one.” This last result, as a special case of de Finetti’s Theorem, is equivalent to the *strong law of large numbers*.

De Finetti’s representation theorem has been generalized by seeking more stringent forms of “symmetry” than simple exchangeability, in the process rationalizing sampling models other than the binomial [see Bernardo and Smith (1994, Chapter 4)]. Although these theorems do not hold exactly for infinite sequences, they hold approximately for sufficiently large finite sequences.

The pragmatic value of de Finetti’s theorem depends on whether it is easier to assess the left-hand side of (1.8), which involves only observable quantities, or instead, the integrand on the right-hand side of (1.8), which involves two distributions and the mathematical fiction . Most statisticians think in terms of the right-hand side. Frequentists implicitly do so with a degenerate distribution for that in effect treats as a constant, and Bayesians do so with a nondegenerate “prior” distribution for . What is important to note here, however, is the isomorphism de Finetti’s theorem suggests between two worlds, one involving only observables and the other involving the parameter . De Finetti put parameters in their proper perspective: (i) They are mathematical constructs that provide a convenient index for a probability distribution, and (ii) they induce conditional independence for a sequence of observables.

Exercise 1.1 (Let’s make a deal) Consider the television game show “Let’s Make a Deal” in which host Monty Hall asks contestants to choose the prize behind one of three curtains. Behind one curtain lies the grand prize; the other two curtains conceal only relatively small gifts. Assume Monty knows what is behind every curtain. Once the contestant has made a choice, Monty Hall reveals what is behind one of the two curtains that were not chosen. Having been shown one of the lesser prizes, the contestant is offered a chance to switch curtains. Should the contestant switch?

Solution

Let denote which curtain hides the grand prize. Let \hat{C} denote the curtain the contestant chooses first, and let C denote the curtain Monty shows the contestant. Assume $\Pr(C = i) = 1/3$, $i = 1, 2, 3$, $\Pr(\hat{C} = i | C = j) = 1/3$, $i = 1, 2, 3$, and that C and \hat{C} are independent. Without loss of generality, suppose $i = 1$ and $j = 2$. Then use Bayes’ theorem for events to compute the numerator and denominator of the following ratio:

$$\begin{aligned} \frac{\Pr(C = 3 | \hat{C} = 2, C = 1)}{\Pr(C = 1 | \hat{C} = 2, C = 1)} &= \frac{\Pr(M=2, \hat{C}=1 | C=3) \Pr(C=3)}{\Pr(M=2, \hat{C}=1) \Pr(C=1)} \\ &= \frac{\Pr(C = 3 | \hat{C} = 1, C = 2)}{\Pr(C = 1 | \hat{C} = 1, C = 2)} \\ &= \frac{\Pr(C = 2 | \hat{C} = 1, C = 3) \Pr(\hat{C} = 1 | C = 3)}{\Pr(C = 2 | \hat{C} = 1, C = 1) \Pr(\hat{C} = 1 | C = 1)} \\ &= \frac{\Pr(C = 2 | \hat{C} = 1, C = 3)}{\Pr(C = 2 | \hat{C} = 1, C = 1)} \end{aligned} \tag{1.13}$$

The numerator of the last line of (1.13) is one because Monty has no choice but to choose $\hat{c} = 2$ when $\hat{c} = 1$ and $c = 3$. The denominator of (1.13), however, is ambiguous because when $\hat{c} = 1$ and $c = 1$, Monty can choose either $\hat{c} = 2$ or $\hat{c} = 3$. The problem formulation does not contain information on Monty's choice procedure in this case. But since this probability must be less than or equal to one, ratio (1.13) can never be less than one. Unless $\Pr(\hat{c} = 2 | \hat{c} = 1, c = 1) = 1$, the contestant is better off switching curtains. If $\Pr(\hat{c} = 2 | \hat{c} = 1, c = 1) = \Pr(\hat{c} = 3 | \hat{c} = 1, c = 1) = 1/2$, then the contestant doubles the probability of winning the grand prize by switching.

Exercise 1.2 (Making Dutch book) Consider a horse race involving N horses. Suppose a bettor's beliefs are such that he believes the probability of horse i winning is p_i , where $p_1 + p_2 + \dots + p_N = 1$. Show how to make Dutch book with such an individual.

Solution

Consider a bet with this person of $1/p_i$ dollars that pays one dollar if horse i wins, and place such a bet on each of the N horses. Then you are guaranteed winning one dollar (since one of the horses has to win) and earning a profit of $1 - (p_1 + p_2 + \dots + p_N) = 0$.

Exercise 1.3 (Independence and exchangeability) Suppose $\pi = [\pi_1 \ \pi_2 \ \dots \ \pi_T]'$ $\sim (0_T, \Sigma)$, where $\Sigma = (1 - \alpha) I_T + \alpha \mathbf{1} \mathbf{1}'$ is positive definite for some scalar α and $\mathbf{1}$ is a $T \times 1$ vector with each element equal to unity. Let $(\pi_t) \ (t = 1, 2, \dots, T)$ be a permutation of $\{\pi_1, \pi_2, \dots, \pi_T\}$ and suppose $[\pi(\pi_1), \pi(\pi_2), \dots, \pi(\pi_T)] = \pi$, where π is a $T \times T$ selection matrix such that, for $t = 1, 2, \dots, T$, row t in π consists of all zeros except column $\pi(t)$, which is unity. Show that these beliefs are exchangeable.

Solution

Note that $\pi' = \pi$ and $\pi = \pi$. Then, $\pi \sim (0_T, \Omega)$, where

$$\begin{aligned} \Omega &= \Sigma \pi' \\ &= [(1 - \alpha) I_T + \alpha \mathbf{1} \mathbf{1}'] \pi' \\ &= (1 - \alpha) \pi' + \alpha \mathbf{1} \mathbf{1}' \pi' \\ &= (1 - \alpha) \pi + \alpha \mathbf{1} \mathbf{1}' \\ &= \Sigma \end{aligned}$$

Hence, beliefs regarding $\pi_t \ (t = 1, 2, \dots, T)$ are exchangeable. Despite this exchangeability, it is interesting to note that if $\alpha \neq 0$, $\pi_t \ (t = 1, 2, \dots, T)$ are *not independent*.

Exercise 1.4 (Predicted probability of success of a Bernoulli random variable) Suppose a researcher makes a coherent probability assignment to an infinite sequence $\pi_t \ (t = 1, 2, 3, \dots)$ of exchangeable Bernoulli random variables. Given an observed sequence of T trials with s successes, find the probability that the next outcome, π_{T+1} , is $T+1$.

Solution

Applying the definition of conditional probability and then Theorem 1.2 to both the numerator and denominator yields

$$\begin{aligned}
 \Pr(T_{+1} = t_{+1} | T = t) &= \frac{\Pr(T = t, T_{+1} = t_{+1})}{\Pr(T = t)} \\
 &= \frac{\int_0^1 (r+y_{T+1})(1-r)^{(T+1-r-y_{T+1})} (r)^t (1-r)^{1-t} dr}{\int_0^1 r(1-r)^{(T-r)} (r)^t (1-r)^{1-t} dr} \\
 &= \frac{\int_0^1 (y_{T+1})(1-r)^{(1-Y_{T+1})} (r)^t (1-r)^{1-t} dr}{\int_0^1 (r)^t (1-r)^{1-t} dr} \\
 &= \int_0^1 (y_{T+1})(1-r)^{(1-y_{T+1})} (r)^t (1-r)^{1-t} dr,
 \end{aligned} \tag{1.14}$$

where

$$(r | t) = \frac{(r)^t (1-r)^{1-t}}{(r)^t (1-r)^{1-t}} \tag{1.15}$$

Therefore $\Pr(T_{+1} = t_{+1} | T = t)$ is simply

$$(r | t) \text{ if } T_{+1} = 1,$$

or

$$1 - (r | t) \text{ if } T_{+1} = 0$$

The simplicity of this exercise hides its importance because it demonstrates most of the essential operations that characterize the Bayesian approach to statistics. First, the existence of the density $(r | t)$ is a result of Theorem 1.2, *not* an assumption. Second, the updating of prior beliefs captured in (1.15) amounts to nothing more than Bayes' theorem. Third, although t ($= 1, 2, \dots$) are independent conditional on T , unconditional on T they are dependent. Finally, the parameter r is merely a mathematical entity indexing the integration in (1.14). Its "real-world existence" is a question only of metaphysical importance.

Exercise 1.5 (Independence and conditional independence) Consider three events i ($= 1, 2, 3$), where $\Pr(i) = p_i$, $i = 1, 2, 3$. Show that the following statements are totally unrelated: (a) i_1 and i_2 are independent and (b) i_1 and i_2 are conditionally independent given i_3 .

Solution

There are $2^3 = 8$ possible three-element strings that can occur when considering i ($= 1, 2, 3$) and their complements \bar{i} ($= 1, 2, 3$). This leaves assessment of $7 = 8 - 1$ probabilities since the eighth is determined by the adding-up condition. These can be assessed in terms of the following probabilities: $\Pr(i_1 \cap i_2) = p_{12}$, $\Pr(i_1 \cap i_3) = p_{13}$,

$\Pr(2 \cap 3) = p_{23}$, and $\Pr(1 \cap 2 \cap 3) = p_{123}$. Independence of 1 and 2 places a restriction on $\Pr(1 \cap 2)$, namely $p_{12} = p_1 p_2$. Conditional independence places a restriction on the remaining probabilities p_{13} , p_{23} , p_3 , and p_{123} . To see this note $\Pr(1 \cap 2 | 3) = p_{123} / p_3$ by simply expressing the conditional as the joint divided by the marginal, and conditional independence implies $\Pr(1 \cap 2 | 3) = \Pr(1 | 3) \Pr(2 | 3) = (p_{13} / p_3)(p_{23} / p_3)$. Putting these equalities together implies $p_{123} = p_{13} p_{23} p_3$. Note that the restrictions implied by independence and conditional independence share no common probabilities.

2

Bayesian inference

In this chapter we extend Chapter 1 to cover the case of random variables. By *Bayesian inference* we mean the updating of prior beliefs into posterior beliefs conditional on observed data. This chapter covers a variety of standard sampling situations in which prior beliefs are sufficiently regular that the updating can proceed in a fairly mechanical fashion. Details of point estimation, interval estimation, hypothesis testing, and prediction are covered in subsequent chapters. We remind the reader that the definitions of many common distributions are provided in the Appendix to this book. Further details on the underlying probability theory are available in Chapters 1 and 2 of Poirier (1995).

One of the appealing things about Bayesian analysis is that it requires only a few general principles that are applied over and over again in different settings. Bayesians begin by writing down a joint distribution of all quantities under consideration (except known constants). Quantities to become known under sampling are denoted by the T -dimensional vector y , and remaining unknown quantities by the K -dimensional vector $\theta \in \Theta \subseteq \mathcal{R}^K$. Unless noted otherwise, we treat θ as a continuous random variable. Working in terms of densities, consider

$$p(y, \theta) = p(\theta)p(y|\theta) = p(y)p(\theta|y), \quad (2.1)$$

where $p(\theta)$ is the *prior density* and $p(\theta|y)$ is the *posterior density*. Viewing $p(y|\theta)$ as a function of θ for known y , any function proportional to it is referred to as a *likelihood function*. We will denote the likelihood function as $L(\theta)$. Unless noted otherwise, we will work with $L(\theta) = p(y|\theta)$ and thus include the integrating constant for $y|\theta$ in our description of the likelihood. We also note that

$$p(y) = \int_{\Theta} p(\theta)L(\theta)d\theta \quad (2.2)$$

is the *marginal density of the observed data* (also known as the *marginal likelihood*).

From (2.1) Bayes' *theorem for densities* follows immediately:

$$p(\theta|y) = \frac{p(\theta)L(\theta)}{p(y)} \propto p(\theta)L(\theta). \quad (2.3)$$

The shape of the posterior can be learned by plotting the right-hand side of (2.3) when $k = 1$ or 2 . Obtaining moments or quantiles, however, requires the integrating constant (2.2). Fortunately, in some situations the integration in (2.2) can be performed analytically, in which case the updating of prior beliefs $p(\theta)$ in light of the data y to obtain the posterior beliefs $p(\theta|y)$ is straightforward. These situations correspond to cases where $p(\theta)$ and $L(\theta)$ belong to the *exponential family* of densities (see Exercise 2.13). In this case the prior density can be chosen so that the posterior density falls within the same elementary family of distributions as the prior. These families are called *conjugate families*.

The denominator in (2.3) serves as a factor of proportionality (not involving θ) that ensures that the posterior density integrates to unity. To simplify much of the analysis that follows, we calculate a posterior density by dropping all factors of proportionality from the prior density and the likelihood function, concentrating attention on the resulting posterior kernel, and then compute the required posterior integrating constant at the end. This works particularly well when using easily recognized conjugate families. Note also that this implies that when considering experiments employing the same prior, and that yield proportional likelihoods for the observed data, identical posteriors will emerge. This reflects the important fact that Bayesian inference is consistent with the *likelihood principle* [see Berger and Wolpert (1988)].

In most practical situations not all elements of θ are of direct interest. Let $\theta_1 \in \Theta_1$, $\theta_2 \in \Theta_2$, and $\theta = [\theta_1, \theta_2] \in \Theta_1 \times \Theta_2$ be partitioned into *parameters of interest* θ_1 and *nuisance parameters* θ_2 not of direct interest. For example, θ_1 may be the mean and θ_2 the variance of some sampling distribution. Nuisance parameters are well named for frequentists, because dealing with them in a general setting is one of the major problems frequentist researchers face. In contrast, Bayesians have a universal approach to eliminating nuisance parameters from the problem: They are integrated out of the posterior density, yielding the marginal posterior density for the parameters of interest, that is,

$$p(\theta_1|y) = \int_{\Theta_2} p(\theta_1, \theta_2|y) d\theta_2, \quad \theta_1 \in \Theta_1. \quad (2.4)$$

Many of the following exercises involve particular distributions. The Appendix of this book contains definitions and properties of many common distributions. It is worth noting that there are two common parameterizations of the gamma distribution and we use both in this chapter (see Appendix Definition 2).

Exercise 2.1 (Conjugate Bernoulli analysis) Given the parameter θ , where $0 < \theta < 1$, consider T iid Bernoulli random variables Y_t ($t = 1, 2, \dots, T$), each with p.m.f.

$$p(y_t|\theta) = \begin{cases} \theta & \text{if } y_t = 1, \\ 1 - \theta & \text{if } y_t = 0. \end{cases} \quad (2.5)$$

The likelihood function is

$$L(\theta) = \theta^m (1 - \theta)^{T-m}, \quad (2.6)$$

where $m = T\bar{y}$ is the number of successes (i.e., $y_t = 1$) in T trials. Suppose prior beliefs concerning θ are represented by a beta distribution with p.d.f.

$$p_B(\theta|\underline{\alpha}, \underline{\delta}) = [\mathcal{B}(\underline{\alpha}, \underline{\delta})]^{-1} \theta^{\underline{\alpha}-1} (1 - \theta)^{\underline{\delta}-1}, \quad 0 < \theta < 1, \quad (2.7)$$

where $\underline{\alpha} > 0$ and $\underline{\delta} > 0$ are known, and $\mathcal{B}(\underline{\alpha}, \underline{\delta}) = \Gamma(\underline{\alpha})\Gamma(\underline{\delta})/\Gamma(\underline{\alpha} + \underline{\delta})$ is the beta function defined in terms of the gamma function $\Gamma(\alpha) = \int_0^\infty t^{\alpha-1} \exp(-t) dt$. This class of priors can represent a wide range of prior opinions. Find the posterior density of θ .

Solution

The denominator (2.2) of posterior (2.3) density is easy to compute. Define

$$\bar{\alpha} = \underline{\alpha} + m, \quad (2.8)$$

$$\bar{\delta} = \underline{\delta} + T - m, \quad (2.9)$$

and consider

$$\begin{aligned} p(y) &= \int_0^1 [\mathcal{B}(\underline{\alpha}, \underline{\delta})]^{-1} \theta^{\underline{\alpha}-1} (1 - \theta)^{\underline{\delta}-1} \theta^m (1 - \theta)^{T-m} d\theta \\ &= \left[\frac{\mathcal{B}(\bar{\alpha}, \bar{\delta})}{\mathcal{B}(\underline{\alpha}, \underline{\delta})} \right] \int_0^1 [\mathcal{B}(\bar{\alpha}, \bar{\delta})]^{-1} \theta^{\bar{\alpha}-1} (1 - \theta)^{\bar{\delta}-1} d\theta \\ &= \left[\frac{\mathcal{B}(\bar{\alpha}, \bar{\delta})}{\mathcal{B}(\underline{\alpha}, \underline{\delta})} \right], \end{aligned} \quad (2.10)$$

where the integral in (2.10) equals unity because the integrand is a beta p.d.f. for θ . From (2.3) and (2.8)–(2.10) it follows that the posterior density of θ is

$$\begin{aligned} p(\theta|y) &= \frac{[\mathcal{B}(\underline{\alpha}, \underline{\delta})]^{-1} \theta^{\underline{\alpha}-1} (1 - \theta)^{\underline{\delta}-1} \theta^m (1 - \theta)^{T-m}}{\mathcal{B}(\bar{\alpha}, \bar{\delta})/\mathcal{B}(\underline{\alpha}, \underline{\delta})} \\ &= [\mathcal{B}(\bar{\alpha}, \bar{\delta})]^{-1} \theta^{\bar{\alpha}-1} (1 - \theta)^{\bar{\delta}-1}, \quad 0 < \theta < 1. \end{aligned} \quad (2.11)$$

Therefore, because posterior density (2.11) is itself a beta p.d.f. with parameters $\bar{\alpha}$ and $\bar{\delta}$ given by (2.8) and (2.9), it follows that the conjugate family of prior distributions for a Bernoulli likelihood is the beta family of p.d.f.s

Exercise 2.2 (Application of Bayes' theorem) A laboratory blood test is 95 percent effective in detecting a certain disease when it is, in fact, present. However, the test also yields a “false positive” result for one percent of the healthy people tested. If 0.1 percent of the population actually has the disease, what is the probability that a person has the disease given that her test result is positive?

Solution

Let D denote the presence of the disease, D^c denote its absence, and $+$ denote a positive test result. Then $\Pr(+|D) = .95$, $\Pr(+|D^c) = .01$, and $P(D) = .001$. Then according to Bayes' theorem

$$P(D|+) = \frac{P(D)P(+|D)}{P(+)} = \frac{.001(.95)}{.001(.95) + .999(.01)} = .0868.$$

Exercise 2.3 (Conjugate normal analysis with unknown mean and known variance)

Given $\theta = [\theta_1 \ \theta_2]' \in \mathcal{R} \times \mathcal{R}^+$, consider a random sample Y_t ($t = 1, 2, \dots, T$) from a $N(\theta_1, \theta_2^{-1})$ population. For reasons that will become clear as we proceed, it is convenient to work in terms of θ_2 , the reciprocal of the variance (called the *precision*). (In later exercises, however, particularly those in the computational chapters, we will work directly with the error variance. Both approaches are commonly employed in the literature).

Assume θ_2 is known. Suppose prior beliefs for θ_1 are represented by the normal distribution

$$\theta_1 \sim N(\underline{\mu}, \underline{h}^{-1}), \quad (2.12)$$

where $\underline{\mu}$ and $\underline{h} > 0$ are given. Find the posterior density of θ_1 and marginal likelihood $p(y)$.

Solution

For later notational convenience, let

$$h = [\theta_2^{-1}/T]^{-1} = T\theta_2, \quad (2.13)$$

$$\bar{h} = \underline{h} + h, \quad (2.14)$$

and

$$\bar{\mu} = \bar{h}^{-1}(\underline{h}\underline{\mu} + h\bar{y}). \quad (2.15)$$

It is useful to employ two identities. The first identity is

$$\sum_{t=1}^T (y_t - \theta_1)^2 = \sum_{t=1}^T (y_t - \bar{y})^2 + T(\bar{y} - \theta_1)^2 = \nu s^2 + T(\bar{y} - \theta_1)^2, \quad (2.16)$$

for all θ_1 , where

$$\nu = T - 1 \quad (2.17)$$

and

$$s^2 = \nu^{-1} \sum_{t=1}^T (y_t - \bar{y})^2. \quad (2.18)$$

The second identity is

$$\underline{h}(\theta_1 - \underline{\mu})^2 + h(\bar{y} - \theta_1)^2 = \bar{h}(\theta_1 - \bar{\mu})^2 + (\underline{h}^{-1} + h^{-1})^{-1}(\bar{y} - \underline{\mu})^2, \quad (2.19)$$

for all θ_1 , \underline{h} , and h .

Now we apply Bayes' theorem to find the posterior density of θ_1 . Using identities (2.13) and (2.16), we write the likelihood function as

$$\begin{aligned}
 L(\theta_1) &= \prod_{t=1}^T \phi(y_t | \theta_1, \theta_2^{-1}) \\
 &= (2\pi\theta_2^{-1})^{-T/2} \exp\left(-\frac{\theta_2}{2} \sum_{t=1}^T (y_t - \theta_1)^2\right) \\
 &= (2\pi\theta_2^{-1})^{-T/2} \exp\left(-\frac{h}{2T} [\nu s^2 + T(\bar{y} - \theta_1)^2]\right) \\
 &= c_1 \phi(\bar{y} | \theta_1, h^{-1}),
 \end{aligned} \tag{2.20}$$

where

$$c_1 = (2\pi)^{-\nu/2} T^{(-1/2)} \theta_2^{\nu/2} \exp\left(-\frac{1}{2} \theta_2 \nu s^2\right) \tag{2.21}$$

does *not* depend on θ_1 . Note that the factorization in (2.20) demonstrates that \bar{y} is a sufficient statistic for θ_1 . Also note that density $\phi(\bar{y} | \theta_1, h^{-1})$ corresponds to the sampling density of the sample mean, given θ_1 .

Using identity (2.19) and factorization (2.20), the numerator of (2.3) is

$$\begin{aligned}
 p(\theta_1) L(\theta_1) &= \phi(\theta_1 | \underline{\mu}, \underline{h}^{-1}) c_1 \phi(\bar{y} | \theta_1, h^{-1}) \\
 &= c_1 (2\pi \underline{h}^{-1})^{-1/2} \exp\left(-\frac{1}{2} [\underline{h}(\theta_1 - \underline{\mu})^2 + h(\bar{y} - \theta_1)^2]\right) \\
 &= c_1 (2\pi \underline{h}^{-1})^{-1/2} \exp\left(-\frac{1}{2} [\bar{h}(\theta_1 - \bar{\mu})^2 + (\underline{h}^{-1} + h^{-1})^{-1}(\bar{y} - \underline{\mu})^2]\right) \\
 &= c_1 (2\pi)^{(1/2)} [\underline{h} \bar{h}^{-1} (\underline{h}^{-1} + h^{-1})]^{1/2} \phi(\bar{y} | \underline{\mu}, \underline{h}^{-1} + h^{-1}) \phi(\theta_1 | \bar{\mu}, \bar{h}^{-1}).
 \end{aligned} \tag{2.22}$$

Bayes' theorem tells us that the posterior distribution of θ_1 is proportional to (2.22). Thus, all terms not involving θ_1 that enter (2.22) multiplicatively are absorbed in the normalizing constant, allowing us to focus on the posterior kernel. When looking at the final line of (2.22), we immediately see that

$$p(\theta_1 | y) = \phi\left(\theta_1 | \bar{\mu}, \bar{h}^{-1}\right). \tag{2.23}$$

The interpretations of quantities (2.14) and (2.15) are now clear from (2.23): They are the posterior precision and posterior mean, respectively. Note that it is the additivity of precisions in (2.14) that motivates working with precisions rather than variances. Because posterior density (2.23) and prior density (2.12) are both members of the normal family, it follows that the conjugate prior for the case of random sampling from a normal population with *known* variance is itself a normal density.

The marginal density of the data [i.e., the denominator of (2.3)] can be obtained from (2.22):

$$\begin{aligned}
 p(y) &= \int_{-\infty}^{\infty} p(\theta_1) L(\theta_1) d\theta_1 \\
 &= c_1 (2\pi)^{(1/2)} [\underline{h} \bar{h}^{-1} (\underline{h}^{-1} + h^{-1})]^{1/2} \phi(\bar{y} | \underline{\mu}, \underline{h}^{-1} + h^{-1}) \int_{-\infty}^{\infty} \phi(\theta_1 | \bar{\mu}, \bar{h}^{-1}) d\theta_1 \\
 &= c_1 (2\pi)^{(1/2)} [\underline{h} \bar{h}^{-1} (\underline{h}^{-1} + h^{-1})]^{1/2} \phi(\bar{y} | \underline{\mu}, \underline{h}^{-1} + h^{-1}).
 \end{aligned} \tag{2.24}$$

Alternatively, we can represent the regression model in the following way. The sampling assumptions regarding y_t imply $y = \iota_T \theta_1 + \epsilon$, where $y = [y_1 \ y_2 \ \cdots \ y_T]'$, ι_T is a $T \times 1$ vector of ones, and $\epsilon \sim N(0, \theta_2^{-1} I_T)$. Similarly, the prior for θ_1 in (2.12) can be written as $\theta_1 = \underline{\mu} + \eta$, where $\eta \sim N(0, \underline{h}^{-1})$. Substituting the equation for our prior into the equation for y gives $y = \iota_T \underline{\mu} + \epsilon + \iota_T \eta$, which implies $y \sim N(\iota_T \underline{\mu}, [\theta_2^{-1} I_T + \underline{h}^{-1} \iota_T \iota_T'])$, or equivalently,

$$p(y) = \phi(y | \iota_T \underline{\mu}, [\theta_2^{-1} I_T + \underline{h}^{-1} \iota_T \iota_T']). \tag{2.25}$$

At first glance, it is not at all obvious that the expressions in (2.24) and (2.25) are in agreement, though if we have done our derivations correctly, they must agree. To show that these expressions are, in fact, the same, we expand (2.25). To fix ideas, we focus on reproducing the portion of (2.24) involving the normal density for \bar{y} . The remaining terms in (2.24) will also fall out of this derivation, though we suppress them here for clarity.

The kernel of (2.25) involves a term of the form

$$\exp \left[-[1/2] (y - \iota_T \underline{\mu})' [\theta_2^{-1} I_T + \underline{h}^{-1} \iota_T \iota_T']^{-1} (y - \iota_T \underline{\mu}) \right].$$

To simplify this quadratic form, we use the result

$$\begin{aligned}
 [\theta_2^{-1} I_T + \underline{h}^{-1} \iota_T \iota_T']^{-1} &= \theta_2 \left[I_T - [\underline{h}^{-1} / (\theta_2^{-1} + T \underline{h}^{-1})] \iota_T \iota_T' \right] \\
 &= h T^{-1} \left[I_T - [\underline{h}^{-1} / (T h^{-1} + T \underline{h}^{-1})] \iota_T \iota_T' \right] \\
 &= h T^{-1} \left[I_T - (\beta / T) \iota_T \iota_T' \right],
 \end{aligned}$$

where $\beta \equiv \underline{h}^{-1} / (\underline{h}^{-1} + h^{-1})$, and, again, we have made use of the definition of h in (2.13). [The expression for this inverse is actually a commonly used formula in longitudinal data models with random effects, as the covariance matrix in such models has an identical structure to the covariance matrix in (2.25)].

Using this result, we can write the quadratic form in our exponential kernel as

$$\exp \left[-[1/2] (h T^{-1}) (y - \iota_T \underline{\mu})' [I_T - (\beta / T) \iota_T \iota_T'] (y - \iota_T \underline{\mu}) \right].$$

With a little work, one can show that this quadratic form can be expressed as

$$(y - \iota_T \underline{\mu})' [I_T - (\beta / T) \iota_T \iota_T'] (y - \iota_T \underline{\mu}) = \sum_t (y_t - \underline{\mu})^2 - (\beta / T) \left[\sum_t (y_t - \underline{\mu}) \right]^2.$$

Following (2.16) [and in the spirit of getting (2.25) to look like (2.24)], we can also write

$$\sum_t (y_t - \underline{\mu})^2 = \nu s^2 + T(\bar{y} - \underline{\mu})^2,$$

and similarly,

$$\sum_t (y_t - \underline{\mu}) = T(\bar{y} - \underline{\mu}).$$

Putting these results together gives

$$(y - \iota_T \underline{\mu})' [I_T - (\beta/T) \iota_T \iota_T'] (y - \iota_T \underline{\mu}) = \nu s^2 + T(\bar{y} - \underline{\mu})^2 - \beta T(\bar{y} - \underline{\mu})^2.$$

The kernel of the density in (2.25) can thus be expressed as

$$\begin{aligned} & \exp \left(-(1/2) [T^{-1} h] [\nu s^2 + T(1 - \beta)(\bar{y} - \underline{\mu})^2] \right) \\ &= \exp \left(-(1/2) \theta_2 \nu s^2 \right) \exp \left(-(1/2) h(1 - \beta)(\bar{y} - \underline{\mu})^2 \right). \end{aligned}$$

Note that the first term on the right-hand side of this equation reproduces the corresponding term in the definition of c_1 in (2.21). As for the last term, by using our definition of β , it follows that

$$h(1 - \beta) = h \left(\frac{h^{-1} + \underline{h}^{-1}}{h^{-1} + \underline{h}^{-1}} - \frac{\underline{h}^{-1}}{h^{-1} + \underline{h}^{-1}} \right) = (h^{-1} + \underline{h}^{-1})^{-1}.$$

Thus, we are indeed left with a term of the form $\phi(\bar{y} | \underline{\mu}, h^{-1} + \underline{h}^{-1})$, as in (2.24). Including the normalizing constant of (2.25) into this derivation will show that (2.24) and (2.25) are identical.

Exercise 2.4 (Conjugate normal analysis with unknown mean and variance) Consider Exercise 2.3, but with the precision θ_2 also unknown. Suppose the joint prior distribution for $\theta = [\theta_1 \ \theta_2]'$ is the normal-gamma distribution, denoted $\theta \sim NG(\underline{\mu}, \underline{q}, \underline{s}^{-2}, \underline{\nu})$ [see Appendix Definition 5 for details] with density

$$\begin{aligned} f_{NG}(\theta | \underline{\mu}, \underline{q}, \underline{s}^{-2}, \underline{\nu}) &= \phi(\theta_1 | \underline{\mu}, \theta_2^{-1} \underline{q}) f_\gamma(\theta_2 | \underline{s}^{-2}, \underline{\nu}) \\ &= \left((2\pi\theta_2^{-1})^{-1/2} \exp \left[-\frac{1}{2} \theta_2 \underline{q}^{-1} (\theta_1 - \underline{\mu})^2 \right] \right) \\ &\quad \times \left[\left(\left[\frac{2}{\underline{\nu} \underline{s}^2} \right]^{\underline{\nu}/2} \Gamma(\underline{\nu}/2) \right)^{-1} \theta_2^{(\underline{\nu}-2)/2} \exp \left(-\frac{1}{2} \theta_2 \underline{\nu} \underline{s}^2 \right) \right] \\ &\propto \theta_2^{(\underline{\nu}-1)/2} \exp \left(-\frac{1}{2} \theta_2 [\underline{\nu} \underline{s}^2 + \underline{q}^{-1} (\theta_1 - \underline{\mu})^2] \right), \end{aligned} \tag{2.26}$$

where $\underline{\mu} \in \mathcal{R}$, \underline{q} , \underline{s}^2 , and $\underline{\nu}$ are known positive constants, and “ \times ” has been used to denote multiplication in multiline equations. Find the posterior density of θ .

Solution

Dropping irrelevant constants not depending on θ_1 and θ_2 , (2.3) implies posterior density $p(\theta|y)$ is proportional to the product of the kernels (2.26) and (2.20):

$$\begin{aligned}
 p(\theta|y) &\propto p(\theta)L(\theta) \\
 &\propto [\phi(\theta_1|\underline{\mu}, \theta_2^{-1}\underline{q})f_\gamma(\theta_2|\underline{s}^{-2}, \underline{\nu})] [c_1(\theta_2)\phi(\bar{y}|\theta_1, h^{-1})] \\
 &\propto \theta_2^{(\underline{\nu}-1)/2} \exp\left(-\frac{\theta_2}{2}[\underline{\nu}\underline{s}^2 + \underline{q}^{-1}(\theta_1 - \underline{\mu})^2]\right) \theta_2^{T/2} \exp\left(-\frac{\theta_2}{2}[\underline{\nu}s^2 + T(\bar{y} - \theta_1)^2]\right) \\
 &\propto \theta_2^{(\bar{\nu}-1)/2} \exp\left(-\frac{\theta_2}{2}[\underline{\nu}\underline{s}^2 + \underline{\nu}s^2 + \underline{q}^{-1}(\theta_1 - \underline{\mu})^2 + T(\bar{y} - \theta_1)^2]\right),
 \end{aligned} \tag{2.27}$$

where

$$\bar{\nu} = \underline{\nu} + T. \tag{2.28}$$

Using identity (2.19) with $\underline{h} = \underline{q}^{-1}$ and $h = T$, we can write the last two terms in the square brackets in (2.27) as

$$\underline{q}^{-1}(\theta_1 - \underline{\mu})^2 + T(\bar{y} - \theta_1)^2 = \bar{q}^{-1}(\theta_1 - \bar{\mu})^2 + (\underline{q} + T^{-1})^{-1}(\bar{y} - \underline{\mu})^2, \tag{2.29}$$

where

$$\bar{q} = (\underline{q}^{-1} + T)^{-1} \tag{2.30}$$

and

$$\bar{\mu} = \bar{q}(\underline{q}^{-1}\underline{\mu} + T\bar{y}). \tag{2.31}$$

Then by letting

$$\bar{s}^2 = [\bar{\nu}]^{-1} [\underline{\nu}\underline{s}^2 + \underline{\nu}s^2 + (\underline{q} + T^{-1})^{-1}(\bar{y} - \underline{\mu})^2] \tag{2.32}$$

$$= [\bar{\nu}]^{-1} [\underline{\nu}\underline{s}^2 + \underline{\nu}s^2 + \underline{q}^{-1}\bar{q}T(\bar{y} - \underline{\mu})^2], \tag{2.33}$$

it follows from (2.27)–(2.33) that posterior density (2.27) can be written

$$p(\theta|y) \propto \theta_2^{(\bar{\nu}-1)/2} \exp\left(-\frac{\theta_2}{2}[\bar{q}^{-1}(\theta_1 - \bar{\mu})^2 + \bar{\nu}\bar{s}^2]\right) \tag{2.34}$$

$$\propto \theta_2^{1/2} \exp\left[-\frac{\theta_2}{2}\bar{q}^{-1}(\theta_1 - \bar{\mu})^2\right] \theta_2^{(\bar{\nu}-2)/2} \exp\left[-\frac{\theta_2}{2}\bar{\nu}\bar{s}^2\right]. \tag{2.35}$$

Comparing (2.26) and (2.35), it is seen that posterior density $p(\theta|y)$ corresponds to the kernel of a $NG(\bar{\mu}, \bar{q}, \bar{s}^{-2}, \bar{\nu})$ distribution, with updating formulas (2.28)–(2.33). Because both (2.26) and (2.35) are proportional to normal-gamma densities, prior density (2.26) is the natural conjugate prior for random sampling from a normal population with unknown mean and variance.

Exercise 2.5 (Properties of a normal-gamma distribution) Consider the normal-gamma prior density in Exercise 2.4 for θ_1 and θ_2 with hyperparameters $\underline{\mu}$, \underline{q} , \underline{s}^{-2} , and $\underline{\nu} > 2$.

- (a) Are θ_1 and θ_2 independent?
- (b) Is θ_1 mean independent of θ_2 (i.e., is the mean of θ_1 independent of θ_2)?
- (c) Are θ_1 and θ_2 uncorrelated?

Solution

- (a) No, because $\theta_1|\theta_2 \sim N(\underline{\mu}, q/\theta_2)$ clearly depends on θ_2 .
- (b) Yes, because $E(\theta_1|\theta_2) = \underline{\mu}$ does not depend on θ_2 .
- (c) Yes, because mean independence implies zero covariance.

Exercise 2.6 (Fictitious sample interpretation of natural conjugate priors) Natural conjugate priors have the desirable feature that prior information can be viewed as “fictitious sample information” in that it is combined with the sample in exactly the same way that additional sample information would be combined. The only difference is that the prior information is “observed” in the mind of the researcher, not in the real world. To clarify this point, reconsider Exercises 2.1, 2.3, and 2.4.

- (a) Using the setup of Exercise 2.1, show that a beta prior distribution with parameters $\underline{\alpha}$ and $\underline{\delta}$ can be interpreted as the information contained in a sample of size $\underline{T} = \underline{\alpha} + \underline{\delta} - 2$ with $\underline{\alpha} - 1$ successes from the Bernoulli process of interest. (Of course if $\underline{\alpha}$ and $\underline{\delta}$ are not integers, then this interpretation must be loosely made.) This does not mean that the researcher has actually observed such a sample, but rather only that the researcher imagines his or her prior information as being from such a sample.
- (b) Using the setup of Exercise 2.3 (known variance) show that a normal prior distribution (2.12) can be interpreted in terms of equivalent sample information.

Solution

- (a) Such an interpretation follows directly from the form of the Bernoulli likelihood in (2.6). In this case, the likelihood for $\underline{\alpha} - 1$ successes in \underline{T} trials is

$$\begin{aligned} L(\theta) &= \theta^{\underline{\alpha}-1} (1 - \theta)^{\underline{T} - (\underline{\alpha}-1)} \\ &= \theta^{\underline{\alpha}-1} (1 - \theta)^{\underline{\delta}-1}, \end{aligned}$$

using the definition of \underline{T} . This is proportional to the beta prior in (2.7), and thus the prior can be given this fictitious sample interpretation.

- (b) Define

$$\underline{T} \equiv \frac{\theta_2^{-1}}{\underline{h}^{-1}} = \frac{\underline{h}}{\theta_2}, \quad (2.36)$$

and write the prior variance as

$$\underline{h}^{-1} = \frac{\theta_2^{-1}}{\underline{T}}. \quad (2.37)$$

Now, consider the likelihood for θ_1 obtained from a sample of size \underline{T} from the population distribution $N(\theta_1, \theta_2^{-1})$ yielding a sample mean $\underline{\mu}$. With θ_2 known, from (2.20) the likelihood for this “sample” is proportional to

$$L(\theta_1) \propto \exp\left(-\frac{\theta_2 T}{2}(\theta_1 - \underline{\mu})^2\right),$$

which forms the kernel of a normal prior for θ_1 with mean $\underline{\mu}$ and variance $T^{-1}\theta_2^{-1} = \underline{h}^{-1}$, as in (2.12).

Similarly, using (2.14), we can define

$$\bar{T} \equiv \frac{\theta_2^{-1}}{\bar{h}^{-1}} = \frac{\bar{h}}{\theta_2} = \underline{T} + T, \quad (2.38)$$

and note that posterior mean (2.15) can be written

$$\bar{\mu} = \frac{\underline{T}\underline{\mu} + T\bar{y}}{\bar{T}}. \quad (2.39)$$

Given this interpretation, (2.38) and (2.39) can be viewed as formulas for pooling the information from the actual sample and the fictitious prior sample.

This updating can also be extended to the more general case of Exercise 2.4. Let

$$\underline{T} = \frac{1}{\underline{q}} \quad (2.40)$$

represent the prior sample size and let

$$\bar{T} = \frac{1}{\bar{q}} = \underline{T} + T \quad (2.41)$$

represent the “total” sample size. Then again posterior mean (2.31) can be written as the weighted average (2.39).

Exercise 2.7 (Conjugate normal analysis with known mean and unknown variance) Consider Exercise 2.3 but with the mean θ_1 known and the population precision $\theta_2 \equiv \sigma^{-2}$ unknown. Suppose a prior density of the form $\sigma^{-2} \sim \gamma(\underline{s}^{-2}, \underline{\nu})$ is employed. Find the posterior for σ^{-2} .

Solution

Using the identity in (2.16) gives the likelihood function

$$L(\sigma^{-2}) = (2\pi)^{-T/2}(\sigma^{-2})^{T/2} \exp\left(-\frac{1}{2}\sigma^{-2}[\nu s^2 + T(\bar{y} - \theta_1)^2]\right),$$

and the prior density is

$$p(\sigma^{-2}) \propto (\sigma^{-2})^{(\underline{\nu}-2)/2} \exp\left(-\frac{1}{2}\sigma^{-2}\underline{\nu}\underline{s}^2\right).$$

Using Bayes' theorem gives the posterior density

$$\begin{aligned} p(\sigma^{-2}|y) &\propto p(\sigma^{-2})L(\sigma^{-2}) \\ &\propto (\sigma^{-2})^{(\underline{\nu}+T-2)/2} \exp\left(-\frac{\sigma^{-2}}{2}[\underline{\nu}\underline{s}^2 + \nu s^2 + T(\bar{y} - \theta_1)^2]\right) \\ &\propto (\sigma^{-2})^{(\bar{\nu}-2)/2} \exp\left(-\frac{\sigma^{-2}}{2}\bar{\nu}\bar{s}^2\right). \end{aligned}$$

Therefore, $\sigma^{-2}|y \sim \gamma([\bar{s}]^{-2}, \bar{\nu})$, with $\bar{\nu}$ defined in (2.28) and $\bar{s}^2 = \bar{\nu}^{-1}[\underline{\nu}\underline{s}^2 + \nu s^2 + T(\bar{y} - \theta_1)^2]$.

Exercise 2.8 (Marginal distribution of θ_1 when $\theta_1, \theta_2 \sim NG$) Consider the normal-gamma prior density of (2.26). Find the marginal distribution of θ_1

Solution

The marginal p.d.f. of θ_1 is obtained by integrating out θ_2 from the joint p.d.f. (2.26). Note that the last line of (2.26) is the kernel of a $\gamma(a, b)$ distribution for θ_2 given all other quantities, where

$$\begin{aligned} a &= \frac{\underline{\nu} + 1}{\underline{\nu}\underline{s}^2} \left[1 + \frac{1}{\underline{\nu}} \left(\frac{(\theta_1 - \underline{\mu})^2}{\underline{s}^2 \underline{q}} \right) \right]^{-1}, \\ b &= \underline{\nu} + 1. \end{aligned}$$

Thus, integrating (2.26) with respect to θ_2 is proportional to the integrating constant of a $\gamma(a, b)$ distribution, namely

$$p(\theta_1 | \underline{\mu}, \underline{q}, \underline{s}^2, \underline{\nu}) \propto \Gamma(b/2) \left(\frac{2a}{b} \right)^{b/2} \propto \left[1 + \frac{1}{\underline{\nu}} \left(\frac{(\theta_1 - \underline{\mu})^2}{\underline{s}^2 \underline{q}} \right) \right]^{-(\underline{\nu}+1)/2}, \quad (2.42)$$

which is the kernel of a $t(\underline{\mu}, \underline{s}^2 \underline{q}, \underline{\nu})$ distribution.

Exercise 2.9 (Conditional distribution of $\theta_2|\theta_1$ when $\theta_1, \theta_2 \sim NG$) Consider the normal-gamma prior of (2.26). Find the conditional distribution of θ_2 given θ_1 .

Solution

Consider (2.26). Discarding any constants of proportionality (including factors involving θ_1) yields

$$p(\theta_2 | \theta_1, \underline{\mu}, \underline{q}, \underline{s}^2, \underline{\nu}) \propto \theta_2^{[(\underline{\nu}+1)-2]/2} \exp\left(-\frac{1}{2}\theta_2(\underline{\nu}+1) \left[\frac{\underline{\nu}\underline{s}^2 + \underline{q}^{-1}(\theta_1 - \underline{\mu})^2}{\underline{\nu}+1} \right]\right),$$

which is recognized as the kernel of a $\gamma[(\underline{\nu}+1)/[\underline{\nu}\underline{s}^2 + \underline{q}^{-1}(\theta_1 - \underline{\mu})^2], \underline{\nu}+1]$ distribution.

Exercise 2.10 (Conjugate exponential analysis) Suppose Y_t ($t = 1, 2, \dots, T$) is a random sample from an exponential distribution $f_{EXP}(y_t|\theta) = \theta \exp(-\theta y_t)$, which has mean θ^{-1} , and that the prior distribution of θ is the Gamma distribution $G(\underline{\alpha}, \underline{\beta})$, where $\underline{\alpha} > 0$ and $\underline{\beta} > 0$. (See Appendix Definition 2 for this parameterization of the gamma density.) Find the posterior distribution of θ .

Solution

The likelihood function is

$$L(\theta) = \prod_{t=1}^T f_{EXP}(y_t|\theta) = \prod_{t=1}^T \theta \exp(-\theta y_t) = \theta^T \exp(-T\bar{y}\theta).$$

Define $\bar{\alpha} = \underline{\alpha} + T$ and $\bar{\beta} = (\underline{\beta}^{-1} + T\bar{y})^{-1}$. Using (2.3) gives the posterior density

$$\begin{aligned} p(\theta|y) &\propto p_G(\theta|\underline{\alpha}, \underline{\beta}) L(\theta) \\ &\propto \theta^{\underline{\alpha}-1} \exp(-\theta/\underline{\beta}) \theta^T \exp(-T\bar{y}\theta) \\ &\propto \theta^{T+\underline{\alpha}-1} \exp[-(\underline{\beta}^{-1} + T\bar{y})\theta] \\ &= \theta^{\bar{\alpha}-1} \exp(-\theta/\bar{\beta}). \end{aligned}$$

Therefore, $\theta|y \sim G(\bar{\alpha}, \bar{\beta})$.

Exercise 2.11 (Conjugate uniform analysis) Suppose Y_t ($t = 1, 2, \dots, T$) is a random sample from a $U(0, \theta)$ distribution, where θ is unknown, and suppose the prior distribution for θ is Pareto with parameters $\underline{\gamma} > 0$ and $\underline{\lambda} > 0$ (see Appendix Definition 11). Find the posterior distribution of θ .

Solution

Let $\bar{\gamma} = \max\{\underline{\gamma}, y_1, y_2, \dots, y_T\}$ and $\bar{\lambda} = \underline{\lambda} + T$. The likelihood function is $L(\theta) = \theta^{-T}$ for $\theta \geq \max\{y_1, y_2, \dots, y_T\}$, and the prior density for θ is

$$f_{Pa}(\theta|\underline{\gamma}, \underline{\lambda}) = \frac{\lambda \underline{\gamma}^{\underline{\lambda}}}{\theta^{\underline{\lambda}+1}} I(\theta > \underline{\gamma}) \propto \theta^{-(\underline{\lambda}+1)} I(\theta > \underline{\gamma}).$$

Using (2.3) gives the posterior density

$$p(\theta|y) \propto \left(\frac{\lambda \underline{\gamma}^{\underline{\lambda}}}{\theta^{\underline{\lambda}+1}} \right) \theta^{-T} I(\theta > \bar{\gamma}) \propto \theta^{-(\bar{\lambda}+1)} I(\theta > \bar{\gamma}),$$

which forms the kernel of a Pareto density with parameters $\bar{\gamma}$ and $\bar{\lambda}$.

Exercise 2.12 (Conjugate poisson analysis) Suppose Y_t ($t = 1, 2, \dots, T$) is a random sample from a Poisson distribution (see Appendix Definition 8) with mean θ and that the prior distribution of θ is the gamma distribution $G(\underline{\alpha}, \underline{\beta})$, where $\underline{\alpha} > 0$ and $\underline{\beta} > 0$. Find the posterior distribution of θ .

Solution

The likelihood function is

$$L(\theta) = \prod_{t=1}^T f_{Po}(y_t|\theta) = \prod_{t=1}^T \frac{\theta^{y_t} \exp(-\theta)}{y_t!} = \frac{\theta^{T\bar{y}} \exp(-T\theta)}{\prod_{t=1}^T y_t!}.$$

Define $\bar{\alpha} = \underline{\alpha} + T\bar{y}$ and $\bar{\beta} = (\underline{\beta}^{-1} + T)^{-1}$. Using (2.3) gives the posterior density

$$\begin{aligned} p(\theta|y) &\propto f_G(\theta|\underline{\alpha}, \underline{\beta}) L(\theta) \\ &\propto \theta^{\underline{\alpha}-1} \exp(-\theta/\underline{\beta}) \theta^{T\bar{y}} \exp(-T\theta) \\ &\propto \theta^{[T\bar{y}+\underline{\alpha}-1]} \exp[-(\underline{\beta}^{-1} + T)\theta] \\ &\propto \theta^{\bar{\alpha}-1} \exp(-\theta/\bar{\beta}). \end{aligned}$$

Therefore, $\theta|y \sim G(\bar{\alpha}, \bar{\beta})$.

***Exercise 2.13 (Conjugate exponential family analysis)** Suppose Y_t ($t = 1, 2, \dots, T$) comprise a random sample from an *exponential family* distribution with density

$$p(y_t|\theta) = a(\theta)b(y_t) \exp \left[\sum_{j=1}^J c_j(\theta)d_j(y_t) \right], \quad (2.43)$$

where $\theta = [\theta_1 \ \theta_2 \ \dots \ \theta_k]'$ and $a(\theta)$, $b(y)$, $c_j(\theta)$, and $d_j(y)$ ($j = 1, 2, \dots, J$) are suitably chosen functions. Consider a prior density of the form

$$p(\theta) = p(\theta|\underline{\alpha}, \underline{\delta}) = [a(\theta)]^{\underline{\alpha}} g(\underline{\alpha}, \underline{\delta}) \exp \left[\sum_{j=1}^J c_j(\theta)\underline{\delta}_j \right], \quad (2.44)$$

where $\underline{\alpha} > 0$ and $\underline{\delta} = [\underline{\delta}_1 \ \underline{\delta}_2 \ \dots \ \underline{\delta}_J]'$ are given and $g(\underline{\alpha}, \underline{\delta})$ is chosen so that (2.44) integrates to unity with respect to θ .

(a) Show that the posterior distribution is of the form $p(\theta|\bar{\alpha}, \bar{\delta})$, where

$$\bar{\alpha} = \underline{\alpha} + T, \quad (2.45)$$

$$\hat{d}_j \equiv \hat{d}_j(y) = \sum_{t=1}^T d_j(y_t), \quad j = 1, 2, \dots, J, \quad (2.46)$$

$$\bar{\delta}_j = \underline{\delta}_j + \hat{d}_j(y), \quad j = 1, 2, \dots, J. \quad (2.47)$$

(b) Consider a mixture prior of the form

$$g(\theta) = \sum_{m=1}^M \underline{w}_m p(\theta|\underline{\alpha}_m, \underline{\delta}_m), \quad (2.48)$$

where $\underline{w}_m \geq 0$ ($m = 1, 2, \dots, M$) are known weights satisfying $\sum_m \underline{w}_m = 1$, $p(\theta|\underline{\alpha}_m, \underline{\delta}_m)$ retains the form of (2.44) for each m , and $\underline{\alpha}_m, \underline{\delta}_m$ ($m = 1, 2, \dots, M$) are given. Find the posterior density based on this prior.

Solution

(a) The likelihood function corresponding to (2.43) is

$$\begin{aligned}
 L(\theta) &= \prod_{t=1}^T a(\theta) b(y_t) \exp \left[\sum_{j=1}^J c_j(\theta) d_j(y_t) \right] \\
 &= [a(\theta)]^T b_T(y) \exp \left[\sum_{t=1}^T \sum_{j=1}^J c_j(\theta) d_j(y_t) \right] \\
 &\propto [a(\theta)]^T \exp \left[\sum_{j=1}^J c_j(\theta) \hat{d}_j(y) \right],
 \end{aligned} \tag{2.49}$$

where $b_T(y) = \prod_{t=1}^T b(y_t)$ and $\hat{d}_j(y)$ is defined in (2.46). Using Bayes' Theorem, the posterior density of θ is

$$\begin{aligned}
 p(\theta|y) &\propto p(\theta) L(\theta) \\
 &\propto [a(\theta)]^{\underline{\alpha}} g(\underline{\alpha}, \underline{\delta}) \exp \left[\sum_{j=1}^J c_j(\theta) \underline{\delta}_j \right] [a(\theta)]^T b_T(y) \exp \left[\sum_{j=1}^J c_j(\theta) \hat{d}_j(y) \right] \\
 &\propto [a(\theta)]^{\underline{\alpha}+T} \exp \left[\sum_{j=1}^J c_j(\theta) [\underline{\delta}_j + \hat{d}_j(y)] \right] \\
 &\propto [a(\theta)]^{\bar{\alpha}} \exp \left[\sum_{j=1}^J c_j(\theta) \bar{\delta}_j \right],
 \end{aligned}$$

with $\bar{\alpha}$ and $\bar{\delta}_j$ defined in (2.45) and (2.47), respectively. Therefore, the posterior density is of the form $p(\theta|\bar{\alpha}, \bar{\delta})$.

(b) Using mixture prior (2.48) together with likelihood (2.49) yields the posterior density

$$\begin{aligned}
 p(\theta|y) &\propto p(\theta) L(\theta) \\
 &= \left[\sum_{m=1}^M \underline{w}_m [a(\theta)]^{\underline{\alpha}_m} g(\underline{\alpha}_m, \underline{\delta}_m) \exp \left[\sum_{j=1}^J c_j(\theta) \underline{\delta}_{mj} \right] \right] \\
 &\quad \times [a(\theta)]^T b_T(y) \exp \left[\sum_{j=1}^J c_j(\theta) \hat{d}_j(y) \right] \\
 &= \sum_{m=1}^M \underline{w}_m [a(\theta)]^{\underline{\alpha}_m+T} g(\underline{\alpha}_m, \underline{\delta}_m) \exp \left[\sum_{j=1}^J c_j(\theta) [\underline{\delta}_{mj} + \hat{d}_j(y)] \right] \\
 &= \sum_{m=1}^M \underline{w}_m [a(\theta)]^{\bar{\alpha}_m} g(\underline{\alpha}_m, \underline{\delta}_m) \exp \left[\sum_{j=1}^J c_j(\theta) \bar{\delta}_{mj} \right],
 \end{aligned}$$

where $\bar{\alpha}_m$ and $\bar{\delta}_{mj}$ are defined analogous to (2.45)–(2.47). This shows that the posterior is a mixture of the same form as the prior. The normalizing constant will serve to update the component weights ω_m .

***Exercise 2.14 (Conjugate multivariate normal analysis)** Consider a random sample Y_t ($t = 1, 2, \dots, T$) from a $N(\mu, \Sigma)$ distribution, where μ is $M \times 1$ and Σ is an $M \times M$ positive definite matrix. Define

$$\bar{y} = \frac{1}{T} \sum_{t=1}^T y_t$$

and

$$S = \sum_{t=1}^T (y_t - \bar{y})(y_t - \bar{y})'.$$

Suppose \underline{m} and $\underline{\mu}$ are both $M \times 1$, $\underline{T} > 0$ and $\underline{\omega} > M$ are both scalars, and \underline{S} is an $M \times M$ positive definite matrix. Consider the natural conjugate priors for the following three cases.

(a) Suppose μ is unknown, Σ^{-1} is known, and the prior distribution for μ is multivariate normal with prior density $p(\mu) = \phi(\mu | \underline{\mu}, \underline{\Sigma})$. Find the posterior distribution for μ .

(b) Suppose μ is known, Σ^{-1} is unknown, and the prior for Σ^{-1} is the Wishart distribution (see Appendix Definition 6), $\Sigma^{-1} \sim W(\underline{S}, \underline{\omega})$. Find the posterior distribution for Σ^{-1} .

(c) Suppose both μ and Σ^{-1} are unknown with prior distribution

$$\begin{aligned} p(\mu, \Sigma^{-1}) &= p(\mu | \Sigma^{-1}) p(\Sigma^{-1}) \\ &= \phi(\mu | \underline{\mu}, \underline{T}^{-1} \underline{\Sigma}) f_W(\Sigma^{-1} | \underline{S}, \underline{\omega}). \end{aligned}$$

Find the posterior distribution for μ and Σ^{-1} .

Solution

The likelihood function is

$$\begin{aligned} L(\mu, \Sigma^{-1}) &= \prod_{t=1}^T (2\pi)^{-M/2} |\Sigma^{-1}|^{(1/2)} \exp \left[-\frac{1}{2} (y_t - \mu)' \Sigma^{-1} (y_t - \mu) \right] \quad (2.50) \\ &= (2\pi)^{-TM/2} |\Sigma^{-1}|^{(T/2)} \exp \left[-\frac{1}{2} \sum_{t=1}^T (y_t - \mu)' \Sigma^{-1} (y_t - \mu) \right] \\ &= (2\pi)^{-TM/2} |\Sigma^{-1}|^{(T/2)} \exp \left[-\frac{1}{2} \sum_{t=1}^T \text{tr} [\Sigma^{-1} (y_t - \mu)(y_t - \mu)'] \right] \\ &= (2\pi)^{-TM/2} |\Sigma^{-1}|^{(T/2)} \exp \left[-\frac{1}{2} \text{tr} \left(\Sigma^{-1} \sum_{t=1}^T (y_t - \mu)(y_t - \mu)' \right) \right], \end{aligned}$$

where “tr” denotes the trace operator. Noting the identity

$$\sum_{t=1}^T (y_t - \mu)(y_t - \mu)' = S + T(\bar{y} - \mu)(\bar{y} - \mu)',$$

we can write likelihood (2.50) as

$$L(\mu, \Sigma^{-1}) = (2\pi)^{-TM/2} |\Sigma^{-1}|^{(T/2)} \exp \left[-\frac{1}{2} \text{tr} (\Sigma^{-1} [S + T(\bar{y} - \mu)(\bar{y} - \mu)']) \right]. \quad (2.51)$$

(a) With Σ known, combining the multivariate normal prior density $p(\mu) = \phi(\mu | \underline{\mu}, \underline{\Sigma})$ with likelihood (2.51) yields

$$\begin{aligned} p(\mu | y) &\propto \exp \left[-\frac{1}{2} (\mu - \underline{\mu})' \underline{\Sigma}^{-1} (\mu - \underline{\mu}) \right] \\ &\times \exp \left[-\frac{1}{2} \text{tr} (\Sigma^{-1} [S + T(\bar{y} - \mu)(\bar{y} - \mu)']) \right] \\ &\propto \exp \left[-\frac{1}{2} ((\mu - \underline{\mu})' \underline{\Sigma}^{-1} (\mu - \underline{\mu}) + (\mu - \bar{y})' (T \Sigma^{-1}) (\mu - \bar{y})) \right]. \end{aligned} \quad (2.52)$$

By completing the square on μ , that is, by writing

$$\begin{aligned} &(\mu - \underline{\mu})' \underline{\Sigma}^{-1} (\mu - \underline{\mu}) + (\mu - \bar{y})' (T \Sigma^{-1}) (\mu - \bar{y}) \\ &= (\mu - \bar{\mu})' (\underline{\Sigma}^{-1} + T \Sigma^{-1}) (\mu - \bar{\mu}) + (\underline{\mu} - \bar{y})' [\underline{\Sigma}^{-1} (\underline{\Sigma}^{-1} + T \Sigma^{-1})^{-1} T \Sigma^{-1}] (\underline{\mu} - \bar{y}), \end{aligned} \quad (2.53)$$

where

$$\bar{\mu} = (\underline{\Sigma}^{-1} + T \Sigma^{-1})^{-1} (\underline{\Sigma}^{-1} \underline{\mu} + T \Sigma^{-1} \bar{y}), \quad (2.54)$$

the posterior kernel (2.52) simplifies to

$$p(\mu | y) \propto \exp \left[-\frac{1}{2} (\mu - \bar{\mu})' (\underline{\Sigma}^{-1} + T \Sigma^{-1}) (\mu - \bar{\mu}) \right], \quad (2.55)$$

which is immediately recognized as a multivariate normal kernel. Therefore,

$$\mu | y \sim N(\bar{\mu}, [\underline{\Sigma}^{-1} + T \Sigma^{-1}]^{-1}). \quad (2.56)$$

(b) When μ is known, combining the Wishart prior $W(\underline{S}, \underline{\omega})$ for Σ^{-1} with likelihood (2.51) yields the posterior distribution for Σ^{-1} :

$$\begin{aligned} p(\Sigma^{-1} | y) &\propto |\Sigma^{-1}|^{(\underline{\omega} - M - 1)/2} \exp \left[-\frac{1}{2} \text{tr}(\underline{S}^{-1} \Sigma^{-1}) \right] \\ &\times \left[|\Sigma^{-1}|^{(T/2)} \exp \left(-\frac{1}{2} \text{tr}([S + T(\bar{y} - \mu)(\bar{y} - \mu)'] \Sigma^{-1}) \right) \right] \\ &\propto |\Sigma^{-1}|^{(T + \underline{\omega} - M - 1)/2} \exp \left[-\frac{1}{2} \text{tr}([\underline{S}^{-1} + S + T(\bar{y} - \mu)(\bar{y} - \mu)'] \Sigma^{-1}) \right]. \end{aligned} \quad (2.57)$$

The density given in the last line of (2.57) is the kernel of a Wishart distribution $W(\bar{S}^{-1}, \bar{\omega})$, where

$$\bar{\omega} = \underline{\omega} + T$$

and

$$\bar{S} = \underline{S}^{-1} + S + T(\bar{y} - \mu)(\bar{y} - \mu)'. \quad (2.58)$$

(c) The posterior for μ and Σ^{-1} is

$$p(\mu, \Sigma^{-1} | y) \propto |\underline{T}\Sigma^{-1}|^{(1/2)} \exp \left[-\frac{1}{2}(\mu - \underline{\mu})' \underline{T}\Sigma^{-1}(\mu - \underline{\mu}) \right] \quad (2.59)$$

$$\times |\Sigma^{-1}|^{(\underline{\omega} - M - 1)/2} \exp \left[-\frac{1}{2} \text{tr}(\underline{S}^{-1} \Sigma^{-1}) \right] \quad (2.60)$$

$$\begin{aligned} & \times \left(|\Sigma^{-1}|^{(T/2)} \exp \left[-\frac{1}{2} \text{tr}([S + T(\bar{y} - \mu)(\bar{y} - \mu)'] \Sigma^{-1}) \right] \right) \\ & \propto |\Sigma^{-1}|^{(1/2)} \exp \left[-\frac{1}{2}(\mu - \underline{\mu})' \underline{T}\Sigma^{-1}(\mu - \underline{\mu}) \right] \quad (2.61) \\ & \times \exp \left[-\frac{1}{2}(\mu - \bar{y})' T \Sigma^{-1}(\mu - \bar{y}) \right] \\ & \times \left(|\Sigma^{-1}|^{(T + \underline{\omega} - M - 1)/2} \exp \left[-\frac{1}{2} \text{tr}([\underline{S}^{-1} + S] \Sigma^{-1}) \right] \right). \end{aligned}$$

Analogous to (2.53) and (2.54) we have

$$\begin{aligned} & (\mu - \underline{\mu})' \underline{T}\Sigma^{-1}(\mu - \underline{\mu}) + (\mu - \bar{y})' (T \Sigma^{-1})(\mu - \bar{y}) \quad (2.62) \\ & = (\mu - \bar{\mu})' [(\underline{T} + T) \Sigma^{-1}](\mu - \bar{\mu}) + (\underline{\mu} - \bar{y})' [\underline{T} T (\underline{T} + T)^{-1} \Sigma^{-1}](\underline{\mu} - \bar{y}), \end{aligned}$$

where

$$\bar{\mu} = \frac{\underline{T}\mu + T\bar{y}}{\underline{T} + T}. \quad (2.63)$$

Using (2.60) and (2.61) simplifies posterior (2.59) to

$$\begin{aligned} p(\mu, \Sigma^{-1} | y) & \propto |\bar{T}\Sigma^{-1}|^{(1/2)} \exp \left[-\frac{1}{2}(\mu - \bar{\mu})' \bar{T}\Sigma^{-1}(\mu - \bar{\mu}) \right] \quad (2.64) \\ & \times \left(|\Sigma^{-1}|^{(\bar{\omega} - M - 1)/2} \exp \left[-\frac{1}{2} \text{tr}(\tilde{S}^{-1} \Sigma^{-1}) \right] \right), \end{aligned}$$

where $\bar{T} = \underline{T} + T$, $\bar{\omega} = T + \underline{\omega}$, and

$$\tilde{S} = \underline{S}^{-1} + S + \frac{\underline{T}T}{\underline{T} + T}(\underline{\mu} - \bar{y})(\underline{\mu} - \bar{y})'. \quad (2.65)$$

These provide updating equations for the hyperparameters of our prior.

Exercise 2.15 (Identification) Let $\theta = [\theta_1 \ \theta_2]' \in \mathcal{R}^2$. Consider a prior density $p(\theta_1, \theta_2)$. Suppose the likelihood function for the observed data y depends on θ_1 and not θ_2 , that is,

$$L(\theta_1, \theta_2) = L(\theta_1).$$

In other words, θ_2 is *not identified*.

- (a) Find the conditional posterior density of θ_2 given θ_1 .
 (b) Find the marginal posterior density of θ_2 .
 (c) Do you learn anything about θ_2 ? Discuss.

Solution

- (a) The joint posterior density

$$p(\theta_1, \theta_2 | y) = \frac{p(\theta_1)p(\theta_2|\theta_1)L(\theta_1)}{p(y)}$$

implies the marginal posterior density

$$p(\theta_1 | y) = \frac{p(\theta_1)L(\theta_1)}{p(y)},$$

and hence the conditional posterior density is

$$\begin{aligned} p(\theta_2 | \theta_1, y) &= \frac{p(\theta_1, \theta_2 | y)}{p(\theta_1 | y)} \\ &= \frac{[p(\theta_1)p(\theta_2|\theta_1)L(\theta_1)]/[p(y)]}{[p(\theta_1)L(\theta_1)]/[p(y)]} \\ &= p(\theta_2 | \theta_1). \end{aligned}$$

- (b) The marginal posterior density of θ_2 is

$$\begin{aligned} p(\theta_2 | y) &= \int_{\mathcal{R}} p(\theta_1, \theta_2 | y) d\theta_1 \\ &= \int_{\mathcal{R}} \frac{p(\theta_1)p(\theta_2|\theta_1)L(\theta_1)}{p(y)} d\theta_1 \\ &= \int_{\mathcal{R}} p(\theta_2 | \theta_1)p(\theta_1 | y) d\theta_1 \\ &= E_{\theta_1 | y}[p(\theta_2 | \theta_1)], \end{aligned}$$

and thus $p(\theta_2 | y) = p(\theta_2)$ if θ_1 and θ_2 are independent *a priori*. In the case of prior dependence, however, $p(\theta_2 | y) \neq p(\theta_2)$.

- (c) From (a) we see that conditional on θ_1 we do not learn about θ_2 – the conditional posterior is identical to the conditional prior. Noting that $p(\theta_2) = E_{\theta_1}[p(\theta_2 | \theta_1)]$, we see from (b) that we learn marginally about θ_2 , provided θ_1 and θ_2 are *dependent*.

Point estimation

Frequentist point-estimation techniques are typically based on the choice of an intuitively sensible criteria (e.g., method of moments) and the statistical properties of the resulting estimators are addressed after the fact. In contrast, Bayesian point estimation begins by announcing a criterion for determining what constitutes a good point estimate, and a method for producing an “optimal” point estimate given the data at hand is then derived.

The Bayesian begins by specifying a *loss (cost) function* $C(\hat{\theta}, \theta)$, that is, a nonnegative function satisfying $C(\theta, \theta) = 0$, which measures the consequences of using $\hat{\theta} = \hat{\theta}(Y)$ (i.e., a particular function of the data) as an estimate when the “state of nature” is θ . Usually, $C(\hat{\theta}, \theta)$ is a nondecreasing function of the *sampling error* $\hat{\theta} - \theta$. Frequentists sometimes argue that the need to specify a loss function is one of the shortcomings of the Bayesian approach. However, there is usually a frequentist analogue to the Bayesian choice of loss function. For instance, the familiar concept of mean squared error (defined in the following) simply corresponds to a particular type of loss function, namely quadratic loss.

A good estimate is one that “minimizes” $C(\hat{\theta}, \theta)$ in some sense, but its randomness must first be eliminated. From the frequentist sampling theory point of view, θ is non-stochastic but $C(\hat{\theta}, \theta)$ is nonetheless stochastic because the estimator $\hat{\theta} = \hat{\theta}(Y)$ is a random variable (i.e., $\hat{\theta}$ depends on the random variable Y). An obvious way to circumscribe the randomness of $C(\hat{\theta}, \theta)$ is to focus attention on its expected value, assuming it exists. Frequentists consider the *risk function*, that is, the nonstochastic function

$$R(\hat{\theta}|\theta) = E_{Y|\theta}[C(\hat{\theta}(Y), \theta)], \quad (3.1)$$

where the expectation (assumed to exist) is taken with respect to the sampling probability function $p(y|\theta)$.

In contrast, the Bayesian perspective is entirely *ex post* (i.e., it conditions on the observed data y).¹ That is, the Bayesian uses $\hat{\theta} = \hat{\theta}(y)$ as a point estimate of the unknown parameter θ . Unlike the frequentist approach, no role is provided for data that could have been observed but were not observed. Since θ is unknown, the Bayesian perspective suggests treating it as a random variable with a distribution reflecting all the information at hand. Such information is fully contained in the posterior distribution of θ : It reflects both prior and sample information.

In contrast to the frequentist approach, the randomness in loss function $C(\hat{\theta}, \theta)$ from the Bayesian perspective arises because θ is an unknown random variable. From the frequentist perspective, although θ is unknown, it is treated as a fixed constant. The *ex ante* perspective of the frequentist view implies that $C(\hat{\theta}, \theta)$ is random because $\hat{\theta}$ is viewed as a random variable with a sampling distribution in repeated samples. The Bayesian solution to the randomness of the loss function is similar to the frequentist solution: Take its expectation before minimization. The expectation, however, is with respect to the posterior distribution $\theta|y$, and not the sampling distribution $y|\theta$ used to obtain the risk function. The Bayesian prescription is equivalent to the principle usually advocated for economic agents acting in a world of uncertainty: Using all available information, choose actions so as to maximize expected utility, or equivalently, minimize expected loss. The prescription is formalized in the following definition.

Definition 3.1 Given the posterior density $p(\theta|y)$ and the cost function $C(\hat{\theta}, \theta)$, the *Bayes estimate* (i.e., the Bayesian point estimate) is defined as the solution (assuming the expectation exists) to the following problem:

$$\min_{\hat{\theta}} c(\hat{\theta}) = E_{\theta|y}[C(\hat{\theta}, \theta)], \quad (3.2)$$

where

$$E_{\theta|y}[C(\hat{\theta}, \theta)] = \int_{\Theta} C(\hat{\theta}, \theta) p(\theta|y) d\theta. \quad (3.3)$$

Most importantly, note that the posterior expectation in (3.3) removes θ from the criterion function (3.2), unlike the case of risk function (3.1). Also note that if the researcher is interested only in a subset θ_1 of the parameter vector $\theta = [\theta_1, \theta_2]'$, where $\Theta = \Theta_1 \times \Theta_2$ and the elements of θ_2 are regarded as nuisance parameters, then this preference can be reflected in the loss function specification: $C(\hat{\theta}, \theta) = C(\hat{\theta}_1, \theta_1)$. In this case the expected loss in (3.3) reduces to

$$\begin{aligned} E_{\theta|y}[C(\hat{\theta}, \theta)] &= \int_{\Theta_1} \int_{\Theta_2} C(\hat{\theta}, \theta) p(\theta|y) d\theta_2 d\theta_1 \\ &= \int_{\Theta_1} C(\hat{\theta}_1, \theta_1) \left[\int_{\Theta_2} p(\theta_1, \theta_2|y) d\theta_2 \right] d\theta_1 \\ &= \int_{\Theta_1} C(\hat{\theta}_1, \theta_1) p(\theta_1|y) d\theta_1. \end{aligned} \quad (3.4)$$

Thus, nuisance parameters are simply marginalized out of the problem.

¹We remind the reader that we are using capital (small) letters to denote a random variable (its realization). Hence, Y is a random variable and y a particular data realization.

Definition 3.2 Consider the case of a single parameter of interest θ . Let c , c_1 , and c_2 be known constants. The loss (cost) function

$$C(\hat{\theta}, \theta) = c(\hat{\theta} - \theta)^2 \quad (3.5)$$

is known as a *quadratic loss function*. The loss function

$$C(\hat{\theta}, \theta) = \begin{cases} c_1|\hat{\theta} - \theta| & \text{if } \hat{\theta} \leq \theta \\ c_2|\hat{\theta} - \theta| & \text{if } \hat{\theta} > \theta \end{cases}, \quad (3.6)$$

is known as an *asymmetric linear loss function*. In the case $c_1 = c_2$, (3.6) is known as a *symmetric linear loss function*. Given constants $c > 0$ and $d > 0$, the loss function

$$C(\hat{\theta}, \theta) = \begin{cases} c & \text{if } |\hat{\theta} - \theta| > d \\ 0 & \text{if } |\hat{\theta} - \theta| \leq d \end{cases}, \quad (3.7)$$

is known as an *all-or-nothing loss function* over $|\hat{\theta} - \theta| \leq d$.

When there are several parameters of interest, the most popular loss functions are the *weighted squared error* generalization of (3.5):

$$C(\hat{\theta}, \theta) = (\hat{\theta} - \theta)'Q(\hat{\theta} - \theta), \quad (3.8)$$

where Q is a positive definite matrix and the all-or-nothing loss function (3.7).

Bayesian point estimates corresponding to the loss functions defined in Definition 3.2 are given in the following exercises.

Exercise 3.1 (Point estimation under quadratic loss) Let $\hat{\theta}$ denote the Bayesian point estimates obtained by solving (3.2) for quadratic loss function (3.5). Show that $\hat{\theta} = E(\theta|y)$.

Solution

Consider

$$\begin{aligned} (\hat{\theta} - \theta)^2 &= [\hat{\theta} - E(\theta|y) + E(\theta|y) - \theta]^2 \\ &= [\hat{\theta} - E(\theta|y)]^2 + 2[\hat{\theta} - E(\theta|y)][E(\theta|y) - \theta] + [E(\theta|y) - \theta]^2. \end{aligned} \quad (3.9)$$

Note that only the last two terms of (3.9) involve θ . Then

$$\begin{aligned} c(\hat{\theta}) &= E_{\theta|y}[c(\hat{\theta} - \theta)^2] \\ &= c \left[[\hat{\theta} - E(\theta|y)]^2 + 2[\hat{\theta} - E(\theta|y)]E_{\theta|y}[E(\theta|y) - \theta] + E_{\theta|y}[E(\theta|y) - \theta]^2 \right] \\ &= c[\hat{\theta} - E(\theta|y)]^2 + 0 + c\text{Var}(\theta|y), \end{aligned} \quad (3.10)$$

which is minimized by setting the first term to zero [i.e., $\hat{\theta} = E(\theta|y)$].

Exercise 3.2 (Point estimation under asymmetric linear loss) Let $\hat{\theta}$ denote the Bayesian point estimates obtained by solving (3.2) for asymmetric linear loss function (3.6). Show that $\hat{\theta} = \zeta_q$, where $q = c_1/(c_1 + c_2)$ and ζ_q denotes the q th quantile of the posterior density $p(\theta|y)$.

Solution

The expected posterior loss is

$$\begin{aligned} E_{\theta|y}[C(\hat{\theta}, \theta)] &= c_2 \int_{-\infty}^{\hat{\theta}} (\hat{\theta} - \theta)p(\theta|y)d\theta + c_1 \int_{\hat{\theta}}^{\infty} -(\hat{\theta} - \theta)p(\theta|y)d\theta \\ &= c_2 \hat{\theta} P(\hat{\theta}|y) - c_2 \int_{-\infty}^{\hat{\theta}} \theta p(\theta|y)d\theta - c_1 \hat{\theta} [1 - P(\hat{\theta}|y)] + c_1 \int_{\hat{\theta}}^{\infty} \theta p(\theta|y)d\theta, \end{aligned} \quad (3.11)$$

where $P(\cdot)$ denotes the c.d.f. corresponding to $p(\cdot)$. Differentiating (3.11) with respect to $\hat{\theta}$ yields

$$\begin{aligned} \frac{\partial E_{\theta|y}[C(\hat{\theta}, \theta)]}{\partial \hat{\theta}} &= c_2 P(\hat{\theta}|y) + c_2 \hat{\theta} p(\hat{\theta}|y) - c_2 \hat{\theta} p(\hat{\theta}|y) \\ &\quad - c_1 [1 - P(\hat{\theta}|y)] + c_1 \hat{\theta} p(\hat{\theta}|y) - c_1 \hat{\theta} p(\hat{\theta}|y) \\ &= -c_1 + (c_1 + c_2) P(\hat{\theta}|y). \end{aligned} \quad (3.12)$$

Equating (3.12) to zero and solving for $\hat{\theta}$ yields

$$\hat{\theta} = P_{\theta|y}^{-1} \left(\frac{c_1}{c_1 + c_2} \right),$$

that is, the $c_1/(c_1 + c_2)$ posterior quantile of θ .

Exercise 3.3 (Point estimation under all-or-nothing loss) Let $\hat{\theta}$ denote the Bayesian point estimate obtained by solving (3.2) for all-or-nothing loss function (3.7). Show that $\hat{\theta}$ is the center of an interval of width $2d$ having maximum posterior probability.

Solution

The expected posterior loss is

$$E_{\theta|y}[C(\hat{\theta}, \theta)] = \int_{\hat{\theta}-d}^{\hat{\theta}+d} cp(\theta|y)d\theta = c[P(\hat{\theta} + d|y) - P(\hat{\theta} - d|y)]. \quad (3.13)$$

Minimizing (3.13) corresponds to picking $\hat{\theta}$ to be the center of an interval of width $2d$ having maximum posterior probability. Differentiating (3.13) with respect to $\hat{\theta}$ yields

$$\frac{\partial E_{\theta|y}[C(\hat{\theta}, \theta)]}{\partial \hat{\theta}} = c [p(\hat{\theta} + d|y) - p(\hat{\theta} - d|y)]. \quad (3.14)$$

Equating (3.14) to zero and solving for $\hat{\theta}$ implies that the end points of this interval have equal posterior density. Also note that, as $d \rightarrow 0$, $\hat{\theta}$ becomes the mode of $p(\theta|y)$.

Exercise 3.4 (Point estimation under quadratic loss for conjugate Bernoulli analysis)

Consider the Bernoulli sampling problem considered in Exercise 2.1 with posterior p.d.f. given by (2.11). Find the Bayesian point estimate of θ under quadratic loss.

Solution

Exercise 3.1 implies that the Bayesian point estimate is the posterior mean:

$$\hat{\theta} = \frac{\bar{\alpha}}{\bar{\alpha} + \bar{\delta}} = \frac{m + \underline{\alpha}}{T + \underline{\alpha} + \underline{\delta}} = \frac{T\bar{y} + \underline{T}(\underline{T}^{-1}\underline{\alpha})}{T + \underline{T}},$$

where the second equality follows by using the definitions in (2.8) and (2.9). In the final equality, $\underline{T} \equiv \underline{\alpha} + \underline{\delta}$. Thus, the posterior mean can be interpreted as a weighted average of the sample mean $\bar{y} = T^{-1}m$ and the effective prior mean $\underline{T}^{-1}\underline{\alpha}$.

Exercise 3.5 (Point estimation under asymmetric linear loss for conjugate Bernoulli analysis)

Consider the Bernoulli sampling problem considered in Exercise 2.1 with posterior p.d.f. given by (2.11). Find the Bayesian point estimate of θ under asymmetric linear loss function (3.6).

Solution

Exercise 3.2 implies that the Bayesian point estimate of θ is the q th posterior quantile where $q = c_1/(c_1 + c_2)$, that is, the value $\hat{\theta}$ that solves

$$\frac{c_1}{c_1 + c_2} = \int_0^{\hat{\theta}} \frac{\theta^{\bar{\alpha}-1}(1-\theta)^{\bar{\delta}-1}}{\mathcal{B}(\bar{\alpha}, \bar{\delta})} d\theta.$$

If $c_1 = c_2$ then $\hat{\theta}$ equals the posterior median.

Exercise 3.6 (Point estimation under all-or-nothing loss for conjugate Bernoulli analysis)

Consider the Bernoulli sampling problem considered in Exercise 2.1 with posterior p.d.f. (2.11). Find the Bayesian point estimate of θ under all-or-nothing loss function (3.7) as $d \rightarrow 0$.

Solution

Exercise 3.3 implies that, under loss function (3.7), the Bayesian point estimate of θ is the posterior mode, which if $\bar{\alpha} > 1$ and $\bar{\delta} > 1$ is found by differentiating the logarithm of (2.11) with respect to θ , equating the result to zero, and solving the resulting equation to yield

$$\hat{\theta} = \frac{\bar{\alpha} - 1}{\bar{\alpha} + \bar{\delta} - 2} = \frac{m + \underline{\alpha} - 1}{T + \underline{\alpha} + \underline{\delta} - 2}.$$

If $\bar{\alpha} = \bar{\delta} = 1$, then the posterior density is uniform and the posterior mode is not unique. If either $\bar{\alpha} < 1$ or $\bar{\delta} < 1$, then there is no posterior mode since the posterior density has asymptote approaching infinity.

Exercise 3.7 (Point estimation for conjugate normal analysis) Consider the normal sampling problem in Exercise 2.3. Find the Bayesian point estimate $\hat{\theta}_1$ of θ_1 under loss functions (3.5)–(3.7). For loss function (3.7), consider the case as $d \rightarrow 0$.

Solution

Since the posterior distribution is normal, the posterior mean and mode are equal. Exercises 3.4 and 3.6 imply that this common value is the Bayesian point estimate under (3.5) and (3.7). Under loss function (3.6), Exercise 3.5 implies that the Bayesian point estimate $\hat{\theta}_1$ of θ_1 is given by the q th posterior quantile, $q = c_1/(c_1 + c_2)$, that is,

$$\hat{\theta}_1 = \bar{\mu} + \bar{\sigma} \Phi^{-1} \left(\frac{c_1}{c_1 + c_2} \right),$$

where $\bar{\mu}$ and $\bar{\sigma}$ are the posterior mean and standard deviation of θ_1 and $\Phi^{-1}(\cdot)$ is the inverse of the standard normal c.d.f. Note that when $c_1 = c_2$, $\Phi^{-1}(1/2) = 0$ and $\hat{\theta}_1$ reduces to simply the posterior mean-median-mode $\bar{\mu}$.

Exercise 3.8 (Point estimation under weighted squared error loss) Suppose $\theta \in \mathcal{R}^K$. Consider weighted squared error loss function (3.8) and suppose its posterior expected value exists. Show that the Bayesian point estimate $\hat{\theta}$ of θ is the posterior mean.

Solution

Proceeding as in Exercise 3.1, consider

$$\begin{aligned} C(\hat{\theta}, \theta) &= (\hat{\theta} - \theta)' Q (\hat{\theta} - \theta) \\ &= [\hat{\theta} - E(\theta|y) + E(\theta|y) - \theta]' Q [\hat{\theta} - E(\theta|y) + E(\theta|y) - \theta] \\ &= [\hat{\theta} - E(\theta|y)]' Q [\hat{\theta} - E(\theta|y)] + 2[\hat{\theta} - E(\theta|y)]' Q [E(\theta|y) - \theta] \\ &\quad + [E(\theta|y) - \theta]' Q [E(\theta|y) - \theta]. \end{aligned} \tag{3.15}$$

Noting that only the last two terms of (3.15) involve θ , we see that the posterior expected loss of (3.15) is

$$\begin{aligned} E_{\theta|y}[C(\hat{\theta}, \theta)] &= [\hat{\theta} - E(\theta|y)]' Q [\hat{\theta} - E(\theta|y)] + 2[\hat{\theta} - E(\theta|y)]' Q E_{\theta|y}[E(\theta|y) - \theta] \\ &\quad + E_{\theta|y}([E(\theta|y) - \theta]' Q [E(\theta|y) - \theta]) \\ &= [\hat{\theta} - E(\theta|y)]' Q [\hat{\theta} - E(\theta|y)] + 0 + E_{\theta|y}([E(\theta|y) - \theta]' Q [E(\theta|y) - \theta]). \end{aligned} \tag{3.16}$$

Picking $\hat{\theta} = E(\theta|y)$ minimizes (3.16). Note that $\hat{\theta}$ does *not* depend on Q .

Exercise 3.9 (Point estimation under alternative squared error loss) Consider the weighted squared error loss function $C(\hat{\theta}, \theta) = w(\theta)(\theta - \hat{\theta})^2$, where $w(\theta) > 0$ and θ is a single unknown parameter. Show that the Bayesian point estimate corresponding to $C(\hat{\theta}, \theta)$ is given by

$$\hat{\theta} = \frac{E_{\theta|y}[\theta w(\theta)]}{E_{\theta|y}[w(\theta)]}. \tag{3.17}$$

Compare the role of $w(\theta)$ to that of the prior for θ .

Solution

Minimizing expected posterior loss requires solving

$$\begin{aligned}\frac{\partial E_{\theta|y}[C(\hat{\theta}, \theta)]}{\partial \hat{\theta}} &= E_{\theta|y} \left(\frac{\partial [C(\hat{\theta}, \theta)]}{\partial \hat{\theta}} \right) \\ &= E_{\theta|y}[2w(\theta)(\hat{\theta} - \theta)] \\ &= 2 \left(\hat{\theta} E_{\theta|y}[w(\theta)] - E_{\theta|y}[\theta w(\theta)] \right) = 0,\end{aligned}$$

provided one can differentiate under the integral sign. The solution to this equation follows immediately:

$$\hat{\theta} = \frac{E_{\theta|y}[\theta w(\theta)]}{E_{\theta|y}[w(\theta)]},$$

which is equivalent to (3.17).

It is also worth noting

$$E_{\theta|y}[C(\hat{\theta}, \theta)] = \int_{\Theta} w(\theta)(\hat{\theta} - \theta)^2 \left[\frac{p(\theta)L(\theta)}{p(y)} \right] d\theta = \int_{\Theta} (\hat{\theta} - \theta)^2 \left[\frac{w(\theta)p(\theta)L(\theta)}{p(y)} \right] d\theta,$$

and thus it is seen that minimizing expected posterior loss $C(\hat{\theta}, \theta) = w(\theta)(\hat{\theta} - \theta)^2$ using prior $p(\theta)$ is equivalent to minimizing expected quadratic loss using the prior

$$w(\theta)p(\theta) / \left[\int_{\Theta} w(\theta)p(\theta) d\theta \right].$$

Exercise 3.10 (Point estimation under nonconjugate normal analysis) Suppose $Y|\theta \sim N(\theta, 1)$ and $p(\theta) \propto \theta \exp(-\theta^2/2)$. Find the Bayesian point estimate of θ under quadratic loss.

Solution

Given $Y|\theta \sim N(\theta, 1)$ and $p(\theta) \propto \theta \exp(-\theta^2/2)$, the Bayesian point estimate of θ under quadratic loss is the posterior mean $E(\theta|y)$ corresponding to the posterior density

$$p(\theta|y) \propto \theta \exp\left(-\frac{\theta^2}{2}\right) \exp\left(-\frac{(y-\theta)^2}{2}\right) \propto \theta \exp\left(-\frac{(y-\theta)^2 + \theta^2}{2}\right).$$

Using identity (2.19) from Chapter 2 with $\underline{\mu} = 0$, $\theta_1 = \theta$, $\bar{y} = y$, $\underline{h} = 1$, and $h = 1$ implies

$$(y - \theta)^2 + \theta^2 = 2 \left(\theta - \frac{y}{2} \right)^2 + \frac{y^2}{2}$$

and

$$p(\theta|y) \propto \theta \exp \left[-\frac{1}{2(1/2)} \left(\theta - \frac{y}{2} \right)^2 \right]. \quad (3.18)$$

Using the results of Exercise 3.1, we know that the posterior mean of (3.18) is the minimizer of

$$\int (\hat{\theta} - \theta)^2 p(\theta|y) dy = \int \theta(\hat{\theta} - \theta)^2 \tilde{p}(\theta|y) dy,$$

where, from (3.18), $\tilde{p}(\theta|y)$ is the kernel of an $N(y/2, 1/2)$ density. In this alternate form, the results of Exercise 3.9 show

$$\hat{\theta} = \frac{E[\theta^2]}{E[\theta]},$$

where the expectation is taken with respect to θ following an $N(y/2, 1/2)$ density. This produces

$$\hat{\theta} = \frac{E(\theta^2)}{E(\theta)} = \frac{(1/2) + (y/2)^2}{y/2} = \frac{y^2 + 2}{2y}.$$

Exercise 3.11 (Point estimation under linex loss: Zellner [1986b]) Given a scalar parameter θ , consider the univariate *linex loss function*

$$C(\hat{\theta}, \theta) = \exp[a(\hat{\theta} - \theta)] - a(\hat{\theta} - \theta) - 1, \quad (3.19)$$

where $a \neq 0$. Its name arises from the fact that, if $a < 0$, $C(\hat{\theta}, \theta)$ rises almost linearly for $\hat{\theta} > \theta$ and almost exponentially for $\hat{\theta} < \theta$. For small values of $|a|$, the loss function is almost symmetric. For $a = 1$, the loss function is quite asymmetric with overestimation being more costly than underestimation. Find the Bayesian point estimate of θ .

Solution

Given the linex loss function (3.19), the posterior expected loss is

$$\begin{aligned} E_{\theta|y}[C(\hat{\theta}, \theta)] &= E_{\theta|y} \left(\exp[a(\hat{\theta} - \theta)] - a(\hat{\theta} - \theta) - 1 \right) \\ &= \exp(a\hat{\theta}) M_{\theta|y}(-a) - a[\hat{\theta} - E(\theta|y)] - 1, \end{aligned} \quad (3.20)$$

where $M_{\theta|y}(-a) = E_{\theta|y}[\exp(-a\theta)]$, the moment-generating function of the posterior distribution of θ , is assumed to exist.

Minimizing (3.20) with respect to $\hat{\theta}$ suggests equating

$$\frac{\partial E_{\theta|y}[C(\hat{\theta}, \theta)]}{\partial \hat{\theta}} = a \exp(a\hat{\theta}) M_{\theta|y}(-a) - a \quad (3.21)$$

to zero, yielding

$$\begin{aligned} \hat{\theta}_1 &= -a^{-1} \ln (E_{\theta|y}[\exp(-a\theta)]) \\ &= -a^{-1} \ln(M_{\theta|y}(-a)). \end{aligned} \quad (3.22)$$

4

Frequentist properties of Bayesian estimators

As noted in Chapter 3, frequentists evaluate point estimators in terms of minimizing *ex ante* risk function (3.1). In contrast, Bayesians focus on minimizing *ex post* posterior loss (3.3). By construction Bayesian point estimates are optimal from the *ex post* standpoint. In general they also have good risk properties. In this chapter we view the Bayesian point estimate $\hat{\theta} = \hat{\theta}(y)$ from an *ex ante* perspective, that is, we consider the Bayesian point estimator $\hat{\theta} = \hat{\theta}(Y)$.

Minimum risk estimators do not exist in general because (3.1) depends on θ and so any estimator that minimizes (3.1) will also depend on θ . Often other conditions are imposed (e.g., unbiasedness) to sidestep the problem. Some statisticians walk a fine line between Bayesian and classical point estimation by taking risk function $R(\hat{\theta}|\theta)$ in (3.1) and “averaging” it over the permissible parameter space Θ to obtain a measure of the “average” performance of the estimator $\hat{\theta}(Y)$. The motivation is clear: To overcome the difficulty that risk functions of different estimators can cross each other at various points in Θ , it is natural to weight the performance of an estimator in a particular region of the parameter space by a weight function. The interpretation of such a function, however, is very close to a prior distribution for θ that weights heavily regions of the parameter space that the researcher feels are most likely to be relevant. Given a positive function $p(\theta)$, whatever its interpretation, consider the problem

$$\min_{\hat{\theta}} \overline{R}(\hat{\theta}), \quad (4.1)$$

where

$$\overline{R}(\hat{\theta}) = \int_{\Theta} R(\hat{\theta}|\theta)p(\theta)d\theta. \quad (4.2)$$

Substituting the risk function (3.1) into (4.2), where the sample space for Y is S_T , yields

$$\bar{R} = \int_{\Theta} \left[\int_{S_T} C[\hat{\theta}(Y), \theta] p(Y|\theta) dY \right] p(\theta) d\theta \quad (4.3)$$

$$= \int_{S_T} \int_{\Theta} C[\hat{\theta}(Y), \theta] p(Y|\theta) p(\theta) d\theta dY \quad (4.4)$$

$$= \int_{S_T} \left[\int_{\Theta} C[\hat{\theta}(Y), \theta] \left(\frac{p(Y|\theta)p(\theta)}{p(Y)} \right) d\theta \right] p(Y) dY, \quad (4.5)$$

provided it is permissible to switch the order of integration. From (4.5) it is clear that the $\hat{\theta}$ that minimizes \bar{R} must also minimize the quantity enclosed in the square brackets in (4.5). This quantity, however, is exactly the expected posterior loss (3.3) corresponding to the prior distribution $p(\theta)$. Hence, a researcher minimizing \bar{R} behaves *as if* solving (3.3) with prior distribution $p(\theta)$. Therefore, any estimator that minimizes weighted risk must be a Bayesian point estimator.

The preceding analysis shows that Bayesian point estimators have desirable frequentist properties. Any Bayesian estimator based on a *proper prior* (i.e., a prior p.d.f. that integrates to unity) for which (4.2) is finite satisfies the minimal frequentist requirement of admissibility (i.e., there does not exist another estimator with lower risk function at all points in the parameter space) – the same cannot be said for maximum likelihood estimators. In this case (4.2) is often referred to as *Bayes risk*. Furthermore, Wald (1950) showed, in most interesting settings, that *all* admissible rules are either Bayes or limits thereof known as generalized Bayes estimators. A *generalized Bayes estimator* is a Bayes estimator based on an *improper prior* (i.e., a prior whose integral is unbounded).¹ In other words, preference for admissible decision rules implies a preference for Bayes rules. Also, if a Bayes estimator is unique, then it must be admissible. In short, Bayes estimators based on proper priors have desirable properties in terms of “final precision” (i.e., *ex post*) by construction and also may have desirable “initial precision” (i.e., *ex ante*) properties. Bayes estimators corresponding to proper priors, however, are almost never unbiased [see Berger (1985, Chapter 8), for details].

***Exercise 4.1 (Sampling properties under quadratic loss in conjugate normal analysis)**

Suppose $Y_t (t = 1, 2, \dots, T)$ is a random sample from an $N(\mu, \sigma^2)$ distribution where σ^2 is known, and the prior distribution for μ is $N(\underline{\mu}, \sigma^2/T)$.

- Consider the quadratic loss function $C(\hat{\mu}, \mu) = (\hat{\mu} - \mu)^2$. Find the Bayesian *point estimate* $\hat{\mu}$ and compute its expected posterior loss.
- Find the expected posterior loss of the *estimate* $\tilde{\mu} = \bar{y}$.
- Compare $\hat{\mu}$ and $\tilde{\mu}$ according to expected posterior loss. Which is smaller?
- Find the mean squared error, $MSE(\hat{\mu})$, where $\hat{\mu} = \hat{\mu}(Y)$ is the Bayesian *point estimator* found in (a).

¹Examples of improper priors will be considered in Chapter 8.

- (e) Find $\text{MSE}(\tilde{\mu})$, where $\tilde{\mu} = \tilde{\mu}(Y) = \bar{Y}$ is the maximum likelihood point estimator in (b).
 (f) Under what conditions will $\text{MSE}(\hat{\mu}) < \text{MSE}(\tilde{\mu})$?

Solution

(a) According to Exercise 3.1 the choice of $\hat{\mu}$ that minimizes expected posterior quadratic loss is $\hat{\mu} = E(\mu|y)$, which in this case is given by (2.39). Evaluated at $\hat{\mu} = E(\mu|y)$, expected posterior quadratic loss is simply the posterior variance, $\text{Var}(\mu|y)$. An expression for this posterior variance is given in (2.23) and is denoted as $\bar{h}^{-1} = (\underline{h} + h)^{-1}$, as defined in (2.14). In terms of the parameterization of this exercise relative to the parameterization used in Exercise 2.3, we have, from (2.12), $\underline{h}^{-1} = \sigma^2/\underline{T}$, and from (2.13), $h = T\sigma^{-2}$. It follows that minimized expected quadratic posterior loss (3.10) becomes

$$c(\hat{\mu}) = \text{Var}(\mu|y) = (\underline{T}\sigma^{-2} + T\sigma^{-2})^{-1} = \sigma^2(\underline{T} + T)^{-1} = \frac{\sigma^2}{\bar{T}}, \quad (4.6)$$

where $\bar{T} = \underline{T} + T$.

(b) Upon adding and subtracting $E(\mu|y)$ in our quadratic loss function, much like Exercise 3.1, plugging $\tilde{\mu} = \bar{y}$ into (3.10) yields

$$c(\tilde{\mu}) = \frac{\sigma^2}{\bar{T}} + [E(\mu|y) - \bar{y}]^2. \quad (4.7)$$

(c) Clearly (4.6) is less than (4.7), which follows by construction since $\hat{\mu}$ minimizes posterior expected loss.

(d) From the *ex ante* standpoint, (2.39) becomes

$$\hat{\mu} = \frac{\underline{T}\mu + T\bar{Y}}{\bar{T}}. \quad (4.8)$$

To calculate $\text{MSE}(\hat{\mu})$, we need to determine its *sampling* mean and variance. From (4.8) the sampling mean and variance are

$$E(\hat{\mu}) = \frac{\underline{T}\mu + TE(\bar{Y})}{\bar{T}} = \frac{\underline{T}\mu + T\mu}{\bar{T}} \quad (4.9)$$

and

$$\text{Var}(\hat{\mu}) = \frac{T^2\text{Var}(\bar{Y})}{\bar{T}^2} = \frac{T^2(\sigma^2/T)}{\bar{T}^2} = \frac{T\sigma^2}{\bar{T}^2}. \quad (4.10)$$

From (4.9) the bias of $\hat{\mu}$ is

$$\text{Bias}(\hat{\mu}) = E(\hat{\mu}) - \mu = \frac{\underline{T}\mu + T\mu - \bar{T}\mu}{\bar{T}} = \frac{\underline{T}(\mu - \mu)}{\bar{T}}.$$

Recall that the MSE of an estimator equals its sampling variance plus its squared bias. Therefore the $\text{MSE}(\hat{\mu})$ is

$$\text{MSE}(\hat{\mu}) = \frac{T\sigma^2}{\bar{T}^2} + \left[\frac{\underline{T}(\mu - \mu)}{\bar{T}} \right]^2 = \frac{T\sigma^2 + \underline{T}^2(\mu - \mu)^2}{\bar{T}^2}. \quad (4.11)$$

(e) It is well known that $\tilde{\mu} = \bar{Y}$ is an unbiased estimator of μ with

$$\text{MSE}(\tilde{\mu}) = \text{Var}(\tilde{\mu}) = \frac{\sigma^2}{T}. \quad (4.12)$$

(f) From (4.11) and (4.12), $\text{MSE}(\hat{\mu}) < \text{MSE}(\tilde{\mu})$ iff

$$\frac{T\sigma^2 + \underline{T}^2(\underline{\mu} - \mu)^2}{\underline{T}^2} < \frac{\sigma^2}{T}, \quad (4.13)$$

or equivalently, iff

$$\left(\frac{\underline{\mu} - \mu}{\sigma/\sqrt{\underline{T}}} \right)^2 < \underline{T}^{-1}(\underline{T} + 2T). \quad (4.14)$$

In other words, the Bayesian point estimator has a smaller MSE than the sample mean provided the prior mean $\underline{\mu}$ is “sufficiently close” to the unknown population mean μ , where closeness is measured relative to the sampling standard deviation σ/\sqrt{T} in \bar{Y} . Therefore, the Bayes estimate wins expected posterior loss comparisons by definition, and sometimes wins MSE comparisons – even to the minimum variance unbiased estimator \bar{Y} .

***Exercise 4.2 (Sampling properties under quadratic loss in conjugate exponential analysis)** Consider a random sample Y_t ($t = 1, 2, \dots, T$) with exponential density

$$p(y_t|\theta) = \theta^{-1}[\exp(-y_t/\theta)], \quad 0 < y_t < \infty, \quad 0 < \theta < \infty. \quad (4.15)$$

Note that \bar{Y} is both the maximum likelihood and the method of moments estimator of θ .

(a) Find $\text{MSE}(\bar{Y})$.

(b) Consider the estimator

$$\hat{\theta} = \frac{T\bar{Y}}{T+1}.$$

Find $\text{MSE}(\hat{\theta})$.

(c) Compare the MSEs from (a) and (b). Which estimator has the lower MSE?

(d) Let $\delta = \theta^{-1}$ and rewrite (4.15) as

$$p(y_t|\delta) = \delta \exp(-\delta y_t), \quad 0 < y_t < \infty, \quad 0 < \delta < \infty. \quad (4.16)$$

Suppose prior beliefs regarding δ have a gamma distribution with hyperparameters $\underline{\alpha}$ and $\underline{\beta}$ [i.e., $\delta \sim G(\underline{\alpha}, \underline{\beta})$, $E(\delta) = \underline{\alpha}\underline{\beta}$, and $\text{Var}(\delta) = \underline{\alpha}\underline{\beta}^2$]. Find the posterior distribution of δ .

(e) Find the posterior distribution of θ corresponding to the answer in part (d).

(f) Consider the loss function $C(\tilde{\theta}, \theta) = (\tilde{\theta} - \theta)^2$. What is the Bayesian point estimate $\tilde{\theta}$ of θ ? Be specific!

(g) Find $\text{MSE}(\tilde{\theta})$. Compare it, in the case $\underline{\alpha} = 1$, to your answers in (a) and (b). Which estimator do you prefer?

(h) In the case $\underline{\alpha} = 1$, compare \bar{Y} , $\hat{\theta}$, and $\tilde{\theta}$ according to expected posterior quadratic loss.

Solution

(a) Since the mean and variance of the exponential p.d.f. in (4.15) are θ and θ^2 (see Appendix Theorem 2), respectively,

$$E(\bar{Y}) = \theta, \quad \text{Var}(\bar{Y}) = \theta^2/T,$$

and hence

$$\text{MSE}(\bar{Y}) = \theta^2/T.$$

(b) Because $E(\hat{\theta}) = T\theta/(T+1)$,

$$\text{Bias}(\hat{\theta}) = \frac{-\theta}{T+1},$$

and

$$\text{Var}(\hat{\theta}) = \frac{T^2\theta^2}{T(T+1)^2} = \frac{T\theta^2}{(T+1)^2},$$

it follows immediately that

$$\text{MSE}(\hat{\theta}) = \frac{T\theta^2 + \theta^2}{(T+1)^2} = \frac{\theta^2}{T+1}.$$

(c) Clearly, $\text{MSE}(\hat{\theta}) < \text{MSE}(\bar{Y})$. Hence, the MLE \bar{Y} is inadmissible under quadratic loss.

(d) The prior p.d.f. for δ is

$$p(\delta) \propto \delta^{(\underline{\alpha}-1)} \exp(-\delta/\underline{\beta}), \quad (4.17)$$

and the likelihood function is

$$L(\delta) = \delta^T \exp\left(-\delta \sum_{t=1}^T y_t\right) = \delta^T \exp(-\delta T\bar{y}). \quad (4.18)$$

Thus,

$$p(\delta|y) \propto p(\delta)L(\delta) \propto \delta^{(\bar{\alpha}-1)} \exp(-\delta/\bar{\beta}), \quad (4.19)$$

where

$$\bar{\alpha} = \underline{\alpha} + T, \quad \bar{\beta} = [\underline{\beta}^{-1} + T\bar{y}]^{-1}.$$

Recognizing (4.19) as the kernel of a gamma density, it follows that

$$\delta|y \sim G(\bar{\alpha}, \bar{\beta}),$$

or

$$\delta|y \sim \gamma(\bar{\mu}, \bar{\nu}),$$

where $\bar{\mu} = \bar{\alpha}\bar{\beta}$ and $\bar{\nu} = 2\bar{\alpha}$. Remember that there are two different parameterizations of the gamma density and here we use both (see Appendix Definition 2).

(e) Noting that $\partial\delta/\partial\theta = -\theta^{-2}$ and using a change-of-variables theorem [see, e.g., Chapter 4 of Poirier (1995)] gives the posterior p.d.f. for θ :

$$p(\theta|y) \propto \left| \frac{\partial\delta}{\partial\theta} \right| p(\delta|y) \propto \theta^{-2} \theta^{1-\bar{\alpha}} \exp[-1/(\bar{\beta}\theta)] \propto \theta^{-(\bar{\alpha}+1)} \exp[-1/(\bar{\beta}\theta)]. \quad (4.20)$$

Recognizing (4.20) as the kernel of an inverted gamma density (see Appendix Theorem 2), it follows that the posterior p.d.f. of θ is inverted gamma, that is,

$$\theta|y \sim IG(\bar{\alpha}, \bar{\beta}).$$

(f) According to Exercise 3.1 the posterior mean minimizes expected posterior quadratic loss, which, in the case of (4.20) [provided $\bar{\alpha} > 1$], is

$$\tilde{\theta} = [\bar{\beta}(\bar{\alpha} - 1)]^{-1} = \frac{\underline{\beta}^{-1} + T\bar{y}}{\underline{\alpha} + T - 1}. \quad (4.21)$$

(g) Viewed *ex ante* the sampling mean and variance of (4.21) are

$$E_{Y|\theta}(\tilde{\theta}) = \frac{\underline{\beta}^{-1} + T\theta}{\underline{\alpha} + T - 1} \quad (4.22)$$

and

$$\text{Var}_{Y|\theta}(\tilde{\theta}) = \frac{T^2(\theta^2/T)}{(\underline{\alpha} + T - 1)^2} = \frac{T\theta^2}{(\underline{\alpha} + T - 1)^2}. \quad (4.23)$$

From (4.22) the bias of $\tilde{\theta}$ is seen to be

$$\text{Bias}(\tilde{\theta}) = \frac{\underline{\beta}^{-1} + (1 - \underline{\alpha})\theta}{\underline{\alpha} + T - 1}.$$

Therefore the $\text{MSE}(\tilde{\theta})$ is

$$\text{MSE}(\tilde{\theta}) = \frac{T\theta^2 + [\underline{\beta}^{-1} + (1 - \underline{\alpha})\theta]^2}{(\underline{\alpha} + T - 1)^2}. \quad (4.24)$$

When $\underline{\alpha} = 1$, (4.24) becomes

$$\text{MSE}(\tilde{\theta}) = \frac{T\theta^2 + \underline{\beta}^{-2}}{T^2} = \frac{\theta^2}{T} + \frac{\underline{\beta}^{-2}}{T^2}.$$

Therefore,

$$\text{MSE}(\hat{\theta}) < \text{MSE}(\bar{Y}) < \text{MSE}(\tilde{\theta}). \quad (4.25)$$

(h) Given $\underline{\alpha} = 1$, the respective posterior expected quadratic losses (3.10) are, provided $\bar{\alpha} > 2$,

$$\begin{aligned} c(\bar{y}) &= E_{\theta|y}(\bar{y} - \theta)^2 \\ &= \text{Var}(\theta|y) + [E(\theta|y) - \bar{y}]^2 \\ &= \text{Var}(\theta|y) + \left(\frac{\underline{\beta}^{-1} + T\bar{y}}{\underline{\alpha} + T - 1} - \bar{y} \right)^2 \\ &= \text{Var}(\theta|y) + [(\underline{\beta}T)^{-1}]^2 \end{aligned} \quad (4.26)$$

(where we have plugged in $\underline{\alpha} = 1$),

$$\begin{aligned} c(\hat{\theta}) &= E_{\theta|y}(\hat{\theta} - \theta)^2 \\ &= \text{Var}(\theta|y) + \left(\frac{\underline{\beta}^{-1} + T\bar{y}}{\underline{\alpha} + T - 1} - \frac{T\bar{y}}{T + 1} \right)^2 \\ &= \text{Var}(\theta|y) + \left((\underline{\beta}T)^{-1} + \frac{\bar{y}}{T + 1} \right)^2, \end{aligned} \quad (4.27)$$

and

$$\begin{aligned} c(\tilde{\theta}) &= E_{\theta|y}(\tilde{\theta} - \theta)^2 \\ &= \text{Var}(\theta|y). \end{aligned} \quad (4.28)$$

From (4.26)–(4.28) it is clear that $c(\tilde{\theta}) < c(\bar{y}) < c(\hat{\theta})$, reversing the MSE ranking in (4.25).

Exercise 4.3 (Expected bias of Bayesian point estimator) Let $\hat{\theta} = \hat{\theta}(Y)$ be the Bayesian estimator of a scalar parameter θ under quadratic loss based on a prior density $p(\theta)$ with finite prior mean. Let $b(\theta)$ be the bias of $\hat{\theta}$ as an estimator of θ . *A priori*, can you determine the expected sign of $b(\theta)$? If so, why? If not, why not?

Solution

According to Exercise 3.1, under quadratic loss, $\hat{\theta} = E(\theta|y)$. By definition

$$b(\theta) = E_{Y|\theta}[E(\theta|y) - \theta].$$

Using iterated expectations

$$\begin{aligned} E_{\theta}[b(\theta)] &= E_{\theta} (E_{Y|\theta}[E(\theta|y) - \theta]) \\ &= E_{\theta,Y}[E(\theta|y) - \theta] \\ &= E_Y (E_{\theta|y}[E(\theta|y) - \theta]) \\ &= E_Y (E(\theta|y) - E(\theta|y)) \\ &= 0. \end{aligned} \quad (4.29)$$

Therefore, the expected bias of $\hat{\theta}$ is zero.

Exercise 4.4 (Balanced loss in conjugate normal analysis: Zellner [1994]) Let Y_t ($t = 1, 2, \dots, T$) be a random sample from an $N(\theta, \sigma^2)$ distribution with σ^2 known. Consider the *balanced loss function*

$$\begin{aligned} C(\hat{\theta}, \theta) &= w \left[\frac{1}{T} \sum_{t=1}^T (y_t - \hat{\theta})^2 \right] + (1 - w)(\hat{\theta} - \theta)^2 \\ &= w[\hat{\sigma}^2 + (\hat{\theta} - \bar{y})^2] + (1 - w)(\hat{\theta} - \theta)^2, \end{aligned} \quad (4.30)$$

where $\hat{\sigma}^2 = 1/T \sum_{t=1}^T (y_t - \bar{y})^2$. Suppose your prior distribution for θ is $N(\underline{\mu}, \sigma^2/T)$.

- (a) Find the Bayesian point estimate of θ (and denote it as $\hat{\theta}$).
 (b) Find the risk function for $\hat{\theta}$.

Solution

(a) The posterior for θ was reviewed in the solution to Exercise 4.1. Specifically, borrowing results from (4.6) and (4.8), we obtain

$$\theta|y \sim N(\bar{\mu}, \sigma^2/\bar{T}),$$

where

$$\bar{\mu} = (\underline{T}\underline{\mu} + T\bar{y})/\bar{T}, \quad \bar{T} = \underline{T} + T.$$

The posterior expectation of (4.30) is

$$\begin{aligned} E_{\theta|y}[C(\hat{\theta}, \theta)] &= w[\hat{\sigma}^2 + (\hat{\theta} - \bar{y})^2] + (1 - w)E_{\theta|y}[(\hat{\theta} - \theta)^2] \\ &= w[\hat{\sigma}^2 + (\hat{\theta} - \bar{y})^2] + (1 - w) \left(\text{Var}(\theta|y) + [\hat{\theta} - E(\theta|y)]^2 \right), \end{aligned} \quad (4.31)$$

which is quadratic in θ . Differentiating (4.31) with respect to $\hat{\theta}$, equating the result to zero, and solving for $\hat{\theta}$ yields

$$\begin{aligned} \hat{\theta} &= w\bar{y} + (1 - w)E(\theta|y) \\ &= w\bar{y} + (1 - w) \left(\frac{\underline{T}\underline{\mu} + T\bar{y}}{\underline{T} + T} \right) \\ &= \tau\underline{\mu} + (1 - \tau)\bar{y}, \end{aligned} \quad (4.32)$$

where

$$\tau = \frac{(1 - w)\underline{T}}{\underline{T} + T}.$$

(b) Viewing (4.30) *ex ante* and writing $\hat{\theta} - \bar{Y}$ as

$$\hat{\theta} - \bar{Y} = \tau\underline{\mu} + (1 - \tau)\bar{Y} - \bar{Y} = \tau(\underline{\mu} - \bar{Y}),$$

it follows that

$$E_{Y|\theta}[(\hat{\theta} - \bar{Y})^2] = \tau^2 \left[\frac{\sigma^2}{T} + (\theta - \underline{\mu})^2 \right]. \quad (4.33)$$

Also writing $\hat{\theta} - \theta$ as

$$\hat{\theta} - \theta = \tau\underline{\mu} + (1 - \tau)\bar{Y} - \theta = (1 - \tau) \left[\bar{Y} - \left(\frac{\theta - \tau\underline{\mu}}{1 - \tau} \right) \right],$$

it follows, after some tedious algebra, that

$$E_{Y|\theta}(\hat{\theta} - \theta)^2 = (1 - \tau)^2 \left[\frac{\sigma^2}{T} + \frac{\tau^2(\theta - \underline{\mu})^2}{(1 - \tau)^2} \right]. \quad (4.34)$$

From the definition of the loss function in (4.30) and the results in (4.33) and (4.34), we thus have

$$R(\hat{\theta}|\theta) = w \left(\frac{(T-1)\sigma^2}{T} + \tau^2 \left[\frac{\sigma^2}{T} + (\theta - \underline{\mu})^2 \right] \right) \\ + (1-w)(1-\tau)^2 \left[\frac{\sigma^2}{T} + \frac{\tau^2(\theta - \underline{\mu})^2}{(1-\tau)^2} \right].$$

Exercise 4.5 (Empirical Bayes estimation) Suppose $Y_t|\theta, \tau$ ($t = 1, 2, \dots, T$) $\stackrel{\text{iid}}{\sim} N(\theta, 1)$ and $\theta|\tau^2 \sim N(0, \tau^2)$.

(a) Find the marginal distribution of $y = [y_1 \ y_2 \ \dots \ y_T]'$. (We revert to lowercase letters since this exercise is not concerned with describing sampling properties, but, instead, seeks to “optimally” choose the hyperparameter τ as a function of the realized data.)

(b) The idea of using the marginal distribution in (a) to estimate τ underlies the *empirical Bayes approach*. The related *Type II maximum likelihood estimator* is defined to be the value of τ^2 that maximizes the marginal likelihood in (a). Find the Type II maximum likelihood estimator of τ^2 .

Solution

(a) Write

$$y = \theta \iota_T + \epsilon = [\iota_T \ I_T] \begin{bmatrix} \theta \\ \epsilon \end{bmatrix},$$

where

$$\begin{bmatrix} \theta \\ \epsilon \end{bmatrix} \sim N_{T+1} \left(0_{T+1}, \begin{bmatrix} \tau^2 & \iota_T' \\ \iota_T & I_T \end{bmatrix} \right).$$

Then $y|\tau^2 \sim N_T(0_T, \Sigma)$, where

$$\Sigma = [\iota_T \ I_T] \begin{bmatrix} \tau^2 & \iota_T' \\ \iota_T & I_T \end{bmatrix} \begin{bmatrix} \iota_T' \\ I_T \end{bmatrix} = I_T + \tau^2 \iota_T \iota_T'.$$

(b) Consider

$$\ln[p(y|\tau^2)] = -\frac{T}{2} \ln(2\pi) - \frac{1}{2} \ln |\Sigma| - \frac{1}{2} y' \Sigma^{-1} y,$$

where

$$|\Sigma| = 1 + T\tau^2 \quad \text{and} \quad \Sigma^{-1} = I_T - \left(\frac{\tau^2}{1 + T\tau^2} \right) \iota_T \iota_T'.$$

Let $\theta = \theta(\tau^2) \equiv \tau^2/(1 + T\tau^2)$. Then, we note

$$y' \Sigma^{-1} y = \sum_{t=1}^T (y_t - \theta T \bar{y}) y_t = \sum_{t=1}^T y_t^2 - \theta T^2 \bar{y}^2.$$

It follows that

$$\frac{\partial \ln[p(y|\tau^2)]}{\partial \tau^2} = -\frac{T}{2(1 + T\tau^2)} + \frac{T^2\bar{y}^2}{2(1 + T\tau^2)^2}. \quad (4.35)$$

Equating (4.35) to zero and solving for the empirical Bayes estimator of τ^2 yields

$$\hat{\tau}^2 = \begin{cases} \frac{T\bar{y}^2 - 1}{T} & \text{if } \bar{y}^2 > 1/T, \\ 0 & \text{if } \bar{y}^2 \leq 1/T. \end{cases}$$

Exercise 4.6 (Point estimation under linex loss) Consider a random sample Y_t ($t = 1, 2, \dots, T$) from an $N(\mu, \sigma^2)$ population with σ^2 known. Suppose prior beliefs for θ are represented by the normal distribution $\mu \sim N(\underline{\mu}, \sigma^2/\underline{T})$, where $\underline{\mu}$ and \underline{T} are given. By Exercise 2.6(b), the corresponding posterior density for μ is $N(\bar{\mu}, \sigma^2/\bar{T})$, where $\bar{\mu} = (\underline{T}\underline{\mu} + T\bar{y})/\bar{T}$ and $\bar{T} = \underline{T} + T$. To simplify computations we consider the limiting case $\underline{T} \rightarrow 0$ (discussed further in Chapter 8) in which the prior information is weak compared to the sample information, and the posterior simplifies to $N(\bar{y}, \sigma^2/T)$. Finally, consider the linex loss function (3.19).

(a) Let $\hat{\mu}_1$ be the Bayesian point estimate (3.22) subject to expected posterior linex loss. Find its linex risk.

(b) Let $\hat{\mu}_2 = \bar{y}$. Find its linex risk.

Solution

(a) The posterior moment-generating function corresponding to posterior $N(\bar{y}, \sigma^2/T)$ is $M_{\theta|y}(v) = \exp(\bar{y}v + \sigma^2 v^2/2T)$. Hence, Bayesian point estimate (3.22) becomes

$$\begin{aligned} \hat{\mu}_1 &= -a^{-1} \ln[M_{\theta|y}(-a)] \\ &= -a^{-1} \ln[\exp(-a\bar{y} + \sigma^2 a^2/2T)] \\ &= \bar{y} - \frac{a\sigma^2}{2T}. \end{aligned} \quad (4.36)$$

The sampling distribution of $\hat{\mu}_1$ is normal with mean $\mu - (a\sigma^2/2T)$ and variance σ^2/T . Therefore, viewed *ex ante*, the linex risk corresponding to Bayesian estimate (4.36) is

$$\begin{aligned} R(\hat{\mu}_1|\mu) &= E_{Y|\mu}(\exp[a(\hat{\mu}_1 - \mu)] - a(\hat{\mu}_1 - \mu) - 1) \\ &= \exp(-a\mu) M_{\hat{\mu}_1|\mu}(a) - a[E_{Y|\mu}(\hat{\mu}_1) - \mu] - 1 \\ &= \exp(-a\mu) \exp\left(a\left[\mu - \frac{a\sigma^2}{2T}\right] + \frac{a^2\sigma^2}{2T}\right) - a\left(-\frac{a\sigma^2}{2T}\right) - 1 \\ &= \frac{a^2\sigma^2}{2T} - 1. \end{aligned} \quad (4.37)$$

(b) The linex risk corresponding to $\hat{\mu}_2 = \bar{Y}$ is

$$\begin{aligned}
 R(\hat{\mu}_2|\mu) &= E_{Y|\mu} (\exp[a(\bar{Y} - \mu)] - a(\bar{Y} - \mu) - 1) \\
 &= \exp(-a\mu) M_{\bar{Y}|\mu}(a) - a[E_{Y|\mu}(\bar{Y}) - \mu] - 1 \\
 &= \exp(-a\mu) \exp\left(a\mu + \frac{a^2\sigma^2}{2T}\right) + 0 - 1 \\
 &= \exp\left(\frac{a^2\sigma^2}{2T}\right) - 1.
 \end{aligned} \tag{4.38}$$

Because (4.38) is greater than or equal to (4.37), this shows that $\hat{\mu}_2 = \bar{Y}$ is inadmissible with respect to linex loss.

Exercise 4.7 (Risk, Bayes risk, and admissibility: Berger [1985]) Consider a single data point generated from $Y|\theta \sim N(\theta, 1)$, a prior $\theta \sim N(0, 1)$, and the loss function

$$C(\hat{\theta}, \theta) = (\hat{\theta} - \theta)^2 \exp(3\theta^2/4). \tag{4.39}$$

- (a) Find the Bayesian point estimate $\hat{\theta}$ of θ .
- (b) Compare the risk of $\hat{\theta}$ to the risk of $\tilde{\theta} = Y$.
- (c) Compute the Bayes risk of $\hat{\theta}$.

Solution

(a) Using Exercise 2.3 it follows that the posterior distribution of θ is $N(.5y, .5)$. Exercise 3.9 implies that the Bayesian point estimate that minimizes expected posterior loss is

$$\hat{\theta} = \frac{E_{\theta|y}[\theta w(\theta)]}{E_{\theta|y}[w(\theta)]}, \tag{4.40}$$

where $w(\theta) = \exp(3\theta^2/4)$. The denominator in (4.40) is

$$\begin{aligned}
 E_{\theta|y}[\exp(3\theta^2/4)] &= \int_{-\infty}^{\infty} \exp(3\theta^2/4) \frac{1}{\sqrt{2(\pi/2)}} \exp\left[-\frac{1}{2} \frac{(\theta - .5y)^2}{.5}\right] d\theta \\
 &= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2(\pi/2)}} \exp\left[\frac{3\theta^2}{4} - \theta^2 + y\theta - \frac{y^2}{4}\right] d\theta \\
 &= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2(\pi/2)}} \exp\left[-\frac{1}{4}(\theta - 2y)^2 + \frac{3y^2}{4}\right] d\theta \\
 &= 2 \exp\left(\frac{3y^2}{4}\right) \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi(2)}} \exp\left[-\frac{1}{(2)(2)}(\theta - 2y)^2\right] d\theta \\
 &= 2 \exp(3y^2/4),
 \end{aligned} \tag{4.41}$$

where the integral in the second-to-last line in (4.41) is unity because the integrand corresponds to an $N(2y, 2)$ density. Repeating the procedure it follows that the numerator in (4.40) is

$$E_{\theta|y}[\theta \exp(3\theta^2/4)] = 4y \exp(3y^2/4). \tag{4.42}$$

Therefore, the ratio of (4.42) to (4.41) yields the Bayesian point estimate $\hat{\theta} = 2y$.

(b) The risk corresponding to $\hat{\theta}$ is

$$\begin{aligned}
 R(\hat{\theta}|\theta) &= E_{Y|\theta}[\exp(3\theta^2/4)(2Y - \theta)^2] \\
 &= \exp(3\theta^2/4) E_{Y|\theta} [4(Y - (\theta/2))^2] \\
 &= 4 \exp(3\theta^2/4) [\text{Var}(Y|\theta) + ((\theta/2) - \theta)^2] \\
 &= (4 + \theta^2) \exp(3\theta^2/4).
 \end{aligned} \tag{4.43}$$

Similarly, the risk corresponding to $\tilde{\theta}$ is

$$\begin{aligned}
 R(\tilde{\theta}|\theta) &= E_{Y|\theta}[\exp(3\theta^2/4)(Y - \theta)^2] \\
 &= \exp(3\theta^2/4) \text{Var}(Y|\theta) \\
 &= \exp(3\theta^2/4).
 \end{aligned} \tag{4.44}$$

Comparing (4.43) to (4.44) we see that the Bayesian point estimator is inadmissible under loss function (4.39) despite the fact that the prior $\theta \sim N(0, 1)$ is proper.

(c) The counterintuitive result in (b) can be explained by noting that the Bayes risk is

$$\begin{aligned}
 E_{\theta}[R(\hat{\theta}|\theta)] &= \int_{-\infty}^{\infty} (4 + \theta^2) \exp(3\theta^2/4) \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}\theta^2\right) d\theta \\
 &= \int_{-\infty}^{\infty} (4 + \theta^2) \frac{1}{\sqrt{2\pi}} \exp\left(\frac{1}{4}\theta^2\right) d\theta \rightarrow \infty,
 \end{aligned} \tag{4.45}$$

which diverges since the integrand approaches ∞ as $|\theta| \rightarrow \infty$. Note that in this case it is not permissible to switch the order of integration in (4.3)–(4.5).

Exercise 4.8 (Likelihood principle) Consider two researchers, A and B . Each researcher must estimate the same unknown parameter $\theta < 0$ subject to the same prior $p(\theta)$ and the same loss function $C(\hat{\theta}, \theta)$. Researcher A observes a random variable X having the gamma distribution $G(3, \theta^{-1})$ with density

$$p(x|\theta) = \left(\frac{\theta^3}{\Gamma(3)}\right) x^2 \exp(-\theta x), \quad x > 0.$$

Researcher B observes a random variable Y having the Poisson distribution $Po(2\theta)$ with mass function

$$p(y|\theta) = \frac{(2\theta)^y \exp(-2\theta)}{y!}, \quad y = 0, 1, 2, 3, \dots$$

In collecting one observation each, researcher A observes $X = 2$ and researcher B observes $Y = 3$. Compare and contrast Bayesian and maximum likelihood estimation for each researcher.

Solution

The respective likelihood functions are

$$L_A(\theta; x = 2) = \left(\frac{\theta^3}{\Gamma(3)}\right) 2^2 \exp(-2\theta) \propto \theta^3 \exp(-2\theta)$$

and

$$L_B(\theta; y = 3) = \frac{(2\theta)^3 \exp(-2\theta)}{3!} \propto \theta^3 \exp(-2\theta).$$

Since these likelihood functions are proportional, it follows from the likelihood principle (see Chapter 2) that the evidence in the data about θ is the same in both data sets. Because the priors and loss functions are the same, from the Bayesian standpoint the two researchers would obtain the same posterior distributions and the same point estimates. From the standpoint of maximum likelihood estimation, since the likelihoods are proportional, the maximum likelihood point *estimates* would be the same ($\theta = 1.5$). However, the maximum likelihood *estimators* would be very different. In the case of researcher A the sampling distribution would be continuous, whereas in the case of researcher B the sampling distribution would be discrete.

Interval estimation

From the Bayesian standpoint, given a region $C \subset \Theta$ and the data y , it is meaningful to ask the following: What is the *probability* that θ lies in C ? The answer is direct:

$$1 - \alpha \equiv \Pr(\theta \in C|y) = \int_C p(\theta|y)d\theta, \quad (5.1)$$

where $0 < \alpha < 1$ is defined implicitly in (5.1). The region C is known as a $1 - \alpha$ *Bayesian credible region*. There is no need to introduce the additional frequentist concept of “confidence.”

This chapter focuses on the case in which the posterior probability content is first set at some preassigned α (say $\alpha = .10, .05$, or $.01$) and then the “smallest” credible region that attains posterior probability content of $1 - \alpha$ is sought. This leads to the following definition.

Definition 5.1 Let $p(\theta|y)$ be a posterior density function. Let $\Theta^* \subset \Theta$ satisfying:

- (a) $\Pr(\theta \in \Theta^*|y) = 1 - \alpha$,
- (b) for all $\theta_1 \in \Theta^*$ and $\theta_2 \notin \Theta^*$, $p(\theta_1|y) \geq p(\theta_2|y)$.

Then Θ^* is defined to be a *highest posterior density (HPD) region of content* $(1 - \alpha)$ for θ .

Given a probability of content $1 - \alpha$, the HPD region Θ^* has the smallest possible volume in the parameter space Θ of any $1 - \alpha$ Bayesian credible region. If $p(\theta|y)$ is not uniform over Θ , then the HPD region of content $1 - \alpha$ is unique. Hereafter, we focus on the case in which there is a single parameter of interest and all other parameters have been integrated out of the posterior.

Constructing an HPD interval is conceptually straightforward. Part (b) of Definition 5.1 implies that, if $[a, b]$ is an HPD interval, then $p(a|y) = p(b|y)$. This suggests a graphical approach in which a horizontal line is gradually moved vertically downward across the posterior density, and where it intersects the posterior density, the corresponding abscissa

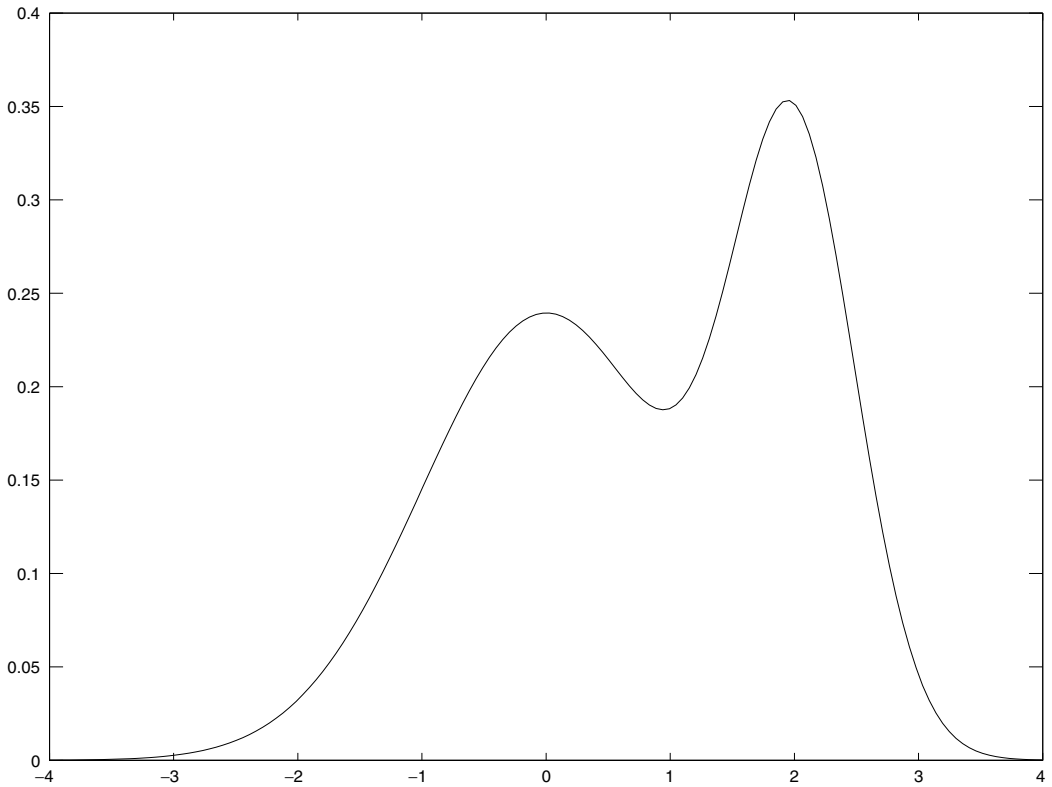


Figure 5.1 — A Hypothetical Posterior Density.

values are noted, and the posterior is integrated between these points. Once the desired posterior probability is reached, the process stops. If the posterior density is symmetric and unimodal with mode θ^m , then the resulting $1-\alpha$ HPD interval is of the form $[\theta^m - \delta, \theta^m + \delta]$, for suitable δ , and it cuts off equal probability $\alpha/2$ in each tail.

In (5.1) we started with the region and then found its posterior probability content. In Definition 5.1 we started with the posterior probability content and then found the smallest region with that content. From a pure decision theoretic perspective, it is interesting to start with a cost function measuring the undesirability of large α and large volume of a region, and then to pick both α and C to minimize expected posterior cost [see Casella, Hwang, and Robert (1993)]. For a thorough discussion of both Bayesian and frequentist interval estimation, see Casella and Berger (2002).

Exercise 5.1 (Graphical determination of an HPD interval) Consider the posterior density in Figure 5.1. Use the graphical approach described earlier to obtain three sets of HPD intervals.

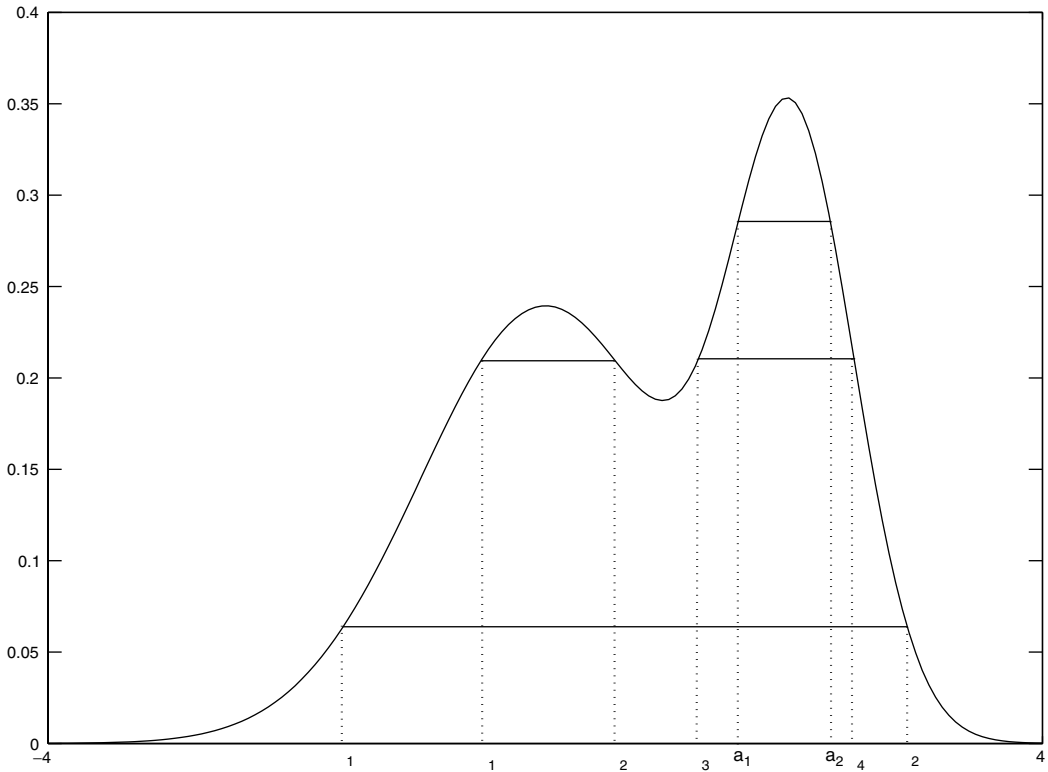


Figure 5.2 — Three different HPD intervals.

Solution

In Figure 5.2, we present three different HPD intervals associated with the posterior in 5.1. The highest line in the figure produces the interval $[a_1, a_2]$. This interval is narrow and contains rather low posterior probability. If this process is repeated a second time the horizontal line produces the disjoint intervals $[b_1, b_2]$ and $[b_3, b_4]$, which contain more posterior probability, but still relatively little. If these two intervals, however, contain the desired posterior probability, then the process stops and the HPD region consists of the union of these two disjoint intervals. If a higher posterior content is desired, the process repeats again until, say, the third horizontal line that produces the interval $[c_1, c_2]$. If even more posterior content is desired, the process can be repeated again.

Exercise 5.2 (Calculating a posterior interval probability) Consider a random sample of size $T = 2$ from an exponential distribution with mean θ^{-1} and a gamma prior distribution $G(\underline{\alpha}, \underline{\beta})$ with $\underline{\alpha} = 2$ and $\underline{\beta} = 4$. Suppose the resulting sample mean is $\bar{y} = .125$. Find the posterior probability of the credible interval $[3.49, 15.5]$.

Solution

From Exercise 2.10 it follows that the posterior distribution of θ is $G(\bar{\alpha}, \bar{\beta})$, where $\bar{\alpha} = 2 + 2 = 4$ and $\bar{\beta} = [4^{-1} + 2(.125)]^{-1} = 2$. $\bar{\beta} = 2$ implies that the gamma distribution reduces to a chi-square distribution with eight degrees of freedom [see Poirier (1995, p. 100)]. Consulting a standard chi-square table [e.g., Poirier (1995, p. 662)], we see that $\Pr(\chi^2(8) \leq 3.49) = .10$ and $\Pr(\chi^2(8) \leq 15.5) = .95$. Therefore, the posterior probability of $[3.49, 15.5]$ is $.95 - .10 = .85$.

Exercise 5.3 (HPD intervals under normal random sampling with known variance)

Consider a random sample Y_1, Y_2, \dots, Y_T from an $N(\theta_1, \theta_2^{-1})$ distribution with *known* variance θ_2^{-1} and an $N(\underline{\mu}, \theta_2^{-1}/\underline{T})$ prior for the unknown mean θ_1 .

(a) Find the $1 - \alpha$ HPD interval for θ_1 .

(b) Suppose $\alpha = .05$, $\theta_2 = .01$, $\underline{\mu} = 8$, $\underline{T} = 5$, $T = 25$, and $\bar{y} = 2$. Find the $1 - \alpha$ HPD interval for θ_1 .

(c) Consider the limiting prior as $\underline{T} \rightarrow 0$ in Exercise 4.6. Find the $1 - \alpha$ HPD interval for θ_1 .

(d) Suppose $\alpha = .05$, $\theta_2 = .01$, $T = 25$, and $\bar{y} = 2$. Find a 95 percent *confidence* interval and interpret your result.

Solution

(a) From Exercise 2.6(b) it follows that the posterior density for θ_1 is

$$p(\theta_1|y) = \phi(\theta_1; \bar{\mu}, \theta_2^{-1}/\bar{T}),$$

where \bar{T} and $\bar{\mu}$ are given by (2.38) and (2.39), respectively. Since the posterior density is symmetric and unimodal with mode $\bar{\mu}$, the resulting HPD interval of content $1 - \alpha$ cuts off equal probability $\alpha/2$ in each tail. Thus, a $1 - \alpha$ HPD interval for θ_1 is

$$\bar{\mu} - z_{\alpha/2}(\theta_2^{-1}/\bar{T})^{1/2} < \theta_1 < \bar{\mu} + z_{\alpha/2}(\theta_2^{-1}/\bar{T})^{1/2}, \quad (5.2)$$

where $z_{\alpha/2}$ is the standard normal critical value that cuts off $\alpha/2$ probability in the right-hand tail. From the standpoint of subjective probability, (5.2) implies that the *posterior probability* of θ_1 lying between $\bar{\mu} \pm z_{\alpha/2}(\theta_2^{-1}/\bar{T})^{1/2}$ is .95.

(b) The given values correspond to an $N(8, 20)$ prior distribution for θ_1 . Using a standard normal table, it follows that $z_{\alpha/2} = 1.96$. Plugging into (2.38) and (2.39) we obtain $\bar{T} = \underline{T} + T = 5 + 25 = 30$ and $\bar{\mu} = (\underline{T}\underline{\mu} + T\bar{y})/\bar{T} = [5(8) + 25(2)]/30 = 3$. Also note that $\theta_2^{-1}/\bar{T} = 100/30 = 10/3$. Hence the posterior distribution for θ_1 is $N(3, 10/3)$ and (5.2) implies the .95 HPD interval $3 \pm 1.96(10/3)^{1/2}$ or $-.58 < \theta_1 < 6.58$.

(c) From Exercise 2.6 it is seen that the posterior hyperparameters simplify to $\bar{\mu} = \bar{y}$ and $\bar{T} = T$. Therefore, HPD interval (5.2) reduces to

$$\bar{y} - z_{\alpha/2}(\theta_2^{-1}/T)^{1/2} < \theta_1 < \bar{y} + z_{\alpha/2}(\theta_2^{-1}/T)^{1/2}, \quad (5.3)$$

which is numerically identical to the standard non-Bayesian confidence interval. The interpretation of (5.3) is, however, in terms of “probability” and not “confidence”: Given $\bar{Y} = \bar{y}$, the *ex post* probability of θ_1 falling in (5.2) is .95. The common misinterpretation of a classical confidence interval in terms of final precision is fortuitously correct in this special case.

(d) Plugging into (5.3) implies the .95 confidence interval

$$2 \pm 1.96(100/25)^{1/2} \quad \text{or} \quad -1.92 < \theta < 5.92.$$

The interpretation of this confidence interval is that it is a realization of a procedure that, upon repeated sampling, yields random intervals $[\bar{Y} - 3.92, \bar{Y} + 3.92]$ that have an *ex ante* sampling probability of .95 of capturing θ_1 . The realized interval $-1.92 < \theta < 5.92$ either does or does not capture the unknown constant θ_1 . The frequentist “confidence” lies in the procedure used to obtain the realized interval.

Exercise 5.4 (HPD intervals under normal random sampling with unknown variance) Consider Exercise 5.3 but with θ_2 *unknown*. Suppose the joint prior distribution for $\theta = [\theta_1 \ \theta_2]'$ is the normal-gamma distribution, denoted $\theta \sim NG(\underline{\mu}, \underline{q}, \underline{s}^{-2}, \underline{\nu})$ with density given in (2.26).

(a) Find the $1 - \alpha$ HPD interval for θ_1 .

(b) Suppose $\underline{\mu} = 8$, $\underline{q} = .20$, $\underline{s}^{-2} = .25$, and $\underline{\nu} = 4$. Find the marginal prior density for θ_1 . How does it compare to the prior in Exercise 5.3(b)?

(c) Suppose $\alpha = .05$, $T = 25$, $\bar{y} = 2$, and $s^2 = 102.5$. Find the $1 - \alpha$ HPD interval for θ_1 .

(d) Using the values in (c), find the $1 - \alpha$ HPD interval for θ_2 .

Solution

(a) The posterior distribution for θ is $NG(\bar{\mu}, \bar{q}, \bar{s}^{-2}, \bar{\nu})$, as given in the solution to Exercise 2.4. By using the solution to Exercise 2.8, it follows immediately that the marginal posterior distribution of θ_1 is $t(\theta_1; \bar{\mu}, \bar{s}^2 \bar{q}, \bar{\nu})$. Hence, analogous to (5.2), a $1 - \alpha$ HPD interval for θ_1 is

$$\bar{\mu} - t_{\bar{\nu}, \alpha/2}(\bar{s}^2 \bar{q})^{1/2} < \theta_1 < \bar{\mu} + t_{\bar{\nu}, \alpha/2}(\bar{s}^2 \bar{q})^{1/2}, \quad (5.4)$$

where $t_{\bar{\nu}, \alpha/2}$ cuts off $\alpha/2$ probability in the right hand tail of a Student t -distribution with $\bar{\nu}$ degrees of freedom. From the standpoint of subjective probability, (5.4) implies that the *posterior probability* of θ_1 lying between $\bar{\mu} \pm t_{\bar{\nu}, \alpha/2}(\bar{s}^2 \bar{q})^{1/2}$ is .95.

(b) Again using the solution to Exercise 2.8, the given hyperparameter values imply that the marginal prior distribution for θ_1 is $t(\theta_1; 8, 20, 4)$. This is similar to the prior in Exercise 5.3: It has the same location but fatter tails.

(c) Plugging the given hyperparameter values into (2.27) and (2.29)–(2.31), we obtain

$$\begin{aligned}\bar{\nu} &= 4 + 25 = 29, \\ \bar{q} &= [.20^{-1} + 25]^{-1} = 1/30, \\ \bar{\mu} &= [(.20)^{-1}(8) + (25)(2)]/30 = 3,\end{aligned}$$

and

$$\bar{s}^2 = 29^{-1}[4(.25) + 24(102.5) + (.20 + 25^{-1})^{-1}(2 - 8)^2] = 90.03.$$

Hence, the posterior distribution for θ_1 is $t(\theta_1; 3, 3.00, 29)$. Using a t -table gives $t_{29}(.975) = 2.045$. Therefore, (5.4) implies the .95 HPD interval

$$3 \pm 2.045[3.00]^{1/2} \quad \text{or} \quad -.54 < \theta_1 < 6.54,$$

similar to the solution of Exercise 5.3(b).

(d) From the definition of the normal-gamma distribution, the marginal posterior distribution for the precision θ_2 is $\gamma(\bar{s}^{-2}, \bar{\nu}) = \gamma(.011, 29)$. Table 5.1 provides values of the p.d.f. and c.d.f. of a $\gamma(.011, 29)$ distribution. From inspection we see that the interval $[a, b]$ with equal p.d.f. ordinates and that contains .95 posterior probability is $[\mathbf{.005736}, \mathbf{.1693}]$.

Table 5.1: p.d.f and c.d.f. values
of a $\gamma(.011, 29)$ random variable.

θ_2	p.d.f.	c.d.f.
.00200	.0018	.0000
.00270	.0405	.0000
.00300	.1136	.0000
.00390	1.212	.0005
.00550	15.55	.0106
.005736	20.15	.0148
.00600	26.21	.0209
.00800	93.59	.1368
.00900	124.4	.2468
.01030	140.9	.4225
.01080	139.1	.4927
.01300	96.14	.7586
.01500	48.75	.9015
.01666	23.02	.9591
.01693	20.15	.9648
.01720	17.50	.9700
.02000	3.466	.9948
.02160	1.213	.9983
.02620	.0405	.9999

Exercise 5.5 (Berger and Wolpert [1988, pp. 5–6]) To further clarify the difference between a Bayesian posterior interval and a confidence interval, consider the following. Given an unknown parameter θ , $-\infty < \theta < \infty$, suppose Y_i ($i = 1, 2$) are iid binary random variables with probabilities of occurrence equally distributed over points of support $\theta - 1$ and $\theta + 1$. That is, $\Pr(Y_i = \theta - 1|\theta) = 1/2$ and $\Pr(Y_i = \theta + 1|\theta) = 1/2$.

Suppose the prior for θ is constant. Such an “improper” prior will be discussed in Chapter 8.

- (a) Suppose we observe $y_1 \neq y_2$. Find the posterior distribution of θ .
- (b) Suppose we observe $y_1 = y_2$. Find the posterior distribution of θ .
- (c) Consider

$$\hat{\theta} = \begin{cases} (1/2)(Y_1 + Y_2) & \text{if } Y_1 \neq Y_2, \\ Y_1 - 1 & \text{if } Y_1 = Y_2. \end{cases} \quad (5.5)$$

In repeated sampling, what is the *ex ante* probability that (5.5) contains θ ? Note that the answer is the same if the second line of (5.5) is changed to $Y_1 + 1$ if $Y_1 = Y_2$.

Solution

(a) If $y_1 \neq y_2$, then one of the values must equal $\theta - 1$ and the other equals $\theta + 1$. *Ex post*, averaging these two values it is *absolutely certain* that $\theta = (1/2)(y_1 + y_2)$; that is, $\Pr[\theta = (1/2)(y_1 + y_2)|y_1, y_2] = 1$.

(b) If $y_1 = y_2 = y$, say, then the common value y is either $\theta - 1$ or $\theta + 1$. Since the prior does not distinguish among values of θ , *ex post*, it is *equally uncertain* whether $\theta = y - 1$ or $\theta = y + 1$; that is, $\Pr(\theta = y - 1|y_1, y_2) = \Pr(\theta = y + 1|y_1, y_2) = 1/2$.

(c) Parts (a) and (b) suggest the *ex post* (i.e., posterior) probability that (5.5) equals θ is either 1 or 1/2, depending on whether $y_1 \neq y_2$ or $y_1 = y_2$. Given the data, we can, of course, determine which posterior probability applies. From the *ex ante* perspective, however, $\Pr(Y_1 \neq Y_2) = \Pr(Y_1 = Y_2) = 1/2$. Therefore, *ex ante*, the probability that (5.5) contains θ is an equally weighted average of our two *ex post* coverage probabilities, and thus the *ex ante* sampling probability of (5.5) containing θ is $(1/2)(1) + (1/2)(1/2) = .75$. The embarrassing question for the pure frequentist is this: Why use the realized value of (5.5) to estimate θ and then report the *ex ante* confidence level 75 percent instead of the appropriate *ex post* measure of uncertainty?

Exercise 5.6 (Conditional frequentist reasoning and ancillarity) The preceding exercise demonstrates the important difference between frequentist *ex ante* and Bayesian *ex post* reasoning. The latter is consistent with the likelihood principle [see Berger and Wolpert (1988)], which, loosely speaking, states that two experiments involving the same unknown parameter θ that give rise to proportional likelihood functions contain the same evidence about θ . Likelihood principle proponents condition on all the data, whereas pure frequentist reasoning averages unconditionally over all the data that could have possibly been observed. *Conditional frequentists* lie somewhere between likelihood principle proponents and pure frequentists, arguing that inference should be conditional on ancillary statistics

(statistics whose distribution does not depend on θ), but otherwise be unconditional. Conditional inference goes a long way toward eliminating some of the embarrassing problems in the pure frequentist approach. Demonstrate this observation using the design of Exercise 5.5 in terms of the statistic $Z = |Y_1 - Y_2|$.

Solution

First, note that Z is ancillary since its p.m.f. is $\Pr(Z = 0) = \Pr(Z = 2) = 1/2$, which does not depend on θ . The coverage probabilities of (5.5), conditioned on Z , are the appealing *ex post* probabilities of .5 and 1, respectively, for $Z = 0$ and $Z = 2$. Therefore, the conditional frequentist inference in this case is identical to the Bayesian posterior probabilities.

Exercise 5.7 (HPD intervals and reparameterization) Given the likelihood $L(\theta)$ and the prior p.d.f. $p(\theta)$, let $\gamma = g(\theta)$ be a one-to-one transformation of θ . Also, let A be a $1 - \alpha$ HPD region for θ and define $B = \{\gamma : \gamma = g(\theta), \theta \in A\}$. Is B a $1 - \alpha$ HPD region for γ ?

Solution

No. Although condition (a) in Definition 5.1 holds, condition (b) does not. To see this consider the following. The fact that A is a $1 - \alpha$ HPD interval for θ implies that if $\theta_1 \in A$ and $\theta_2 \notin A$, then $p(\theta_1|y) \geq p(\theta_2|y)$. By definition of B , $\gamma_1 \in B$ iff $\theta = g^{-1}(\gamma_1) \in A$. Suppose $\gamma_1 \in B$, and $\gamma_2 \notin B$. By a change-of-variables,

$$p(\gamma_1|y) = \left| \frac{\partial g^{-1}(\gamma_1)}{\partial \gamma} \right| p_{\theta|y}[g^{-1}(\gamma_1)],$$

$$p(\gamma_2|y) = \left| \frac{\partial g^{-1}(\gamma_2)}{\partial \gamma} \right| p_{\theta|y}[g^{-1}(\gamma_2)],$$

$\gamma_2 \notin B$ iff $\theta_2 = g^{-1}(\gamma_2) \notin A$. Therefore, although $p_{\theta|y}[g^{-1}(\gamma_1)] \geq p_{\theta|y}[g^{-1}(\gamma_2)]$, the Jacobian terms cannot be ordered. In other words, the posterior probability content is maintained under reparameterization, but the minimal length is not.

Hypothesis testing

Suppose the relevant *decision space* is $D = \{d_1, d_2\}$, where d_j denotes choice of hypothesis H_j ($j = 1, 2$). Let the unsubscripted “d” denote a generic decision (i.e., either d_1 or d_2). Extensions to cases involving more than two hypotheses are straightforward. Let $C(d; \theta) \geq 0$ denote the relevant loss function describing the consequences of decision d when the unknown state of nature is θ . From the subjectivist perspective, $C(d; \theta)$ is random because uncertainty regarding θ is expressed by a distribution function.

As in the case of estimation, the optimal decision d_* in the hypothesis-testing context minimizes expected cost. After observing data y yielding likelihood $L(\theta)$, the relevant expression of uncertainty is the posterior distribution. Therefore, d_* is defined by

$$d_* = \underset{d}{\operatorname{argmin}} [c(d|y)], \quad (6.1)$$

where $c(d|y)$, the *posterior expected cost (loss) of taking decision d*, is

$$c(d|y) = E_{\theta|y}[C(d; \theta)] = E[C(d; \theta)|y]. \quad (6.2)$$

From the Bayesian perspective, a hypothesis is of interest only if the prior distribution assigns it positive probability. Therefore, assume

$$\pi_j = \Pr(H_j) = \Pr(\theta \in \Theta_j) > 0, \quad j = 1, 2, \quad (6.3)$$

with $\pi_1 + \pi_2 = 1$. The prior probability function [density/(mass function)] can be decomposed as

$$p(\theta) = \begin{cases} \pi_1 p(\theta|H_1) & \text{if } \theta \in \Theta_1, \\ \pi_2 p(\theta|H_2) & \text{if } \theta \in \Theta_2, \end{cases} \quad (6.4)$$

where $p(\theta|H_j)$ is the prior probability function under H_j ($j = 1, 2$).

Under H_j the probability function of the data, marginal of all parameters (known as the marginal likelihood), is

$$p(y|H_j) = \int_{\Theta_j} L(\theta) dP(\theta|H_j) = E_{\theta|H_j}[L(\theta)], \quad j = 1, 2. \quad (6.5)$$

From Bayes' theorem it follows immediately that the posterior probability of H_j is

$$\bar{\pi}_j = \Pr(H_j|y) = \frac{\pi_j p(y|H_j)}{p(y)} \quad j = 1, 2, \quad (6.6)$$

where the marginal probability function of the data (marginal of *both* parameters and hypotheses) is

$$p(y) = \pi_1 p(y|H_1) + \pi_2 p(y|H_2), \quad (6.7)$$

noting, of course, that $\bar{\pi}_1 + \bar{\pi}_2 = 1$.

With these preliminaries out of the way, we can now derive the posterior distribution for θ to be used in (6.2). Under H_j , the posterior probability function of θ is (according to Bayes' theorem)

$$p(\theta|y, H_j) = \frac{p(\theta|H_j)L(\theta)}{p(y|H_j)}, \quad \theta \in \Theta_j, \quad j = 1, 2. \quad (6.8)$$

Using posterior probabilities (6.6), the marginal posterior probability function of θ is

$$p(\theta|y) = \begin{cases} \bar{\pi}_1 p(\theta|y, H_1) & \text{if } \theta \in \Theta_1, \\ \bar{\pi}_2 p(\theta|y, H_2) & \text{if } \theta \in \Theta_2. \end{cases} \quad (6.9)$$

Using (6.9) and defining the *expected posterior loss of decision d given H_j* as

$$c(d|y, H_j) = E_{\theta|y, H_j}[C(d; \theta)], \quad j = 1, 2, \quad (6.10)$$

we can write posterior expected loss (6.2) as

$$c(d|y) = \bar{\pi}_1 c(d|y, H_1) + \bar{\pi}_2 c(d|y, H_2). \quad (6.11)$$

Without loss of generality, assume that the correct decisions yield zero loss:

$$C(d_1; \theta) = 0 \text{ if } \theta \in \Theta_1, \quad (6.12)$$

$$C(d_2; \theta) = 0 \text{ if } \theta \in \Theta_2. \quad (6.13)$$

Under (6.12) and (6.13) it follows from (6.11) that

$$c(d_1|y) = \bar{\pi}_2 c(d_1|y, H_2), \quad (6.14)$$

$$c(d_2|y) = \bar{\pi}_1 c(d_2|y, H_1). \quad (6.15)$$

Therefore, it is optimal to choose H_2 [i.e., $c(d_2|y) < c(d_1|y)$], denoted $d_* = d_2$, iff

$$\frac{\bar{\pi}_2}{\bar{\pi}_1} > \frac{c(d_2|y, H_1)}{c(d_1|y, H_2)}. \quad (6.16)$$

Because $\bar{\pi}_2 = 1 - \bar{\pi}_1$, (6.16) can also be written as

$$d_* = d_2 \text{ iff } \bar{\pi}_1 < \left[1 + \frac{c(d_2|y, H_1)}{c(d_1|y, H_2)} \right]^{-1}. \quad (6.17)$$

The quantities π_2/π_1 and $\bar{\pi}_2/\bar{\pi}_1$ are the *prior odds* and *posterior odds*, respectively, of H_2 versus H_1 . From (6.6) it follows immediately that these two odds are related by

$$\frac{\bar{\pi}_2}{\bar{\pi}_1} = B_{21} \left(\frac{\pi_2}{\pi_1} \right), \quad (6.18)$$

where

$$B_{21} = \frac{p(y|H_2)}{p(y|H_1)} \quad (6.19)$$

is known as the *Bayes factor for H_2 versus H_1* . Bayes factor (6.19) is a ratio of marginal likelihoods (see Equation 6.5). Also note that (6.18) implies that the posterior probabilities are related to prior odds by

$$\bar{\pi}_1 = \left[1 + B_{21} \left(\frac{\pi_2}{\pi_1} \right) \right]^{-1} \quad (6.20)$$

and

$$\bar{\pi}_2 = \left[1 + B_{12} \left(\frac{\pi_1}{\pi_2} \right) \right]^{-1}, \quad (6.21)$$

where $B_{12} = B_{21}^{-1}$. Finally, substituting (6.18) into (6.16) and rearranging, yields

$$d_* = d_2 \text{ if } B_{21} \geq \left[\frac{c(d_2|y, H_1)}{c(d_1|y, H_2)} \right] \left[\frac{\pi_1}{\pi_2} \right]. \quad (6.22)$$

In general expected posterior loss $c(d|y, H_j)$ defined by (6.10) depends on the data y , and hence Bayes factors (6.19) do not serve as complete data summaries. In other words, because the right-hand side of the inequality in (6.22) depends on the data, it cannot be interpreted as a “critical value” analogous to those that appear in frequentist testing. One exception is when both hypotheses are simple (see Exercise 6.1). The following shows another exception.

Suppose the loss resulting from decision d_i when $\theta \in \Theta_j$, $i \neq j$, is constant for all $\theta \in \Theta_j$:

$$C(d_1; \theta) = \bar{C}_1 \text{ for all } \theta \in \Theta_2 \quad (6.23)$$

and

$$C(d_2; \theta) = \bar{C}_2 \text{ for all } \theta \in \Theta_1, \quad (6.24)$$

where \bar{C}_1 and \bar{C}_2 are given constants. Under (6.23) and (6.24), expected posterior losses (6.10) depend neither on the data nor on the prior for θ :

$$\begin{aligned} c(d_i|y, H_j) &= E_{\theta|y, H_j}[C(d; \theta)] \\ &= \int_{\Theta_j} \bar{C}_i p(\theta|y, H_j) d\theta = \bar{C}_i \text{ for } i \neq j. \end{aligned} \quad (6.25)$$

Therefore, decision rule (6.22) reduces to

$$d_* = d_2 \text{ iff } B_{21} > \frac{\pi_1 \bar{C}_2}{\pi_2 \bar{C}_1}. \quad (6.26)$$

The right-hand side of the inequality in (6.26) is a known constant that serves as a *Bayesian critical value*.

How reasonable is the loss structure of (6.23) and (6.24)? Although it holds trivially when both hypotheses are simple, in the case of composite hypotheses, however, its appropriateness is suspect. For example, when testing $H_1 : \theta = 0$ versus $H_2 : \theta > 0$, should not the loss associated with d_1 depend on how far θ under H_2 deviates from $\theta = 0$? Kadane and Dickey (1980) argue in the affirmative.

Exercise 6.1 (Two simple hypotheses and general likelihood) Suppose H_1 and H_2 are both simple hypotheses, with $\Theta_j = \{\theta_j\}$, $j = 1, 2$, so that prior probability function (6.4) reduces to the simple two-point mass function

$$p(\theta) = \begin{cases} \pi_1 & \text{if } \theta = \theta_1, \\ \pi_2 & \text{if } \theta = \theta_2. \end{cases} \quad (6.27)$$

Consider data y yielding likelihood $L(\theta)$. Under what conditions is it optimal to reject H_1 ?

Solution

Posterior probability function (6.9) reduces to the two-point mass function

$$p(\theta|y) = \begin{cases} \bar{\pi}_1 & \text{if } \theta = \theta_1, \\ \bar{\pi}_2 & \text{if } \theta = \theta_2, \end{cases} \quad (6.28)$$

where the posterior probabilities of each hypothesis (see Equation 6.6) are simply

$$\bar{\pi}_j = \frac{\pi_j L(\theta_j)}{p(y)}, \quad j = 1, 2, \quad (6.29)$$

and the marginal p.d.f. (6.7) of the data is

$$p(y) = \pi_1 L(\theta_1) + \pi_2 L(\theta_2). \quad (6.30)$$

Also note that Bayes factor (6.19) reduces to the simple likelihood ratio

$$B_{21} = \frac{L(\theta_2)}{L(\theta_1)}. \quad (6.31)$$

In the case of simple hypotheses

$$c(d_1|y, H_2) = C(d_1; \theta_2), \quad (6.32)$$

$$c(d_2|y, H_1) = C(d_2; \theta_1). \quad (6.33)$$

Therefore, from (6.22) and (6.31)–(6.33) it follows that

$$d_* = d_2 \text{ iff } \frac{L(\theta_2)}{L(\theta_1)} > \left[\frac{C(d_2; \theta_1)}{C(d_1; \theta_2)} \right] \left[\frac{\pi_1}{\pi_2} \right]. \quad (6.34)$$

In other words, choose d_2 iff the likelihood ratio of H_2 versus H_1 is greater than the ratio of the prior expected loss from choosing d_2 relative to the prior expected loss from choosing d_1 .

Exercise 6.2 (Two simple hypotheses and normal likelihoods) Suppose a random sample Y_1, Y_2, \dots, Y_T is available from an $N(\theta, \sigma^2)$ population in which θ is unknown and σ^2 is *known*. Consider Exercise 6.1 where it is desired to test the simple hypothesis $H_1 : \theta = \theta_1$ versus the simple hypothesis $H_2 : \theta = \theta_2$ with $\theta_1 < \theta_2$. Compute (6.34) and simplify your answer.

Solution

Analogous to (2.20) and (2.21), factor the likelihood as

$$L(\theta) = c_1 \phi(\bar{y}|\theta, h^{-1}), \quad (6.35)$$

where $c_1 = (2\pi\sigma^2)^{-(\nu/2)} T^{-(1/2)} \exp(-\nu s^2/[2\sigma^2])$, $h = T/\sigma^2$, and $\nu s^2 = \sum_{t=1}^T (y_t - \bar{y})^2$. Because c_1 does not depend on the unknown θ , (6.35) shows that the sample mean \bar{y} is a sufficient statistic for θ . According to the Likelihood principle (see Chapter 2), no information is lost by ignoring c_1 and merely considering the second factor. Similarly, we can work with $c(d|\bar{y}, H_j)$ instead of $c(d|y, H_j)$. The Bayes factor for H_2 versus H_1 is

$$\begin{aligned} B_{21} &= \frac{L(\theta_2)}{L(\theta_1)} \\ &= \frac{(2\pi\sigma^2/T)^{(-1/2)} \exp[-T(\bar{y} - \theta_2)^2/2\sigma^2]}{(2\pi\sigma^2/T)^{(-1/2)} \exp[-T(\bar{y} - \theta_1)^2/2\sigma^2]} \\ &= \exp\left(-\frac{1}{2}h[(\bar{y} - \theta_2)^2 - (\bar{y} - \theta_1)^2]\right) \\ &= \exp\left(\frac{1}{2}h[2\bar{y}(\theta_2 - \theta_1) + (\theta_1^2 - \theta_2^2)]\right). \end{aligned} \quad (6.36)$$

According to (6.34), H_1 should be rejected iff

$$\exp\left(\frac{1}{2}h[2\bar{y}(\theta_2 - \theta_1) + (\theta_1^2 - \theta_2^2)]\right) > \frac{\bar{C}_{2\pi_1}}{\bar{C}_{1\pi_2}},$$

or equivalently iff

$$\bar{y} > \frac{\theta_1 + \theta_2}{2} + [h(\theta_2 - \theta_1)]^{-1} \ln\left(\frac{\bar{C}_{2\pi_1}}{\bar{C}_{1\pi_2}}\right).$$

***Exercise 6.3 (Two one-sided hypotheses and normal likelihoods)** Consider $H_1 : \theta \leq \theta_*$ versus $H_2 : \theta > \theta_*$ with θ_* a given constant. In this situation, $\Theta_1 = \{\theta : \theta \leq \theta_*\}$ and $\Theta_2 = \{\theta : \theta > \theta_*\}$. A convenient prior distribution for θ is the split (or truncated) normal distribution

$$p(\theta|H_1) = [\Phi(\theta_*|\underline{\mu}_1, \underline{h}_1^{-1})]^{-1} \phi(\theta|\underline{\mu}_1, \underline{h}_1^{-1}), \quad \theta \in \Theta_1, \quad (6.37)$$

and

$$p(\theta|H_2) = [1 - \Phi(\theta_*|\underline{\mu}_2, \underline{h}_2^{-1})]^{-1} \phi(\theta|\underline{\mu}_2, \underline{h}_2^{-1}), \quad \theta \in \Theta_2. \quad (6.38)$$

Suppose a random sample Y_1, Y_2, \dots, Y_T is available from an $N(\theta, \sigma^2)$ population in which θ is unknown and σ^2 is *known*. Under what conditions is it optimal to reject H_1 ?

Solution

Analogous to (2.22), it is straightforward to show that

$$\phi(\theta|\underline{\mu}_j, \underline{h}_j^{-1})L(\theta) = \phi(\theta|\bar{\mu}_j, \bar{h}_j^{-1})\phi(\bar{y}|\underline{\mu}_j, \underline{h}_j^{-1} + h^{-1}), \quad j = 1, 2, \quad (6.39)$$

where

$$\bar{h}_j = \underline{h}_j + h, \quad h = T\sigma^{-2}, \quad (6.40)$$

and

$$\bar{\mu}_j = \frac{\underline{h}_j \underline{\mu}_j + h \bar{y}}{\bar{h}_j}. \quad (6.41)$$

From (6.37)–(6.39) it follows that

$$\begin{aligned} p(y|H_1) &= \int_{\Theta_1} p(\theta|H_1)L(\theta)d\theta \\ &= \left[\frac{\Phi(\theta_*|\bar{\mu}_1, \bar{h}_1^{-1})}{\Phi(\theta_*|\underline{\mu}_1, \underline{h}_1^{-1})} \right] \phi(\bar{y}|\underline{\mu}_1, \underline{h}_1^{-1} + h^{-1}) \end{aligned} \quad (6.42)$$

and

$$\begin{aligned} p(y|H_2) &= \int_{\Theta_2} p(\theta|H_2)L(\theta)d\theta \\ &= \left[\frac{1 - \Phi(\theta_*|\bar{\mu}_2, \bar{h}_2^{-1})}{1 - \Phi(\theta_*|\underline{\mu}_2, \underline{h}_2^{-1})} \right] \phi(\bar{y}|\underline{\mu}_2, \underline{h}_2^{-1} + h^{-1}). \end{aligned} \quad (6.43)$$

From (6.39) it also follows that the conditional posterior densities under each hypothesis are

$$p(\theta|y, H_1) = [\Phi(\theta_*|\bar{\mu}_1, \bar{h}_1^{-1})]^{-1} \phi(\theta|\bar{\mu}_1, \bar{h}_1^{-1}), \quad \theta \in \Theta_1, \quad (6.44)$$

and

$$p(\theta|y, H_2) = [1 - \Phi(\theta_*|\bar{\mu}_2, \bar{h}_2^{-1})]^{-1} \phi(\theta|\bar{\mu}_2, \bar{h}_2^{-1}), \quad \theta \in \Theta_2. \quad (6.45)$$

The hyperparameters $\bar{\mu}_j$ and \bar{h}_j of conditional posterior densities (6.44) and (6.45) are updated (see Equations 6.40 and 6.41) in the usual fashion from the prior hyperparameter values $\underline{\mu}_j$ and \underline{h}_j , $j = 1, 2$, because (6.42) and (6.43), like (6.37) and (6.38), form yet another conjugate prior for a random sample from an $N(\theta, \sigma^2)$ distribution with σ^2 known. The Bayes factor B_{21} for H_2 versus H_1 is simply the ratio of (6.43) to (6.42).

The unconstrained split-normal density implied by (6.37) and (6.38) is quite flexible but fairly “parameter-rich”: It requires specification of five hyperparameters $\underline{\mu}, \underline{h}_1, \underline{\mu}_2, \underline{h}_2$, and $\underline{\pi}_1 = \Pr(H_1) = 1 - \underline{\pi}_2$. In the special case in which

$$\underline{\mu} = \underline{\mu}_1 = \underline{\mu}_2, \quad (6.46)$$

$$\underline{h} = \underline{h}_1 = \underline{h}_2, \quad (6.47)$$

$$\underline{\pi}_1 = \Phi(\theta_* | \underline{\mu}, \underline{h}^{-1}), \quad (6.48)$$

and

$$\underline{\pi}_2 = 1 - \Phi(\theta_* | \underline{\mu}, \underline{h}^{-1}), \quad (6.49)$$

$p(\theta) = \phi(\theta | \underline{\mu}, \underline{h}^{-1})$, as in the familiar normal estimation problem. The case in which (6.46) holds with

$$\underline{\mu} = \theta_* \quad (6.50)$$

is often attractive. Prior specification (6.37) and (6.38) with all of (6.46)–(6.49) holding [i.e., $p(\theta) = \phi(\theta | \theta_*, \underline{h}^{-1})$] is attractive in many settings and requires elicitation of only \underline{h} . Restrictions (6.46)–(6.49) impose continuity on the prior $p(\theta)$ and reduce the number of hyperparameters to be specified to two ($\underline{\mu}$ and \underline{h}). Under (6.46)–(6.49), the Bayes factor of H_2 versus H_1 simplifies to

$$B_{21} = \left[\frac{1 - \Phi(\theta_* | \bar{\mu}, \bar{h}^{-1})}{\Phi(\theta_* | \bar{\mu}, \bar{h}^{-1})} \right] \left[\frac{\underline{\pi}_1}{\underline{\pi}_2} \right], \quad (6.51)$$

where $\bar{\mu}$ and \bar{h} are given by the common values across $j = 1, 2$ of (6.40) and (6.41). Bayes factor (6.51) implies (see Equation 6.20) that the posterior probability $\bar{\pi}_1 = \Pr(H_1 | y)$ equals simply

$$\bar{\pi}_1 = \Phi(\theta_* | \bar{\mu}, \bar{h}^{-1}).$$

Under “all-or-nothing” loss, (6.51) implies that decision rule (6.26) reduces to reject H_1 iff

$$\bar{y} > \theta_* + (\underline{h}/h) (\theta_* - \underline{\mu}) + \gamma h^{-(1/2)}, \quad (6.52)$$

where

$$\gamma \equiv -[1 + (\underline{h}/h)]^{(1/2)} \Phi^{-1} \left[\left(1 + \frac{\bar{C}_2}{\bar{C}_1} \right)^{-1} \right]. \quad (6.53)$$

The second term in (6.52) serves as an additive adjustment to θ_* : requiring \bar{y} to be larger to reject H_1 when the prior mean $\underline{\mu}$ lies to the left of θ_* . Under (6.50) this adjustment term vanishes.

The constant γ defined by (6.53) plays the role of the tabulated critical value used in frequentist testing. However, it is an explicit function of the relative cost of a type I and type II error. Usually, the problem is posed so that the ratio \bar{C}_2/\bar{C}_1 is greater than unity. Letting

$$\alpha = \left(1 + \frac{\bar{C}_2}{\bar{C}_1} \right)^{-1},$$

we see in this case that $\Phi^{-1}(\alpha) < 0$ and hence $\gamma > 0$. The quantity α is like the level of significance in a frequentist test. Provided the prior is informative (i.e., $0 < \underline{h} < \infty$), γ is a function of sample size T through $h = T/\sigma^2$ and the first factor in (6.53) is greater than unity. As prior information becomes diffuse (i.e., $\underline{h} \rightarrow 0$), the second term in (6.52) vanishes independently of (6.50) holding, and the first factor in (6.51) approaches unity. Therefore, if prior information is diffuse and $\bar{C}_2/\bar{C}_1 = 19$, then $\alpha = .05$, $\gamma = -\Phi^{-1}(.05) = 1.645$, and (6.52) is identical to the rejection region of a frequentist one-sided test at the .05 level of significance.

Exercise 6.4 (Simple null, two-sided alternative, and general likelihood) Consider $H_1 : \theta = \theta_*$ versus $H_2 : \theta \neq \theta_*$, where θ_* is a given constant. In this case $\Theta_1 = \{\theta_*\}$ and $\Theta_2 = \{\theta : \theta \in \Theta \text{ and } \theta \neq \theta_*\}$. Suppose prior probability function (6.4) is of the mixed form

$$p(\theta) = \begin{cases} \pi_1 & \text{if } \theta = \theta_*, \\ \pi_2 p(\theta|H_2) & \text{if } \theta \in \Theta_2, \end{cases} \quad (6.54)$$

with a “spike” of height π_1 at $\theta = \theta_*$. Consider data y yielding likelihood $L(\theta)$. Under what conditions is it optimal to reject H_1 ?

Solution

The posterior probability function is also of mixed form:

$$p(\theta|y) = \begin{cases} \bar{\pi}_1 & \text{if } \theta = \theta_* \\ \bar{\pi}_2 p(\theta|y, H_2) & \text{if } \theta \in \Theta_2, \end{cases} \quad (6.55)$$

where

$$\bar{\pi}_1 = \Pr(H_1|y) = \frac{\pi_1 L(\theta_*)}{p(y)}, \quad (6.56)$$

$$\bar{\pi}_2 = \Pr(H_2|y) = \frac{\pi_2 p(y|H_2)}{p(y)}, \quad (6.57)$$

$$p(y) = \pi_1 L(\theta_*) + \pi_2 p(y|H_2), \quad (6.58)$$

and $p(y|H_2)$ is given by (6.5). Also note that Bayes factor (6.19) yields

$$B_{21} = \frac{\int_{\Theta_2} p(\theta|H_2) L(\theta) d\theta}{L(\theta_*)}. \quad (6.59)$$

Compared to the numerator in (6.19), the numerator in (6.59) marginalizes the likelihood $L(\theta)$ using the prior density $p(\theta|H_2)$. Also, the right-hand side of the inequality in decision rule (6.22) depends on data, and so B_{21} does not fully summarize the effect of the data on the optimal decision.

Exercise 6.5 (Simple null hypothesis, two-sided alternative, and normal likelihoods)

Consider Exercise 6.4. Suppose the prior density $p(\theta)$ is mixed prior density (6.54) with

$$p(\theta|H_2) = \phi(\theta|\underline{\mu}, \underline{h}^{-1}). \quad (6.60)$$

Also, suppose a random sample Y_1, Y_2, \dots, Y_T is available from an $N(\theta, \sigma^2)$ population in which θ is unknown and σ^2 is *known*. Under what conditions is it optimal to reject H_1 ?

Solution

From (6.43) and (2.23), and also noting that the integral over Θ_2 is equivalent to the integral over $-\infty$ to ∞ , we have that

$$p(y|H_2) = \phi(\bar{y}|\underline{\mu}, \underline{h}^{-1} + h^{-1}) \quad (6.61)$$

and the Bayes factor for H_2 versus H_1 is

$$\begin{aligned} B_{21} &= \frac{\phi(\bar{y}|\underline{\mu}, \underline{h}^{-1} + h^{-1})}{\phi(\bar{y}|\theta_*, h^{-1})} \\ &= \left[\frac{h^{-1}}{\underline{h}^{-1} + h^{-1}} \right]^{(1/2)} \exp \left[-\frac{1}{2} ((\underline{h}^{-1} + h^{-1})^{-1}(\bar{y} - \underline{\mu})^2 - h(\bar{y} - \theta_*)^2) \right]. \end{aligned} \quad (6.62)$$

By using identity (2.19) with $\theta_1 = \theta_*$, it follows from (6.62) that

$$B_{21} = (\underline{h}/\bar{h})^{(1/2)} \exp \left(-\frac{1}{2} [\underline{h}(\theta_* - \underline{\mu})^2 - \bar{h}(\theta_* - \bar{\mu})^2] \right), \quad (6.63)$$

where $\bar{h} = \underline{h} + h$ is the posterior precision. Then (6.26) implies that $d_* = d_2$ iff

$$\ln(B_{21}) > \ln \left(\frac{\bar{C}_{2\pi_1}}{\bar{C}_{1\pi_2}} \right), \quad (6.64)$$

or equivalently, by using (6.63),

$$|\bar{h}^{(1/2)}(\bar{\mu} - \theta_*)| > \left(\underline{h}(\theta_* - \underline{\mu})^2 + 2 \ln \left[\left(\frac{\bar{C}_{2\pi_1}}{\bar{C}_{1\pi_2}} \right) (\underline{h}/\bar{h})^{(1/2)} \right] \right), \quad (6.65)$$

with

$$\bar{\mu} - \theta_* = \frac{\underline{h}(\underline{\mu} - \theta_*) + h(\bar{y} - \theta_*)}{\bar{h}}. \quad (6.66)$$

Exercise 6.6 (Special cases of Exercise 6.5) Consider (6.65).

- (a) What happens as $\underline{h} \rightarrow 0$?
- (b) What happens when $\underline{\mu} = \theta_*$?
- (c) Analyze $\bar{\pi}_1$ when $\underline{\mu} = \theta_*$.

Solution

(a) As $\underline{h} \rightarrow 0$, inequality (6.65) becomes $\sqrt{\bar{h}}|\bar{y} - \theta_*| > \infty$, which cannot be satisfied. Hence, as the prior becomes less precise, H_1 is *never* rejected.

(b) When $p(\theta|H_2)$ is “centered” over H_1 (i.e., $\underline{\mu} = \theta_*$), then (6.65) reduces to reject H_1 iff

$$|\bar{h}^{(1/2)}(\bar{\mu} - \theta_*)| > (\underline{h}/\bar{h})^{(1/2)} 2 \ln \left[\left(\frac{\bar{C}_2 \pi_1}{\bar{C}_1 \pi_2} \right) \right], \quad (6.67)$$

where the left-hand side of (6.67) is the familiar frequentist test statistic. The right-hand side of (6.67) serves as a “critical value.” Not only does this Bayesian “critical value” depend on the relative expected costs of type I and type II errors, but more importantly it is an increasing function of sample size T through $\bar{h} = \underline{h} + (T/\sigma^2)$. This is reminiscent of the common dictum shared by all good frequentist statisticians that the level of significance should be chosen to be a decreasing function of sample size. However, because the level of significance is usually chosen on the basis of “convention,” this warning is seldom operationalized.

(c) Using (6.63) and (6.20) yields

$$\bar{\pi}_1 = \left[1 + \left(\frac{\pi_1}{1 - \pi_1} \right)^{-1} \left(\frac{\underline{h}\sigma^2}{\underline{h}\sigma^2 + T} \right)^{(1/2)} \exp \left[-\frac{1}{2} \left(\frac{T}{\underline{h}\sigma^2 + T} \right) z^2 \right] \right]^{-1}, \quad (6.68)$$

where $z = |\bar{h}^{(1/2)}(\bar{y} - \theta_*)|$. From (6.68) it is seen that the frequentist test statistic z can be large (suggesting to a frequentist with a fixed level of significance that H_1 should be rejected), while at the same time the posterior probability of H_1 is also close to unity. In fact, letting $T \rightarrow \infty$ in such a way that z (which also depends on T) is held fixed, it is easy to see that $\bar{\pi}_1 \rightarrow 1$. In other words, for any large finite value of the test statistic z , there is a sample size T for which the posterior probability $\bar{\pi}_1$ can be made arbitrarily close to unity.

Exercise 6.7 (Relationship among posterior probabilities under different prior probabilities) Consider J hypotheses $H_j : \theta_j \in \Theta_j$ ($j = 1, 2, \dots, J$). Define

$$m_j(y) = \frac{p_j(y|H_j)}{\sum_{i=1}^J p_i(y|H_i)}. \quad (6.69)$$

Show that $m_j(y)$ ($j = 1, 2, \dots, J$) are all that are required to summarize the data to compute posterior probabilities for *any* set of given prior probabilities.

Solution

In general,

$$\bar{\pi}_j = \frac{\pi_j p_j(y|H_j)}{\sum_{i=1}^J \pi_i p_i(y|H_i)} = \left[\sum_{i=1}^J \left(\frac{\pi_i}{\pi_j} \right) B_{ij} \right]^{-1}, \quad j = 1, 2, \dots, J. \quad (6.70)$$

Because $B_{ij} = m_i(y)/m_j(y)$ ($j = 1, 2, \dots, J$), it follows from (6.70) that $\{m_j(y)\}_{j=1}^J$ are all that are required to summarize the data to compute posterior probabilities for *any* set of given prior probabilities π_j ($j = 1, 2, \dots, J$) via (6.70).

Exercise 6.8 (Savage–Dickey density ratio) Consider the likelihood $L(\theta)$ and the hypotheses $H_1 : \theta_2 = 0$ versus $H_2 : \theta_2 \neq 0$, where $\theta = [\theta'_1, \theta'_2]'$. Let $g(\theta_1, \theta_2) = g(\theta_1|\theta_2)g(\theta_2)$ be a continuous density defined over \mathcal{R}^k and consider the prior densities $p(\theta_1|H_1) = g(\theta_1|\theta_2 = 0)$ and $p(\theta_1, \theta_2|H_2) = g(\theta_1, \theta_2)$. Show that the Bayes factor for H_1 versus H_2 can be written as the ratio (known as the *Savage–Dickey density ratio*) of the marginal *posterior* density of θ_2 to the marginal *prior* density of θ_2 , each evaluated at $\theta_2 = 0$, and computed using $g(\theta_1, \theta_2)$ as the prior, that is,

$$B_{12} = \frac{p(\theta_2 = 0|H_2, y)}{p(\theta_2 = 0|H_2)}, \quad (6.71)$$

where

$$\begin{aligned} p(\theta_1, \theta_2|y, H_2) &= \frac{p(\theta_1, \theta_2|H_2)L(\theta_1, \theta_2)}{p(y|H_2)} \\ &= \frac{g(\theta_1, \theta_2)L(\theta_1, \theta_2)}{\int_{\Theta} g(\theta_1, \theta_2)L(\theta_1, \theta_2) d\theta_1 d\theta_2}. \end{aligned} \quad (6.72)$$

Solution

Integrating (6.72) with respect to θ_1 yields

$$\begin{aligned} p(\theta_2|H_2, y) &= \int_{\Theta_1} p(\theta_1, \theta_2|H_2, y) d\theta_1 \\ &= \int_{\Theta_1} \left[\frac{g(\theta_1|\theta_2)g(\theta_2)L(\theta_1, \theta_2)}{\int_{\Theta} g(\theta_1, \theta_2)L(\theta_1, \theta_2) d\theta_1 d\theta_2} \right] d\theta_1. \end{aligned} \quad (6.73)$$

Dividing (6.73) by $p(\theta_2|H_2)$ and evaluating both at $\theta_2 = 0$ implies

$$\begin{aligned} \frac{p(\theta_2 = 0|y, H_2)}{p(\theta_2 = 0|H_2)} &= \int_{\Theta_1} \left[\frac{g(\theta_1|\theta_2 = 0)g(\theta_2 = 0)L(\theta_1, \theta_2 = 0)}{g(\theta_2 = 0) \int_{\Theta} g(\theta_1, \theta_2)L(\theta_1, \theta_2) d\theta_1 d\theta_2} \right] d\theta_1 \\ &= \int_{\Theta_1} \left[\frac{p(\theta_1|H_1)L(\theta_1, \theta_2 = 0)}{\int_{\Theta} g(\theta_1, \theta_2)L(\theta_1, \theta_2) d\theta_1 d\theta_2} \right] d\theta_1 \\ &= \frac{p(y|H_1)}{p(y|H_2)} = B_{12}. \end{aligned} \quad (6.74)$$

Exercise 6.9 (Generalized Savage–Dickey density ratio) Consider Exercise 6.8. Without assuming any restriction on prior density $p(\theta|H_1)$ as done in Exercise 6.8, find a c such that

$$B_{12} = c \left(\frac{p(\theta_2 = 0|H_2, y)}{p(\theta_2 = 0|H_2)} \right).$$

Solution

Consider

$$\begin{aligned}
B_{12} &= \frac{\int_{\Theta_1} L(\theta_1, \theta_2 = 0) p(\theta_1 | H_1) d\theta_1}{p(y | H_2)} \\
&= p(\theta_2 = 0 | y, H_2) \left[\int_{\Theta_1} \frac{L(\theta_1, \theta_2 = 0) p(\theta_1 | H_1)}{p(\theta_2 = 0 | y, H_2) p(y | H_2)} d\theta_1 \right] \\
&= p(\theta_2 = 0 | y, H_2) \left[\int_{\Theta_1} \frac{L(\theta_1, \theta_2 = 0) p(\theta_1 | H_1)}{\left(\frac{p(\theta_1, \theta_2 = 0 | y, H_2)}{p(\theta_1 | \theta_2 = 0, y, H_2)} \right) p(y | H_2)} d\theta_1 \right] \\
&= p(\theta_2 = 0 | y, H_2) \left[\int_{\Theta_1} \frac{L(\theta_1, \theta_2 = 0) p(\theta_1 | H_1) p(\theta_1 | \theta_2 = 0, y, H_2)}{\left(\frac{L(\theta_1, \theta_2 = 0) p(\theta_1, \theta_2 = 0 | y, H_2)}{p(y | H_2)} \right) p(y | H_2)} d\theta_1 \right] \\
&= \left(\frac{p(\theta_2 = 0 | y, H_2)}{p(\theta_2 = 0 | H_2)} \right) \left[\int_{\Theta_1} \left(\frac{p(\theta_1 | H_1)}{p(\theta_1 | \theta_2 = 0, H_2)} \right) p(\theta_1 | \theta_2 = 0, y, H_2) d\theta_1 \right] \\
&= \left(\frac{p(\theta_2 = 0 | y, H_2)}{p(\theta_2 = 0 | H_2)} \right) E_{\theta_1 | \theta_2 = 0, y, H_2} \left[\frac{p(\theta_1 | H_1)}{p(\theta_1 | \theta_2 = 0, H_2)} \right].
\end{aligned}$$

So,

$$c = E_{\theta_1 | \theta_2 = 0, y, H_2} \left[\frac{p(\theta_1 | H_1)}{p(\theta_1 | \theta_2 = 0, H_2)} \right]$$

gives the desired result.

Prediction

Prediction provides discipline and pragmatic importance to empirical research. Suppose a deity told you the values of all unknown parameters in your model so that estimation and hypothesis testing became moot. What would you do with your model? The obvious answer is to use it for what it is intended to do: Make *ex ante* probability statements about future observables.

Suppose the information set consists of the union of past data $Y = y$, yielding the parametric likelihood function $L(\theta)$, and other information in the form of a prior distribution $p(\theta)$. The sampling distribution of an out-of-sample Y_* (possibly a vector) given $Y = y$ and θ would be an acceptable predictive distribution if θ were known, but without knowledge of θ it cannot be used. Then the *Bayesian predictive probability distribution* $p(y_*|y)$ is

$$\begin{aligned}
 p(y_*|y) &= \frac{p(y_*, y)}{p(y)} \\
 &= \int_{\Theta} \frac{p(y_*, y, \theta)}{p(y)} d\theta \\
 &= \int_{\Theta} p(y_*|y, \theta) \left[\frac{p(\theta)p(y|\theta)}{p(y)} \right] d\theta \\
 &= \int_{\Theta} p(y_*|y, \theta)p(\theta|y) d\theta \\
 &= E_{\theta|y}[p(y_*|y, \theta)].
 \end{aligned} \tag{7.1}$$

In other words, the Bayesian predictive probability function is the posterior expectation of $p(y_*|y, \theta)$. If the past and future are independent, conditional on θ (as in random sampling), then $p(y_*|y, \theta) = p(y_*|\theta)$.

Given the predictive distribution, point prediction proceeds analogously to point estimation, as discussed in Chapter 3. Letting $C(\hat{Y}_*, Y_*)$ denote a *predictive cost (loss) function*

measuring the performance of a predictor \hat{Y}_* of Y_* , we define the optimal point predictor \hat{Y}_{**} to be the solution

$$\hat{Y}_{**} = \underset{\hat{Y}_*}{\operatorname{argmin}} \left(E_{Y_*|Y} [C(\hat{Y}_*, Y_*)] \right). \quad (7.2)$$

For example, if Y_* is a scalar and predictive loss is quadratic in *prediction (forecast) error* $\hat{Y}_* - Y_*$ [i.e., $C(\hat{Y}_*, Y_*) = (\hat{Y}_* - Y_*)^2$], then the optimal point estimate according to (7.2) is

$$\hat{Y}_{**} = E(Y_*|Y). \quad (7.3)$$

Results for generating point predictions analogous to those for point estimates found in Chapter 3 can be derived for other familiar loss functions. Similarly, prediction (forecast) intervals for Y_* can be constructed from $p(y_*|y)$ similarly to the HPD intervals for θ from the posterior density $p(\theta|y)$ in Chapter 5.

In effect, the predictive density described in (7.1) treats all parameters as nuisance parameters and integrates them out of the predictive problem. A similar strategy is used when adding parametric hypotheses to the analysis. Suppose that the permissible parameter space Θ is partitioned into J mutually exclusive subsets Θ_j ($j = 1, 2, \dots, J$) such that $\Theta = \Theta_1 \cup \Theta_2 \cup \dots \cup \Theta_J$. This partitioning gives rise to J hypotheses: $H_j : \theta \in \Theta_j$ ($j = 1, 2, \dots, J$). Also suppose that the likelihood $L(\theta)$, prior probabilities $\pi_j = \Pr(H_j)$, and conditional prior p.d.f.s $p(\theta|H_j)$ ($j = 1, 2, \dots, J$) are given. Conditional on hypothesis H_j and given data y leading to the posterior p.d.f. $p(\theta|y, H_j)$, the j th conditional predictive density of Y_* is

$$p(y_*|y, H_j) = \int_{\Theta} p(y_*|\theta, y, H_j) p(\theta|y, H_j) d\theta. \quad (7.4)$$

Using the posterior probabilities $\bar{\pi}_j = \Pr(H_j|y)$ ($j = 1, 2, \dots, J$), the unconditional (marginal) predictive density of Y_* is the mixture

$$\begin{aligned} p(y_*|y) &= \sum_{j=1}^J p(y_*, H_j|y) \\ &= \sum_{j=1}^J \bar{\pi}_j p(y_*|y, H_j). \end{aligned} \quad (7.5)$$

Predictive probability function (7.5) is the basis for constructing prediction intervals and point forecasts unconditional on the competing hypotheses. For example, under quadratic loss the optimal Bayesian point estimate is the predictive mean

$$E(y_*|y) = \sum_{j=1}^J \bar{\pi}_j E(y_*|y, H_j). \quad (7.6)$$

Note that (7.6) is a weighted average of the optimal point forecasts $E(y_*|y, H_j)$ under each hypothesis. Furthermore, the weights in this case, the posterior probabilities

$\bar{\pi}_j$ ($j = 1, 2, \dots, J$) of the respective hypotheses, have an intuitive appeal: The forecasts of more probable hypotheses *a posteriori* receive more weight. The “optimality” of these weights is self-evident to the Bayesian: Given the likelihood, hypotheses, and priors, (7.5) is the unique predictive distribution for future observables.

Exercise 7.1 (Predictive p.m.f. for conjugate Bernoulli analysis) Consider the Bernoulli sampling problem in Exercise 2.1 with the beta prior $\theta \sim \beta(\underline{\alpha}, \underline{\delta})$. Find the predictive p.m.f. for the out-of-sample value y_* .

Solution

According to Exercise 2.1 the posterior is $\theta|y \sim \beta(\bar{\alpha}, \bar{\delta})$, where $\bar{\alpha} = \underline{\alpha} + m$, $\bar{\delta} = \underline{\delta} + T - m$, and m is the number of successes in T observations. Hence the predictive p.m.f. is

$$\begin{aligned}
 p(y_*|y) &= \int_0^1 \theta^{y_*} (1 - \theta)^{(1-y_*)} [\mathcal{B}(\bar{\alpha}, \bar{\delta})]^{-1} \theta^{\bar{\alpha}-1} (1 - \theta)^{(\bar{\delta}-1)} d\theta \\
 &= [\mathcal{B}(\bar{\alpha}, \bar{\delta})]^{-1} \int_0^1 \theta^{(y_* + \bar{\alpha}-1)} (1 - \theta)^{(1-y_* + \bar{\delta}-1)} d\theta \\
 &= \left[\frac{\mathcal{B}(\bar{\alpha} + y_*, \bar{\delta} + 1 - y_*)}{\mathcal{B}(\bar{\alpha}, \bar{\delta})} \right] \\
 &\quad \times \int_0^1 [\mathcal{B}(\bar{\alpha} + y_*, \bar{\delta} + 1 - y_*)]^{-1} \theta^{(y_* + \bar{\alpha}-1)} (1 - \theta)^{(1-y_* + \bar{\delta}-1)} d\theta \\
 &= \left[\frac{\Gamma(\bar{\alpha} + y_*) \Gamma(\bar{\delta} + 1 - y_*)}{\Gamma(\bar{\alpha} + \bar{\delta} + 1)} \right] \left[\frac{\Gamma(\bar{\alpha} + \bar{\delta})}{\Gamma(\bar{\alpha}) \Gamma(\bar{\delta})} \right], \quad y_* = 0, 1,
 \end{aligned} \tag{7.7}$$

where the beta and gamma functions used in this question are discussed in Exercise 2.1 and the integral in the third line in (7.7) is unity because the integrand is the density of a $\beta(\bar{\alpha} + y_*, \bar{\delta} + 1 - y_*)$ distribution. Predictive p.m.f. (7.7) is known as a *beta-binomial* p.m.f. [see, e.g., Poirier (1995, p. 116)]. It has mean $\bar{\alpha}/(\bar{\alpha} + \bar{\delta})$ and variance $[\bar{\alpha}\bar{\delta}(\bar{\alpha} + \bar{\delta} + 1)]/[(\bar{\alpha} + \bar{\delta})^2(\bar{\alpha} + \bar{\delta} + 1)]$.

Exercise 7.2 (Predictive p.m.f. for conjugate Poisson analysis) Suppose Y_t ($t = 1, 2, \dots, T$) is a random sample from a Poisson distribution with mean θ and that the prior distribution of θ is the gamma distribution $G(\underline{\alpha}, \underline{\beta})$, where $\underline{\alpha} > 0$ and $\underline{\beta} > 0$. Find the predictive density for an out-of-sample value y_* .

Solution

From Exercise 2.12 we know $\theta|y \sim G(\bar{\alpha}, \bar{\beta})$, where $\bar{\alpha} = \underline{\alpha} + T\bar{y}$ and $\bar{\beta} = (\underline{\beta}^{-1} + T)^{-1}$.

Therefore, predictive p.m.f. (7.1) becomes

$$\begin{aligned}
 p(y_*|y) &= \int_0^\infty p(y_*|\theta)p(\theta|y)d\theta \tag{7.8} \\
 &= \int_0^\infty \left[\frac{\theta^{y_*} \exp(-\theta)}{y_*!} \right] \left[\left(\bar{\beta}^{\bar{\alpha}} \Gamma(\bar{\alpha}) \right)^{-1} \theta^{\bar{\alpha}-1} \exp(-\theta/\bar{\beta}) \right] d\theta \\
 &= \left(y_*! \bar{\beta}^{\bar{\alpha}} \Gamma(\bar{\alpha}) \right)^{-1} \int_0^\infty \theta^{(y_*+\bar{\alpha}-1)} \exp[-\theta/(\bar{\beta}^{-1}+1)^{-1}] d\theta \\
 &= \left(\frac{(\bar{\beta}^{-1}+1)^{-(\bar{\alpha}+y_*)} \Gamma(\bar{\alpha}+y_*)}{y_*! \bar{\beta}^{\bar{\alpha}} \Gamma(\bar{\alpha})} \right) \\
 &\quad \times \int_0^\infty \left((\bar{\beta}^{-1}+1)^{-(\bar{\alpha}+y_*)} \Gamma(\bar{\alpha}+y_*) \right)^{-1} \theta^{(\bar{\alpha}+y_*-1)} \exp[-\theta/(\bar{\beta}^{-1}+1)^{-1}] d\theta \\
 &= \left(\frac{(\bar{\beta}^{-1}+1)^{-(\bar{\alpha}+y_*)} \Gamma(\bar{\alpha}+y_*)}{y_*! \bar{\beta}^{\bar{\alpha}} \Gamma(\bar{\alpha})} \right), \quad y_* = 0, 1, 2, \dots,
 \end{aligned}$$

where the integral in the second-to-last line is unity because it is the density of a $G(\bar{\alpha} + y_*, [\bar{\beta}^{-1} + 1]^{-1})$ distribution. Predictive p.m.f. (7.8) is known as the Poisson-gamma p.m.f.

Exercise 7.3 (Predictive density for conjugate normal analysis with known variance)

Consider Exercise (2.6b). Suppose σ^2 is known and that the prior distribution for μ is $N(\underline{\mu}, \sigma^2/\underline{T})$. Find the predictive density of an out-of-sample y_* .

Solution

Under this prior, the results of Exercise 2.6(b) in conjunction with (2.25) show that the posterior distribution for μ is $N(\bar{\mu}, \sigma^2/\bar{T})$, where $\bar{\mu} = \bar{T}^{-1}(\underline{T}\underline{\mu} + T\bar{y})$ and $\bar{T} = \underline{T} + T$. Noting that

$$(y_* - \mu)^2 + \bar{T}(\mu - \bar{\mu})^2 = (1 + \bar{T})(\mu - c)^2 + \bar{T}(1 + \bar{T})^{-1}(y_* - \bar{\mu})^2, \tag{7.9}$$

where $c = (1 + \bar{T})^{-1}(\bar{T}\bar{\mu} + y_*)$, we can obtain the predictive density (7.1) as

$$\begin{aligned}
 p(y_*|y) &= \int_{\mathcal{R}} p(y_*|y, \mu)p(\mu|y)d\mu \tag{7.10} \\
 &= \int_{\mathcal{R}} \phi(y_*|\mu, \sigma^2)\phi(\mu|\bar{\mu}, \sigma^2/\bar{T})d\mu \\
 &= \phi\left(y_*|\bar{\mu}, \sigma^2[1 + \bar{T}^{-1}]\right) \int_{\mathcal{R}} \phi(\mu|c, \sigma^2[1 + \bar{T}]^{-1}) d\mu \\
 &= \phi\left(y_*|\bar{\mu}, \sigma^2[1 + \bar{T}^{-1}]\right), \quad -\infty < y_* < \infty.
 \end{aligned}$$

Exercise 7.4 (Predictive point and interval estimation for conjugate normal analysis with known variance) Consider Exercise 7.3.

- (a) Find the predictive point forecast of y_* under quadratic loss.
 (b) Find a $1 - \alpha$ forecast interval for y_* .

Solution

- (a) Under quadratic loss, the Bayesian point forecast of y_* is the predictive mean $\bar{\mu}$.
 (b) A $1 - \alpha$ prediction (forecast) interval for y_* is given by

$$\bar{S}_* = \left\{ y_* : y_* \in \mathcal{R} \text{ and } \bar{\mu} - z_{\alpha/2}\sigma[1 + \bar{T}^{-1}]^{(1/2)} < y_* < \bar{\mu} + z_{\alpha/2}\sigma[1 + \bar{T}^{-1}]^{(1/2)} \right\}, \quad (7.11)$$

where $z_{\alpha/2}$ satisfies $\Phi(z_{\alpha/2}) - \Phi(-z_{\alpha/2}) = 1 - \alpha$.

Exercise 7.5 (Predictive density for conjugate multivariate normal analysis with known covariance matrix) Consider the multivariate normal sampling distribution in Exercise 2.14. Suppose Σ is known and that the prior distribution for μ is $N_m(\underline{\mu}, \Sigma/\underline{T})$. Find the predictive density of an out-of-sample value y_* .

Solution

According to Exercise 2.14, the posterior distribution of μ is $N(\bar{\mu}, \bar{T}^{-1}\Sigma)$, where $\bar{\mu}$ and \bar{T} are defined as in the solution to part (c). By using the identity

$$\begin{aligned} & (y_* - \mu)' \Sigma^{-1} (y_* - \mu) + \bar{T}(\mu - \bar{\mu})' \Sigma^{-1} (\mu - \bar{\mu}) \\ &= (1 + \bar{T})(\mu - c)' \Sigma^{-1} (\mu - c) + \bar{T}(1 + \bar{T})^{-1} (y_* - \bar{\mu})' \Sigma^{-1} (y_* - \bar{\mu}), \end{aligned} \quad (7.12)$$

in the same manner as in Exercise 7.3, and with c defined accordingly, the predictive density is

$$\begin{aligned} p(y_*|y) &= \int_{\mathcal{R}^m} p(y_*|y, \mu) p(\mu|y) d\mu \\ &= \int_{\mathcal{R}^m} \phi(y_*|\mu, \Sigma) \phi(\mu|\bar{\mu}, \Sigma/\bar{T}) d\mu \\ &= \phi(y_*|\bar{\mu}, [1 + \bar{T}^{-1}]\Sigma) \int_{\mathcal{R}^m} \phi(\mu|c, [1 + \bar{T}]^{-1}\Sigma) d\mu \\ &= \phi(y_*|\bar{\mu}, [1 + \bar{T}^{-1}]\Sigma), \quad y_* \in \mathcal{R}^m. \end{aligned} \quad (7.13)$$

Exercise 7.6 (Predictive estimation: conjugate multivariate normal analysis with known covariance matrix) Consider Exercise 7.5.

- (a) Find the predictive point forecast of y_* under quadratic loss.
 (b) Find a $1 - \alpha$ forecast region for y_* .

Solution

(a) Under quadratic loss, the Bayesian point forecast of y_* is the predictive mean $\bar{\mu}$.

(b) A $1 - \alpha$ predictive (forecast) ellipsoid for y_* is given by

$$\bar{S}_{m*} = \left\{ y_* : y_* \in \mathcal{R}^m \text{ and } (y_* - \bar{\mu})' \Sigma^{-1} (y_* - \bar{\mu}) \leq c_\alpha [1 + \bar{T}^{-1}] \right\}, \quad (7.14)$$

where c_α cuts off α probability in the right-hand tail of the $\chi^2(m)$ distribution. Note that, in Exercises 7.3 through 7.6, frequentist results emerge as $\underline{T} \rightarrow 0$.

Exercise 7.7 (Predictive density for conjugate normal analysis with unknown mean and variance) Consider Exercise 7.3 with *both* μ and σ^2 unknown. Suppose the prior distribution for μ and σ^{-2} is the natural conjugate normal-gamma prior $NG(\underline{\mu}, \underline{T}^{-1}, \underline{s}^{-2}, \underline{\nu})$ introduced in Exercise 2.4. Find the predictive density of an out-of-sample y_* .

Solution

The posterior distribution for μ and σ^{-2} is $NG(\bar{\mu}, \bar{T}^{-1}, \bar{s}^{-2}, \bar{\nu})$, where $\bar{\mu}$, \bar{T} , \bar{s}^{-2} , and $\bar{\nu}$ are defined analogously to (2.31), (2.30), (2.32), and (2.28), respectively. From Exercise 7.3 we know that the predictive density given σ^{-2} is

$$\begin{aligned} p(y_* | y, \sigma^{-2}) &= \int_{\mathcal{R}} p(y_* | y, \mu, \sigma^{-2}) p(\mu | y, \sigma^{-2}) d\mu \\ &= \phi(y_* | \bar{\mu}, \sigma^2 [1 + \bar{T}^{-1}]). \end{aligned} \quad (7.15)$$

Proceeding as in Exercise 2.8, we have that the predictive density *unconditional* of σ^{-2} is

$$\begin{aligned} p(y_* | y) &= \int_0^\infty \phi(y_* | \bar{\mu}, \sigma^2 [1 + \bar{T}^{-1}]) f_\gamma(\sigma^{-2} | \bar{s}^{-2}, \bar{\nu}) d\sigma^{-2} \\ &= \int_0^\infty \left((2\pi\sigma^2 [1 + \bar{T}^{-1}])^{-(1/2)} \exp \left[-\frac{(y_* - \bar{\mu})^2}{2\sigma^2 [1 + \bar{T}^{-1}]} \right] \right) \\ &\quad \times \left(\left[\left(\frac{2}{\bar{\nu}\bar{s}^2} \right)^{(\bar{\nu}/2)} \Gamma(\bar{\nu}/2) \right]^{-1} \sigma^{-2(\bar{\nu}-2)/2} \exp \left(-\frac{1}{2} \sigma^{-2} \bar{\nu}\bar{s}^2 \right) \right) d\sigma^{-2} \\ &= f_t(y_* | \bar{\mu}, \bar{s}^2 [1 + \bar{T}^{-1}], \bar{\nu}), \end{aligned} \quad (7.16)$$

a univariate Student t density with mean $\bar{\mu}$, scale parameter $\bar{s}^2 [1 + \bar{T}^{-1}]$, and degrees of freedom parameter $\bar{\nu}$ (see Appendix Definition 4).

Exercise 7.8 (Predictive estimation: conjugate normal analysis with unknown mean and variance) Consider Exercise 7.7.

(a) Find the predictive point forecast of y_* under quadratic loss.

(b) Find a $1 - \alpha$ forecast interval for y_* .

Solution

- (a) Under squared error loss, the Bayesian point forecast of y_* is the predictive mean $\bar{\mu}$.
 (b) A $1 - \alpha$ predictive (forecast) interval for y_* is given by

$$\bar{S}_* = \left\{ y_* : \bar{\mu} - t_{[(\alpha/2), \bar{v}]} \bar{s} [1 + \bar{T}^{-1}]^{1/2} < y_* < \bar{\mu} + t_{[(\alpha/2), \bar{v}]} \bar{s} [1 + \bar{T}^{-1}]^{1/2} \right\}, \quad (7.17)$$

where $t_{[(\alpha/2), \bar{v}]}$ cuts off $(\alpha/2)$ probability in the right-hand tail of a Student t -distribution with \bar{v} degrees of freedom.

Exercise 7.9 (Conjugate multivariate normal analysis with unknown mean and covariance matrix) Consider Exercise 7.5 with both μ and Σ unknown. Suppose the prior distribution for μ and Σ^{-1} is the natural conjugate normal-Wishart prior $NW_m(\underline{\mu}, \underline{T}, \underline{S}, \underline{\omega})$ given in Exercise 2.14(c).

- (a) Find the joint posterior distribution of μ and Σ^{-1} .
 (b) Find the predictive point forecast for an out-of-sample y_* under quadratic loss.
 (c) Find the posterior predictive density for y_* and calculate a $1 - \alpha$ forecast interval for y_* .

Solution

(a) The prior specification implies $\mu | \Sigma^{-1} \sim N(\underline{\mu}, \underline{T}^{-1} \Sigma)$ and $\Sigma^{-1} \sim W_m(\underline{S}, \underline{\omega})$. Thus, from Exercise 2.14, the posterior distribution for μ and Σ^{-1} is $NW_m(\bar{\mu}, \bar{T}, \bar{S}, \bar{\omega})$, where $\bar{\mu} = \bar{T}^{-1}(\underline{T}\underline{\mu} + T\bar{y})$, $\bar{T} = \underline{T} + T$, and

$$\bar{S}^{-1} = \underline{S}^{-1} + S + \frac{TT}{\underline{T} + T} (\underline{\mu} - \bar{y})(\underline{\mu} - \bar{y})', \quad (7.18)$$

$$\bar{\omega} = \underline{\omega} + T; \quad (7.19)$$

that is, $\mu | \Sigma^{-1}, y \sim N_m(\bar{\mu}, \bar{T}^{-1} \Sigma)$ and $\Sigma^{-1} | y \sim W_m(\bar{S}, \bar{\omega})$.

- (b) Under quadratic loss, the optimal point predictor of Y_* is $\bar{\mu}$.
 (c) By letting $\Sigma^{-1} > 0$ denote that Σ^{-1} is positive definite, the predictive density is

$$\begin{aligned} p(y_* | y) &= \int_{\Sigma^{-1} > 0} \int_{\mathcal{R}^m} p(y_* | y, \Sigma^{-1}, \mu) p(\mu, \Sigma^{-1} | y) d\mu d\text{vech}(\Sigma^{-1}) \\ &= \int_{\Sigma^{-1} > 0} \left[\int_{\mathcal{R}^m} \phi(y_* | \mu, \Sigma) \phi_m(\mu | \bar{\mu}, \bar{T}^{-1} \Sigma) d\mu \right] f_W(\Sigma^{-1} | \bar{S}, \bar{\omega}) d\text{vech}(\Sigma^{-1}) \\ &= \int_{\Sigma^{-1} > 0} \phi(y_* | \bar{\mu}, [1 + \bar{T}^{-1}] \Sigma) f_W(\Sigma^{-1} | \bar{S}, \bar{\omega}) d\text{vech}(\Sigma^{-1}) \\ &= t(y_* | \bar{\mu}, [1 + \bar{T}^{-1}] \bar{S}, \bar{\omega}). \end{aligned}$$

A $1 - \alpha$ prediction (forecast) ellipsoid for y_* is given by

$$\bar{S}_* = \left\{ y_* : (y_* - \bar{\mu})' \bar{S}^{-1} (y_* - \bar{\mu}) \leq c_\alpha [1 + \bar{T}^{-1}] \right\},$$

where c_α cuts off α probability in the right-hand tail of the $F(m, \bar{\omega} - m)$ distribution.

Exercise 7.10 (Predictive density for conjugate exponential analysis) Suppose, as in Exercise 2.10, Y_t ($t = 1, 2, \dots, T$) is a random sample from an exponential distribution with mean θ^{-1} and that the prior distribution of θ is the gamma distribution $G(\underline{\alpha}, \underline{\beta})$, where $\underline{\alpha} > 0$ and $\underline{\beta} > 0$. Find the predictive density for an out-of-sample value y_* .

Solution

From Exercise 2.10, we know $\theta|y \sim G(\bar{\alpha}, \bar{\beta})$, where $\bar{\alpha} = \underline{\alpha} + T$ and $\bar{\beta} = (\underline{\beta}^{-1} + T\bar{y})^{-1}$. Therefore, predictive density (7.1) becomes

$$p(y_*|y) = \int_0^\infty p(y_*|\theta)p(\theta|y)d\theta \quad (7.20)$$

$$\begin{aligned} &= \int_0^\infty [\theta \exp(-\theta y_*)] \left[(\bar{\beta} \Gamma(\bar{\alpha}))^{-1} \theta^{\bar{\alpha}-1} \exp(-\theta/\bar{\beta}) \right] d\theta \\ &= (\bar{\beta} \Gamma(\bar{\alpha}))^{-1} \int_0^\infty [\theta^{(\bar{\alpha}+1)-1} \exp[-\theta(y_* + \bar{\beta}^{-1})]] d\theta \\ &= \left(\frac{(\bar{\beta}^{-1} + y_*)^{-(\bar{\alpha}+1)} \Gamma(\bar{\alpha} + 1)}{\bar{\beta} \Gamma(\bar{\alpha})} \right) \quad (7.21) \\ &\quad \times \int_0^\infty \left((\bar{\beta}^{-1} + y_*)^{-(\bar{\alpha}+1)} \Gamma(\bar{\alpha} + 1) \right)^{-1} \theta^{(\bar{\alpha}+1)-1} \exp[-\theta/(y_* + \bar{\beta}^{-1})^{-1}] d\theta \\ &= \left(\frac{(\bar{\beta}^{-1} + y_*)^{-(\bar{\alpha}+1)} \Gamma(\bar{\alpha} + 1)}{\bar{\beta} \Gamma(\bar{\alpha})} \right), \quad y_* > 0 \end{aligned}$$

where the integral in the second-to-last line is unity because it is the density of a $G(\bar{\alpha} + 1, [\bar{\beta}^{-1} + y_*]^{-1})$ density. The density in the last line of (7.21) is known as the gamma-gamma p.d.f.

Choice of prior

Most researchers find the logic of Bayesian analysis compelling, once a prior has been specified. The stage of the process where most frequentists can be found circling the wagons is when the prior is chosen.

The pure subjectivist engaged in personal research needs only to elicit the prior that reflects his or her subjective prior beliefs. Usually the likelihood is parameterized to facilitate thinking in terms of θ , and so subject matter considerations should suggest plausible values of θ . Elicitation techniques are nicely surveyed by Garthwaite, Kadane, and O'Hagan (2005). These techniques tend to advocate thinking in terms of beliefs concerning future observables and backing out the implied beliefs regarding the hyperparameters of a conjugate prior for the unobserved parameters. Unlike most researchers, economists seem equally adept at thinking in terms of observables or unobservable parameters. Perhaps this is because the econometrician is both the statistician and the substantive field expert.

But why should prior beliefs conform to the conjugate prior form? One reason is that natural conjugate priors have an interpretation in terms of a prior fictitious sample from the same process that gives rise to the likelihood function. This corresponds to organizing prior beliefs by viewing the observable world through the same parametric window used for viewing the data.

Public research, however, inevitably requires prior sensitivity analysis and the elicitation of a family of priors likely to interest a wide range of readers. This requires the researcher to wear various hats representing interesting professional positions in prior space. Objective Bayesians seek techniques that deliver a single compelling prior that warrants universal appeal to researchers as a reference standard. Because terms like “objective” and “subjective” are quite value loaded, we simply adopt Lindley's (2004) view that “objectivity is merely subjectivity when nearly everyone agrees.”

Objective Bayesians evoke general rules or principles to determine priors in a wide variety of settings. Kass and Wasserman (1996) survey numerous formal rules that have been

suggested for choosing a prior. Many of these rules reflect the desire to “let the data speak for themselves.” These priors are intended to lead to proper posteriors dominated by the data. They also serve as benchmarks for posteriors derived from ideal subjective considerations. At first, many of these priors were also motivated on simplicity grounds. But as problems were discovered, and other issues were seen to be relevant, derivation of such priors became more complicated, possibly even more so than a legitimate attempt to elicit an actual subjective prior.

Technically, objective priors are only positive functions to be formally used in Bayes’ theorem to obtain objective posteriors that, for a given model, describe whatever the data have to say about some particular quantity of interest, either a function of the model’s parameters or a function of future observations. In other words, objective priors are those for which the contributions of the data are *posterior dominant* for the quantity of interest. The resulting objective posteriors are intended to describe the inferential content of the data for scientific communication.

Early interpretations of letting the data speak for themselves sought “noninformative” priors, but the inherent ambiguity of the term led to many candidates. Even in the simple Bernoulli case there were four legitimate candidates [see Geisser (1984)]. In nested hypothesis-testing situations, there arose questions such as: What does it mean to be “equally noninformative” in spaces of different dimensions? To sidestep such annoying questions, researchers have explored other interpretations of letting the data speak for themselves.

One interpretation is to use frequentist reasoning and require posterior probabilities to agree with sampling probabilities. For example, this led researchers to seek “matching” priors that result in credibility regions having approximately the same frequentist coverage (i.e., confidence level) as posterior coverage. But problems with multiparameter cases and nuisance parameters have plagued the approach. Another appeal to frequentist reasoning is to rationalize maximum likelihood estimates in a Bayesian framework by appropriate choice of prior distribution and loss function, specifically a uniform prior and an all-or-nothing loss function. But in what parameterization should one be uniform? This question can also be rephrased: In what parameterization should we wish to be “noninformative”?

To overcome parameterization problems, Jeffreys (1961) sought a general rule for choosing a prior so that the same posterior inferences were obtained regardless of the parameterization chosen. Jeffreys (1961) makes a general argument (but with numerous qualifications) in favor of choosing a prior proportional to the square root of the information matrix, that is,

$$p(\theta) \propto |\mathcal{I}(\theta)|^{1/2},$$

where

$$\mathcal{I}(\theta) \equiv E_{Y|\theta} \left[-\frac{\partial^2 \ln L(\theta)}{\partial \theta \partial \theta'} \right]$$

is the *information matrix of the sample*. This prior has the desirable feature that if the model is reparameterized by a one-to-one transformation, say $\alpha = h(\theta)$, then choosing the prior

$$p(\alpha) \propto \left| E_{Y|\alpha} \left[-\frac{\partial^2 \ln L(\alpha)}{\partial \alpha \partial \alpha'} \right] \right|^{1/2}$$

leads to identical posterior inferences as using $p(\theta)$. Such priors are said to follow *Jeffreys' rule*. There is a fair amount of agreement that such priors may be reasonable in one-parameter problems, but substantially less agreement in multiple-parameter problems. For example, assuming the mean and variance are independent in the Neyman and Scott (1948) problem, gives an inconsistent Bayesian estimator of the variance under Jeffreys' prior.

Kass and Wasserman (1996, p. 1343) argue that Jeffreys' viewpoint evolved toward seeking priors chosen by convention, rather than as unique representations of "ignorance." If the information matrix is difficult to compute (as is often the case with standard econometric models), then Jeffreys' prior is difficult to compute. Also, Jeffreys emphasized the distinction between estimation and testing problems and advocated different priors for each. This set the stage for a result that has plagued subsequent attempts to define "objective" priors where the intentions of the researcher matter.

Usually, Jeffreys' rule and other formal rules reviewed by Kass and Wasserman (1996) lead to *improper priors*, that is, priors that integrate to infinity rather than unity. When blindly plugged into Bayes' theorem as a prior they can lead to proper posterior densities, but not always. Furthermore, improper priors can dominate the data and lead to improper posteriors, inadmissible Bayesian point estimators, incoherent inferences, and *marginalization paradoxes*. See Kass and Wasserman (1996, pp. 1356–1359) for details.

Bernardo (1979) suggested a method for constructing *reference priors*, which was subsequently refined in Berger and Bernardo (1989, 1992). These priors are intended to serve as "defaults," that is, choices that can be made automatically without any contemplation of their suitability in a particular problem. Reference priors maximize the missing information in the experiment as measured by the Kullback–Leibler distance between the posterior and prior density as the sample size approaches infinity in a suitable fashion. In the one-dimensional case, the method leads to Jeffreys' prior. Reference priors deal with marginalization paradoxes. In regular continuous multiparameter problems, they usually lead to the solutions that Jeffreys suggested using ad hoc arguments rather than his general multivariate rule. Reference priors preserve desirable features of the Jeffreys' prior such as the invariance property, but they often avoid paradoxical results produced by Jeffreys' prior in multiparameter settings. For example, under the reference prior the Bayesian estimator of the variance in Neyman and Scott (1948) is consistent. Moreover, reference analysis can deal with nonregular cases that cause problems for other methods. A recent survey of reference analysis was done by Bernardo (2004). Kass and Wasserman (1996, p. 1362) describe reference priors as "the default among the defaults."

Objective priors often have properties that seem rather non-Bayesian. Most objective priors depend on some or all of the following: (a) the form of the likelihood, (b) the sample size, (c) an expectation with respect to the sampling distribution, (d) the parameters of

interest, and (e) whether the researcher is engaging in estimating, testing, or predicting. The dependency in (c) of Jeffreys' prior on a sampling theory expectation makes it sensitive to a host of problems related to the likelihood principle. In light of (d), an objective prior can depend on *subjective* choices such as which are the parameters of interest and which are nuisance parameters. Different quantities of interest require different objective priors that cannot be combined in a coherent manner. Catalogues of objective prior distributions can be found in Kass and Wasserman (1996) and Yang and Berger (1996).

Improper priors cause more problems in hypothesis testing and model selection than in estimation or prediction. The reason is that improper priors involve arbitrary constants that are not restricted by the necessity of the prior to integrate to unity. Although these constants cancel when computing the posterior density, Bayes factors involve a ratio of arbitrary constants. One idea [Berger and Pericchi (1996)] is to use enough data to update an improper prior into a proper posterior and to use the latter to compute a Bayes factor using the remaining data. Of course, there are many such potential training samples. So, try all possible ways and then somehow combine (via, e.g., arithmetic or geometric average) the different Bayes factors. Amazingly, sometimes this results in a legitimate Bayes factor corresponding to a proper prior called the intrinsic prior.

Particularly in high-dimensional problems, the effects of a prior can be subtle: It may have little posterior influence on some functions of the data and have an overwhelming influence on other functions. The message from this and similar examples is that improper priors must be used with care when the dimension of the parameter space is large. Of course, that does not imply that proper priors are necessarily better in these problems. The remarks by Efron (1973) emphasize the practical difficulties with diffuse proper priors that may accompany theoretical difficulties found with improper priors.

Choice of prior is as much an art as it is a science. In the exercises that follow we attempt to illustrate some of the analytical points just made. No exercises on formal elicitation are included, but in subsequent chapters the reader will see how we address the issue of prior choice in the context of specific empirical applications. Our advice is to *use objective priors only with great care, and never alone*. We recommend including objective priors in the class of priors over which a *sensitivity analysis* is performed.

Exercise 8.1 (Bernoulli sampling distribution) Consider a random sample from a Bernoulli distribution with p.m.f. given in (2.5). Find Jeffreys' prior.

Solution

The log likelihood is

$$\begin{aligned}\ln L(\theta) &= \sum_{t=1}^Y [y_t \ln(\theta) + (1 - y_t) \ln(1 - \theta)] \\ &= x \ln(\theta) + (T - x) \ln(1 - \theta),\end{aligned}$$

where the number of successes, $x \equiv y_1 + y_2 + \cdots + y_T$, has a binomial distribution with p.m.f.

$$p(x|T, \theta) = \frac{T!}{x!(T-x)!} \theta^x (1-\theta)^{T-x}$$

and mean $T\theta$. Note that x is a sufficient statistic for θ . The Hessian is

$$\frac{\partial^2 \ln L(\theta)}{\partial \theta^2} = -\frac{x}{\theta^2} - \frac{T-x}{(1-\theta)^2},$$

and the information matrix is

$$\mathcal{I}(\theta) = E_{X|\theta} \left[-\frac{\partial^2 \ln L(\theta)}{\partial \theta^2} \right] = \frac{T\theta}{\theta^2} + \frac{T-T\theta}{(1-\theta)^2} = \frac{T}{\theta} + \frac{T}{(1-\theta)} = \frac{T}{\theta(1-\theta)}.$$

Therefore, Jeffreys' prior is

$$p(\theta) \propto \theta^{-1/2} (1-\theta)^{-1/2}. \quad (8.1)$$

This is recognized as a $\beta(1/2, 1/2)$ distribution. This is a rare case in which the Jeffreys' prior is proper.

Exercise 8.2 (Negative binomial sampling distribution) In contrast to the random variable X in the preceding exercise, consider a random variable Z denoting the number of failures before the α th success. Then Z has a negative binomial distribution with p.m.f.

$$p(z|\alpha, \theta) = \frac{\Gamma(\alpha+z)}{z! \Gamma(\alpha)} \theta^\alpha (1-\theta)^z, \quad z = 0, 1, 2, \dots, \quad (8.2)$$

where α is known, θ is unknown, and $E(Z|\alpha, \theta) = \alpha(1-\theta)/\theta$. Suppose $\alpha = x$, the realized number of successes in Exercise 8.1. Find Jeffrey's prior for θ .

Solution

The log likelihood is

$$\ln L(\theta) = \ln(\Gamma[x+z]) - \ln(z!) - \ln(\Gamma[x]) + x \ln(\theta) + z \ln(1-\theta).$$

The Hessian is

$$\frac{\partial^2 \ln L(\theta)}{\partial \theta^2} = -\frac{x}{\theta^2} - \frac{z}{(1-\theta)^2},$$

and the information matrix is

$$\mathcal{I}(\theta) = E_{Z|\theta} \left[-\frac{\partial^2 \ln L(\theta)}{\partial \theta^2} \right] = \frac{x}{\theta^2} + \frac{x(1-\theta)/\theta}{(1-\theta)^2} = x \left[\frac{1}{\theta^2} + \frac{1}{\theta(1-\theta)} \right] = \frac{x}{\theta^2(1-\theta)}.$$

Therefore, Jeffreys' prior is

$$p(\theta) \propto \theta^{-1} (1-\theta)^{-1/2}, \quad (8.3)$$

which is improper. Therefore, the Jeffreys' prior differs in the binomial and negative binomial cases, even though their likelihood functions are proportional. Thus, use of the

Jeffreys' prior can violate the likelihood principle. The sampling expectation in the definition is the culprit.

Exercise 8.3 (Kass and Wasserman [1996, Section 4.3]) Aside from violating the likelihood principle, sample space-dependent priors can lead to situations where the posterior depends on the order the data are received. Yet for a fixed prior, we get the same posterior no matter what order the data are processed, assuming independence.

(a) Suppose that y_1 is the number of successes (which in this example denotes the “heads” outcome) in n tosses of a coin with success probability θ . Find the posterior under Jeffreys' prior.

(b) Suppose that we flip the coin until another head appears, and suppose that this takes $Y_2 = y_2$ tosses. Using the posterior from (a) as a prior, find the corresponding posterior given y_1 and y_2 .

(c) Now suppose the experiment is performed in reverse order. Beginning with Jeffreys' prior

$$p(\theta) \propto \theta^{-1}(1 - \theta)^{-1/2}$$

for the negative binomial (Exercise 8.2), update sequentially on Y_2 then Y_1 to find the posterior.

Solution

(a) Using Bayes' theorem, (8.1) and the results of Exercise 8.1, we obtain the posterior

$$p_1(\theta|y_1) \propto \theta^{y_1-1/2}(1 - \theta)^{n-y_1-1/2}.$$

(b) In terms of the negative binomial of Exercise 8.2, and (8.2) specifically, we have $\alpha = 1$ and $z = y_2 - 1$. Putting this together with the result in part (a), we obtain

$$p_2(\theta|y_1, y_2) \propto \theta^{y_1+1-1/2}(1 - \theta)^{n-y_1+y_2-1-1/2}.$$

(c) Using Jeffreys' prior for the negative binomial together with the data y_2 yields a posterior of the form

$$p_3(\theta|y_2) \propto \theta^{-1+1}(1 - \theta)^{y_2-1-1/2}.$$

Using this as a prior for θ and combining this prior with the binomial likelihood for the y_1 data gives

$$p_4(\theta|y_1, y_2) \propto \theta^{y_1-1+1}(1 - \theta)^{n-y_1+y_2-1-1/2}.$$

So, the posterior in (c) is different from the posterior in (b).

Exercise 8.4 (Poisson sampling distribution) Suppose Y_t ($t = 1, 2, \dots, T$) is a random sample from a Poisson distribution with mean θ . Find Jeffreys' prior.

Solution

Using the likelihood function given in the solution to Exercise 2.12, the Hessian is

$$\frac{\partial^2 \ln L(\theta)}{\partial \theta^2} = -\frac{T\bar{y}}{\theta^2},$$

and the information matrix is

$$\mathcal{I}(\theta) = E_{\bar{Y}|\theta} \left[-\frac{\partial^2 \ln L(\theta)}{\partial \theta^2} \right] = \frac{T}{\theta}.$$

Therefore, Jeffreys' prior is

$$p(\theta) \propto \theta^{-1/2}.$$

Exercise 8.5 (Exponential sampling distribution and reparameterization) Suppose Y_t ($t = 1, 2, \dots, T$) are iid random variables from an exponential distribution with mean θ .

(a) Derive Jeffreys' prior for θ .

(b) Derive Jeffreys' prior for $\alpha = \theta^{-1}$.

(c) Find the posterior density of θ corresponding to the prior density in (a). Be specific in noting the family to which it belongs.

(d) Find the posterior density of α corresponding to the prior density in (b). Be specific in noting the family to which it belongs.

Solution

(a) The sampling density

$$p(y_t|\theta) = \theta^{-1} \exp(-y_t/\theta), \quad 0 < y_t < \infty,$$

implies the log likelihood

$$\ln L(\theta) = -T \ln(\theta) - T\bar{y}\theta^{-1}$$

and the derivatives

$$\begin{aligned} \frac{\partial \ln L(\theta)}{\partial \theta} &= -\frac{T}{\theta} + \frac{T\bar{y}}{\theta^2}, \\ \frac{\partial^2 \ln L(\theta)}{\partial \theta^2} &= \frac{T}{\theta^2} - \frac{2T\bar{y}}{\theta^3}. \end{aligned}$$

Hence, we have the information matrix [recalling $E(\bar{Y}) = \theta$]

$$\mathcal{I}(\theta) = E_{\bar{Y}|\theta} \left[-\frac{\partial^2 \ln L(\theta)}{\partial \theta^2} \right] = -\frac{T}{\theta^2} + \frac{2T}{\theta^2} = \frac{T}{\theta^2}.$$

Therefore, Jeffreys' prior is

$$p(\theta) \propto \theta^{-1}.$$

(b) Similarly,

$$\ln L(\alpha) = T \ln(\alpha) - T\bar{y}\alpha,$$

with derivatives

$$\begin{aligned}\frac{\partial \ln L(\alpha)}{\partial \alpha} &= \frac{T}{\alpha} - T\bar{y}, \\ \frac{\partial^2 \ln L(\alpha)}{\partial \alpha^2} &= -\frac{T}{\alpha^2}.\end{aligned}$$

Note that observed information is constant in this case. Therefore, Jeffreys' prior is

$$p(\alpha) \propto \alpha^{-1}.$$

(c) The posterior density of θ corresponding to the prior density in (a) is

$$\begin{aligned}p(\theta|y) &\propto \theta^{-1} \theta^{-T} \exp(-T\bar{y}\theta^{-1}) \\ &= \theta^{-(T+1)} \exp(-T\bar{y}\theta^{-1}),\end{aligned}$$

which is recognized to be the kernel of an inverted gamma density (see Appendix Theorem 2), specifically $IG(T, [T\bar{y}]^{-1})$.

(d) The posterior density of α corresponding to the prior density in (b) is

$$\begin{aligned}p(\alpha|y) &\propto \alpha^{-1} \alpha^T \exp(-T\bar{y}\alpha) \\ &\propto \alpha^{T-1} \exp(-T\bar{y}\alpha),\end{aligned}$$

which is recognized to be the kernel of a gamma density (see Appendix Theorem 2), specifically $G(T, [T\bar{y}]^{-1})$. The invariance property of Jeffreys' prior refers to the fact that the posteriors in (c) and (d) are related by a simple change of variable.

Exercise 8.6 (Normal sampling distribution with unknown mean and variance) Consider a random sample from an $N(\mu, \sigma^2)$ distribution with $\theta = [\mu \ \sigma^2]'$ unknown. Find the prior that is proportional to the square root of the information matrix.

Solution

It is straightforward to show that the Hessian matrix is

$$\frac{\partial^2 \ln L(\theta)}{\partial \theta \partial \theta'} = \begin{bmatrix} -T/\sigma^2 & -(1/\sigma^4) \sum_{t=1}^T (y_t - \mu) \\ -(1/\sigma^4) \sum_{t=1}^T (y_t - \mu) & [T/(2\sigma^4)] - (1/\sigma^6) \sum_{t=1}^T (y_t - \mu)^2 \end{bmatrix},$$

and the information matrix is

$$\mathcal{I}(\theta) = E_{Y|\theta} \left[-\frac{\partial^2 \ln L(\theta)}{\partial \theta \partial \theta'} \right] = \begin{bmatrix} T/\sigma^2 & 0 \\ 0 & T/(2\sigma^4) \end{bmatrix}.$$

Therefore, $|\mathcal{I}(\theta)|^{1/2} = T/[\sqrt{2}\sigma^3]$, yielding the prior

$$p(\theta) \propto \sigma^{-3}.$$

This result led Jeffreys to modify his general rule. He argued that in this case the prior for each parameter should be treated separately and then multiplied. In the case of known σ^2 , Jeffreys' prior for μ is $p(\mu) \propto c$, where c is an arbitrary constant. In the case of known μ , Jeffreys' prior for σ^2 is $p(\sigma^2) \propto \sigma^{-2}$. Therefore, in the case of unknown μ and σ^2 , Jeffreys recommended the commonly used prior $p(\theta) \propto \sigma^{-2}$.

Exercise 8.7 (Hypothesis testing with improper priors) Problems arise in testing $H_1 : \theta_1 \in \Theta_1$ versus $H_2 : \theta_2 \in \Theta_2$ if improper priors such as $p(\theta_i|H_i) \propto c_i$, $i = 1, 2$, are used. Demonstrate this fact.

Solution

Consider the Bayes factor

$$B_{12} = \frac{\int_{\Theta_1} p(\theta_1|H_1)L(\theta_1) d\theta_1}{\int_{\Theta_2} p(\theta_2|H_2)L(\theta_2) d\theta_2} = \frac{c_1 \int_{\Theta_1} L(\theta_1) d\theta_1}{c_2 \int_{\Theta_2} L(\theta_2) d\theta_2}.$$

Even if the integrals converge, B_{12} remains indeterminate because c_1/c_2 is a ratio of arbitrary constants.

Exercise 8.8 (Bayes factor with a diffuse proper prior) We have noted earlier that Bayes factors are arbitrary when improper priors are used (e.g., Exercise 8.7) but that this difficulty does not occur with proper priors. This may lead the unwary reader to think that employing a proper prior with a very large variance overcomes the former problem, while at the same time serving to let the data speak for themselves. This exercise shows that such reasoning is incorrect.

Suppose Y_t ($t = 1, 2, \dots, T$), given θ , are iid $N(0, 1)$ under hypothesis H_1 and $N(\theta, 1)$ under hypothesis H_2 . Assume the prior $\theta|H_2 \sim N(0, v)$, $v > 0$. Find the Bayes factor B_{21} and discuss its behavior for large v .

Solution

The problem can be embedded in Exercise 2.3 and its notation. Under H_1 there are no unknown parameters, and so the likelihood and marginal likelihood functions are the same. In particular,

$$p(y|H_1) = L(\theta) = (2\pi)^{-T/2} \exp \left[-(1/2) \sum_{t=1}^T y_t^2 \right].$$

Under H_2 , $p(y|H_2)$ can be obtained from (2.21) (with $\theta_2 = 1$) and (2.23) (with $\underline{\mu} = 0$):

$$\begin{aligned} p(y|H_2) &= \left[(2\pi)^{-\nu/2} T^{-1/2} \exp(-\nu s^2/2) \right] \left[(2\pi[v + T^{-1}])^{-1/2} \exp \left(-\frac{\bar{y}^2}{2(v + T^{-1})} \right) \right] \\ &= (2\pi)^{-T/2} T^{-1/2} (v + T^{-1})^{-1/2} \exp \left[-(1/2) \left(\sum_{t=1}^T y_t^2 - T\bar{y}^2 + \frac{\bar{y}^2}{v + T^{-1}} \right) \right]. \end{aligned}$$

The Bayes factor B_{21} is

$$\begin{aligned} B_{21} &= \frac{p(y|H_2)}{p(y|H_1)} \\ &= T^{-1/2}(v + T^{-1})^{-1/2} \exp \left[-(1/2) \left(\sum_{t=1}^T y_t^2 - T\bar{y}^2 + \frac{\bar{y}^2}{v + T^{-1}} - \sum_{t=1}^T y_t^2 \right) \right] \\ &= T^{-1/2}(v + T^{-1})^{-1/2} \exp \left[\left(\frac{vT}{1 + vT} \right) \frac{T\bar{y}^2}{2} \right]. \end{aligned}$$

For moderate T and large v ,

$$B_{21} \approx v^{-1/2} T^{-1/2} \exp(T\bar{y}/2).$$

Therefore, B_{21} depends heavily on the arbitrarily chosen “large” value of v .

***Exercise 8.9 (Strong inconsistency of improper priors)** Consider a random sample Y_1, Y_2, \dots, Y_T from an $N(\theta, 1)$ distribution. Let A be the event that $|\bar{Y}| \geq |\theta|$, where \bar{Y} is the sample mean. Suppose the prior for θ is $p(\theta) \propto \text{constant}$. Let $\Phi(c|\mu, \sigma^2)$ denote the normal c.d.f. at c with mean μ and variance σ^2 . That is, $\Phi(c|\mu, \sigma^2) \equiv \Pr(z \leq c)$, where $z \sim N(\mu, \sigma^2)$.

(a) From the standpoint of the sampling model, show that

$$\Pr(A|\theta) = (1/2) + \Phi(-2|\theta|\sqrt{T}|0, 1).$$

(b) From the standpoint of the posterior distribution, show that

$$\Pr(A|y) = (1/2) - \Phi(-2|\bar{y}|\sqrt{T}|0, 1).$$

(c) Are (a) and (b) consistent?

Solution

(a) We know that $\bar{Y}|\theta \sim N(\theta, T^{-1})$. Let $X = |\bar{Y}|$. Then the *sampling c.d.f.* of X is

$$\begin{aligned} F_X(c) &= \Pr(X \leq c) = \Pr(|\bar{Y}| \leq c) = \Pr(-c \leq \bar{Y} \leq c) \\ &= \Phi(c|\theta, T^{-1}) - \Phi(-c|\theta, T^{-1}). \end{aligned} \tag{8.4}$$

Therefore,

$$\begin{aligned} \Pr(A|\theta) &= \Pr(X \geq |\theta|) \\ &= 1 - F_X(|\theta|) \\ &= 1 - \Phi(|\theta| \mid \theta, T^{-1}) + \Phi(-|\theta| \mid \theta, T^{-1}). \end{aligned} \tag{8.5}$$

If $\theta > 0$ then $|\theta| = \theta$, $\Phi(|\theta| \mid \theta, T^{-1}) = (1/2)$, and $\Phi(-|\theta| \mid \theta, T^{-1}) = \Phi(-2\theta\sqrt{T}|0, 1) = \Phi(-2|\theta|\sqrt{T}|0, 1)$. Hence, (8.5) becomes

$$\Pr(A|\theta) = (1/2) + \Phi(-2|\theta|\sqrt{T}|0, 1). \tag{8.6}$$

If $\theta \leq 0$ then $|\theta| = -\theta$, $\Phi(|\theta| \mid \theta, T^{-1}) = \Phi(-2\theta\sqrt{T}|0, 1) = \Phi(2|\theta|\sqrt{T}|0, 1)$, and $\Phi(-|\theta| \mid \theta, T^{-1}) = (1/2)$. Hence, (8.5) becomes

$$\begin{aligned}\Pr(A|\theta) &= (3/2) - \Phi(2|\theta|\sqrt{T}|0, 1) \\ &= (3/2) - [1 - \Phi(-2|\theta|\sqrt{T}|0, 1)] \\ &= (1/2) + \Phi(-2|\theta|\sqrt{T}|0, 1).\end{aligned}\tag{8.7}$$

Since (8.6) and (8.7) are the same, their common value holds for all θ .

(b) Under the uniform prior for θ , we know that $\theta|y \sim N(\bar{y}, T^{-1})$. Let $X = |\theta|$. Then, similar to (8.4), the *posterior c.d.f.* of X is

$$F_{X|y}(c) = \Phi(c|\bar{y}, T^{-1}) - \Phi(-c|\bar{y}, T^{-1}).\tag{8.8}$$

Therefore,

$$\Pr(A|y) = \Pr(|\theta| \leq |\bar{y}| \mid y) = F_X(|\bar{y}|) = \Phi(|\bar{y}| \mid \bar{y}, T^{-1}) - \Phi(-|\bar{y}| \mid \bar{y}, T^{-1}).\tag{8.9}$$

If $\bar{y} > 0$ then $|\bar{y}| = \bar{y}$, $\Phi(|\bar{y}| \mid \bar{y}, T^{-1}) = (1/2)$, and $\Phi(-|\bar{y}| \mid \bar{y}, T^{-1}) = \Phi(-2\bar{y}\sqrt{T}|0, 1) = \Phi(-2|\bar{y}|\sqrt{T}|0, 1)$. Hence, (8.9) becomes

$$\Pr(A|y) = (1/2) - \Phi(-2|\bar{y}|\sqrt{T}|0, 1).\tag{8.10}$$

If $\bar{y} \leq 0$ then $|\bar{y}| = -\bar{y}$, $\Phi(|\bar{y}| \mid \bar{y}, T^{-1}) = \Phi(-2\bar{y}\sqrt{T}|0, 1) = \Phi(2|\bar{y}|\sqrt{T}|0, 1)$, and $\Phi(-|\bar{y}| \mid \bar{y}, T^{-1}) = (1/2)$. Hence, (8.9) becomes

$$\begin{aligned}\Pr(A|y) &= \Phi(2|\bar{y}|\sqrt{T}|0, 1) - (1/2) \\ &= [1 - \Phi(-2|\bar{y}|\sqrt{T}|0, 1)] - (1/2) \\ &= (1/2) - \Phi(-2|\bar{y}|\sqrt{T}|0, 1).\end{aligned}\tag{8.11}$$

Since (8.10) and (8.11) are the same, their common value holds for all \bar{y} .

(c) No. According to (a), $\Pr(A|\theta) > 1/2$ for all θ and hence $\Pr(A) > 1/2$. However, according to (b), $\Pr(A|\bar{y}) < 1/2$ for all \bar{y} and hence $\Pr(A) < 1/2$. Therefore, the two are inconsistent (known as *strong inconsistency* in the literature). This demonstrates one problem with improper priors and is related to incoherence. See Kass and Wasserman (1996, Section 4.2) for a discussion of further problems.

Asymptotic Bayes

As seen in previous chapters, the Bayesian approach typically involves the evaluation of multidimensional integrals such as those appearing in posterior expectations, marginal likelihoods, and predictive densities. In nonconjugate situations, these integrals cannot be calculated analytically. In future chapters we present questions relating to various methods for evaluating such integrals using posterior simulation. However, posterior simulation can be computationally demanding and, hence, Bayesians sometimes use asymptotic approximations. Such approximations are the focus of this chapter.

In most situations the prior p.d.f. does not depend on sample size T , and so in large samples, the likelihood eventually dominates the prior density over their common support. This suggests that consideration of the behavior of the posterior density as $T \rightarrow \infty$ may provide useful analytical approximations when T is in fact large. Such "Bayesian asymptotics" differ, however, from frequentist asymptotics in an important way. The thing being approximated in Bayesian asymptotics is conceptually well defined in finite samples, namely the posterior p.d.f. Nuisance parameters can be addressed in a straightforward manner by marginalizing them out of the posterior to obtain the exact marginal posterior p.d.f. for the parameters of interest. In general, there is no frequentist classical sampling distribution of estimators in finite samples that is free of nuisance parameters.

In the same way that sampling distributions of maximum likelihood estimators (MLEs) in regular situations are asymptotically normal, posterior densities are also approximately normal. This result is stated informally in the following theorem. See Bernardo and Smith (1994, pp. 285–297) and the references cited therein for more details. Also see Kwan (1998) for discussion of cases in which Bayesian and frequentist asymptotic inferences differ because the asymptotic convergence is nonuniform (e.g., unit root models).

Theorem 9.1 Under suitable regularity conditions, the posterior density can be approximated as $T \rightarrow \infty$ by $\phi_k \left(\theta | \hat{\theta}_{\text{ML}}, \left[\mathcal{I} \left(\hat{\theta}_{\text{ML}} \right) \right]^{-1} \right)$, where $\hat{\theta}_{\text{ML}}$ is the MLE and $\mathcal{I}(\cdot)$ is the information matrix.

While not providing a formal statement and proof of Theorem 9.1, we note that the prior, $p(\theta)$, must provide positive density to a neighborhood of the true θ_0 appearing in frequentist asymptotics, and the prior density should not depend on T . Also, the parameter vector needs to be identified. The approximation itself in Theorem 9.1 does not depend on the prior and is essentially numerically (but not conceptually) identical to the approximation offered by the asymptotic distribution theory for maximum likelihood. Berger (1985, p. 224) argues informally that the quality of the approximation in Theorem 9.1 can usually be improved in finite samples by the following approximations that reflect the observed data and the prior.

Corollary The following approximations offer increasing improvements on the approximation in Theorem 9.1.

(a) $\phi_k \left(\theta | \hat{\theta}_{\text{ML}}, \left[-\overline{\mathcal{H}} \left(\hat{\theta}_{\text{ML}} \right) \right]^{-1} \right)$, where $\overline{\mathcal{H}}(\theta)$ is the Hessian of the log posterior

$$\overline{\mathcal{H}}(\theta) = \frac{\partial^2 \ln [p(\theta)]}{\partial \theta \partial \theta'} + \frac{\partial^2 \ln L(\theta)}{\partial \theta \partial \theta'}, \quad (9.1)$$

(b) $\phi_k \left(\theta | \hat{\theta}, \left[-\overline{\mathcal{H}} \left(\hat{\theta} \right) \right]^{-1} \right)$, where $\hat{\theta}$ is the posterior mode,

(c) $\phi_k(\theta | E[\theta | y], \text{Var}[\theta | y])$.

Any of these normal approximations can be used to provide approximations to, for example, posterior means and variances.

There are other, more accurate, approximations for posterior features of interest of which those based on *Laplace's approximation* are the most popular. To explain the use of Laplace's approximation, we begin at a high level of generality. The appearance of integrals is ubiquitous in Bayesian analysis, and often the integrals involved do not have closed-form expressions. Thus convenient approximations to such integrals are desirable. Consider the integral

$$J = \int_{\Theta} b(\theta) \exp[-Th(\theta)] d\theta, \quad (9.2)$$

where θ is $K \times 1$, $\Theta = \mathcal{R}^K$, $-h(\cdot)$ is a twice-differentiable function having a unique maximum at $\hat{\theta}$, $\partial h(\theta) / \partial \theta|_{\theta=\hat{\theta}} = 0$, $H(\theta) = \frac{\partial^2 h(\theta)}{\partial \theta \partial \theta'} > 0$, with $b(\cdot)$ is continuous in a neighborhood of $\hat{\theta}$, with $b(\hat{\theta}) \neq 0$. Expanding $h(\theta)$ as a second-order Taylor series around $\hat{\theta}$ and dropping the remainder yields an approximation to $\exp[-Th(\theta)]$ proportional to a

normal density. When integrated against this normal density, the $O(T^{-1/2})$ terms of the expansion of $b(\theta)$ and $h(\theta)$ vanish and Laplace's approximation to (9.2) is

$$\hat{J} = (2\pi)^{\frac{K}{2}} b(\hat{\theta}) \left| TH(\hat{\theta}) \right|^{-\frac{1}{2}} \exp \left[-Th(\hat{\theta}) \right], \quad (9.3)$$

which satisfies $J = \hat{J} [1 + O(T^{-1})]$. The assumption $\Theta = \mathcal{R}^K$ can usually be satisfied by reparameterization.

Often ratios of integrals such as (9.2) are encountered. Consider

$$r = \frac{J_N}{J_D} = \frac{\int_{\mathcal{R}^{K_N}} b_N(\theta_N) \exp[-Th_N(\theta_N)] d\theta_N}{\int_{\mathcal{R}^{K_D}} b_D(\theta_D) \exp[-Th_D(\theta_D)] d\theta_D}, \quad (9.4)$$

where for $i = N, D$, θ_i is $K_i \times 1$, $-h_i(\cdot)$ is a twice-differentiable function having a unique maximum at $\hat{\theta}_i$, $\frac{\partial h_i(\theta_i)}{\partial \theta_i} \Big|_{\theta_i=\hat{\theta}_i} = 0$, $H_i(\theta) = \frac{\partial^2 h_i(\theta_i)}{\partial \theta_i \partial \theta_i^T} > 0$, and $b_i(\cdot)$ is continuous in a neighborhood of $\hat{\theta}_i$, with $b_i(\hat{\theta}_i) \neq 0$. If $h_D(\theta) = h_N(\theta)$, then (9.4) is said to be in *standard form*. Applying Laplace's method separately to the numerator and denominator in (9.4) yields

$$\hat{r} = \frac{\hat{J}_N}{\hat{J}_D} = \frac{(2\pi)^{\frac{K_N}{2}} b_N(\hat{\theta}_N) \left| TH_N(\hat{\theta}_N) \right|^{-\frac{1}{2}} \exp \left[-Th_N(\hat{\theta}_N) \right]}{(2\pi)^{\frac{K_D}{2}} b_D(\hat{\theta}_D) \left| TH_D(\hat{\theta}_D) \right|^{-\frac{1}{2}} \exp \left[-Th_D(\hat{\theta}_D) \right]}. \quad (9.5)$$

Once again, $r = \hat{r} [1 + O(T^{-1})]$.

Posterior moments are an example of ratios such as (9.4). Suppose $g : \mathcal{R}^K \rightarrow \mathcal{R}$ and consider

$$\begin{aligned} G \equiv E[g(\theta) | y] &= \int_{\mathcal{R}^K} g(\theta) p(\theta | y) d\theta \\ &= \frac{\int_{\mathcal{R}^K} g(\theta) p(\theta) L(\theta) d\theta}{\int_{\mathcal{R}^K} p(\theta) L(\theta) d\theta}. \end{aligned} \quad (9.6)$$

A sufficient condition for the existence of G is that all integrals required are defined on bounded regions and have integrands of bounded variation. There are many ways of applying Laplace's method, yielding different estimates with relative error of $O(T^{-1})$. The choice of $b_i(\cdot)$ and $h_i(\cdot)$ ($i = N, D$) affects the Laplace estimate of G .

If $g(\theta) > 0$, then we can use the *fully exponential method* of Tierney and Kadane (1986) to get an easy approximation with an improved relative error of $O(T^{-2})$. G in (9.6) is a special case of (9.4) with $\theta = \theta_N = \theta_D$, $K = K_N = K_D$, $b_N(\theta) = b_D(\theta) = 1$,

$$h_D(\theta) = -(1/T) (\ln[p(\theta)] + \ln[L(\theta)]),$$

$$h_N(\theta) = h_D(\theta) - (1/T) \ln[g(\theta)],$$

$\hat{\theta}_N = \operatorname{argmax}_{\theta} \{-h_N(\theta)\}$, and $\hat{\theta}_D = \operatorname{argmax}_{\theta} \{-h_D(\theta)\}$ is the posterior mode. Writing (9.6) as

$$G = \frac{\int_{\mathcal{R}^K} \exp[-Th_N(\theta)] d\theta}{\int_{\mathcal{R}^K} \exp[-Th_D(\theta)] d\theta} \quad (9.7)$$

$$= \frac{\int_{\mathcal{R}^K} \exp\left[-T\left(h_N(\hat{\theta}_N) - \frac{1}{2}(\theta - \hat{\theta}_N)' \left[-H_N(\hat{\theta}_N)^{-1}\right]^{-1}(\theta - \hat{\theta}_N) + R_N\right)\right] d\theta}{\int_{\mathcal{R}^K} \exp\left[-T\left(h_D(\hat{\theta}_D) - \frac{1}{2}(\theta - \hat{\theta}_D)' \left[-H_D(\hat{\theta}_D)^{-1}\right]^{-1}(\theta - \hat{\theta}_D) + R_D\right)\right] d\theta},$$

where R_N and R_D denote Taylor series remainders, we obtain the Tierney–Kadane estimate

$$\hat{G}_{\text{TK}} = \frac{|H_N(\hat{\theta}_N)|^{-\frac{1}{2}} \exp[-Th_N(\hat{\theta}_N)]}{|H_D(\hat{\theta}_D)|^{-\frac{1}{2}} \exp[-Th_D(\hat{\theta}_D)]}, \quad (9.8)$$

and $E[g(\theta)|y] = \hat{G}_{\text{TK}}[1 + O(T^{-2})]$.

In the full exponential parameterization the $O(T^{-1})$ terms in the numerator and denominator cancel, resulting in the improved $O(T^{-2})$ approximation. Note, however, that (9.8) is *not* invariant to reparameterizations. Illustrative examples suggest that the reparameterizations that lead to locally uniform noninformative prior representations typically improve the accuracy of approximations. Hence, we recommend reparameterizing so that the new Hessian matrices are constant over \mathcal{R}^K . If $g(\theta) \leq 0$, the fully exponential method cannot be applied directly. Tierney, Kass, and Kadane (1989) discuss two alternatives for this case.

In the following exercises we investigate these approximations in the context of simple problems in which analytical expressions for all steps are available and hence for which the quality of the approximations can be analytically studied.

Exercise 9.1 (Posterior means) Identify (9.6) in each of the following cases:

- (a) $K = 1$ and $g(\theta) = \theta^r$.
- (b) A is a subset of Θ and $g(\theta) = I_A(\theta)$ is an indicator function equal to unity when $\theta \in A$ and equal to zero when $\theta \notin A$.
- (c) $g(\theta) = p(y_*|y, \theta)$ for some out-of-sample value y_* .
- (d) $g(\theta) = \exp(v'\theta)$, where v is a given $K \times 1$ vector of constants.
- (e) $\alpha = [\beta' \theta']'$ and $g(\theta) = p(\beta|\theta, y)$.

Solution

- (a) G is the r th posterior moment of θ around the origin.
- (b) $G = \Pr(A|y) = \Pr(\theta \in A|y)$.

- (c) G is the predictive density $p(y_*|y)$.
 (d) G is the posterior moment-generating function of θ evaluated at v .
 (e) G is the marginal posterior density $p(\beta|y)$.

Exercise 9.2 (Laplace approximation of a gamma integrating constant) Suppose $\theta|y \sim G(\bar{\alpha}, \bar{\beta})$ as in Exercise 2.10. The kernel of this distribution is $\theta^{\bar{\alpha}-1} \exp(-\theta/\bar{\beta})$. This exercise explores the use of the Laplace approximation to estimate the integrating constant $c = \beta^{\bar{\alpha}} \Gamma(\bar{\alpha})$ of the gamma density. Discuss the Laplace approximation using (9.2) with $b(\theta) = \theta^{\bar{\alpha}-1}$ and $h(\theta) = -\theta/[T\bar{\beta}]$.

Solution

Since $-h(\theta) = \theta/[T\bar{\beta}]$ does not have an interior maximum over $\theta > 0$, the Laplace method cannot be used in this parameterization.

Exercise 9.3 (Laplace approximation of a gamma integrating constant, 2) Consider Exercise 9.2. Write the kernel in the full exponential form $\exp[-Th(\theta)]$ and find the Laplace approximation \hat{c}_1 of c .

Solution

Straightforwardly,

$$h(\theta) = -T^{-1}[(\bar{\alpha} - 1) \ln(\theta) - \theta/\bar{\beta}].$$

Equating

$$\frac{\partial h(\theta)}{\partial \theta} = -T^{-1} \left[\frac{\bar{\alpha} - 1}{\theta} - \bar{\beta}^{-1} \right]$$

implies a mode at $\hat{\theta} = \bar{\beta}(\bar{\alpha} - 1)$. Note that

$$\exp[-Th(\hat{\theta})] = \exp \left[(\bar{\alpha} - 1) \ln \hat{\theta} - \hat{\theta}/\bar{\beta} \right] = [\bar{\beta}(\bar{\alpha} - 1)]^{\bar{\alpha}-1} \exp[-(\bar{\alpha} - 1)].$$

Furthermore, the Hessian is

$$H(\theta) = \frac{\partial^2 h(\theta)}{\partial \theta^2} = \frac{\bar{\alpha} - 1}{T\theta^2}$$

and when evaluated at the mode equals

$$H(\theta) = \left[T\bar{\beta}^2(\bar{\alpha} - 1) \right]^{-1}.$$

Therefore, using (9.3) with $b(\theta) = 1$ yields the Laplace approximation

$$\begin{aligned} \hat{c}_1 &= \sqrt{2\pi} |TH(\hat{\theta})|^{-1/2} \exp[-Th(\hat{\theta})] \\ &= \sqrt{2\pi} \bar{\beta}(\bar{\alpha} - 1)^{1/2} [\bar{\beta}(\bar{\alpha} - 1)]^{\bar{\alpha}-1} \exp[-(\bar{\alpha} - 1)] \\ &= \bar{\beta}^{\bar{\alpha}} \left\{ [2\pi(\bar{\alpha} - 1)]^{1/2} [\bar{\alpha} - 1]^{\bar{\alpha}-1} \exp[-(\bar{\alpha} - 1)] \right\}. \end{aligned} \tag{9.9}$$

It turns out that the term in curly braces in the last line of (9.9) is Stirling's approximation to $(\bar{\alpha} - 1)! = \Gamma(\bar{\alpha})$. Therefore, \hat{c}_1 is a sensible estimate of $c = \bar{\beta}^{\bar{\alpha}} \Gamma(\bar{\alpha})$.

Exercise 9.4 (Laplace approximation of a gamma integrating constant, 3) Consider Exercise 9.3. Because $0 < \theta < \infty$, there is reason to believe the result in this case can be improved by transforming $\gamma \in \mathcal{R}$, where, say, $\gamma = \ln(\theta)$. Find the Laplace approximation \hat{c}_2 of c using the parameterization in terms of γ .

Solution

Noting that $d\gamma = \theta^{-1} d\theta$, we obtain the integral of the transformed kernel,

$$\begin{aligned} \int_0^\infty \theta^{\bar{\alpha}-1} \exp(-\theta/\bar{\beta}) d\theta &= \int_{-\infty}^\infty \exp[\gamma(\bar{\alpha} - 1)] \exp[-\exp(\gamma)/\bar{\beta}] \exp(\gamma) d\gamma \\ &= \int_{-\infty}^\infty \exp[\gamma\bar{\alpha} - \exp(\gamma)/\bar{\beta}] d\gamma \\ &= \int_{-\infty}^\infty \exp[-Th(\gamma)] d\gamma, \end{aligned}$$

where

$$h(\gamma) = -T^{-1}[\gamma\bar{\alpha} - \bar{\beta}^{-1} \exp(\gamma)].$$

Equating

$$\frac{\partial h(\gamma)}{\partial \gamma} = -T^{-1}[\bar{\alpha} - \bar{\beta}^{-1} \exp(\gamma)]$$

to zero implies a mode at

$$\hat{\gamma} = \ln(\bar{\beta}\bar{\alpha}).$$

Note that this value of $\hat{\gamma}$ implies $\hat{\theta} = \bar{\beta}\bar{\alpha}$, which is different than the mode in Exercise 9.3. Also note that

$$\exp[-Th(\hat{\gamma})] = \exp[\bar{\alpha} \ln(\bar{\beta}\bar{\alpha}) - \bar{\alpha}] = \bar{\beta}^{\bar{\alpha}} \bar{\alpha}^{\bar{\alpha}} \exp(-\bar{\alpha}).$$

The Hessian is

$$H(\gamma) = \frac{\partial^2 h(\gamma)}{\partial \gamma^2} = T^{-1} \bar{\beta}^{-1} \exp(\gamma),$$

and when evaluated at the mode it equals

$$H(\hat{\gamma}) = T^{-1} \bar{\alpha}.$$

Therefore, using (9.3) with $b(\cdot) = 1$ yields the Laplace approximation

$$\begin{aligned} \hat{c}_2 &= \sqrt{2\pi} |TH(\hat{\theta})|^{-1/2} \exp[-Th(\hat{\theta})] \\ &= \sqrt{2\pi} \bar{\alpha}^{-1/2} [\bar{\beta}\bar{\alpha}]^{\bar{\alpha}} \exp[-\bar{\alpha}] \\ &= \bar{\beta}^{\bar{\alpha}} \left\{ [2\pi\bar{\alpha}]^{1/2} \bar{\alpha}^{\bar{\alpha}} \exp[-\bar{\alpha}] \right\} \bar{\alpha}^{-1}. \end{aligned} \tag{9.10}$$

Table 9.1:
Comparison of two Laplace approximations of a gamma integrating constant.

$\bar{\alpha}$	$\Gamma(\bar{\alpha})$	$\hat{c}_1/[\bar{\beta}^{\bar{\alpha}}]$	$\hat{c}_2/[\bar{\beta}^{\bar{\alpha}}]$	Absolute Percentage Error	
				\hat{c}_1	\hat{c}_2
2	1	.922	.960	7.8	4.0
3	2	1.919	1.945	4.0	2.7
4	6	5.836	5.877	2.7	2.1
5	24	23.506	23.604	2.1	1.7
6	120	118.019	118.346	1.7	1.4
7	720	710.078	711.485	1.4	1.2
8	5,040	4,980.396	4,987.799	1.2	1.0
9	40,320	39,902.395	39,948.541	1.0	.9
10	362,880	359,536.870	359,869.560	.9	.8

Again, the term in curly braces in the final expression of (9.10) is Stirling's approximation to $\bar{\alpha}! = \Gamma(\bar{\alpha} + 1)$. Hence, $\{\cdot\} \bar{\alpha}^{-1} = (\bar{\alpha} - 1)\Gamma(\bar{\alpha})$. Therefore, \hat{c}_2 is also a sensible estimate of $c_2 = \bar{\beta}^{\bar{\alpha}} \Gamma(\bar{\alpha})$.

Exercise 9.5 (Comparison of Laplace approximations of a gamma integrating constant) Consider Exercises 9.3 and 9.4. Compare the relative approximation errors

$$|\hat{c}_i - c|/c, \quad i = 1, 2,$$

for $\bar{\alpha} = \underline{\alpha} + T = 2, 3, 4, \dots, 10$.

Solution

In calculating the percentage errors the value of $\bar{\beta}$ drops out, and the comparison amounts to how well do $\hat{c}_i/[\bar{\beta}^{\bar{\alpha}}]$, $i = 1, 2$, approximate $\Gamma(\bar{\alpha})$. Straightforward computation yields the percentage errors in Table 9.1. Although the percentage errors are impressively small for both estimates, clearly the reparameterization to the form $\gamma = \ln(\theta)$ results in an improved Laplace approximation.

Exercise 9.6 (Laplace approximation of the posterior mean of an exponential random variable) Suppose $Y_t|\theta$, $t = 1, 2, \dots, T$ are iid exponential random variables with mean θ . Consider Jeffreys' prior $p(\theta) = \theta^{-1}$. According to Exercise 8.5(c), the corresponding posterior distribution is the inverted gamma $IG(T, [T\bar{y}]^{-1})$ with posterior density

$$p(\theta|y) = c^{-1} \theta^{-(T+1)} \exp(-T\bar{y}\theta^{-1}),$$

$c = \Gamma(T)[T\bar{y}]^{-T}$, and posterior mean equaling $E(\theta|y) = T\bar{y}/(T-1)$. Write the integrand of the posterior mean in fully exponential form $\exp[-Th(\theta)]$ and find the Laplace approximation to the posterior mean.

Solution

The posterior mean can be written

$$\begin{aligned} G = E(\theta|y) &= \int_0^\infty \theta c^{-1} \theta^{-(T+1)} \exp(-T\bar{y}\theta^{-1}) d\theta \\ &= \int_0^\infty \exp[-T(T^{-1} \ln(c) + \ln \theta + \bar{y}\theta^{-1})] d\theta \\ &= \int_0^\infty [-Th(\theta)] d\theta, \end{aligned}$$

where $h(\theta) = T^{-1} \ln(c) + \ln(\theta) + \bar{y}\theta^{-1}$. Equating the derivative

$$\frac{\partial h(\theta)}{\partial \theta} = \frac{1}{\theta} - \frac{\bar{y}}{\theta^2}$$

to zero and solving yields the mode $\hat{\theta} = \bar{y}$. The Hessian is

$$H(\theta) = -\frac{1}{\theta^2} + \frac{2\bar{y}}{\theta^3}.$$

The Laplace approximation to $G = E(\theta|y)$ requires

$$\begin{aligned} \exp[-Th(\hat{\theta})] &= \exp[-\ln(c) - T \ln(\hat{\theta}) - T\bar{y}/\hat{\theta}] \\ &= [\Gamma(T)]^{-1} T^T \bar{y}^T \bar{y}^{-T} \exp(-T) \\ &= [\Gamma(T)]^{-1} T^T \exp(-T) \end{aligned}$$

and

$$TH(\hat{\theta}) = -\frac{T}{\hat{\theta}^2} + \frac{2T\bar{y}}{\hat{\theta}^3} = \frac{T}{\bar{y}^2}.$$

Therefore, using (9.3) with $b(\cdot) = 1$ yields the Laplace approximation

$$\begin{aligned} \hat{G}_L &= (2\pi)^{1/2} |TH(\hat{\theta})|^{-1/2} \exp[-Th(\hat{\theta})] \\ &= (2\pi)^{1/2} T^{-1/2} \bar{y} [\Gamma(T)]^{-1} T^T \exp(-T) \\ &= \bar{y} [\Gamma(T)]^{-1} \left\{ (2\pi T)^{1/2} T^T \exp(-T) \right\} T^{-1} \\ &= [\Gamma(T)]^{-1} \left\{ (2\pi T)^{1/2} T^T \exp(-T) \right\} T^{-2} (T-1) \frac{T\bar{y}}{T-1}. \end{aligned} \tag{9.11}$$

Again, the term in curly braces is Stirling's approximation to $T! = \Gamma(T+1)$. Hence, $\{\cdot\}T^{-1} = \Gamma(T)$. Therefore, \hat{G}_L is a sensible estimate of $E(\theta|y)$.

Exercise 9.7 (Tierney–Kadane approximation: exponential posterior mean: Achcar and Smith [1990]) Consider Exercise 9.6. Using the fully exponential method, find the Tierney–Kadane approximation to this posterior mean.

Solution

A random sample of size T from an exponential distribution with mean θ yields the log likelihood

$$\ln L(\theta) = -T \ln(\theta) - T\bar{y}\theta^{-1}.$$

The posterior mean is a ratio of integrals as in (9.6) with $g(\theta) = \theta$. To employ the fully exponential version of Tierney–Kadane, we need to express the integrands in the denominator and numerator of (9.6) in the forms $\exp[-Th_D(\theta)]$ and $\exp[-Th_N(\theta)]$, respectively. Thus we construct

$$\begin{aligned} h_D(\theta) &= -T^{-1}[\ln[p(\theta)] + \ln L(\theta)] \\ &= -T^{-1}[-(T+1)\ln(\theta) - T\bar{y}/\theta] \end{aligned} \quad (9.12)$$

and

$$h_N(\theta) = -T^{-1}[-T \ln(\theta) - T\bar{y}/\theta]. \quad (9.13)$$

Equating the derivatives

$$\frac{\partial h_D(\theta)}{\partial \theta} = -T^{-1} \left[-\frac{T+1}{\theta} + \frac{T\bar{y}}{\theta^2} \right] \quad (9.14)$$

and

$$\frac{\partial h_N(\theta)}{\partial \theta} = -T^{-1} \left[-\frac{T}{\theta} + \frac{T\bar{y}}{\theta^2} \right] \quad (9.15)$$

to zero and solving yields the modes $\hat{\theta}_D = T\bar{y}/(T+1)$ and $\hat{\theta}_N = \bar{y}$. The corresponding Hessians for the denominator and numerator are

$$H_D(\theta) = -T^{-1} \left[\frac{T+1}{\theta^2} - \frac{2T\bar{y}}{\theta^3} \right] \quad (9.16)$$

and

$$H_N(\theta) = -T^{-1} \left[\frac{T}{\theta^2} - \frac{2T\bar{y}}{\theta^3} \right], \quad (9.17)$$

respectively. The Tierney–Kadane approximation to $G = E(\theta|y)$ requires evaluating (9.12), (9.13), (9.16), and (9.17) at the posterior modes to obtain

$$\begin{aligned} \exp[-Th_D(\hat{\theta}_D)] &= \exp[-(T+1)\ln(\hat{\theta}_D) - T\bar{y}/\hat{\theta}_D] \\ &= \left(\frac{T\bar{y}}{T+1} \right)^{-(T+1)} \exp[-(T+1)], \end{aligned} \quad (9.18)$$

$$\begin{aligned} \exp[-Th_N(\hat{\theta}_N)] &= \exp[-T \ln(\hat{\theta}_N) - T\bar{y}/\hat{\theta}_N] \\ &= \bar{y}^{-T} \exp[-T], \end{aligned} \quad (9.19)$$

$$\begin{aligned}
 TH_D(\hat{\theta}_D) &= - \left[\frac{T+1}{\hat{\theta}_D^2} - \frac{2T\bar{y}}{\hat{\theta}_D^3} \right] \\
 &= \frac{(T+1)^3}{T^2\bar{y}^2},
 \end{aligned} \tag{9.20}$$

and

$$\begin{aligned}
 TH_N(\hat{\theta}_N) &= - \left[\frac{T}{\hat{\theta}_N^2} - \frac{2T\bar{y}}{\hat{\theta}_N^3} \right] \\
 &= \frac{T}{\bar{y}^2}.
 \end{aligned} \tag{9.21}$$

Substituting (9.18)–(9.21) into (9.8) yields the Tierney–Kadane approximation

$$\hat{G}_{\text{TK}} = \frac{T^{T-(1/2)} \exp(1)\bar{y}}{(T+1)^{T-(1/2)}} = \left[\frac{T^{T-(1/2)}(T-1) \exp(1)}{(T+1)^{T-(1/2)T}} \right] \frac{T\bar{y}}{T-1}, \tag{9.22}$$

where the term outside of the square brackets of (9.22) equals $E(\theta|y)$.

Exercise 9.8 (Tierney–Kadane approximation: exponential posterior mean, 2: Achcar and Smith [1990]) Consider again Exercise 9.7. Because $0 < \theta < \infty$, there is reason to believe your result can be improved by transforming to $\gamma \in \mathcal{R}$, where, say, $\gamma = \ln(\theta)$. Proceeding as in Exercise 9.7, but parameterizing in terms of γ , we can show that the Laplace approximation of $E(\theta|y) = T\bar{y}/(T-1)$ is

$$\hat{G}_{\text{TK}} = \frac{(T-1)^{T-(3/2)} \exp(1)\bar{y}}{(T)^{T-(3/2)}} = \left[\frac{(T-1)^{T-(1/2)} \exp(1)}{T^{T-(1/2)}} \right] \frac{T\bar{y}}{T-1}, \tag{9.23}$$

where, again, the term appearing outside the square brackets of (9.23) equals $E(\theta|y)$. Note that the percentage errors $|\hat{G} - G|/G$ of (9.11), (9.22), and (9.23) depend only on the sample size T and not the data. Compare these relative errors for $T = 2, 3, 4, \dots, 10, 15, 20, 25$. Which estimator do you prefer?

Solution

Straightforward computation yields the percentage errors in Table 9.2. Two comments are worth noting. First, the simple Laplace approximation (9.11) is decidedly the worst of the three. At first this might be surprising since it uses the exact posterior density including its integrating constant [i.e., the denominator in (9.7)] and uses a Laplace approximation of the numerator. But this is precisely why the two Tierney–Kadane estimates (9.22) and (9.23), which use a Laplace approximation in *both* the numerator and denominator, dominate. As noted in the text, the $O(T^{-1})$ errors of the numerator and denominator cancel, leaving a multiplicative error of order $O(T^{-2})$ compared to the $O(T^{-1})$ error of (9.11). Second, as seen in the solution to Exercise 9.5, and specifically in Table 9.1, by reparameterizing from $\theta \in \mathcal{R}^+$ to $\gamma \in \mathcal{R}$, the quality of the Tierney–Kadane approximation improves. Therefore, (9.23) is the clear winner.

Table 9.2: Laplace and two Tierney–Kadane approximations for exponential mean.

T	Absolute Percentage Errors		
	(9.11)	(9.22)	(9.23)
2	52.0	26.0	3.9
3	35.2	11.7	1.4
4	26.5	6.6	.7
5	21.3	4.3	.4
6	17.8	3.0	.3
7	15.3	2.2	.2
8	13.4	1.7	.1
9	11.9	1.3	.1
10	10.7	1.1	.0
15	7.2	.0	.0
20	5.4	.0	.0
25	4.3	.0	.0

Exercise 9.9 (Approximate Bayes factors) In Chapter 6 we considered Bayes factors that are ratios of marginal likelihoods. Approximation of these ratios differs importantly from the preceding discussion in that the integrals of the ratio are usually of different dimensions [i.e., $K_N \neq K_D$ in (9.7)]. In what follows we switch notation from K_N to K_D to K_1 and K_2 , and similarly for other subscripted quantities. Given hypotheses H_i , $i = 1, 2$, and prior densities $p(\theta_i|H_i) = p_i(\theta_i)$, $i = 1, 2$, consider the marginal likelihoods

$$p(y|H_i) = \int_{\mathcal{R}^{K_i}} p_i(\theta_i) L(\theta_i) d\theta_i, \quad i = 1, 2. \quad (9.24)$$

(a) Using the fully exponential method, find the Laplace approximations $\hat{p}(y|H_i)$, $i = 1, 2$, to the marginal densities in (9.24).

(b) Find $-2 \ln(\hat{B}_{12})$, where

$$\hat{B}_{12} = \frac{\hat{p}(y|H_1)}{\hat{p}(y|H_2)}$$

is an approximation of the Bayes factor

$$B_{12} = \frac{p(y|H_1)}{p(y|H_2)}.$$

Solution

(a) Using the fully exponential method, the Laplace approximations to (9.24) are

$$\hat{p}(y|H_i) = (2\pi)^{\frac{K_i}{2}} \left| TH_i \left(\hat{\theta}_i \right) \right|^{\frac{1}{2}} \exp[-Th_i(\hat{\theta}_i)], \quad (9.25)$$

where

$$h_i(\hat{\theta}_i) = -\frac{1}{T} \ln \left[p_i(\hat{\theta}_i) L(\hat{\theta}_i) \right],$$

$$H_i(\hat{\theta}_i) = \frac{\partial^2 h_i(\theta_i)}{\partial \theta_i \partial \theta_i'} \Big|_{\theta_i = \hat{\theta}_i},$$

and

$$\hat{\theta}_i = \operatorname{argmax}_{\theta_i} \{-h_i(\theta_i)\}$$

is the posterior mode under H_i .

(b) From (9.25) it follows that

$$\begin{aligned} -2 \ln [\hat{B}_{12}] &= -2 \ln \left[(2\pi)^{\frac{K_1}{2}} p_1(\hat{\theta}_1) \left| TH_1(\hat{\theta}_1) \right|^{-\frac{1}{2}} L(\hat{\theta}_1) \right] \\ &\quad - (-2) \ln \left[(2\pi)^{\frac{K_2}{2}} p_2(\hat{\theta}_2) \left| TH_2(\hat{\theta}_2) \right|^{-\frac{1}{2}} L(\hat{\theta}_2) \right] \\ &= -2[\ln L(\hat{\theta}_1) - \ln L(\hat{\theta}_2)] + (K_1 - K_2) \ln(T) + c_{12}, \end{aligned} \quad (9.26)$$

where

$$\begin{aligned} c_{12} &= -2 \ln \left[(2\pi)^{\frac{K_1}{2}} p_1(\hat{\theta}_1) \left| H_1(\hat{\theta}_1) \right|^{-\frac{1}{2}} \right] \\ &\quad + 2 \ln \left[(2\pi)^{\frac{K_2}{2}} p_2(\hat{\theta}_2) \left| H_2(\hat{\theta}_2) \right|^{-\frac{1}{2}} \right] + O(T^{-1}) \\ &= (K_2 - K_1) \ln(2\pi) + 2 \ln \left[\frac{|H_1(\hat{\theta}_1)|}{|H_2(\hat{\theta}_2)|} \right] + 2 \ln \left[\frac{p_2(\hat{\theta}_2)}{p_1(\hat{\theta}_1)} \right] + O(T^{-1}) \end{aligned} \quad (9.27)$$

is of $O(1)$. Posterior modes $\hat{\theta}_1$ and $\hat{\theta}_2$ can be replaced by their corresponding MLEs, $\hat{\theta}_1^{\text{ML}}$ and $\hat{\theta}_2^{\text{ML}}$, without any deterioration in the order of the approximation error.

Schwarz (1978) suggested a *consistent* model selection criterion with all quantities in (9.26) evaluated at $\hat{\theta}_1^{\text{ML}}$ and $\hat{\theta}_2^{\text{ML}}$ and $c_{12} = 0$. In general, however, setting $c_{12} = 0$ in (9.26) results in a relative error of size $O(1)$. Kass and Wasserman (1995) and Poirier (1996) note choices of prior in which the relative error is only $O(T^{-1/2})$. Since c_{12} involves differences in prior densities at the respective posterior modes, priors have more effect on posterior odds analysis than on point estimates. In other words, the prior matters to $O(1)$ for testing, but it matters less for estimation since posterior modes and MLEs differ by only $O(T^{-1})$.

Exercise 9.10 (Approximate marginal posteriors) Let $\theta = [\theta_1 \ \theta_2]'$. Consider the normal-gamma posterior distribution

$$\begin{aligned} p(\theta | \bar{\mu}, \bar{q}, \bar{s}^{-2}, \bar{\nu}) &= \phi(\theta_1 | \bar{\mu}, \theta_2^{-1} \bar{q}) \gamma(\theta_2 | \bar{s}^{-2}, \bar{\nu}) \\ &= \left((2\pi \theta_2^{-1} \bar{q})^{-1/2} \exp \left[-\frac{1}{2} \theta_2 \bar{q}^{-1} (\theta_1 - \bar{\mu})^2 \right] \right) \\ &\quad \times \left(\left[\left(\frac{2}{\bar{\nu} \bar{s}^2} \right)^{\bar{\nu}/2} \Gamma(\bar{\nu}/2) \right]^{-1} \theta_2^{(\bar{\nu}-2)/2} \exp[-(1/2) \theta_2 \bar{\nu} \bar{s}^2] \right) \\ &\propto \theta_2^{(\bar{\nu}-1)/2} \exp \left(-\frac{1}{2} \theta_2 [\bar{\nu} \bar{s}^2 + \bar{q}^{-1} (\theta_1 - \bar{\mu})^2] \right). \end{aligned}$$

Write the marginal posterior density of θ_1 as

$$\begin{aligned} p(\theta_1 | y) &= \int_{-\infty}^{\infty} p(\theta_1, \theta_2 | \bar{\mu}, \bar{q}, \bar{s}^2, \bar{\nu}) \, d\theta_2 \tag{9.28} \\ &= \frac{\int_{-\infty}^{\infty} p(\theta_1, \theta_2 | \bar{\mu}, \bar{q}, \bar{s}^2, \bar{\nu}) L(\theta_1, \theta_2) \, d\theta_2}{\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} p(\theta_1, \theta_2 | \bar{\mu}, \bar{q}, \bar{s}^2, \bar{\nu}) L(\theta_1, \theta_2) \, d\theta_1 \, d\theta_2} \\ &= \frac{\int_{-\infty}^{\infty} \phi(\theta_1 | \bar{\mu}, \bar{s}^2) f_G(\theta_2 | \bar{s}^{-2}, \bar{\nu}) \, d\theta_2}{\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \phi(\theta_1 | \bar{\mu}, \bar{s}^2) f_G(\theta_2 | \bar{s}^{-2}, \bar{\nu}) \, d\theta_1 \, d\theta_2}, \end{aligned}$$

where $f_G(\theta_2 | \cdot, \cdot)$ denotes the gamma density (see Appendix Definition 2). Write both integrands appearing in (9.28) in fully exponential form and apply Laplace approximations to both the numerator (N) and denominator (D) to obtain an approximate marginal posterior density for θ_1 . Tierney and Kadane (1986, p. 84) note that this approximation results in a multiplicative error of $O(T^{-3/2})$ in $T^{-1/2}$ neighborhoods of $\hat{\theta}_{D1}$.

Solution

Define

$$h_D(\theta_1, \theta_2) = \frac{1}{2T} \left(\theta_2 [\bar{q}^{-1} (\theta_1 - \bar{\mu})^2 + \bar{\nu} \bar{s}^2] - (\bar{\nu} - 1) \ln(\theta_2) \right),$$

and for given θ_1 define $h_N(\theta_2) = h_D(\theta_1, \theta_2)$. Then, canceling constants in (9.28), we can rewrite the kernels of the integrands so that (9.28) becomes

$$p(\theta_1 | y) = \frac{\int_{-\infty}^{\infty} \exp[-T h_N(\theta_2)] \, d\theta_2}{\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \exp[-T h_D(\theta_1, \theta_2)] \, d\theta_1 \, d\theta_2}. \tag{9.29}$$

The following derivatives are straightforward to obtain:

$$\begin{aligned}\frac{\partial h_D(\theta_1, \theta_2)}{\partial \theta_1} &= \frac{1}{T} \theta_2 \bar{q}^{-1} (\theta_1 - \bar{\mu}), \\ \frac{\partial h_D(\theta_1, \theta_2)}{\partial \theta_2} &= \frac{1}{2T} (\bar{\nu} \bar{s}^2 + \bar{q}^{-1} (\theta_1 - \bar{\mu})^2 - (\bar{\nu} - 1) \theta_2^{-1}), \\ \frac{\partial^2 h_D(\theta_1, \theta_2)}{\partial \theta_1^2} &= \frac{\theta_2}{T \bar{q}}, \\ \frac{\partial^2 h_D(\theta_1, \theta_2)}{\partial \theta_2^2} &= \frac{\bar{\nu} - 1}{2T \theta_2^2}, \\ \frac{\partial^2 h_D(\theta_1, \theta_2)}{\partial \theta_1 \partial \theta_2} &= \frac{\theta_1 - \bar{\mu}}{T \bar{q}}, \\ \frac{\partial h_N(\theta_2)}{\partial \theta_2} &= \frac{1}{2T} (\bar{\nu} \bar{s}^2 + \bar{q}^{-1} (\theta_1 - \bar{\mu})^2 - (\bar{\nu} - 1) \theta_2^{-1}), \\ \frac{\partial^2 h_N(\theta_2)}{\partial \theta_2^2} &= \frac{\bar{\nu} - 1}{2T \theta_2^2}.\end{aligned}$$

Equating first-order partial derivatives to zero and solving yields the denominator and numerator modes:

$$\begin{aligned}\hat{\theta}_{D1} &= \bar{\mu}, \\ \hat{\theta}_{D2} &= \frac{\bar{\nu} - 1}{\bar{\nu} \bar{s}^2}, \\ \hat{\theta}_{N2} &= \frac{\bar{\nu} - 1}{\bar{\nu} \bar{s}^2 + \bar{q}^{-1} (\theta_1 - \bar{\mu})^2 - (\bar{\nu} - 1) \theta_2^{-1}}.\end{aligned}$$

The Laplace approximations require

$$\exp[-Th_D(\hat{\theta}_{D1}, \hat{\theta}_{D2})] = \exp\left[-\frac{1}{2}(\bar{\nu} - 1)\right] (\bar{\nu} - 1)^{(\bar{\nu}-1)/2} [\bar{\nu} \bar{s}^2]^{-(\bar{\nu}-1)/2}, \quad (9.30)$$

$$\exp[-Th_N(\hat{\theta}_{N2})] = \exp\left[-\frac{1}{2}(\bar{\nu} - 1)\right] (\bar{\nu} - 1)^{(\bar{\nu}-1)/2} [\bar{\nu} \bar{s}^2 + \bar{q}^{-1} (\theta_1 - \bar{\mu})^2]^{-(\bar{\nu}-1)/2}, \quad (9.31)$$

$$|TH_D(\hat{\theta}_{D1}, \hat{\theta}_{D2})| = \frac{\bar{\nu} \bar{s}^2}{2 \bar{q} T^2}, \quad (9.32)$$

$$|TH_N(\hat{\theta}_{N2})| = \frac{[\bar{\nu} \bar{s}^2 + \bar{q}^{-1} (\theta_1 - \bar{\mu})^2]^2}{2T(\bar{\nu} - 1)}. \quad (9.33)$$

Substituting (9.29)–(9.32) into (9.8) and simplifying yields the approximation

$$p(\theta_1|y) \propto (\bar{\nu}\bar{s}^2\bar{q})^{-1/2} \left[1 + \bar{\nu}^{-1} \left(\frac{\theta_1 - \bar{\mu}}{\bar{s}^2\bar{q}} \right)^2 \right]^{-(\bar{\nu}+1)/2}. \quad (9.34)$$

Since we know in this conjugate problem that the marginal posterior of θ_1 is a t -distribution with location $\bar{\mu}$, scale $(\bar{s}^2\bar{q})^{1/2}$, and $\bar{\nu}$ degrees of freedom, it is seen from (9.34) that the method produces the exact kernel of the marginal posterior in this case! By varying θ_1 the marginal density can be traced out. However, it needs to be scaled to integrate to unity.

The linear regression model

The linear regression model is the workhorse of econometrics. In addition to being important in its own right, this model is an important component of other, more complicated models. The linear regression model posits a linear relationship between the dependent variable y_i and a $1 \times k$ vector of explanatory variables, x_i , where $i = 1, \dots, N$ indexes the relevant observational unit (e.g., individual, firm, time period, etc.). In matrix notation, the linear regression model can be written as

$$y = X\beta + \epsilon, \quad (10.1)$$

where $y = [y_1 \ y_2 \ \dots \ y_N]'$ is an N vector,

$$X = \begin{bmatrix} x_1 \\ \vdots \\ x_N \end{bmatrix}$$

is an $N \times k$ matrix, and $\epsilon = [\epsilon_1 \ \epsilon_2 \ \dots \ \epsilon_N]'$ is an N vector of errors. Assumptions about ϵ and X define the likelihood function. Unless explicitly noted otherwise, the questions in this chapter assume that ϵ and X satisfy what are often referred to as the classical assumptions. With regards to the explanatory variables, we assume that either they are not random or, if they are random variables, it is acceptable to proceed conditionally on them. That is, the distribution of interest is $y|X$ [see Poirier (1995), pp. 448–458 for a more detailed discussion of the random explanatory variable case]. For simplicity of notation, we will not include X as a conditioning argument in most equations and leave this source of conditioning implicit. Finally, in this chapter and those that follow we typically reserve capital letters such as X to denote matrices and lowercase letters such as y and x_i to denote scalar or vector quantities. When the distinction between scalars and vectors is important, this will be made clear within the exercise. We also drop the convention of using capital letters

to denote random variables, since our focus throughout the remainder of the book is almost exclusively *ex post*. This notation is primarily needed when considering *ex ante* properties and their contrast with those obtained from the *ex post* viewpoint (e.g., in Chapter 4). These issues are not taken up in the following chapters and thus this notational distinction is decidedly less critical.

With regard to the errors in the regression model, we assume they are independently normally distributed with mean zero and common variance σ^2 . That is,

$$\epsilon \sim N(0_N, \sigma^2 I_N), \quad (10.2)$$

where 0_N is an N vector of zeroes, I_N is the $N \times N$ identity matrix, and $N(\cdot, \cdot)$ denotes the multivariate normal distribution (see Appendix Definition 3). In this chapter we choose to work with the error precision, $h \equiv \sigma^{-2}$, and, thus, the normal linear regression model depends on the parameter vector $[\beta' h']'$. By using the properties of the multivariate normal distribution, it follows that $p(y|\beta, h) = \phi(y; X\beta, h^{-1}I_N)$, and thus the likelihood function is given by

$$L(\beta, h) = \frac{h^{\frac{N}{2}}}{(2\pi)^{\frac{N}{2}}} \exp \left[-\frac{h}{2} (y - X\beta)' (y - X\beta) \right]. \quad (10.3)$$

In an analogous fashion to expressions given in Exercises 2.3 and 2.14, it is often convenient to write the likelihood function in terms of ordinary least squares (OLS) quantities

$$\hat{\beta} = (X'X)^{-1} X'y$$

and

$$SSE = (y - X\hat{\beta})' (y - X\hat{\beta}).$$

This can be done by using the fact that

$$(y - X\beta)' (y - X\beta) = SSE + (\beta - \hat{\beta})' X'X (\beta - \hat{\beta})$$

and thus

$$L(\beta, h) = \frac{h^{\frac{N}{2}}}{(2\pi)^{\frac{N}{2}}} \exp \left[-\frac{h}{2} \left\{ SSE + (\beta - \hat{\beta})' X'X (\beta - \hat{\beta}) \right\} \right]. \quad (10.4)$$

Exercise 10.1 (Posterior distributions under a normal-gamma prior) For the normal linear regression model under the classical assumptions, use a normal-Gamma prior [i.e., the prior for β and h is $NG(\underline{\beta}, Q, \underline{s}^{-2}, \nu)$; see Appendix Definition 5]. Derive the posterior for β and h and thus show that the normal-gamma prior is a conjugate prior for this model.

Solution

Using Bayes' theorem and the properties of the normal-gamma density, we have

$$p(\beta, h|y) \propto p(\beta, h) L(\beta, h) \\ = \phi(\beta; \underline{\beta}, h^{-1}\underline{Q}) f_{\gamma}(h|\underline{s}^{-2}, \underline{\nu}) \phi(y; X\beta, h^{-1}I_N),$$

where $f_{\gamma}()$ denotes the gamma density (see Appendix Definition 2) and $\phi()$ the multivariate normal p.d.f. This question can be solved by plugging in the forms for each of the densities in these expressions, rearranging them (using some theorems in matrix algebra), and recognizing that the result is the kernel of normal-gamma density. (Details are provided in the following.) Since the posterior and prior are both normal-gamma, conjugacy is established. The steps are elaborated in the remainder of this solution.

Begin by writing out each density (ignoring integrating constants not involving the parameters) and using the expression for the likelihood function in (10.4):

$$p(\beta, h|y) \propto \left\{ h^{\frac{k}{2}} \exp \left[-\frac{h}{2} (\beta - \underline{\beta})' \underline{Q}^{-1} (\beta - \underline{\beta}) \right] \right\} \\ \left\{ h^{\frac{\underline{\nu}-2}{2}} \exp \left[-\frac{h\underline{\nu}s^2}{2} \right] \right\} \\ \times \left\{ h^{\frac{N}{2}} \exp \left[-\frac{h}{2} \left\{ SSE + (\beta - \hat{\beta})' X'X (\beta - \hat{\beta}) \right\} \right] \right\} \\ = h^{\frac{\bar{\nu}+k-2}{2}} \exp \left[-\frac{h}{2} \left(\underline{\nu}s^2 + SSE + (\beta - \underline{\beta})' \underline{Q}^{-1} (\beta - \underline{\beta}) \right. \right. \\ \left. \left. + (\beta - \hat{\beta})' X'X (\beta - \hat{\beta}) \right) \right],$$

where $\bar{\nu} = \underline{\nu} + N$. In this expression, β enters only in the terms in the exponent,

$$(\beta - \underline{\beta})' \underline{Q}^{-1} (\beta - \underline{\beta}) + (\beta - \hat{\beta})' X'X (\beta - \hat{\beta}).$$

Rewrite this term as

$$(\beta - \underline{\beta})' \underline{Q}^{-1} (\beta - \underline{\beta}) + (\beta - \hat{\beta})' X'X (\beta - \hat{\beta}) = (\hat{\beta} - \underline{\beta})' X'X \bar{Q} \underline{Q}^{-1} (\hat{\beta} - \underline{\beta}) \\ + (\beta - \bar{\beta})' \bar{Q}^{-1} (\beta - \bar{\beta}),$$

where

$$\bar{Q} = (\underline{Q}^{-1} + X'X)^{-1}$$

and

$$\bar{\beta} = \bar{Q} (\underline{Q}^{-1} \underline{\beta} + X'X \hat{\beta}).$$

Thus, β enters only through the term $(\beta - \bar{\beta})' \bar{Q}^{-1} (\beta - \bar{\beta})$ and we can establish that the kernel of $\beta|y, h$ is given by

$$p(\beta|y, h) \propto \exp \left[-\frac{h}{2} (\beta - \bar{\beta})' \bar{Q}^{-1} (\beta - \bar{\beta}) \right].$$

Since this is the kernel of a normal density we have established that $\beta|y, h \sim N(\bar{\beta}, h^{-1} \bar{Q}^{-1})$.

We can derive $p(h|y)$ by using the fact that

$$p(h|y) = \int p(\beta, h|y) d\beta = \int p(h|y) p(\beta|y, h) d\beta.$$

Since p.d.f.s integrate to one we can integrate out the component involving normal density for $p(\beta|y, h)$ and we are left with

$$\begin{aligned} p(h|y) &\propto h^{\frac{\bar{\nu}-2}{2}} \exp \left[-\frac{h}{2} \left\{ \underline{\nu} s^2 + SSE + (\hat{\beta} - \underline{\beta})' X' X \bar{Q} \underline{Q}^{-1} (\hat{\beta} - \underline{\beta}) \right\} \right] \\ &\equiv h^{\frac{\bar{\nu}-2}{2}} \exp \left[-\frac{h}{2} \bar{\nu} \bar{s}^2 \right] \end{aligned}$$

with

$$\bar{\nu} \bar{s}^2 = \underline{\nu} s^2 + SSE + (\hat{\beta} - \underline{\beta})' X' X \bar{Q} \underline{Q}^{-1} (\hat{\beta} - \underline{\beta}).$$

We recognize the final expression for $p(h|y)$ as the kernel of a gamma density, and thus we have established that $h|y \sim \gamma(\bar{s}^{-2}, \bar{\nu})$.

Since $\beta|y, h$ is normal and $h|y$ is gamma, it follows immediately that the posterior for β and h is $NG(\bar{\beta}, \bar{Q}, \bar{s}^{-2}, \bar{\nu})$. Since prior and posterior are both normal-gamma, conjugacy has been established.

Exercise 10.2 (A partially informative prior) Suppose you have a normal linear regression model with *partially informative* natural conjugate prior where prior information is available only on $J \leq k$ linear combinations of the regression coefficients and the prior for h is the standard noninformative one: $p(h) \propto 1/h$. Thus, $R\beta|h \sim N(r, h^{-1} \underline{V}_r)$, where R is a known $J \times k$ matrix with $\text{rank}(R) = J$, r is a known J vector, and \underline{V}_r is a $J \times J$ positive definite matrix. Show that the posterior is given by

$$\beta, h|y \sim NG(\tilde{\beta}, \tilde{V}, \tilde{s}^{-2}, \tilde{\nu}),$$

where

$$\tilde{V} = (R' \underline{V}_r^{-1} R + X' X)^{-1},$$

$$\tilde{\beta} = \tilde{V} (R' \underline{V}_r^{-1} \underline{\beta} + X' X \hat{\beta}),$$

$$\tilde{\nu} = N,$$

and

$$\overline{\nu s^2} = \underline{\nu s^2} + \left(\tilde{\beta} - \hat{\beta} \right)' X' X \left(\tilde{\beta} - \hat{\beta} \right) + \left(R\hat{\beta} - r \right)' \underline{V}_r^{-1} \left(R\hat{\beta} - r \right).$$

Solution

Partition $X = [X_1 \ X_2]$ and $R = [R_1 \ R_2]$, where X_1 is $N \times (k - J)$, X_2 is $N \times J$, R_1 is $J \times (k - J)$, and R_2 is $J \times J$. Partition $\beta = [\beta_1' \ \beta_2']'$ conformably. The linear regression model in (10.1) can be written as

$$y = Z\gamma + \epsilon,$$

where $Z = XA^{-1}$ and $\gamma = A\beta$ for any $k \times k$ nonsingular matrix A . If we take

$$A = \begin{bmatrix} I_{k-J} & 0_{(k-J) \times J} \\ R_1 & R_2 \end{bmatrix}$$

then

$$\begin{aligned} Z &= XA^{-1} = [Z_1, Z_2] \\ &= [X_1 - X_2 R_2^{-1} R_1, X_2 R_2^{-1}] \end{aligned}$$

and $\gamma = [\gamma_1' \ \gamma_2']' = [\beta_1' \ (R\beta)']'$. In words, we have transformed the model so that an informative prior is employed for $\gamma_2 = R\beta$ whereas a noninformative prior is employed for γ_1 . The remainder of the proof is essentially the same as for Exercise 10.1. Assuming a normal-gamma natural conjugate prior for γ and h leads to a normal-gamma posterior. Taking noninformative limiting cases for the priors for γ_1 and h and then transforming back to the original parameterization (e.g., using $\beta = A^{-1}\gamma$ and $X = ZA$) yields the expressions given in the question.

Exercise 10.3 (Ellipsoid bound theorem) Consider the normal linear regression model with natural conjugate prior, $\beta, h \sim NG(\underline{\beta}, \underline{Q}, \underline{s}^{-2}, \underline{\nu})$. Show that if $\underline{\beta} = 0$ then, for any choice of \underline{Q} , the posterior mean $\overline{\beta}$ must lie in the ellipsoid:

$$\left(\overline{\beta} - \frac{\hat{\beta}}{2} \right)' X' X \left(\overline{\beta} - \frac{\hat{\beta}}{2} \right) \leq \frac{\hat{\beta}' X' X \hat{\beta}}{4}.$$

Solution

When $\underline{\beta} = 0$, the formula for the posterior mean $\overline{\beta}$ is given in Exercise 10.1 as

$$\overline{\beta} = \overline{Q} X' X \hat{\beta}, \tag{10.5}$$

where

$$\overline{Q} = (\underline{Q}^{-1} + X' X)^{-1}. \tag{10.6}$$

Multiplying (10.5) by \bar{Q}^{-1} (using the expression in Equation 10.6) we obtain

$$(\underline{Q}^{-1} + X'X) \bar{\beta} = X'X\hat{\beta},$$

rearranging and multiplying both sides of the equation by $\bar{\beta}'$ yields

$$\bar{\beta}' \underline{Q}^{-1} \bar{\beta} = \bar{\beta}' X'X (\hat{\beta} - \bar{\beta}),$$

and since \underline{Q}^{-1} is positive semidefinite, it follows that $\bar{\beta}' \underline{Q}^{-1} \bar{\beta} \geq 0$ and hence

$$\bar{\beta}' X'X (\hat{\beta} - \bar{\beta}) \geq 0. \quad (10.7)$$

But the ellipsoid bound equation given in the question can be rearranged to be identical to (10.7). In particular, expanding the quadratic form on the left-hand side of the ellipsoid bound equation yields

$$\bar{\beta}' X'X \bar{\beta} - \frac{\hat{\beta}'}{2} X'X \bar{\beta} - \bar{\beta}' X'X \frac{\hat{\beta}}{2} + \frac{\hat{\beta}'}{2} X'X \frac{\hat{\beta}}{2} \leq \frac{\hat{\beta}' X'X \hat{\beta}}{4},$$

which implies

$$\bar{\beta}' X'X \bar{\beta} - \bar{\beta}' X'X \hat{\beta} \leq 0,$$

or

$$\bar{\beta}' X'X (\hat{\beta} - \bar{\beta}) \geq 0.$$

Thus, (10.7) is equivalent to the ellipsoid bound theorem.

You may wish to read Leamer (1982) or Poirier (1995, pp. 526–537) for more details about ellipsoid bound theorems.

***Exercise 10.4 (Problems with Bayes factors using noninformative priors: Leamer [1978, p. 111])** Suppose you have two normal linear regression models:

$$M_j : y = X_j \beta_j + \epsilon_j,$$

where $j = 1, 2$, X_j is an $N \times k_j$ matrix of explanatory variables, β_j is a k_j vector of regression coefficients, and ϵ_j is an N vector of errors distributed as $N(0_N, h_j^{-1} I_N)$. If natural conjugate priors are used for both models [i.e., $\beta_j, h_j | M_j \sim NG(\underline{\beta}_j, \underline{Q}_j, \underline{s}_j^{-2}, \underline{\nu}_j)$], then the posterior is $\beta_j, h_j | y, M_j \sim NG(\bar{\beta}_j, \bar{Q}_j, \bar{s}_j^{-2}, \bar{\nu}_j)$ (where $\bar{\beta}_j, \bar{Q}_j, \bar{s}_j^{-2}$, and $\bar{\nu}_j$ are as given in the solution to Exercise 10.1) and the Bayes factor comparing M_2 to M_1 is given by

$$BF_{21} = \frac{c_2 \left(\frac{|\bar{Q}_2|}{|\underline{Q}_2|} \right)^{\frac{1}{2}} (\bar{\nu}_2 \bar{s}_2^2)^{-\frac{\bar{\nu}_2}{2}}}{c_1 \left(\frac{|\bar{Q}_1|}{|\underline{Q}_1|} \right)^{\frac{1}{2}} (\bar{\nu}_1 \bar{s}_1^2)^{-\frac{\bar{\nu}_1}{2}}},$$

where

$$c_j = \frac{\Gamma\left(\frac{\bar{\nu}_j}{2}\right) \left(\bar{\nu}_j \bar{s}_j^2\right)^{\frac{\bar{\nu}_j}{2}}}{\Gamma\left(\frac{\nu_j}{2}\right) \pi^{\frac{N}{2}}}.$$

(a) Consider a noninformative prior created by letting $\underline{\nu}_j \rightarrow 0$, $\underline{Q}_j^{-1} = cI_{k_j}$, and letting $c \rightarrow 0$ for $j = 1, 2$. Show that the Bayes factor reduces to

$$\begin{cases} 0 & \text{if } k_2 > k_1, \\ \left[\frac{|X_2'X_2|}{|X_1'X_1|} \right]^{-\frac{1}{2}} \left(\frac{SSE_2}{SSE_1} \right)^{-\frac{N}{2}} & \text{if } k_2 = k_1, \\ \infty & \text{if } k_2 < k_1. \end{cases}$$

The case $k_1 = k_2$ corresponds to two nonnested models with the same number of explanatory variables.

(b) Consider a noninformative prior created by setting $\underline{\nu}_j \rightarrow 0$, $\underline{Q}_j^{-1} = (c^{1/k_j}) I_{k_j}$ and letting $c \rightarrow 0$ for $j = 1, 2$. Show that the Bayes factor reduces to

$$\left[\frac{|X_2'X_2|}{|X_1'X_1|} \right]^{-\frac{1}{2}} \left(\frac{SSE_2}{SSE_1} \right)^{-\frac{N}{2}}.$$

(c) Consider a noninformative prior created by setting $\underline{\nu}_j \rightarrow 0$, $\underline{Q}_j^{-1} = (c^{1/k_j}) X_j'X_j$ and letting $c \rightarrow 0$ for $j = 1, 2$. Show that the Bayes factor reduces to

$$\left(\frac{SSE_2}{SSE_1} \right)^{-\frac{N}{2}}.$$

Solution

In all cases, if $\underline{\nu}_1 = \underline{\nu}_2 \rightarrow 0$ at the same rate then $c_1 = c_2$ and these integrating constants cancel out in the Bayes factor. Furthermore, under the various assumptions about \underline{Q}_j^{-1} it can be seen that

$$\bar{\nu}_j \bar{s}_j^2 = SSE_j,$$

where $\bar{\nu}_j = N$ and $\bar{Q}_j = (X_j'X_j)^{-1}$ for $j = 1, 2$. Thus, in all cases, the Bayes factor reduces to

$$BF_{21} = \left(\frac{|\underline{Q}_2^{-1}|}{|\underline{Q}_1^{-1}|} \right)^{\frac{1}{2}} \left[\frac{|X_2'X_2|}{|X_1'X_1|} \right]^{-\frac{1}{2}} \left(\frac{SSE_2}{SSE_1} \right)^{-\frac{N}{2}}.$$

In part (a) $|\underline{Q}_j^{-1}| = c^{k_j}$ and hence $(|\underline{Q}_2^{-1}|/|\underline{Q}_1^{-1}|) = c^{k_2-k_1}$. If $k_1 = k_2$ then $(|\underline{Q}_2^{-1}|/|\underline{Q}_1^{-1}|) = 1$ for all c . If $k_1 > k_2$ then $(|\underline{Q}_2^{-1}|/|\underline{Q}_1^{-1}|) \rightarrow \infty$ as $c \rightarrow 0$. If $k_1 < k_2$ then $(|\underline{Q}_2^{-1}|/|\underline{Q}_1^{-1}|) \rightarrow 0$ as $c \rightarrow 0$. Thus, the result in part (a) is established.

In part (b) $|\underline{Q}_j^{-1}| = c$ and hence $\left(|\underline{Q}_2^{-1}|/|\underline{Q}_1^{-1}|\right) = 1$ for all c regardless of what k_j is. Thus, the result in part (b) is established.

In part (c) $|\underline{Q}_j^{-1}| = c |X_j' X_j|$ and hence $\left(|\underline{Q}_2^{-1}|/|\underline{Q}_1^{-1}|\right) = |X_2' X_2|/|X_1' X_1|$ for all c regardless of what k_j is. Using this result, simplifies the Bayes factor to the expression given in part (c).

Exercise 10.5 (Multicollinearity) Consider the normal linear regression model with natural conjugate prior: $NG(\underline{\beta}, \underline{Q}, \underline{s}^{-2}, \underline{\nu})$. Assume in addition that $Xc = 0$ for some non-zero vector of constants c . Note that this is referred to as a case of *perfect multicollinearity*. It implies that the matrix X is not of full rank and $(X'X)^{-1}$ does not exist.

(a) Show that, despite this pathology, the posterior exists if \underline{Q} is positive definite.

(b) Define

$$\alpha = c' \underline{Q}^{-1} \beta.$$

Show that, given h , the prior and posterior distributions of α are identical and equal to

$$N(c' \underline{Q}^{-1} \underline{\beta}, h^{-1} c' \underline{Q}^{-1} c).$$

Hence, although prior information can be used to surmount the problems caused by perfect multicollinearity, there are some combinations of the regression coefficients about which (conditional) learning does not occur.

Solution

(a) The solution to this question is essentially the same as to Exercise 10.1. The key thing to note is that, since \underline{Q} is positive definite, \underline{Q}^{-1} exists and, hence, $\overline{\underline{Q}}^{-1}$ exists despite the fact that $X'X$ is rank deficient. In the same manner, it can be shown that the posterior for $\underline{\beta}$ and h is $NG(\overline{\underline{\beta}}, \overline{\underline{Q}}, \overline{\underline{s}}^{-2}, \overline{\underline{\nu}})$ with

$$\overline{\underline{Q}} = (\underline{Q}^{-1} + X'X)^{-1},$$

$$\overline{\underline{\beta}} = \overline{\underline{Q}} (\underline{Q}^{-1} \underline{\beta} + X'y),$$

$$\overline{\underline{s}}^2 = \underline{\nu} \underline{s}^2 + (y - X\overline{\underline{\beta}})'(y - X\overline{\underline{\beta}}) + (\overline{\underline{\beta}} - \underline{\beta})' \underline{Q}^{-1} (\overline{\underline{\beta}} - \underline{\beta}),$$

and $\overline{\underline{\nu}} = N + \underline{\nu}$.

(b) The properties of the normal-gamma distribution imply

$$\beta|h \sim N(\underline{\beta}, h^{-1} \underline{Q}).$$

The properties of the normal distribution imply

$$\alpha|h \equiv c' \underline{Q}^{-1} \beta|h \sim N(c' \underline{Q}^{-1} \underline{\beta}, h^{-1} c' \underline{Q}^{-1} c),$$

which establishes that the prior has the required form.

Using the result from part (a) gives the relevant posterior of the form

$$\beta|y, h \sim N(\bar{\beta}, h^{-1}\bar{Q}),$$

which implies

$$\alpha|y, h \sim N(c'\underline{Q}^{-1}\bar{\beta}, h^{-1}c'\underline{Q}^{-1}\bar{Q}\underline{Q}^{-1}c).$$

The mean can be written as

$$\begin{aligned} c'\underline{Q}^{-1}\bar{\beta} &= c'(\bar{Q}^{-1} - X'X)\bar{Q}(\underline{Q}^{-1}\underline{\beta} + X'y) \\ &= c'\underline{Q}^{-1}\underline{\beta} + c'X'y - c'X'X\bar{Q}(\underline{Q}^{-1}\underline{\beta} + X'y) \\ &= c'\underline{Q}^{-1}\underline{\beta}, \end{aligned}$$

since $Xc = 0$.

The variance can be written as

$$\begin{aligned} h^{-1}c'\underline{Q}^{-1}\bar{Q}\underline{Q}^{-1}c &= h^{-1}c'\underline{Q}^{-1}\left[(\underline{Q}^{-1} + X'X)^{-1}\underline{Q}^{-1}\right]c \\ &= h^{-1}c'\underline{Q}^{-1}\left[I - (\underline{Q}^{-1} + X'X)^{-1}X'X\right]c \\ &= h^{-1}c'\underline{Q}^{-1}c, \end{aligned}$$

since $Xc = 0$. Note that this derivation uses a standard theorem in matrix algebra that says, if G and H are $k \times k$ matrices and $(G + H)^{-1}$ exists, then

$$(G + H)^{-1}H = I_k - (G + H)^{-1}G.$$

Combining these derivations, we see that the posterior is also

$$\alpha|y, h \sim N(c'\underline{Q}^{-1}\underline{\beta}, h^{-1}c'\underline{Q}^{-1}c),$$

as required,

Basics of Bayesian computation

Bayesian econometrics typically involves extensive use of *posterior simulation*. Posterior simulation is discussed in detail in, for example, Gelman, Carlin, Stern, and Rubin (1995), Geweke (1999), Carlin and Louis (2000), Chib (2001), Geweke and Keane (2001), Koop (2003), Lancaster (2004), and Geweke (2005). The reader unfamiliar with the basic ideas and terminology of posterior computation is referred to these references for background reading. Some of the most important methods of posterior simulation are Monte Carlo integration, importance sampling, Gibbs sampling and the Metropolis–Hastings algorithm. The latter two of these are examples of *Markov chain Monte Carlo* (MCMC) algorithms. In this chapter we use simple distributions and commonly employed models to solve questions that illustrate how these and other algorithms are implemented.

Before we start the questions, we provide a brief description of notation and the basic ideas underlying each algorithm. We let $p(\theta|y)$ denote the posterior distribution for the parameters of a model θ . Suppose we can take iid draws from $p(\theta|y)$, which we denote by $\theta^{(r)}$ for $r = 1, \dots, R$. Let $g(\theta)$ be a function of the model's parameters. The weak law of large numbers implies

$$\hat{g} = \frac{\sum_{r=1}^R g(\theta^{(r)})}{R}$$

converges weakly in probability to $E[g(\theta)|y]$ as R goes to infinity (provided $E[g(\theta)|y]$ exists). Thus, estimates of any function of the model's parameters can be calculated by randomly drawing from the posterior and averaging functions of the draws. This strategy is referred to as *Monte Carlo integration*.

It is rarely possible to do Monte Carlo integration since directly drawing from the posterior cannot easily be done except in a few special cases. An alternative algorithm, *importance sampling*, involves drawing from a more convenient density $q(\theta)$, called the

importance function. If the importance function has support that includes the support of the posterior, then

$$\hat{g} = \frac{\sum_{r=1}^R w(\theta^{(r)}) g(\theta^{(r)})}{\sum_{r=1}^R w(\theta^{(r)})},$$

where

$$w(\theta^{(r)}) = \frac{p(\theta^{(r)}|y)}{q(\theta^{(r)})}$$

converges to $E[g(\theta)|y]$ as R goes to infinity (provided $E[g(\theta)|y]$ exists). Thus taking weighted averages of iid draws from the importance function allows for the posterior expectation of any function of the model parameters to be estimated.

Gibbs sampling is a popular algorithm commonly used when the conditional posterior distributions are easy to work with. We illustrate Gibbs sampling for the case where the parameter vector is broken into two blocks: $\theta' = [\theta'_1 \theta'_2]$, although the algorithm generalizes easily to the case of many blocks (including the case where some of the blocks contain latent data). Gibbs sampling produces a sequence of draws $\theta^{(r)}$ for $r = 1, \dots, R$ with the property that

$$\hat{g} = \frac{\sum_{r=1}^R g(\theta^{(r)})}{R}$$

converges to $E[g(\theta)|y]$ as R goes to infinity (provided $E[g(\theta)|y]$ exists). Gibbs sampling produces this sequence by iteratively drawing from the posterior conditional distributions. That is, $\theta_2^{(r)}$ is drawn from $p(\theta_2|y, \theta_1^{(r-1)})$ and $\theta_1^{(r)}$ is drawn from $p(\theta_1|y, \theta_2^{(r)})$. It can be shown that, under weak conditions, this sequence of draws converges to a sequence of draws from the joint posterior. Note that the Gibbs sampler requires a choice of starting value $\theta_1^{(0)}$. To eliminate dependence on this starting value, it is common to discard the initial R_0 draws (referred to as burn-in draws) from the Gibbs sampler.

The *Metropolis–Hastings algorithm* is really an entire class of algorithms. Like Monte Carlo integration and Gibbs sampling it produces a sequence of draws $\theta^{(r)}$ for $r = 1, \dots, R$ with the property that

$$\hat{g} = \frac{\sum_{r=1}^R g(\theta^{(r)})}{R}$$

converges to $E[g(\theta)|y]$ as R goes to infinity. It involves drawing from a convenient density akin to the importance function, referred to as the *candidate generating density*. Candidate draws of $\theta^{(r)}$ are either accepted or rejected with a certain probability referred to as an *acceptance probability*. If they are rejected, then $\theta^{(r)}$ is set to $\theta^{(r-1)}$. The acceptance probability depends on the precise form of the algorithm [see Chib and Greenberg (1995), Geweke (1999), or Koop (2003)]. Two common variants are the *independence chain* and the *random walk chain* Metropolis–Hastings algorithms. These algorithms can be combined so that, for instance, it is common to use a Metropolis-within-Gibbs algorithm.

Many of the following questions relate to the normal linear regression model with a natural conjugate prior. Chapter 10 defines notation for this model and provides the relevant posterior formulas. Other exercises provided in this chapter describe procedures for determining whether mistakes have been made when programming a posterior simulator and for checking the credibility of the model's assumptions. Finally, other valuable yet less commonly used algorithms such as acceptance/rejection sampling are described and applied in the last set of exercises in this chapter.

In some cases, solutions to computer problems will be provided by offering sketches of programs. Complete programs in Matlab are also included on the Web site associated with this book. Other relevant programming languages (e.g., GAUSS) have a similar structure to Matlab, so Matlab code should be useful even to readers who do not use this program. Note that posterior simulation algorithms (and artificial data sets) depend on random number generators and, thus, differ from run to run. Thus, your results will not be exactly the same as those presented throughout this book. For short runs (e.g., in Exercise 11.3 with $R = 10$) your results could be very different (although the general pattern of results should not differ substantially from ours).

11.1 Monte Carlo integration

Exercise 11.1 (Numerical standard error of Monte Carlo estimate) Monte Carlo integration can be used to estimate $E[\theta|y]$ as $\hat{\theta} = \sum_{r=1}^R \theta^{(r)} / R$. For finite R this procedure provides us with an estimate of the posterior mean of θ . Use a central limit theorem to derive the standard error (known as the *numerical standard error*, or NSE) associated with the estimate that can be used to evaluate the accuracy of the estimate for any given R .

Solution

Using the central limit theorem for iid random variables we have

$$\sqrt{R}(\hat{\theta} - \theta) \sim N(0, \sigma^2)$$

as $R \rightarrow \infty$, where $\sigma^2 = \text{Var}(\theta|y)$. The latter quantity can itself be estimated using Monte Carlo integration [e.g., $\widehat{\text{Var}}(\theta|y) \approx (1/R) \sum_r (\theta^{(r)} - \hat{\theta})^2$], and we shall denote this estimate by $\hat{\sigma}^2$. We can use this asymptotic result to obtain an approximate one for finite R :

$$\hat{\theta} \stackrel{\text{approx}}{\sim} N\left(\theta, \frac{\hat{\sigma}^2}{R}\right).$$

The numerical standard error is

$$\text{NSE} = \frac{\hat{\sigma}}{\sqrt{R}}.$$

Exercise 11.2 (Drawing from standard distributions) Simulation-based inference via the Metropolis–Hastings algorithm or Gibbs sampler requires the researcher to be able to draw from standard distributions. In this exercise we discuss how Matlab can be used to obtain draws from a variety of standard continuous distributions. Specifically, we obtain draws from the uniform, normal, Student t , beta, exponential, and chi-square distributions using Matlab (see the Appendix for definitions of these distributions). This exercise is designed to be illustrative – Matlab is capable of generating variates from virtually any distribution that an applied researcher will encounter (and the same applies to other relevant computer languages such as GAUSS).

Using Matlab, obtain sets of 10, 100, and 100,000 draws from the uniform, standard normal, Student $t(3)$ [denoted $t(0, 1, 3)$ in the notation of the Appendix], $\text{beta}(3,2)$ [denoted $\beta(3,2)$ in the Appendix], exponential with mean 5 [denoted $\text{exponential}(5)$ in the following], and $\chi^2(3)$ distributions. For each sample size calculate the mean and standard deviation and compare these quantities to the known means and standard deviations from each distribution.

Solution

Our Matlab program for implementing this exercise is provided on the Web site associated with this book. Table 11.1 gives a summary of these results.

Table 11.1: Simulation results.

	Uniform		Standard Normal	
	Mean	Std.	Mean	Std.
True Value	.500	$(1/\sqrt{12})$	0	1
$N = 10$.482	.314	.204	.870
$N = 100$.489	.303	.059	1.06
$N = 100,000$.500	.289	−.001	.999
	Student $t(3)$		Beta(3,2)	
	Mean	Std.	Mean	Std.
True Value	0	$\sqrt{3}$.6	.2
$N = 10$.002	.920	.616	.181
$N = 100$	−.167	1.46	.585	.193
$N = 100,000$.003	1.70	.600	.199
	Exponential(5)		$\chi^2(3)$	
	Mean	Std.	Mean	Std.
True Value	5	5	3	$\sqrt{6}$
$N = 10$	3.42	3.35	2.98	2.55
$N = 100$	4.61	4.75	2.85	2.43
$N = 100,000$	5.01	5.02	3.00	2.45

For the largest sample size, $N = 100,000$, our simulated means and standard deviations are virtually identical to the known means and standard deviations. It is also worth noting that the exponential(5) and $\chi^2(3)$ distributions are equivalent to the $G(1, 5)$ and $G(3/2, 2)$ distributions, respectively (see the Appendix). The Matlab routine for drawing from the flexible gamma distribution is “gamrnd,” thus providing an alternate way for obtaining draws from the exponential and chi-square distributions.

Exercise 11.3 (Monte Carlo integration in a regression problem) (a) Generate an artificial data set of size $N = 100$ from the normal linear regression model, as described in Exercise 10.1, with an intercept and one other explanatory variable. Set the intercept (β_1) to 0, the slope coefficient (β_2) to 1.0, and h to 1.0. Generate the explanatory variable by taking random draws from the $U(0, 1)$ distribution.

(b) Calculate the posterior mean and standard deviation for the slope coefficient β_2 for this data set using a normal-gamma prior with $\underline{\beta} = [0 \ 1]'$, $\underline{V} = I_2$, $\underline{s}^{-2} = 1$, and $\underline{\nu} = 1$.

(c) Calculate the Bayes factor comparing the model $M_1 : \beta_2 = 0$ with $M_2 : \beta_2 \neq 0$.

(d) Carry out a prior sensitivity analysis by setting $\underline{V} = cI_2$ and repeating parts (b) and (c) for values of $c = 0.01, 1.0, 100.0$, and 1×10^6 . How sensitive is the posterior to changes in prior information? How sensitive is the Bayes factor? (See Exercises 6.7, 8.7, and 8.8 for related discussion.)

(e) Repeat part (b) using Monte Carlo integration for various values of R . How large does R have to be before you reproduce the results of the previous parts to two decimal places?

(f) Calculate the NSEs associated with the posterior mean of the slope coefficient for the models. Does the NSE seem to give a reliable guide to the accuracy of the approximation provided by Monte Carlo integration?

Solution

(a) A program to generate artificial data as described in this exercise has the following form:

- Step 1: Set N , β_1 , β_2 , and h to the required values (i.e., 100, 0, 1, and, 1 respectively).
- Step 2: Take a draw of the explanatory variable x_i from the $U(0, 1)$ distribution.
- Step 3: Take a draw of the error ϵ_i from the $N(0, 1)$.
- Step 4: Set the dependent variable equal to $\beta_1 + \beta_2 x_i + \epsilon_i$.
- Step 5: Repeat Steps 2 through 4 N times and save all the values of the dependent and explanatory variables.

(b and c) A program to calculate the posterior mean and standard deviation of the regression coefficients involves first setting the prior hyperparameters to their required values and then evaluating the formulas in Chapter 10, Exercise 10.1. A program to calculate the marginal likelihood evaluates the formula from Chapter 10 (see, e.g., the formula for the Bayes factor in Exercise 10.4 and remember that the Bayes factor is the ratio of marginal likelihoods). A simple way to calculate the Bayes factor comparing the two models is

simply to run the program with the artificial data set and then run it again with the only explanatory variable being an intercept. Again, the programs are provided on the Web site associated with this book.

The answer to part (d) provides empirical results.

(d) The prior sensitivity analysis can be done by running the program for $c = 0.01, 1.0, 100.0, 1 \times 10^6$. To calculate the Bayes factor the program can be run twice for each value of c , once with the entire artificial data set (i.e., $k = 2$) and once with X containing only an intercept (i.e., $k = 1$). Our results will differ from yours since you will generate a different artificial data set. Tables 11.2A and 11.2B present posterior means, standard deviations and Bayes factors. The results for $c = 1$ complete the answer to part (b).

Table 11.2A: Results of prior sensitivity analysis for β_2 .

	Posterior Mean	Posterior Std. Dev.
$c = 0.01$	1.015	0.079
$c = 1.0$	1.066	0.289
$c = 100$	1.074	0.312
$c = 10^6$	1.074	0.312

Table 11.2B: Bayes factors for $\beta_2 = 0$.

	Bayes Factor
$c = 0.01$	4.7×10^{-7}
$c = 1.0$	0.008
$c = 100$	0.096
$c = 10^6$	9.251

These tables show some interesting patterns. Note, first, that the posterior mean and standard deviation are more robust to changes in the prior than are the Bayes factors (a standard finding in empirical Bayesian work). In fact, except for the case where the prior is extremely informative (as in $c = 0.01$), the posterior means and standard deviations are virtually the same for all prior choices. The Bayes factors vary from strong evidence against $\beta_2 = 0$ through strong evidence in favor of $\beta_2 = 0$. This motivates the common belief that prior elicitation is more important when the researcher is interested in model comparison than when he or she is interested in estimating parameters in a single model. In general, as the prior for β_2 gets more noninformative (i.e., c gets larger), the Bayes factor indicates more support for the restricted model (i.e., $\beta_2 = 0$). It can be shown that a noninformative prior (i.e., $c \rightarrow \infty$) will yield an infinite Bayes factor and, thus, the restricted model will always be supported, regardless of the data evidence. This is referred to as *Bartlett's paradox* [see Poirier (1995), page 390].

Second, our prior is centered over the value used to generate the artificial data. Thus, prior and likelihood information are not in conflict. In this case, having a more informative

prior (i.e., setting c to a small value) will ensure more accurate estimation (i.e., the posterior standard deviation of β_2 decreases as c decreases). The reader interested in further investigation might want to experiment with different priors that are in conflict with the likelihood (e.g., what happens if $\underline{\beta} = [0 \ 0]'$?, what happens if $\underline{\beta} = [0 \ -1]'$?, etc.).

(e and f) Monte Carlo integration for β can be done in two ways. The researcher can directly draw from the posterior for β using the multivariate t -distribution (see the Appendix to this book). Alternatively, the researcher can draw from the posterior of h (which is gamma) and then draw from $p(\beta|y, h)$ (which is normal). This technique is referred to as the *method of composition*. Here we adopt the first approach. This requires computer code for randomly drawing from the t -distribution. This is available in many places. For instance, James LeSage's Econometrics Toolbox [see LeSage (1999)] provides Matlab code for drawing from many common distributions including the t -distribution.

The structure of computer code that does Monte Carlo integration to carry out posterior inference on β is as follows:

- Step 1: Do all the preliminary things to create all the variables used in the Monte Carlo procedure i.e., load the data, specify prior hyperparameters, and evaluate the posterior mean, scale matrix, and degrees of freedom for the multivariate t posterior [see code for part (c)].
- Step 2: Draw from the multivariate t posterior for β .
- Step 3: Repeat Step 2 R times and average the draws of β_2 (which gives you the Monte Carlo estimate of the posterior mean) and average the draws of β_2^2 [which gives you the Monte Carlo estimate of $E(\beta_2^2|y)$, which, along with the posterior mean, can be used to calculate the posterior variance and standard deviation of the slope parameter].

Matlab code for performing these calculations and those for part (f) are provided on the Web site associated with this book. The formula for the numerical standard error is derived in Exercise 11.1. As described there, this depends only on the Monte Carlo estimate of $\text{Var}(\beta_2|y)$ (which was called $\hat{\sigma}^2$ in Exercise 11.1) and R . These are readily available in the Monte Carlo integration procedure [see code for part (e)].

Table 11.3 presents the posterior mean and standard deviation for β_2 calculated analytically [see part (b)] and using Monte Carlo integration with different numbers of replications. In addition, for Monte Carlo estimates we present the NSEs for $E(\beta_2|y)$.

Table 11.3: Posterior results for β_2 .

	Posterior Mean	Posterior Std. Dev.	NSE for $E(\beta_2 y)$
Analytical	1.066	0.289	—
$R = 10$	1.121	0.341	0.108
$R = 100$	1.029	0.289	0.029
$R = 1,000$	1.069	0.292	0.009
$R = 10,000$	1.067	0.288	0.002
$R = 1,000,000$	1.066	0.289	3×10^{-3}

The row labeled “Analytical” presents the correct posterior mean and standard deviation (with numbers taken from Table 11.2 with $c = 1$). Since Monte Carlo integration depends on random number generation, no two Monte Carlo procedures will yield exactly the same results. Hence, your numbers may be different from ours. However, the following patterns should be evident. First, as the number of replications increases, the approximation error associated with the Monte Carlo integration gets smaller. Second, it seems that roughly 1,000 replications are required to obtain results accurate to two decimal places. However, third, a better way of gauging the accuracy of the Monte Carlo approximation is through the numerical standard error, which does seem very reliable (in the sense that, for each value of R , the Monte Carlo estimate is within one or two numerical standard errors of the true posterior mean).

11.2 Importance sampling

Exercise 11.4 (Importance sampling) The purpose of this question is to learn about the properties of importance sampling in a very simple case. Assume you have a model with a single parameter, θ , and its posterior is $N(0, 1)$.

- Write a program that calculates the posterior mean and standard deviation of θ using Monte Carlo integration.
- Write a program that calculates the posterior mean and standard deviation of θ using importance sampling and calculates the mean and standard deviation of the importance sampling weights. Use the $t(0, 1, \nu)$ distribution as the importance function.
- Carry out Monte Carlo integration and importance sampling with $\nu = 2, 5$, and 100 for a given number of replications (e.g., $R = 10$). Compare the accuracy of the estimates across different algorithms and choices for ν . Discuss your results, paying attention to the issue of what happens to the importance sampling weights as ν increases.
- Redo part (c) using the $t(3, 1, \nu)$ as an importance function. Discuss the factors affecting accuracy of importance sampling in light of the fact that this importance function is a poor approximation to the posterior.

Solution

(a and b) The structure of computer code that does Monte Carlo integration and importance sampling to carry out posterior inference on θ is as follows:

- Step 1: Do all the preliminary things to create all the variables used in the Monte Carlo and importance sampling procedures (i.e., specify parameters of the importance function and initialize all the Monte Carlo and importance sampling sums at zero).
- Step 2: Draw θ from the normal posterior (for Monte Carlo integration).
- Step 3: Draw θ from the importance function and calculate the importance sampling weight (and weight squared). Multiply the drawn θ by the importance sampling weight to obtain that draw’s contribution to the numerator of the importance sampling estimator (see, e.g., the introduction to this chapter).

Table 11.4: Monte Carlo and importance sampling results for $N(0, 1)$ posterior.

	Post. Mean	Post. Std. Dev.	Mean Imp. Samp. Wt.	Std. Dev. Imp. Samp. Wt.
Monte Carlo	0.046	0.874	—	—
Importance Sampling				
$t(0, 1, 2)$	-0.110	0.500	0.815	0.473
$t(0, 1, 5)$	0.055	0.721	0.948	0.106
$t(0, 1, 100)$	0.049	0.925	1.004	9.6×10^{-4}
$t(3, 1, 2)$	1.469	0.525	0.373	0.638
$t(3, 1, 5)$	1.576	0.426	0.186	0.439
$t(3, 1, 100)$	1.996	0.368	0.055	0.120

- Step 4: Repeat Steps 2 and 3 R times.
- Step 5: Average the Step 2 draws and draws squared (to produce the Monte Carlo estimate of the posterior mean and standard deviation). Divide the Step 3 draws and draws squared by the sum of the weights (to produce the importance sampling estimate of the posterior mean and standard deviation). Calculate the mean and standard deviation of the weights.

The Matlab code used to perform these steps is provided on the Web site associated with this book.

(c and d) Table 11.4 presents Monte Carlo integration and the importance sampling results for each of the different importance functions specified in the question using $R = 10$. Note that as $R \rightarrow \infty$ the posterior means and standard deviations will be the same for all approaches (i.e., they will equal their true values of 0 and 1). Hence, choosing a relatively small value of R is important to see the effect of the importance function on accuracy of estimation. In an empirical exercise you would choose R to be much larger.

Table 11.4 illustrates some of the issues relating to selection of an importance function. First, note that as ν increases the t -distribution approaches the normal; the $t(0, 1, 100)$ is virtually the same as the true $N(0, 1)$ posterior. Thus, importance sampling is virtually the same as Monte Carlo integration and the importance sampling weights are all virtually 1.0. Second, the $t(0, 1, 5)$ and $t(0, 1, 2)$ importance functions both approximate the posterior well, but have fatter tails. Thus, they yield results that are a bit less accurate than Monte Carlo integration and have importance sampling weights that vary widely across draws. That is, the standard deviation of the importance sampling weights indicates that some draws are receiving much more weight than others (in contrast to the more efficient Monte Carlo integration procedure, which weights all draws equally).

Third, the importance functions with mean 3.0 all approximate the $N(0, 1)$ posterior poorly. Results with these importance functions are way off, indicating that it is important to choose an importance function that approximates the posterior well. Note that the posterior means are all too high. These importance functions are taking almost all of their draws in implausible regions of the parameter space (e.g., most of the importance sampling

draws will be greater than 2.0 whereas the true posterior allocates very little probability to this region). Of course, importance sampling corrects for this by giving little weight to the draws greater than 2.0 and great weight to the (very few) draws less than 2.0 and, as $R \rightarrow \infty$, importance sampling estimates will converge to the true values. But with small R importance sampling results can be misleading and care should be taken (e.g., by looking at numerical standard errors) to make sure R is large enough to yield accurate results. You might wish to experiment to find out how large R has to be to obtain accurate results using the $t(3, 1, \nu)$ importance function. (We find setting $R = 10,000$ yields reasonably accurate results.)

Fourth, of the $t(3, 1, 2)$, $t(3, 1, 5)$, and $t(3, 1, 100)$ importance functions, it is the one with $\nu = 2$ degrees of freedom that seems to be yielding most accurate results. This is because it has fatter tails and, thus, is taking more dispersed draws, which means more draws in regions where the $N(0, 1)$ posterior is appreciable.

Overall, these results suggest that it is important to get an importance function that approximates the posterior reasonably well. In empirical exercises, strategies such as setting the mean and variance of a t -distribution importance function to maximum likelihood or posterior quantities are common. However, as an insurance policy it is common to choose a small degrees-of-freedom parameter to ensure that key regions of the parameter space are covered. A common rule of thumb is that importance functions should have fatter tails than the posterior.

Exercise 11.5 (Importance sampling: prior sensitivity) The purpose of this question is to see how importance sampling can be used to carry out a prior sensitivity analysis.

This question uses the normal linear regression model with natural conjugate prior and a data set. Definitions and notation for the model are given in the introduction to Chapter 10. Generate an artificial data set of size $N = 100$ from the normal linear regression model with an intercept and one other explanatory variable. Set the intercept (β_1) to 0, the slope coefficient (β_2) to 1.0, and h to 1.0. Generate the explanatory variable by taking random draws from the $U(0, 1)$ distribution (although you may wish to experiment with other data sets). Suppose your prior beliefs are reflected in your base prior, which has prior hyperparameter values of $\underline{\beta} = [0 \ 1]'$, $\underline{V} = I_2$, $\underline{s}^{-2} = 1$, and $\underline{\nu} = 1$. However, you wish to carry out a prior sensitivity analysis with respect to the prior mean and standard deviation of the slope coefficient and, hence, also want to consider priors with $\underline{\beta} = [0 \ c]'$ and $\underline{V} = \begin{bmatrix} 1 & 0 \\ 0 & d \end{bmatrix}$ for values of $c = 0, 1$, and 2 and $d = 0.01, 1$, and 100 .

- Calculate the posterior mean and standard deviation for the slope coefficient β_2 for this data set for every one of these priors using analytical results.
- Write a program that does the same things as part (a) using Monte Carlo integration and use this program to produce $R = 100$ draws of β_2 using the base prior.
- Write a program that uses importance sampling to carry out the prior sensitivity analysis, *using only the draws produced in part (b)*. Compare results obtained using this approach to those obtained in part (a). How do your results change when you set $R = 10,000$?

Solution

(a and b) The solution to parts (a) and (b) is given as part of the solution to Exercise 11.3. Empirical results for these parts are given in the following. The answer to part (c) is based on the insight that the posterior for the normal linear regression model with the base prior can be treated as an importance function for the posteriors corresponding to the other priors used in the prior sensitivity analysis. Formally, let M_1 be the normal linear regression model with base prior, M_2 be the normal linear regression model with any other prior, and $\theta = [\beta_1 \ \beta_2 \ h]'$ be the parameter vector (common to both models). Monte Carlo integration using the program written for part (b) will provide draws from $p(\theta|y, M_1)$. If we treat this posterior as an importance function in a Bayesian analysis of M_2 , then the importance sampling weights will be

$$w_r = \frac{p(\theta = \theta^{(r)}|y, M_2)}{p(\theta = \theta^{(r)}|y, M_1)},$$

where $\theta^{(r)}$ for $r = 1, \dots, R$ are draws from $p(\theta|y, M_1)$. In a prior sensitivity analysis, the importance sampling weights simplify considerably since both models have the same likelihood function, which cancels out in the ratio, yielding

$$w_r = \frac{p(\theta = \theta^{(r)}|M_2)}{p(\theta = \theta^{(r)}|M_1)}.$$

In fact, for the particular prior structure in the question, the importance sampling weights simplify even further since the same prior is used for the error precision in each model and β_1 and β_2 are, *a priori*, uncorrelated with one another. Thus, we have

$$w_r = \frac{p(\beta_2 = \beta_2^{(r)}|M_2)}{p(\beta_2 = \beta_2^{(r)}|M_1)},$$

which is the ratio of two t densities.

(c) The sketch outline of the program is the same as for any Monte Carlo integration/importance sampling program (see the solutions to Exercises 11.3 and 11.4) and will not be repeated here. The Matlab code used to answers parts (a), (b), and (c) is provided on the Web site associated with this book.

Table 11.5 presents results using this program. A discussion of this table completes the answer to this question. It is worth noting that we are carrying out a sensitivity analysis over a very wide range of priors for β_2 . To give an idea about the information in the data, note that the OLS estimate of β_2 is 1.074 and its standard error is 0.311. The priors with large prior variances (i.e., $d = 100$) are very noninformative relative to the data, whereas priors with $d = 0.01$ are more informative than the data. Priors with $c = 1.0$ are centered over the true value used to generate the data, whereas priors with $c = 0.0$ and $c = 2.0$ are far away from the true value (at least for priors with informative values for the prior variance of β_2). In the columns of Table 11.5 labeled “Analytical Results” we can see that results are quite sensitive to the prior (especially when $d = 0.01$).

Table 11.5: Results from prior sensitivity analysis.

Hyperparameter		Analytical Results		Importance Sampling Results			
Values							
c	d	$E(\beta_2 y)$	$\sqrt{\text{Var}(\beta_2 y)}$	$R = 100$		$R = 10,000$	
0.0	0.01	0.081	0.089	0.286	0.294	0.116	0.090
0.0	1.0	0.957	0.291	0.963	0.323	0.997	0.306
0.0	100.0	1.073	0.306	1.076	0.344	1.076	0.322
1.0	0.01	1.006	0.084	1.019	0.099	1.008	0.095
1.0	1.0	1.066	0.291	1.068	0.325	1.070	0.305
1.0	100.0	1.074	0.306	1.077	0.344	1.077	0.322
2.0	0.01	1.930	0.088	1.910	0.046	1.923	0.097
2.0	1.0	1.176	0.291	1.174	0.328	1.164	0.307
2.0	100.0	1.076	0.306	1.078	0.344	1.079	0.322

The posterior corresponding to the base prior (with $c = 1.0$ and $d = 1.0$) is used as an importance function for the posteriors corresponding to all the other priors. We expect importance sampling to work best when the former posterior is similar to the latter. Given the wide spread of priors, it might be the case that importance sampling works poorly, especially for the priors with $d = 0.01$. When $R = 100$ replications are used, then importance sampling results can, indeed, be a bit off, especially for the $d = 0.01$ cases. However, when $R = 10,000$ replications are used, the performance of the importance sampling algorithm works much better, even for the extreme cases considered here. Thus, Table 11.5 indicates that importance sampling can be used to carry out prior sensitivity analyses, even when the sensitivity analysis is over a wide range of priors. Of course, for the normal linear regression model with a natural conjugate prior it is not necessary to use importance sampling since analytical results are available. However, for a model that does not allow for analytical posterior results, the strategy outlined in this question – of taking posterior draws from a model based on one prior and then reweighting them using importance sampling – may be a very efficient way of conducting a prior sensitivity analysis.

11.3 Gibbs sampling and the Metropolis–Hastings algorithm

The algorithms presented in Exercises 11.1–11.5 fell under the heading of *noniterative* methods. That is, no requirement was imposed that the algorithms needed to “converge” in order to generate draws from the posterior $p(\theta|y)$ or to accurately characterize posterior moments of the form $E[g(\theta)|y]$. In the following set of exercises, we turn to describe two *iterative* algorithms: the Gibbs sampler and the Metropolis–Hastings algorithm. Since these methods are quite powerful and widely-used, several exercises are designed to illustrate their use in fitting a selection of standard econometric models.

Exercise 11.6 (Gibbs sampling) The *Gibbs sampler* is an indispensable tool in the Bayesian’s toolkit. As discussed in the introduction to this chapter, the Gibbs sampler is an iterative algorithm. Instead of drawing from the joint posterior distribution directly – which typically is not possible in problems of even moderate complexity – the researcher instead draws from the individual *posterior conditional distributions*. This essentially breaks the joint estimation problem into a series of smaller, more manageable problems. Under certain regularity conditions, iteratively sampling from the complete conditional distributions (updating the conditioning to reflect the most recent values obtained from the algorithm) converges to drawing from the joint posterior distribution itself. Here, we do not discuss the theoretical properties of the sampler at length – we invite the reader to see Casella and George (1992) for an excellent introductory reference and Tierney (1994) for a more advanced reference. In this exercise we simply provide some suggestive evidence that the Gibbs sampler works, and specifically, we show that, for the bivariate discrete case, the target distribution is an invariant distribution.

Suppose X and Y are discrete random variables with $X \in \{x_1, x_2, \dots, x_J\}$ and $Y \in \{y_1, y_2, \dots, y_K\}$. Let the joint distribution of X and Y be represented by p , with $p(x_j, y_k) = \Pr(X = x_j, Y = y_k)$.

Consider implementing the Gibbs sampler and suppose that the current value of the chain [denoted (x^t, y^t)] was generated from the joint distribution p . Show that the next value of the chain, (x^{t+1}, y^{t+1}) , obtained via the Gibbs sampler, must also be generated from p . That is, once we have converged to obtain a draw from p , then the subsequent draws in our parameter chain must also be drawn from p .

Solution

Without loss of generality, consider an arbitrary point in the support of the joint posterior distribution, (x_0, y_0) . We note that

$$\begin{aligned} \Pr(x^{t+1} = x_0, y^{t+1} = y_0) &= \sum_{j,k} \left[\Pr(x^{t+1} = x_0, y^{t+1} = y_0 | x^t = x_j, y^t = y_k) \right. \\ &\quad \left. \times \Pr(x^t = x_j, y^t = y_k) \right]. \end{aligned}$$

Now, consider the first term in this summation. Again, without loss of generality, suppose that we first draw from $y|x$ and then, given the realized value of y , say y^* , we draw from $x|y = y^*$. In this case, we can write

$$\begin{aligned} \Pr(x^{t+1} = x_0, y^{t+1} = y_0 | x^t = x_j, y^t = y_k) &= \Pr(y^{t+1} = y_0 | x^t = x_j) \\ &\quad \times \Pr(x^{t+1} = x_0 | y^{t+1} = y_0). \end{aligned}$$

Therefore, substituting this result into the summation, we obtain

$$\begin{aligned} \Pr(x^{t+1} = x_0, y^{t+1} = y_0) &= \Pr(x^{t+1} = x_0 | y^{t+1} = y_0) \\ &\quad \times \sum_{j,k} \Pr(y^{t+1} = y_0 | x^t = x_j) \Pr(x^t = x_j, y^t = y_k). \end{aligned}$$

We are given that the current value of the chain is distributed according to p , and, of course, the complete conditionals are also based upon p . Thus, we can write

$$\begin{aligned}
 \Pr(x^{t+1} = x_0, y^{t+1} = y_0) &= \frac{p(x_0, y_0)}{p(y_0)} \sum_{j,k} \frac{p(x_j, y_0)}{p(x_j)} p(y_k | x_j) p(x_j) \\
 &= \frac{p(x_0, y_0)}{p(y_0)} \sum_{j,k} p(x_j, y_0) p(y_k | x_j) \\
 &= \frac{p(x_0, y_0)}{p(y_0)} \sum_j p(x_j, y_0) \sum_k p(y_k | x_j) \\
 &= \frac{p(x_0, y_0)}{p(y_0)} \sum_j p(x_j, y_0) \\
 &= \frac{p(x_0, y_0)}{p(y_0)} p(y_0) \\
 &= p(x_0, y_0).
 \end{aligned}$$

Thus, the probability that the next value of the chain equals (x_0, y_0) is the actual joint probability of that event.

Exercise 11.7 (Gibbs sampling from the bivariate normal) The purpose of this question is to learn about the properties of the Gibbs sampler in a very simple case.

Assume that you have a model that yields a bivariate normal posterior,

$$\begin{pmatrix} \theta_1 \\ \theta_2 \end{pmatrix} \sim N \left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix} \right),$$

where $|\rho| < 1$ is the (known) posterior correlation between θ_1 and θ_2 .

- Write a program that uses Monte Carlo integration to calculate the posterior means and standard deviations of θ_1 and θ_2 .
- Write a program that uses Gibbs sampling to calculate the posterior means and standard deviations of θ_1 and θ_2 .
- Set $\rho = 0$ and compare the programs from parts (a) and (b) for a given number of replications (e.g., $R = 100$) and compare the accuracy of the two algorithms.
- Repeat part (c) of this question for $\rho = .5, .9, .99$, and $.999$. Discuss how the degree of correlation between θ_1 and θ_2 affects the performance of the Gibbs sampler. Make graphs of the Monte Carlo and Gibbs sampler replications of θ_1 (i.e., make a graph with x axis being replication number and y axis being θ_1). What can the graphs you have made tell you about the properties of Monte Carlo and Gibbs sampling algorithms?
- Repeat parts (c) and (d) using more replications (e.g., $R = 10,000$) and discuss how Gibbs sampling accuracy improves with the number of replications.

Solution

(a and b) A Matlab program for this question is provided on the Web site associated with this book. Monte Carlo and Gibbs sampling algorithms have the same general structure: an introductory part where quantities are initialized (in this question, this means a value for ρ is set and all the sums used to calculate Monte Carlo and Gibbs estimates are initialized at zero), then a loop (in Matlab this is a “for loop”; in other languages it is often called a “do loop”), which repeatedly draws from the posterior (in the case of Monte Carlo integration) or the posterior conditionals (in the case of Gibbs sampling) and sums the draws. Finally, the program ends with the final calculations (e.g., dividing sums by number of replications to get Monte Carlo or Gibbs estimates) and printing out and/or graphing of results.

The Monte Carlo integration algorithm requires the program to take random draws from the bivariate normal posterior given in the question. The Gibbs sampling algorithm requires the posterior conditionals. Using the properties of the multivariate normal distribution (see Theorem 3 of the Appendix) it follows immediately that

$$\theta_1|\theta_2, y \sim N(\rho\theta_2, 1 - \rho^2)$$

and

$$\theta_2|\theta_1, y \sim N(\rho\theta_1, 1 - \rho^2).$$

Of course, in practice there would be no need to do either Gibbs sampling or Monte Carlo integration since the true mean and standard deviation of θ_1 and θ_2 are given in the question. However, the purpose of this question is to compare posterior simulation algorithms and it is useful to know what the true answer should be when making such comparisons.

(c and d) Table 11.6 provides Monte Carlo integration and Gibbs sampler results for θ_1 and θ_2 for all the posteriors specified in the question. With the Gibbs sampler we discard an

Table 11.6: Posterior means and standard deviations of θ_1 and θ_2 with $R = 100$.

	$E(\theta_1 y)$	$\sqrt{\text{Var}(\theta_1 y)}$	$E(\theta_2 y)$	$\sqrt{\text{Var}(\theta_2 y)}$
True Value	0.00	1.00	0.00	1.00
Monte Carlo Integration				
$\rho = 0.00$	0.07	1.02	0.16	1.13
$\rho = 0.50$	0.02	0.89	0.06	0.94
$\rho = 0.90$	0.03	0.95	−0.06	0.97
$\rho = 0.99$	−0.07	1.09	0.06	1.08
$\rho = 0.999$	0.02	1.00	0.02	1.02
Gibbs Sampling				
$\rho = 0.00$	0.14	1.16	0.06	1.09
$\rho = 0.50$	0.19	1.03	0.24	0.99
$\rho = 0.90$	0.42	0.78	0.29	0.83
$\rho = 0.99$	−0.42	0.64	−0.40	0.63
$\rho = 0.999$	0.61	0.17	0.62	0.16

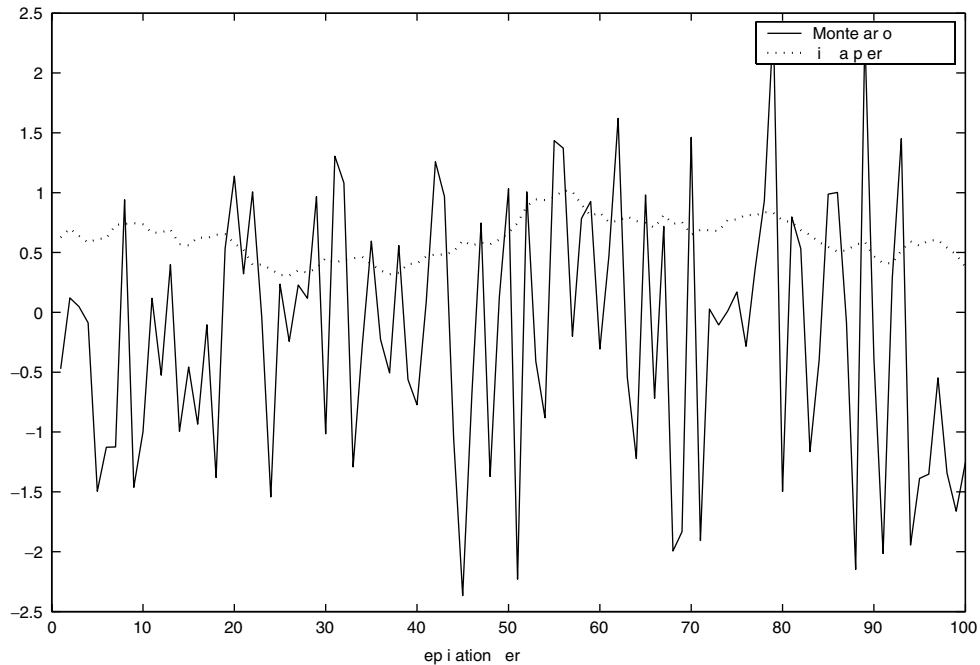


Figure 11.1 — Path of Monte Carlo and Gibbs sampling draws for θ_1 with $\rho = .999$.

initial 100 burn-in replications to remove the effect of initial conditions. We stress that since posterior simulation depends on random number generation, you may get results different from those presented in Table 11.6. That is, a posterior simulation algorithm run twice will not yield the same answer, especially with a small number of replications such as $R = 100$. With these qualifications in mind, the patterns in Table 11.6 are clear. For Monte Carlo integration, the degree of accuracy does not depend upon ρ . But for Gibbs sampling, it does.

When $\rho = 0$, Gibbs sampling is equivalent to Monte Carlo integration since, in this case, θ_1 and θ_2 are independent of one another and drawing from the posterior conditionals is equivalent to drawing from the two marginal distributions, which is equivalent to drawing from the joint distribution (owing to independence). Table 11.6 indicates that, for $\rho = 0$, Gibbs sampling is roughly as accurate as Monte Carlo integration. However, as ρ becomes bigger and bigger, Gibbs sampling becomes less and less accurate. Note, in particular, that the posterior standard deviations become much smaller than their true values when ρ is large.

The reason for the poor performance of Gibbs sampling when θ_1 and θ_2 are highly correlated is that the sequence of Gibbs draws becomes highly correlated as well. This can be seen in Figure 11.1 (or inferred from the equations for the conditional posteriors just given), which plots the draws of θ_1 for $\rho = 0.999$. The Monte Carlo integration sequence is constructed from purely random draws from the posterior, but the Gibbs sampling draws

Table 11.7: Posterior means and standard deviations of θ_1 and θ_2 with $R = 10,000$.

	$E(\theta_1 y)$	$\sqrt{\text{Var}(\theta_1 y)}$	$E(\theta_2 y)$	$\sqrt{\text{Var}(\theta_2 y)}$
True Value	0.00	1.00	0.00	1.00
Monte Carlo Integration				
$\rho = 0.00$	0.01	0.99	-0.01	1.00
$\rho = 0.50$	0.01	1.00	0.02	1.00
$\rho = 0.90$	0.00	1.00	0.00	0.99
$\rho = 0.99$	-0.01	1.00	0.00	1.00
$\rho = 0.999$	-0.02	0.99	-0.02	0.99
Gibbs Sampling				
$\rho = 0.00$	0.00	0.99	0.01	1.00
$\rho = 0.50$	-0.03	0.99	-0.02	1.00
$\rho = 0.90$	0.04	0.99	0.05	1.00
$\rho = 0.99$	-0.07	0.95	-0.07	0.95
$\rho = 0.999$	-0.32	0.87	-0.32	0.87

are highly correlated with one another, resulting in the smooth, trendlike behavior of the Gibbs sequence. Intuitively, R random draws from a distribution are typically going to contain more information than R correlated draws (i.e., if two draws are highly correlated with one another, the second draw will mostly contain the same information as the first), and thus Monte Carlo integration is to be preferred to Gibbs sampling in this special case where the researcher has the choice between approaches.

Graphs of the Gibbs sampler and Monte Carlo integration draws (not presented here) for other values of ρ exhibit the expected patterns. When $\rho = 0$, the graphs of the Gibbs and Monte Carlo draws have the same erratic random pattern. However, as ρ increases, the graphs of Monte Carlo draws remain random, but the Gibbs draws become smoother and smoother (exhibiting greater serial correlation) until we reach the extreme case presented in Figure 11.1.

(e) Table 11.7 is the same as Table 11.6, except that the program was run using $R = 10,000$ draws. With this larger number of draws, Monte Carlo integration is now quite accurate for all posteriors. The performance of the Gibbs sampler is much improved as well. However, for high values of ρ , the accuracy of the Gibbs sampler is still reasonably poor. We ran the program with $R = 100,000$ for the posterior with $\rho = 0.999$, and even with this many draws we found the posterior mean of θ_1 to be 0.09 and the posterior standard deviation of θ_1 to equal 0.93, which are still reasonably far from the true values.

These findings highlight the importance of choosing a sufficient number of draws for the Gibbs sampler. For Monte Carlo integration, a reliable rule of thumb is that $R = 10,000$ will yield an estimate of a parameter's posterior mean with error that is 1% of its posterior standard deviation. This is usually accurate enough for empirical work. For Gibbs sampling, no such rule of thumb exists (as is demonstrated by this question) and careful

use of Gibbs sampling convergence diagnostics is required [see, e.g., Koop (2003), pages 64–68].

Exercise 11.8 (Gibbs sampling in a regression problem) Consider the regression model

$$y_i = x_i\beta + \epsilon_i, \quad \epsilon_i \stackrel{\text{iid}}{\sim} N(0, \sigma^2), \quad i = 1, 2, \dots, n.$$

In this exercise, unlike many of the others appearing to this point, we choose to work with the error variance σ^2 rather than the error precision σ^{-2} . We employ proper priors for these parameters of the forms

$$\begin{aligned} \beta &\sim N(\mu_\beta, V_\beta), \\ \sigma^2 &\sim IG(a, b), \end{aligned}$$

with μ_β , V_β , a , and b given, and $IG(\cdot, \cdot)$ denotes the inverted gamma density (see Appendix Theorem 2 for parameterization and further discussion).

Show how the Gibbs sampler can be used to fit this linear regression model.

Solution

The likelihood function for this regression model is

$$L(\beta, \sigma^2) = (2\pi)^{(-n/2)} (\sigma^2)^{(-n/2)} \exp\left(-\frac{1}{2\sigma^2} (y - X\beta)'(y - X\beta)\right),$$

with

$$y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \quad \text{and} \quad X = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}.$$

By Bayes' theorem, it follows that

$$p(\beta, \sigma^2 | y) \propto L(\beta, \sigma^2) \exp\left(-\frac{1}{2}(\beta - \mu_\beta)' V_\beta^{-1} (\beta - \mu_\beta)\right) (\sigma^2)^{-[a+1]} \exp\left(-\frac{1}{b\sigma^2}\right).$$

To implement the Gibbs sampler we first need to derive the complete posterior distributions for all the parameters of our model. The complete posterior conditional for β follows immediately from an application of standard results (see, e.g., Exercise 10.1), since we are conditioning on the value of σ^2 :

$$\beta | \sigma^2, y \sim N(D_\beta d_\beta, D_\beta),$$

where

$$D_\beta = \left(X'X/\sigma^2 + V_\beta^{-1}\right)^{-1}$$

and

$$d_\beta = X'y/\sigma^2 + V_\beta^{-1}\mu_\beta.$$

For the complete conditional for σ^2 , we have

$$p(\sigma^2|\beta, y) \propto p(\beta, \sigma^2|y) \\ \propto (\sigma^2)^{-[(n/2)+a+1]} \exp\left(-\frac{1}{\sigma^2} \left[b^{-1} + \frac{1}{2}(y - X\beta)'(y - X\beta)\right]\right),$$

so that, using the parameterization described in Appendix Theorem 2,

$$\sigma^2|\beta, y \sim IG\left(\frac{n}{2} + a, \left[b^{-1} + \frac{1}{2}(y - X\beta)'(y - X\beta)\right]^{-1}\right).$$

To implement the Gibbs sampler, we would start with an initial value for σ^2 , and then sample from the posterior conditional $\beta|\sigma^2, y$ described here. Given a draw from this conditional, say β^* , we would then draw a new value of σ^2 from $\sigma^2|\beta = \beta^*, y$. Repeating in this fashion, we will converge to drawing from the joint posterior density.

Exercise 11.9 (Gibbs sampling in a model of groupwise heteroscedasticity) Suppose we generalize the regression model in the previous exercise so that

$$y = X\beta + \epsilon,$$

where

$$E(\epsilon\epsilon') \equiv \Sigma = \begin{bmatrix} \sigma_1^2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ \vdots & \ddots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \sigma_1^2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \sigma_2^2 & 0 & 0 & 0 & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & 0 & \sigma_2^2 & 0 & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & \sigma_J^2 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \sigma_J^2 \end{bmatrix}.$$

That is, we relax the homoscedasticity assumption with one of *groupwise heteroscedasticity* – where we presume there are J different groups with identical variance parameters within each group, but different parameters across groups. Let n_j represent the number of observations belonging to group J .

Given priors of the form

$$\beta \sim N(\mu_\beta, V_\beta), \\ \sigma_j^2 \stackrel{\text{ind}}{\sim} IG(a_j, b_j), \quad j = 1, 2, \dots, J,$$

show how the Gibbs sampler can be used to estimate this model.

Solution

We obtain the following complete posterior conditional for β (since conditioned on all the variance parameters, Σ is known):

$$\beta | \{\sigma_j^2\}_{j=1}^J, y \sim N(D_\beta d_\beta, D_\beta),$$

where

$$D_\beta = \left(X' \Sigma^{-1} X + V_\beta^{-1} \right)^{-1}$$

and

$$d_\beta = X' \Sigma^{-1} y + V_\beta^{-1} \mu_\beta.$$

For the complete posterior for each σ_j^2 , first note that the likelihood function for this problem can be written as

$$L(\beta, \{\sigma_j^2\}) = \prod_{j=1}^J (2\pi)^{-(n_j/2)} (\sigma_j^2)^{-(n_j/2)} \exp \left(-\frac{1}{2\sigma_j^2} (y_j - X_j \beta)' (y_j - X_j \beta) \right),$$

with y_j and X_j being the $n_j \times 1$ outcome vector and $n_j \times k$ design matrix, respectively, for group j . Given this form of the likelihood together with the independent inverse gamma priors for $\{\sigma_j^2\}_{j=1}^J$, similar arguments to those given in the previous exercise show

$$\sigma_j^2 | \Gamma_{-\sigma_j^2}, y \stackrel{\text{ind}}{\sim} IG \left(\frac{n_j}{2} + a_j, \left[b_j^{-1} + \frac{1}{2} (y_j - X_j \beta)' (y_j - X_j \beta) \right]^{-1} \right), \quad j = 1, 2, \dots, J,$$

with Γ denoting all the parameters in our model and Γ_{-x} denoting all parameters other than x .

Thus, to implement the sampler, one can start with an initial condition for Σ , sample β from $\beta | \{\sigma_j^2\}, y$, and then, given the value of β just drawn, sample independently from the posterior conditionals for the σ_j^2 . Iterating over this process will produce a set of draws from the joint posterior density.

Exercise 11.10 (Gibbs sampling in the SUR model) Consider a two-equation version of the seemingly unrelated regression (SUR) model [Zellner (1962)]:

$$y_{i1} = x_{i1} \beta_1 + \epsilon_{i1},$$

$$y_{i2} = x_{i2} \beta_2 + \epsilon_{i2},$$

where

$$\epsilon_i = [\epsilon_{i1} \ \epsilon_{i2}]' \stackrel{\text{iid}}{\sim} N(0, \Sigma), \quad i = 1, 2, \dots, n,$$

x_{i1} and x_{i2} are $1 \times k_1$ and $1 \times k_2$, respectively, and

$$\Sigma \equiv \begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{bmatrix}.$$

Suppose you employ priors of the form

$$\beta = [\beta_1' \ \beta_2']' \sim N(\mu_\beta, V_\beta)$$

and

$$\Sigma^{-1} \sim W(\Omega, \nu),$$

where W denotes a Wishart distribution (see Appendix Definition 6 for a more detailed description of the Wishart).

- (a) Derive a posterior simulator for fitting this two-equation SUR model.
 (b) Conduct an experiment by first generating $n = 2,500$ draws from the model

$$y_{i1} = 1 + .3x_{i1} + \epsilon_{i1},$$

$$y_{i2} = 2 - .7x_{i2} + \epsilon_{i2},$$

with

$$\Sigma = \begin{bmatrix} .5 & -.45\sqrt{.5}\sqrt{.3} \\ -.45\sqrt{.5}\sqrt{.3} & .3 \end{bmatrix}$$

and $x_{ij} \stackrel{\text{iid}}{\sim} N(0, 1)$. Written in this way, the correlation between ϵ_1 and ϵ_2 is $-.45$.

Using the hyperparameters $\mu_\beta = 0_4$, $V_\beta = 1000I_4$, $\Omega = I_4$, and $\nu = 4$, fit the generated data using the results in part (a). Compare parameter posterior means to the parameter values used to generate the data.

Solution

- (a) First, note that the likelihood function can be expressed as

$$L(\beta, \Sigma) = \prod_{i=1}^n (2\pi)^{-1} |\Sigma|^{-1/2} \exp \left(-\frac{1}{2} \sum_{i=1}^n (\tilde{y}_i - \tilde{X}_i \beta)' \Sigma^{-1} (\tilde{y}_i - \tilde{X}_i \beta) \right),$$

where

$$\tilde{y}_i = [y_{i1} \ y_{i2}]', \quad \tilde{X}_i = \begin{bmatrix} x_{i1} & 0 \\ 0 & x_{i2} \end{bmatrix}$$

and \tilde{X}_i is $2 \times (k_1 + k_2)$. Combining this likelihood function with the Wishart prior for Σ^{-1} , we obtain the following posterior conditional distribution (see Exercise 12.3 for detailed derivations):

$$\Sigma^{-1} | \beta, y \sim W \left[\left(\Omega^{-1} + \sum_{i=1}^n (\tilde{y}_i - \tilde{X}_i \beta)(\tilde{y}_i - \tilde{X}_i \beta)' \right)^{-1}, n + \nu \right].$$

Alternatively, stacking the SUR model as

$$y = X\beta + \epsilon,$$

where

$$y = [y'_1 \ y'_2]', \quad X = \begin{bmatrix} X_1 & 0 \\ 0 & X_2 \end{bmatrix}, \quad \epsilon = [\epsilon'_1 \ \epsilon'_2]'$$

and y_j , X_j , and ϵ_j have been stacked over i for $j = 1, 2$, we obtain the following complete posterior conditional for β :

$$\beta | \Sigma, y \sim N(D_\beta d_\beta, D_\beta),$$

where

$$D_\beta = \left(X'(\Sigma^{-1} \otimes I_N)X + V_\beta^{-1} \right)^{-1}$$

and

$$d_\beta = X'(\Sigma^{-1} \otimes I_N)y + V_\beta^{-1}\mu_\beta.$$

We implement our posterior simulator by first sampling from the posterior conditional for β and then sampling from the conditional for Σ^{-1} . We also note that similar derivations would apply for a SUR model with more than two equations.

(b) We fit the model by cycling through the posterior conditionals derived in part (a). For the elements of the parameter vector β , we obtain posterior means equal to 1.00, .294, 1.99, and $-.706$, and the values used to generate the data were 1, .3, 2, and $-.7$, respectively. The true values of σ_1^2 , σ_2^2 , and ρ_{12} were .5, .3, and $-.45$, and we obtain posterior means for these quantities equal to .502, .300, and $-.432$, respectively. Thus, our posterior simulator performs well at recovering the parameters used to generate these data.

Exercise 11.11 (Gibbs sampling in a regression model with an unknown changepoint)

Suppose that the density for a time series y_t , $t = 1, 2, \dots, T$, conditioned on its lags, the model parameters, and other covariates, can be expressed as

$$y_t | \theta_1, \theta_2, \beta_1, \beta_2, \sigma^2, \tau^2, \lambda, x_t \sim \begin{cases} N(\theta_1 + \theta_2 x_t, \sigma^2) & \text{if } t \leq \lambda \\ N(\beta_1 + \beta_2 x_t, \tau^2) & \text{if } t > \lambda. \end{cases}$$

In this model, λ is a *changepoint*: For periods until λ , one regression is assumed to generate y , and following λ , a new regression is assumed to generate y .

Suppose you employ priors of the form

$$\theta = [\theta_1 \ \theta_2]' \sim N(\mu_\theta, V_\theta),$$

$$\beta = [\beta_1 \ \beta_2]' \sim N(\mu_\beta, V_\beta),$$

$$\sigma^2 \sim IG(a_1, a_2),$$

$$\tau^2 \sim IG(b_1, b_2),$$

$$\lambda \sim \text{uniform}\{1, 2, \dots, T-1\}.$$

Note that λ is treated as a parameter of the model, and by placing a uniform prior over the elements $1, 2, \dots, T-1$, a changepoint is assumed to exist.

- (a) Derive the likelihood function for this model.
 (b) Describe how the Gibbs sampler can be employed to estimate the parameters of this model, given these priors.
 (c) Generate 500 observations from the changepoint model

$$y_t | \theta_1, \theta_2, \beta_1, \beta_2, \sigma^2, \tau^2, \lambda, x_t \sim \begin{cases} N(2 + x_t, .2) & \text{if } t \leq 85, \\ N(1.5 + .8x_t, .5) & \text{if } t > 85, \end{cases}$$

where $x_t \stackrel{\text{iid}}{\sim} N(0, 1)$ for simplicity. Select prior hyperparameters as follows: $\mu_\theta = \mu_\beta = 0_2$, $V_\theta = V_\beta = 100I_2$, $a_1 = b_1 = 3$, and $a_2 = b_2 = (1/2)$. Using the generated data and the algorithm from part (b), fit the model and compare point estimates with the parameter values used to generate the data.

Solution

- (a) The likelihood function for this model is

$$L(\theta_1, \theta_2, \beta_1, \beta_2, \sigma^2, \tau^2, \lambda) = \prod_{t \leq \lambda} \phi(y_t; \theta_1 + \theta_2 x_t, \sigma^2) \prod_{t > \lambda} \phi(y_t; \beta_1 + \beta_2 x_t, \tau^2).$$

The joint posterior is proportional to this likelihood times the given priors for the model parameters.

- (b) Conveniently, and much like a Gaussian mixture model (see Chapter 15 for exercises related to mixtures), the data divide into two parts given the changepoint λ . As such, standard results can be used for the linear regression model (see Exercise 11.8) to simulate the regression and variance parameters within each of the two regimes. Specifically, we obtain

$$\theta | \beta, \sigma^2, \tau^2, \lambda, y \sim N(D_\theta d_\theta, D_\theta),$$

where

$$D_\theta = (X'_\theta X_\theta / \sigma^2 + V_\theta^{-1})^{-1} \quad \text{and} \quad d_\theta = X'_\theta y_\theta / \sigma^2 + V_\theta^{-1} \mu_\theta.$$

In these expressions, we have defined

$$X_\theta = X_\theta(\lambda) = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_\lambda \end{bmatrix} \quad \text{and} \quad y_\theta = y_\theta(\lambda) = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_\lambda \end{bmatrix}.$$

Similarly, we obtain the posterior conditional for β :

$$\beta | \theta, \sigma^2, \tau^2, \lambda, y \sim N(D_\beta d_\beta, D_\beta),$$

where

$$D_\beta = (X'_\beta X_\beta / \tau^2 + V_\beta^{-1})^{-1} \quad \text{and} \quad d_\beta = X'_\beta y_\beta / \tau^2 + V_\beta^{-1} \mu_\beta.$$

Again, we define

$$X_\beta = X_\beta(\lambda) = \begin{bmatrix} 1 & x_{\lambda+1} \\ 1 & x_{\lambda+2} \\ \vdots & \vdots \\ 1 & x_T \end{bmatrix} \quad \text{and} \quad y_\beta = y_\beta(\lambda) = \begin{bmatrix} y_{\lambda+1} \\ y_{\lambda+2} \\ \vdots \\ y_T \end{bmatrix}.$$

For the conditionals for the variance parameters, we have

$$\sigma^2 | \beta, \theta, \tau^2, \lambda, y \sim IG \left(\frac{\lambda}{2} + a_1, \left[a_2^{-1} + (1/2) \sum_{t=1}^{\lambda} (y_t - \theta_1 - \theta_2 x_t)^2 \right]^{-1} \right)$$

and

$$\tau^2 | \beta, \theta, \sigma^2, \lambda, y \sim IG \left(\frac{T - \lambda}{2} + b_1, \left[b_2^{-1} + (1/2) \sum_{t=\lambda+1}^T (y_t - \beta_1 - \beta_2 x_t)^2 \right]^{-1} \right).$$

Finally, for the changepoint λ , we obtain

$$p(\lambda | \beta, \theta, \sigma^2, \tau^2, y) \propto \prod_{t \leq \lambda} \phi(y_t; \theta_1 + \theta_2 x_t, \sigma^2) \prod_{t > \lambda} \phi(y_t; \beta_1 + \beta_2 x_t, \tau^2), \quad \lambda = 1, 2, \dots, T-1.$$

This final distribution does not take a standard form. Since λ is discrete valued, however, we can calculate the (unnormalized) density ordinates for $\lambda = 1, 2, \dots, T-1$, and then make the density proper by dividing each ordinate by the sum of the ordinate values. Draws from this discrete distribution can be readily obtained.

(c) The Matlab code used is available on the Web site associated with this book. We run the Gibbs sampler for 1,000 iterations and discard the first 200 of these as the burn-in. In terms of the regression coefficients θ and β , respectively, we obtain posterior means (and their true values) equal to 1.94 (2), .977 (1), 1.51 (1.5), and .803 (.8). For the variance parameters σ^2 and τ^2 , respectively, we obtain posterior means (and true values) equal to .190 (.2) and .487 (.5). Finally, for the changepoint λ , we obtain a posterior mean of 82.87 and standard error equal to 9.09. Thus, all point estimates were quite close to the values used to generate the data.

Exercise 11.12 (Gibbs sampling in a count data model with an unknown changepoint)

Suppose that y_1, y_2, \dots, y_T are a time series of *count* responses, generated as follows:

$$y_t | \gamma, \delta, \lambda \sim \begin{cases} Po(\gamma) & \text{if } t \leq \lambda, \\ Po(\delta) & \text{if } t > \lambda, \end{cases}$$

where $Po(\mu)$ denotes the Poisson distribution with mean μ (see Appendix Definition 8). Like the previous exercise, λ is a *changepoint*; for periods up to and including λ , the counts are generated with mean γ , and for periods after λ , the counts are generated with mean equal to δ .

Suppose you employ priors of the form

$$\begin{aligned}\gamma &\sim G(a_1, a_2), \\ \delta &\sim G(d_1, d_2), \\ \lambda &\sim \text{uniform}\{1, 2, \dots, T - 1\}.\end{aligned}$$

- (a) Derive the likelihood function for this model.
 (b) Describe how the Gibbs sampler can be employed to estimate the parameters of this model.
 (c) On the Web site associated with this book, we provide data on the number of British coal mine disasters over the period 1851–1962. [See Carlin, Gelfand, and Smith (1992), and Gill (2002) for further analysis.] Based on a quick inspection of these data, it appears as if the number of coal mine accidents has decreased over time, perhaps owing to increased safety concerns or technological innovations. Consistent with this idea, use the coal mine data and fit the Poisson model with an unknown changepoint. When does the change occur? What are the expected number of disasters before and after this change?

Solution

- (a) The likelihood function for the model is as follows:

$$L(\gamma, \delta, \lambda) = \prod_{t=1851}^{\lambda} \frac{\gamma^{y_t} \exp(-\gamma)}{y_t!} \prod_{t=\lambda+1}^{1962} \frac{\delta^{y_t} \exp(-\delta)}{y_t!}.$$

Given the priors described, the joint posterior distribution can be expressed as

$$\begin{aligned}p(\gamma, \delta, \lambda | y) &\propto \left[\gamma^{a_1-1} \exp\left(-\frac{\gamma}{a_2}\right) \delta^{d_1-1} \exp\left(-\frac{\delta}{d_2}\right) \right] \\ &\times \prod_{t=1851}^{\lambda} \frac{\gamma^{y_t} \exp(-\gamma)}{y_t!} \prod_{t=\lambda+1}^{1962} \frac{\delta^{y_t} \exp(-\delta)}{y_t!}.\end{aligned}$$

- (b) Collecting all the terms in the posterior involving γ , we obtain

$$p(\gamma | \delta, \lambda, y) \propto \gamma^{a_1 + [\sum_{t=1851}^{\lambda} y_t] - 1} \exp[-\gamma(a_2^{-1} + n_{\lambda})],$$

where $n_{\lambda} \equiv \lambda - 1851 + 1$. We recognize this is the kernel of a

$$G\left(a_1 + \sum_{t=1851}^{\lambda} y_t, [a_2^{-1} + n_{\lambda}]^{-1}\right)$$

density. Similarly, we obtain

$$\delta | \gamma, \lambda, y \sim G\left(d_1 + \sum_{t=\lambda+1}^{1962} y_t, [d_2^{-1} + \tilde{n}_{\lambda}]^{-1}\right),$$

where $\tilde{n}_{\lambda} = 1962 - (\lambda + 1) + 1 = 1962 - \lambda$.

For the posterior conditional for the changepoint λ , we have

$$p(\lambda|\gamma, \delta, y) \propto \prod_{t=1851}^{\lambda} \frac{\gamma^{y_t} \exp(-\gamma)}{y_t!} \prod_{t=\lambda+1}^{1962} \frac{\delta^{y_t} \exp(-\delta)}{y_t!}, \quad \lambda = 1, 2, \dots, T-1.$$

This distribution is discrete and can be drawn from in the same way as described in part (b) of the previous exercise.

(c) Applying the algorithm described in part (b) to our coal mine data, we obtain a posterior mean (and standard deviation) of the changepoint λ Equal to 1890 (2.4). Looking quickly at the raw data, we see that this seems to be a reasonable division. The frequency of no disasters appears to increase significantly after 1890, whereas disasters consistently and frequently took place prior to 1890. Similarly, γ , the expected number of disasters prior to the change point, had a posterior mean of 3.07 and a posterior standard deviation equal to (.303). The expected number of disasters after the changepoint, δ , had a posterior mean of .924 and a standard deviation equal to (.117). These results appear consistent with the raw data and suggest a significant reduction in major coal mine accidents after 1890.

Exercise 11.13 (Error diagnostics with Gibbs sampling) Consider the simple regression model

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \quad \epsilon_i \stackrel{\text{iid}}{\sim} N(0, \sigma^2).$$

As an applied researcher, you are concerned that the assumption of normal errors may not be appropriate for your given application.

(a) Suggest some possible diagnostics that one could use (given simulated output from the posterior) to investigate whether the data are at odds with the normality assumption.

(b) Now, generate 2,000 observations from this simple regression model with $\beta_0 = 1$, $\beta_1 = .5$, and $x_i \stackrel{\text{iid}}{\sim} N(0, 1)$. Do this for three different cases. In the first case, generate the errors from a normal density with $\sigma^2 = .4$. In the second case, generate the errors from a Uniform density over the interval $[-5, 5]$. In the final case, generate the errors from a $\chi^2(5)$ density. (In this final case, you may want to subtract 5 from each simulated error so that the resulting errors are centered over zero.) Fit each model using the Gibbs sampler (see Exercise 11.8) and perform the diagnostics suggested in part (a) to investigate whether normality is appropriate for each data set.

Solution

(a) There are numerous ways one could examine whether the normality assumption is appropriate. An intuitive idea is to compare known features of the normal density to the posterior distribution of those features that are implied by the model. For example, if the errors are truly normal, then the error *skewness*

$$\frac{E(\epsilon^3)}{[E(\epsilon^2)]^{(3/2)}}$$

is equal to zero (i.e., the distribution is symmetric about zero), and the error *kurtosis*

$$\frac{E(\epsilon^4)}{[E(\epsilon^2)]^2}$$

equals three. In terms of our model, we might approximate the first of these quantities as

$$\text{skewness} \approx \frac{(1/n) \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i)^3}{(\sigma^2)^{(3/2)}},$$

and similarly,

$$\text{kurtosis} \approx \frac{(1/n) \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i)^4}{(\sigma^2)^{(2)}}.$$

Both of these last two expressions are functions of the model parameters (and data), and thus we can summarize the posterior distributions associated with these quantities. (For each draw from the posterior, we calculate skewness and kurtosis as shown here. We collect these draws for each iteration of the posterior simulator and then use this collection of values to calculate posterior means, posterior standard deviations, etc.) If we were to find that the posterior distribution of skewness, for example, was centered away from zero, or that the kurtosis posterior placed little mass near 3, we would find evidence against the normality assumption.

Other possibilities for diagnostic checking include the QQ plot [e.g., Lancaster (2004, 148)] or calculating posterior predictive p -values. The posterior predictive p -value essentially quantifies how reliably actual features of the data density y can be replicated by generating artificial data from the predictive density of the model, and then calculating those same features using simulated output from the posterior predictive. See, for example, Koop (2003, pp. 100–104) for more discussion of posterior predictive p -values.

(b) Data are generated as stated in part (b) of this exercise. For each model, we run the Gibbs sampler for 2,000 iterations, discarding the first 200 as the burn-in period. For each of the 1,800 postconvergence draws for each model under consideration, we calculate skewness and kurtosis as in the solution to part (a). We plot posterior densities of these quantities in Figure 11.2.

As you can see from the figure, when the errors are generated from a normal density, the skewness posterior is approximately centered over zero and the kurtosis posterior is concentrated over values near 3. Based on this evidence, we would not be able to conclude that normality was inappropriate, and thus we would likely proceed with the normal model. When the errors are generated from a uniform density, the skewness posterior is again centered near zero (since the errors remain symmetric), but the kurtosis value is clearly bounded away from 3. For the χ^2 case, we see that the skewness posterior is concentrated over positive values [which is sensible given that the χ^2 (5) has a pronounced right tail], with a kurtosis clearly larger than 3. These results suggest that normality is not an appropriate assumption for these data sets. In cases such as these, the researcher may wish to expand

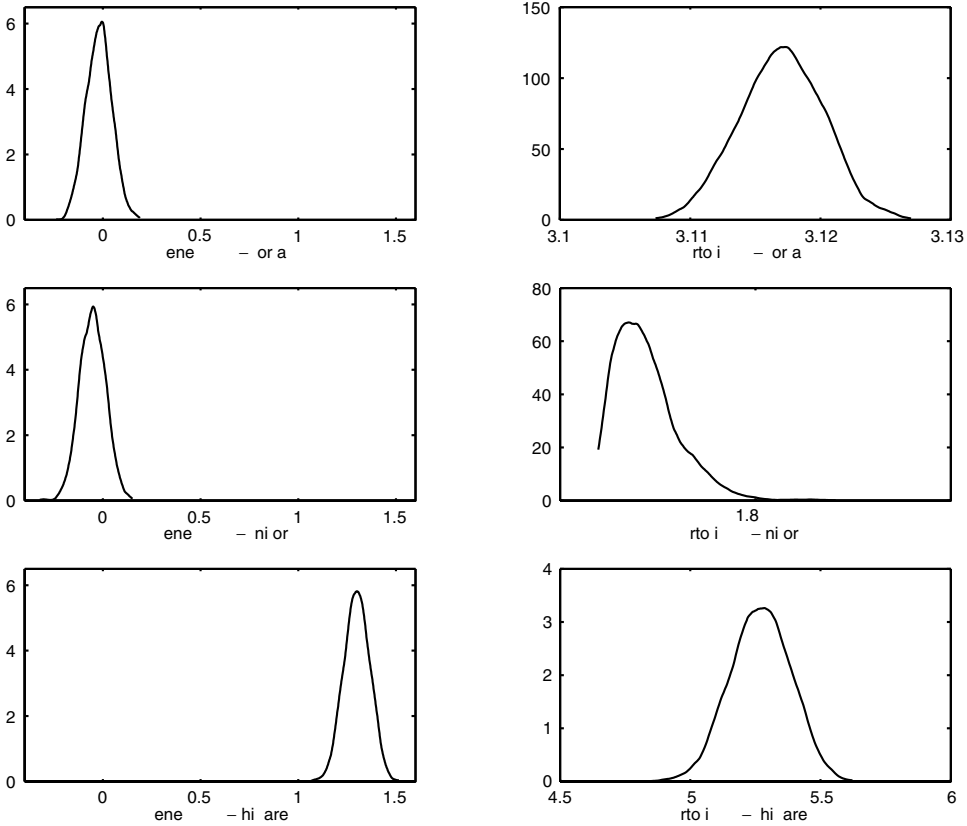


Figure 11.2 — Posterior densities of skewness and kurtosis for three different experimental designs.

the model to include more flexible classes of error distributions. We discuss alternatives for doing this in the exercises of Chapter 15.

Exercise 11.14 (Calculating numerical standard errors with Gibbs sampling) Consider a set of m *correlated* draws from the posterior of a scalar parameter θ , and denote these draws as $\{\theta_i\}_{i=1}^m$. Suppose that $\text{Var}(\theta_i) = \sigma^2 \ \forall i$ and $\text{Cov}(\theta_j, \theta_{j-k}) = \sigma_k \ \forall j, k$. In Gibbs sampling, we typically report a point estimate (posterior mean) of θ of the form

$$\bar{\theta} = \frac{1}{m} \sum_{i=1}^m \theta_i.$$

What is the *numerical standard error* associated with $\bar{\theta}$?

Solution

We note

$$\begin{aligned}
 \text{Var}(\bar{\theta}) &= \frac{1}{m^2} \text{Var}\left(\sum_i \theta_i\right) \\
 &= \frac{1}{m^2} \text{Var}(\theta_1 + [\theta_2 + \theta_3 + \cdots + \theta_m]) \\
 &= \frac{1}{m^2} [\text{Var}(\theta_1) + 2\text{Cov}(\theta_1, \theta_2 + \cdots + \theta_m) + \text{Var}(\theta_2 + \cdots + \theta_m)] \\
 &= \frac{1}{m^2} [\sigma^2 + 2(\sigma_1 + \sigma_2 + \cdots + \sigma_{m-1}) + \text{Var}(\theta_2 + \cdots + \theta_m)].
 \end{aligned}$$

Applying the same type of argument to the sum $\theta_2 + \cdots + \theta_m$ in the last part of this final equality, we obtain

$$\text{Var}(\theta_2 + \cdots + \theta_m) = \sigma^2 + 2(\sigma_1 + \sigma_2 + \cdots + \sigma_{m-2}) + \text{Var}(\theta_3 + \cdots + \theta_m).$$

Continuing in this fashion, and grouping together the resulting terms, we obtain

$$\text{Var}(\bar{\theta}) = \frac{1}{m^2} (m\sigma^2 + 2[(m-1)\sigma_1 + (m-2)\sigma_2 + \cdots + \sigma_{m-1}]).$$

Factoring out an $m\sigma^2$ from this expression, we obtain

$$\begin{aligned}
 \text{Var}(\bar{\theta}) &= \frac{1}{m^2} m\sigma^2 \left(1 + 2 \left[\frac{(m-1)\sigma_1}{m\sigma^2} + \frac{(m-2)\sigma_2}{m\sigma^2} + \cdots + \frac{\sigma_{m-1}}{m\sigma^2} \right] \right) \\
 &= \frac{\sigma^2}{m} \left[1 + 2 \sum_{j=1}^{m-1} \left(\frac{m-j}{m} \right) \frac{\sigma_j}{\sigma^2} \right] \\
 &= \frac{\sigma^2}{m} \left[1 + 2 \sum_{j=1}^{m-1} \left(1 - \frac{j}{m} \right) \rho_j \right],
 \end{aligned}$$

where $\rho_j \equiv \sigma_j/\sigma^2$ is a measure of the correlation (i.e., the *lag correlation*) between θ_i and θ_{i-j} .

This expression provides a way for researchers to calculate *numerical standard errors*, or NSEs, associated with a Gibbs sampling estimate of a posterior moment of interest. Given the derivation here, we have

$$\text{NSE}(\bar{\theta}) = \sqrt{\frac{\sigma^2}{m} \left[1 + 2 \sum_{j=1}^{m-1} \left(1 - \frac{j}{m} \right) \rho_j \right]}.$$

Note that ρ_j can be calculated from the draws themselves by simply computing the correlation coefficient between the vectors $[\theta_1 \ \theta_2 \ \cdots \ \theta_{m-j}]$ and $[\theta_{j+1} \ \theta_{j+2} \ \cdots \ \theta_m]$. In practice, the summation in the NSE calculation is usually “cut off” when the lag correlations fall below a certain threshold, such as .05 or .01. Also note that if the $\rho_j > 0$, as is typically the case in Gibbs sampling, $\text{Var}(\bar{\theta}) > \sigma^2/m$, the variance obtained under iid sampling.

Finally, it is worth mentioning that, unlike frequentist standard errors, the numerical standard errors here can be made arbitrarily small by choosing m sufficiently large (i.e., taking enough draws from the posterior). The degree of precision with which the (finite-sample) posterior mean is estimated is not a function of the size of the data set, but rather, it is a function of the number of simulations taken from the posterior.

Exercise 11.15 (Gibbs sampling in a regression model with inequality constraints: Geweke [1996b]) Consider the regression model

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \epsilon_i, \quad i = 1, 2, \dots, n,$$

where $\epsilon_i \stackrel{\text{iid}}{\sim} N(0, \sigma^2)$.

Let $\beta = [\beta_0 \ \beta_1 \ \cdots \ \beta_k]'$. Suppose that the regression coefficients are known to satisfy constraints of the form

$$a < H\beta < b,$$

where a and b are known $(k+1) \times 1$ vectors of lower and upper limits, respectively, and H is a nonsingular matrix that selects elements of β to incorporate the known constraints. The vectors a and b may also contain elements equal to $-\infty$ or ∞ , respectively, if a particular linear combination of β is not to be bounded from above or below (or both). Thus, in this formulation of the model, we are capable of imposing up to $k+1$ inequality restrictions, but no more.

(a) Reparameterize the model in terms of $\gamma = H\beta$, introduce a prior for γ , and provide a posterior simulator for the reparameterized model.

(b) Consider, as a special case of the regression model in part (a), the Cobb–Douglas production function where

$$\ln y_i = \beta_0 + \beta_1 \ln L_i + \beta_2 \ln K_i + \epsilon_i,$$

with y_i denoting output, L_i denoting labor, and K_i denoting capital. How would you impose that the production function satisfies diminishing returns to scale?

(c) Using the output, capital, and labor data presented on the Web site associated with this book (with 39 annual observations for the cigarette industry from 1958 to 1996), fit the Cobb–Douglas production function without imposing any inequality constraints. To do this, employ a diffuse prior on σ^2 of the form $p(\sigma^2) \propto \sigma^{-2}$ and a normal prior for β of the form $\beta \sim N(0, 10I_3)$. Do the data provide evidence of diminishing returns to scale?

(d) Now, reparameterize the model, as in (b), to restrict the production function to satisfy diminishing returns to scale. Fit the model using the algorithm in (a) under $p(\sigma^2) \propto \sigma^{-2}$. For the regression coefficients, employ independent, truncated $N(0, 10I_3)$ priors, where the regions of truncation for each element of γ are defined as in (b). How do your results compare to those of (c)?

Solution

(a) First, write the regression model as

$$y = X\beta + \epsilon,$$

where

$$y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \quad X = \begin{bmatrix} 1 & x_{11} & \cdots & x_{1k} \\ 1 & x_{21} & \cdots & x_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{nk} \end{bmatrix}, \quad \text{and} \quad \epsilon = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}.$$

Since $\gamma = H\beta$ and H is nonsingular, write $\beta = H^{-1}\gamma$. The reparameterized regression model thus becomes

$$\begin{aligned} y &= XH^{-1}\gamma + \epsilon_i \\ &= \tilde{X}\gamma + \epsilon_i, \end{aligned}$$

with

$$\tilde{X} \equiv XH^{-1} \quad \text{and} \quad a < \gamma < b.$$

In the γ parameterization, note that independent priors can be employed that satisfy the stated inequality restrictions. To this end, we specify independent truncated normal priors for elements of γ of the form

$$p(\gamma_j) \propto \phi(\gamma_j; 0, V_j) I(a_j < \gamma_j < b_j), \quad j = 0, 1, \dots, k,$$

with $I(\cdot)$ denoting the standard indicator function. The posterior distribution is proportional to the product of the prior and likelihood:

$$p(\gamma, \sigma^2 | y) \propto \sigma^{-2} \left[\prod_{j=0}^k p(\gamma_j) \right] \phi(y; \tilde{X}\gamma, \sigma^2 I_n).$$

By using the results of Exercise 11.8 (and appropriately accounting for the region of truncation), it follows that the posterior conditional for $\gamma | \sigma^2, y$ takes the following form:

$$p(\gamma | \sigma^2, y) \propto \phi(\gamma; \bar{\gamma}, \bar{H}) I(a < \gamma < b), \quad (11.1)$$

where

$$\bar{H} = (\tilde{X}'\tilde{X}/\sigma_\epsilon^2 + V^{-1})^{-1}, \quad \bar{\gamma} = \bar{H}(\tilde{X}'y/\sigma_\epsilon^2),$$

and V is the $(k+1) \times (k+1)$ diagonal matrix with j th diagonal element equal to V_j .

The density in (11.1) is a multivariate normal density truncated to the regions defined by a and b . Drawing directly from this multivariate truncated normal is nontrivial in general, but as Geweke [1991, 1996(b)] points out, the posterior conditional distributions $\gamma_j | \gamma_{-j}, \sigma^2, y$ (with γ_{-j} denoting all elements of γ other than γ_j) are univariate truncated normal. Thus, one can change the algorithm of Exercise 11.8 to sample *individually* from each coefficient's conditional posterior (rather than sampling γ in a single step), where each draw is obtained from a univariate truncated normal distribution. Fortunately, such draws can be easily obtained. (See Exercise 11.20 for an efficient procedure for sampling from the univariate truncated normal.)

Let $\Omega = [\bar{H}]^{-1} = \tilde{X}'\tilde{X}/\sigma_\epsilon^2 + V^{-1}$, ω_{ij} denote the (i, j) element of Ω , and $\bar{\gamma}_j$ denote the j th element of $\bar{\gamma}$. Using properties of the conditional normal distribution (see Appendix Theorem 3) and the formula for a partitioned inverse [see Abadir and Magnus (2006, p. 106)], we obtain

$$\gamma_j | \gamma_{-j}, \sigma^2, y \stackrel{\text{ind}}{\sim} TN_{(a_j, b_j)} \left(\bar{\alpha}_j - \omega_{jj}^{-1} \sum_{i \neq j} \omega_{ji}(\gamma_i - \bar{\gamma}_i), \omega_{jj}^{-1} \right), \quad j = 0, 1, \dots, k, \quad (11.2)$$

where $TN_{(a, b)}(\mu, \sigma^2)$ denotes a normal density with mean μ and variance σ^2 truncated to the interval (a, b) .

The derivation of the posterior conditional for the variance parameter utilizes steps similar to those in Exercise 11.8, though, in this case, an improper prior for σ^2 has been employed. It follows that

$$\sigma^2 | \gamma, y \sim IG \left(\frac{n}{2}, \left[\frac{1}{2} (y - \tilde{X}\gamma)'(y - \tilde{X}\gamma) \right]^{-1} \right). \quad (11.3)$$

A Gibbs sampling algorithm for the regression model with linear inequality constraints thus follows by independently sampling from (11.2) for each j and then sampling from (11.3).

(b) It is well known that $\beta_1 + \beta_2$ determines returns to scale. In particular, our production function is characterized by diminishing returns to scale provided $\beta_1 + \beta_2 < 1$.

To impose diminishing returns to scale, yet leave other parameters unrestricted, we can therefore specify

$$\begin{bmatrix} -\infty \\ -\infty \\ -\infty \end{bmatrix} < \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 1 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix} < \begin{bmatrix} \infty \\ \infty \\ 1 \end{bmatrix},$$

and thus a , b , and H are defined accordingly. With these definitions, H^{-1} is simply

$$H^{-1} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & -1 & 1 \end{bmatrix},$$

and \tilde{X} , the transformed X matrix, has columns equal to a constant, $\ln[L/K]$ and $\ln K$. Similarly,

$$\gamma = \begin{bmatrix} \gamma_0 \\ \gamma_1 \\ \gamma_2 \end{bmatrix} = H\beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_1 + \beta_2 \end{bmatrix},$$

and thus γ_2 represents the return to scale in the reparameterized model.

(c and d) On the Web site associated with this book, we provide Matlab code for fitting the production function model without inequality constraints, as in part (c), and also when imposing that the production function exhibits diminishing returns to scale. In both cases, we make use of the Gibbs sampler to fit the model. When no constraints are imposed, we run

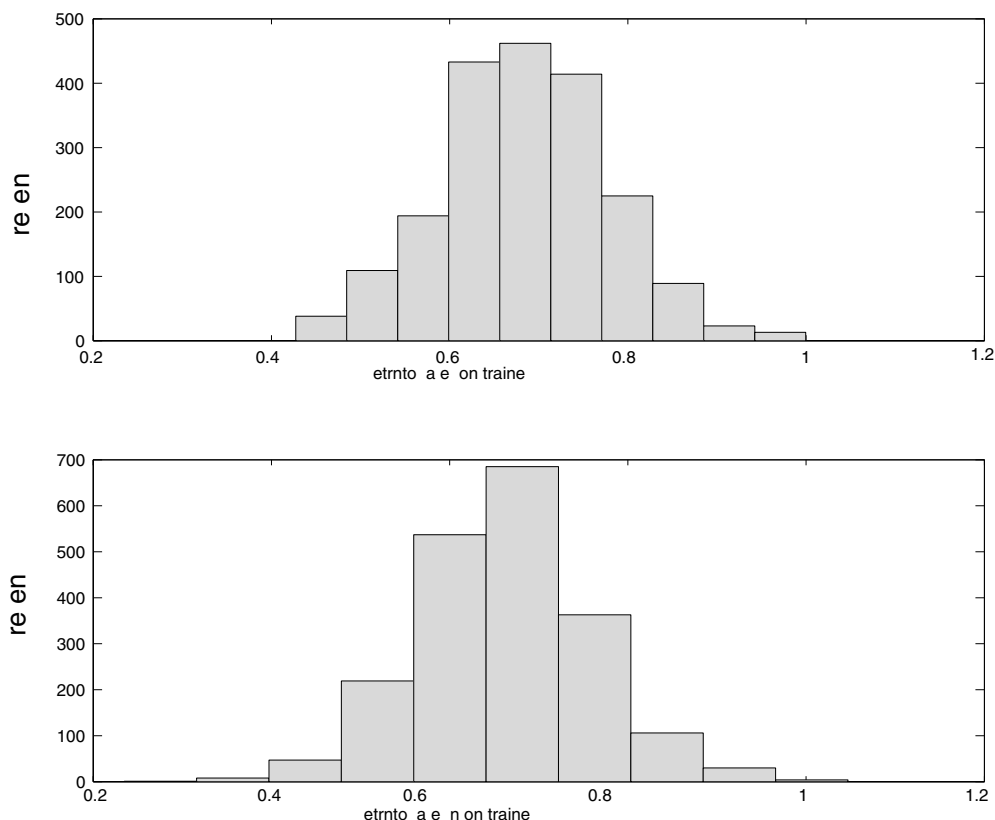


Figure 11.3 — Posterior histograms for return-to-scale parameter in constrained (upper) and unconstrained (lower) models.

the simulator for 3,000 iterations, discarding the first 1,000 as the burn-in. When constraints are imposed [as in part (d)], we run the sampler for 105,000 iterations and discard the first 5,000 as the burn-in. In the latter case, our posterior simulations suffer from high degrees of autocorrelation. We retain every 50th draw from this algorithm and thus obtain a final posterior sample of 2,000 simulations. [Note in this case that γ_0 and γ_1 could be drawn jointly in our posterior simulator to gain efficiency. Our code for this exercise, however, follows the general algorithm in (11.2).] For the constrained simulator, OLS estimates are used as starting values.

Presented in Figure 11.3 are posterior frequencies associated with the return-to-scale parameter in the unconstrained and constrained models. For the model in (c), which does not require a reparameterization, the figure plots a histogram of $\beta_1 + \beta_2$, and for the model in (d), an analogous histogram for γ_2 is provided. In the unconstrained case, which is presented in the lower half of the figure, the return-to-scale posterior places only very small mass on values exceeding one. Thus, our data lend support to the hypothesis of diminishing returns to scale. In the top portion of the figure, which formally imposes diminishing

returns to scale through our prior, no mass is placed to the right of one. The fact that both figures possess similar shapes should not be regarded as surprising, since the data appear to support the hypothesis of diminishing returns to scale.

Exercise 11.16 (Checking for errors in a posterior simulator: Geweke [2004]) Even the most careful researcher can make mistakes when programming a posterior simulator. A reasonably comprehensive way to avoid (or at least significantly limit) such problems is to apply the posterior simulator to a large artificial data set that is generated from the model under consideration. In such a case, the parameters of the data-generating process are known and point estimates (e.g., posterior means) can be compared to the “true” parameter values used to generate the data. If the point estimates are far from the “true” parameters, even under reasonably diffuse priors and with a large data set, this suggests that a programming error may have occurred (or, potentially, this reveals a type of identification problem; see, for example, Exercise 14.2). Conversely, when agreement is reached between estimates and actual values under a variety of different experimental designs (i.e., upon changing the “true” parameters’ values), this provides support, though not proof, that the code is error free.

Geweke (2004) provides a more formal procedure for testing a posterior simulator, and the intent of this exercise is to review and apply this technique. We stress that the method that follows can be used to check the accuracy of *any* posterior simulator and is not limited only to those models that make use of Gibbs sampling. We have chosen to include this exercise in the Gibbs sampling section simply because we employ the Gibbs sampler in our generated-data experiment of part (b).

Let $p(y, \theta)$ be a joint density of observables y and unobservables θ , and consider two alternate strategies for simulating from this joint density.

The first approach, termed the *marginal-conditional simulator*, follows from the method of composition. That is, obtain draws $\theta^{(m)}$ and $y^{(m)}$ from the joint distribution $p(y, \theta)$ by drawing

$$\theta^{(m)} \sim p(\theta), \quad (11.4)$$

$$y^{(m)} \sim p(y|\theta = \theta^{(m)}), \quad (11.5)$$

where $p(\theta)$ denotes the prior and $p(y|\theta)$ the data density.

As an alternative, consider the *successive-conditional simulator*, defined as follows

$$\tilde{y}^{(m)} \sim p(y|\theta = \tilde{\theta}^{(m-1)}), \quad (11.6)$$

$$\tilde{\theta}^{(m)} \sim q(\theta|\tilde{\theta}^{(m-1)}, \tilde{y}^{(m)}). \quad (11.7)$$

Here, q generically represents the *transition kernel* of a posterior simulator. For example, if a Gibbs sampler is employed, then q represents the conditional distributions corresponding to each element of the (potentially blocked) parameter vector. Thus, in the successive-conditional simulator, the spirit is to apply the Gibbs sampler to aid in sampling from $p(y, \theta)$ – first drawing from $y|\theta$, and then drawing from $\theta|y$ (as a posterior simulator is

designed to do). Importantly, the successive-conditional simulator requires a starting value $\tilde{\theta}^{(0)}$, which is obtained from the prior [i.e., $\tilde{\theta}^{(0)} \sim p(\theta)$].

(a) Show how these two distinct simulation approaches can be used to check for errors in a posterior simulator.

(b) Consider the following regression model:

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \quad x_i \stackrel{\text{iid}}{\sim} N(0, 1), \quad \epsilon_i \stackrel{\text{iid}}{\sim} N(0, \sigma^2), \quad i = 1, 2, \dots, 10.$$

Suppose priors of the form $\sigma^2 \sim IG(a, b)$ and $\beta \sim N(\mu_\beta, V_\beta)$ are employed. Following Exercise 11.8, we have the posterior conditional distributions associated with this regression model given as

$$\beta | \sigma^2, y \sim N(D_\beta d_\beta, D_\beta), \quad (11.8)$$

where

$$D_\beta = (X'X/\sigma^2 + V_\beta^{-1})^{-1}, \quad d_\beta = X'y/\sigma^2 + V_\beta^{-1}\mu_\beta,$$

and

$$\sigma^2 | \beta, y \sim IG([10/2] + a, [b^{-1} + .5(y - X\beta)'(y - X\beta)]^{-1}). \quad (11.9)$$

Using the hyperparameters $a = 3$, $b = 1/2$, $\mu_\beta = [1 \ 1]'$, and $V_\beta = .25I_2$, generate 100,000 draws from $p(y, \theta)$ using both the marginal-conditional and successive-conditional simulators. For the latter, use the Gibbs sampler, as described by (11.8) and (11.9), to generate draws from (11.7). Given your solution in (a), provide a test for the correctness of the Gibbs sampler that was used for fitting the regression model.

(c) Now, deliberately include some mistakes into your Gibbs sampling algorithm. Specifically, (1) Replace V_β^{-1} with V_β in the expression for D_β in the posterior conditional for β , (2) assume that the prior mean for β is $[0 \ 0]'$, even though a prior mean of $[1 \ 1]'$ was used, and (3) replace $b^{-1} = (1/2)^{-1}$ with $b = (1/2)$ in the conditional posterior for σ^2 in (11.9). Do the tests derived in part (b) diagnose these mistakes?

Solution

(a) First, note that the marginal-conditional simulator only requires drawing from the prior and data densities. Typically, this is easy to do and easy to program.

Let $g^{(m)} = g(\theta^{(m)}, y^{(m)})$ be the m th value of some function g mapping $\Theta \times Y$ into the real line, which is evaluated using output from the marginal-conditional simulator. Define $\tilde{g}^{(m)} = g(\tilde{\theta}^{(m)}, \tilde{y}^{(m)})$ analogously, noting that \tilde{g} uses output from the successive-conditional simulator.

We can think of g as being a *test function* in the following sense: If both simulators are programmed correctly, both will provide draws from the joint distribution $p(y, \theta)$, and thus moments and other functions of interest should be similar when evaluated using the outputs of both simulators. This suggests that, for a given g , averages of the form $\bar{g} = [1/M_1] \sum_{m=1}^{M_1} g^{(m)}$ and $\tilde{\bar{g}} = [1/M_2] \sum_{m=1}^{M_2} \tilde{g}^{(m)}$ (where M_1 and M_2 denote the number of simulations in the marginal-conditional and successive-conditional simulators) should

be similar. To the extent that \bar{g} and \tilde{g} differ significantly, this provides evidence that a mistake has occurred in the programming of *at least one* of the simulators. In general, the error is more likely to take place in the successive-conditional simulator since the posterior simulator is most difficult to program and the marginal-conditional simulator is comparably easy to implement.

Different choices of g define different test functions. For concreteness here, we can think about comparing moments of θ across our two simulators, and thus choose, for example, $g(\theta, y) = \theta_j$ (for some element θ_j of θ) or $g(\theta, y) = \theta_j^2$. In practice, a variety of moments could be used.

Geweke shows that, provided no programming errors are present in both simulators,

$$\frac{\bar{g} - \tilde{g}}{(M_1^{-1}\hat{\sigma}_g^2 + M_2^{-1}\hat{\tau}_g^2)^{1/2}} \rightarrow N(0, 1), \quad (11.10)$$

where $\hat{\sigma}_g^2 = M_1^{-1} \sum_{m=1}^{M_1} [g^{(m)}]^2 - \bar{g}^2$ and $M_1^{-1}\hat{\sigma}_g^2$ is the variance of \bar{g} . Similarly, $M_2^{-1}\hat{\tau}_g^2$ is the variance of \tilde{g} . For the case of the successive-conditional simulator, the draws are not independent, and thus $M_2^{-1}\hat{\tau}_g^2$ must be calculated using the formula provided in Exercise 11.14.

(b) Code for the marginal-conditional and successive-conditional simulators is available on the Web site associated with this book. For our test functions, we individually compare the means of β_0 , β_1 , and σ^2 across both simulators. When doing so, we calculate test statistics according to (11.10) equal to 1.03, .25, and .75 for the respective means. Thus, our tests do not offer conclusive evidence that errors are present in our code.

(c) When replacing V_β^{-1} with V_β in the definition of D_β , we obtain test statistics equal to -260.3 , -292.5 , and $-.6$ for equality of means associated with β_0 , β_1 , and σ^2 , respectively. When assuming that $\mu_\beta = [0 \ 0]'$, we obtain test statistics equal to 224.67, 206.49, and $-.158$, respectively. Finally, when replacing b^{-1} with b in the posterior conditional for σ^2 , we obtain test statistics equal to $-.30$, $-.275$, and 181, respectively. Thus, when programming errors are present, the tests clearly reveal evidence of an error. Moreover, the pattern of rejection also seems to suggest the posterior conditional where the error was made.

Exercise 11.17 (Metropolis–Hastings algorithm, 1: Gelman, Carlin, Stern, and Louis [1995]) Analogous to the results for the Gibbs sampler derived in Exercise 11.6, this exercise provides some suggestive evidence regarding why the Metropolis–Hastings (M–H) algorithm works. Specifically, we show that, for the case of a discrete valued random variable, the target distribution is an invariant distribution.

Suppose that θ is a discrete-valued random variable with probability mass function $p(\theta)$, and let θ^{t-1} be the current value of the M–H chain, where it is assumed that $\theta^{t-1} \sim p$. Show that $\theta^t \sim p$, when θ^t is obtained according to the M–H algorithm.

Solution

Without loss of generality, let us choose two distinct points in the support of θ , and call them θ^a and θ^b . As discussed at the outset of this chapter, we will define the M–H “acceptance ratio” as

$$r_t(\theta^a|\theta^b) = \frac{p(\theta^a)}{q(\theta^a|\theta^b)} \frac{q(\theta^b|\theta^a)}{p(\theta^b)}.$$

This ratio denotes the probability of accepting the candidate draw θ^a given that the chain is currently at θ^b ; we note, of course, that the draw is automatically accepted if this ratio exceeds one. In this expression $q(\cdot|\cdot)$ denotes the *candidate-generating density* or *proposal density*. In a random walk chain, for example, candidate values might be generated from, say, an $N(\theta^{t-1}, \Sigma)$ distribution, and thus $q(\cdot|\cdot)$ would denote the normal density evaluated at the candidate point θ^a with mean equal to the current value of the chain, $\theta^{t-1} = \theta^b$.

Without loss of generality, let us label these points so that $r_t(\theta^a|\theta^b) > 1$, implying that $r_t(\theta^b|\theta^a) < 1$, since these terms are reciprocals. [The case where $r_t(\theta^a|\theta^b) = 1$ follows similarly].

Now consider

$$\Pr(\theta^{t-1} = \theta^b, \theta^t = \theta^a) = \Pr(\theta^t = \theta^a | \theta^{t-1} = \theta^b) \Pr(\theta^{t-1} = \theta^b).$$

The first term on the right-hand side of the equation represents the probability that the next state of the chain is θ^a given that the chain is currently at θ^b .

To arrive at θ^a at the next iteration, two things must happen: (1) the candidate must be the value drawn from the proposal density, and (2) given that it is drawn, it must be accepted. Our labeling of the ratio r_t implies that θ^a will always be accepted once it is drawn. As a result,

$$\begin{aligned} \Pr(\theta^{t-1} = \theta^b, \theta^t = \theta^a) &= q(\theta^a|\theta^b) \Pr(\theta^{t-1} = \theta^b) \\ &= q(\theta^a|\theta^b) p(\theta^b), \end{aligned}$$

where the last line follows from our assumption that $\theta^{t-1} \sim p$. Now, consider another joint probability

$$\Pr(\theta^{t-1} = \theta^a, \theta^t = \theta^b) = \Pr(\theta^t = \theta^b | \theta^{t-1} = \theta^a) \Pr(\theta^{t-1} = \theta^a).$$

For θ^b to be the next value of the chain, it must be both drawn from the proposal density and accepted once it has been drawn. Thus,

$$\begin{aligned} \Pr(\theta^{t-1} = \theta^a, \theta^t = \theta^b) &= q(\theta^b|\theta^a) r_t(\theta^b|\theta^a) \Pr(\theta^{t-1} = \theta^a) \\ &= q(\theta^b|\theta^a) \left(\frac{p(\theta^b) q(\theta^a|\theta^b)}{q(\theta^b|\theta^a) p(\theta^a)} \right) p(\theta^a) \\ &= q(\theta^a|\theta^b) p(\theta^b) \\ &= \Pr(\theta^{t-1} = \theta^b, \theta^t = \theta^a). \end{aligned}$$

Note that the equality of the joint probabilities implies

$$\Pr(\theta^t = \theta^a | \theta^{t-1} = \theta^b) \Pr(\theta^{t-1} = \theta^b) = \Pr(\theta^{t-1} = \theta^a | \theta^t = \theta^b) \Pr(\theta^t = \theta^b).$$

Summing this equation over θ^a implies that

$$\Pr(\theta^{t-1} = \theta^b) = \Pr(\theta^t = \theta^b),$$

so that the marginals are indeed the same. Since $\theta^{t-1} \sim p$, it follows that $\theta^t \sim p$ as well.

Exercise 11.18 (Metropolis–Hastings algorithm, 2) The purpose of this question is to investigate the properties of two types of M–H algorithms.

Assume you wish to carry out Bayesian inference on a parameter θ with posterior given by

$$p(\theta|y) \propto \exp\left(-\frac{1}{2}|\theta|\right),$$

for $-\infty < \theta < \infty$. This is a special case of the Laplace distribution. For the purposes of this question you do not have to know the integrating constant. Readers who want to know what it is (or want to know more about the Laplace distribution) are referred to Poirier (1995, p. 110). This Laplace distribution has mean 0 and variance 8.

Suppose you cannot find a way of directly drawing from $p(\theta|y)$ and hence wish to use an M–H algorithm.

- Derive an independence chain M–H algorithm using an $N(0, d^2)$ candidate-generating density and write a program that implements this algorithm.
- Derive a random walk chain M–H algorithm using a normal increment and write a program that implements this algorithm.
- Compare the performances of the algorithms developed in parts (a) and (b).

Solution

The introduction to this chapter gives some references for the reader unfamiliar with the M–H algorithm.

(a) All M–H algorithms provide a sequence or “chain” of draws that are then averaged in the same way as Monte Carlo integration draws are to provide estimates of posterior functions of interest (e.g., posterior means). Denote this sequence of draws as $\theta^{(r)}$ for $r = 1, \dots, R$. M–H algorithms proceed by taking candidate draws from a convenient density. Denote a candidate for the r th draw in the sequence by θ^* . The candidate draw is either accepted (i.e., $\theta^{(r)} = \theta^*$) or rejected, in which case the chain remains where it is (i.e., $\theta^{(r)} = \theta^{(r-1)}$). The acceptance probability is chosen to ensure the chain can be treated as draws from the posterior.

The independence chain algorithm chooses a single candidate generating density and simply takes random draws from it. In this case, the acceptance probability has the form [see, e.g., Koop (2003, p. 95)]

$$\alpha(\theta^{(r-1)}, \theta^*) = \min \left[\frac{p(\theta = \theta^*|y) q(\theta = \theta^{(r-1)})}{p(\theta = \theta^{(r-1)}|y) q(\theta = \theta^*)}, 1 \right],$$

where $q(\cdot)$ is the candidate-generating density. If we substitute in the posterior given in the question and the suggested $N(0, d^2)$ candidate-generating density, we obtain

$$\alpha(\theta^{(r-1)}, \theta^*) = \min \left[\exp \left[\frac{1}{2} \left(\left| \theta^{(r-1)} \right| - |\theta^*| - \left\{ \frac{\theta^{(r-1)}}{d} \right\}^2 + \left\{ \frac{\theta^*}{d} \right\}^2 \right) \right], 1 \right]. \quad (11.11)$$

The parameter d should be chosen to optimize the performance of the M–H algorithm. Unfortunately, there is no hard and fast rule as to how this should be done. You may think you want to choose d so as to make the acceptance probability as high as possible. After all, if the candidate-generating density is exactly the same as the posterior, then the acceptance probability is one and the independence chain M–H algorithm is equivalent to Monte Carlo integration. However, a high acceptance probability can also be achieved if the candidate-generating density has too little dispersion. For instance, an extreme case would set d very close to zero. In this case, the acceptance probability is always near one. But this means all draws are virtually identical to one another and nonsensical results will be produced. In practice, it is always sensible using the convergence diagnostics (e.g., the ones developed for the Gibbs sampler also work for the M–H algorithm) to assess computational performance.

A computer program that carries out Bayesian inference using this independence chain M–H algorithm has the following form:

- Step 1: Do all the preliminary things to create all the variables used in the M–H algorithm (i.e., initialize all the M–H sums at zero and specify an initial value, $\theta^{(0)}$).
- Step 2: Draw θ^* from the $N(0, d^2)$ candidate-generating density.
- Step 3: Calculate the acceptance probability $\alpha(\theta^{(r-1)}, \theta^*)$ given in (11.11).
- Step 4: Set $\theta^{(r)} = \theta^*$ with probability $\alpha(\theta^{(r-1)}, \theta^*)$, or else set $\theta^{(r)} = \theta^{(r-1)}$.
- Step 4: Repeat Steps 2, 3, and 4 for $r = 1, \dots, R$.
- Step 5: Average the draws appropriately to produce M–H estimates of whatever posterior functions of interest (e.g., posterior means) are required.

Note that M–H algorithms, like Gibbs sampling algorithms, depend on an initial condition, $\theta^{(0)}$, which must be chosen by the researcher. It is common to discard an initial set of burn-in draws to remove the effects of this choice.

The Matlab program that implements this algorithm is provided on the Web site associated with this book.

(b) The solution to part (b) is the same as part (a), except that the candidate-generating density is different and, as a result, the acceptance probability is different.

The random walk chain M–H algorithm generates candidate draws according to

$$\theta^* = \theta^{(r-1)} + z,$$

where z is called the *increment random variable*. In this question, a normal increment is suggested and, thus, θ^* is drawn from the $N(\theta^{(r-1)}, c^2)$, where c is calibrated to ensure

a reasonable acceptance rate. In contrast to the independence chain M–H algorithm, a rule of thumb is available to provide guidance in choosing c . In particular, a common practice in scalar or very low dimension problems is to choose c so that about 50% of the candidate draws are accepted. As with any posterior simulation algorithm, though, it is always sensible to use appropriate convergence diagnostics.

For the random walk chain M–H algorithm, the acceptance probability is [see, e.g., Koop (2003, p. 97)]

$$\alpha\left(\theta^{(r-1)}, \theta^*\right) = \min \left[\frac{p\left(\theta = \theta^* | y\right)}{p\left(\theta = \theta^{(r-1)} | y\right)}, 1 \right].$$

Plugging in the form given for the posterior, we obtain

$$\alpha\left(\theta^{(r-1)}, \theta^*\right) = \min \left[\exp \left[\frac{1}{2} \left\{ \left| \theta^{(r-1)} \right| - \left| \theta^* \right| \right\} \right], 1 \right]. \quad (11.12)$$

The sketch of the computer program for the random walk chain M–H algorithm is the same as the one given in the solution to part (a), except that the candidate-generating density and acceptance probability are replaced as described here. A Matlab program that has both independence and random walk chain M–H algorithms is provided on the Web site for this book.

(c) Table 11.8 gives the estimated posterior mean and variance of θ for various values of d (for the independence chain M–H algorithm) and c (for the random walk chain M–H algorithm) as well as the proportion of accepted draws. In all cases, we include 10,000 replications (after discarding an initial 100 burn-in replications).

For the independence chain M–H algorithm, we can see that candidate generating densities that are either too dispersed (e.g., $d = 100$) or not dispersed enough ($d = 1.0$) yield poor results. However, the problems associated with being not dispersed enough are worse since, if d is chosen to be too small, then some regions of the parameter space for θ may never be visited with a finite number of draws. Because we know the true variance of θ to be 8, it is likely that setting the variance of the candidate-generating density to be about 8 is sensible. This does seem to be the case.

For the random walk chain M–H algorithm we get a similar pattern of results. If the increment variable has too small a standard deviation, then the chain does not traverse the entire distribution (for finite R) and inaccurate estimates (especially of posterior variances) occur. In Table 11.8, this is revealed in the row for $c = 0.10$, which reveals serious underestimation of the posterior variance. If the standard deviation of the increment variable is too large, then poor results also occur. This is because too many extreme draws, in improbable regions of the posterior, are taken and these are not accepted (e.g., when $c = 100$, then only 3% of the candidate draws are accepted). This results in an inefficient computational algorithm.

Note that the commonly used rule of thumb, which says you should choose c to yield approximately a 50% acceptance rate, does seem to be working well in this case.

Table 11.8: Posterior results for θ for various M–H algorithms.

	Posterior Mean	Posterior Variance	Acceptance Probability
True Value	0.00	8.00	—
Independence Chain M–H Algorithm			
$d = 1.00$	0.16	3.48	0.65
$d = 2.00$	0.04	7.15	0.84
$d = 3.00$	0.01	7.68	0.79
$d = 6.00$	−0.05	8.03	0.49
$d = 20.00$	−0.07	7.87	0.16
$d = 100.00$	0.12	7.41	0.03
Random Walk Chain M–H Algorithm			
$c = 0.10$	0.33	1.13	0.98
$c = 0.50$	0.20	6.87	0.91
$c = 1.00$	−0.08	7.01	0.83
$c = 4.00$	−0.03	7.93	0.52
$c = 10.00$	0.05	7.62	0.28
$c = 100.00$	0.23	6.25	0.03

11.4 Other (noniterative) methods for generating random variates

In the examples considered by Exercises 11.7–11.12 and 11.15, the posterior simulators only involved sampling from well-known distributions. Routines for drawing from such distributions are typically found in most, if not all, statistical software packages, as Exercise 11.2 suggests. In some cases, however, a posterior simulator may require the researcher to generate draws from a nonstandard distribution. Although importance sampling (e.g., Exercises 11.4 and 11.5) and the M–H algorithm (e.g., Exercises 11.17 and 11.18) often provide useful solutions in such a situation, it is certainly possible that simpler or more efficient computational procedures are available. Moreover, it is also useful to look inside the “black box” of existing routines in software packages to see how, in fact, draws from standard distributions are often obtained. In the exercises that follow in the remainder, of this chapter, we describe several alternative approaches for generating variates from nonuniform distributions. The first of these, discussed in the following two exercises, is the *inverse transform method*.

Exercise 11.19 (The inverse transform) Suppose that X is a continuous random variable with distribution function F and density f . Further, assume that the distribution function F can be easily calculated. Let $U \sim U(0, 1)$, a uniform random variable on the unit interval, and define $Y = F^{-1}(U)$.

Derive the distribution of the random variable Y . How can this result help us to generate variates from the density f ?

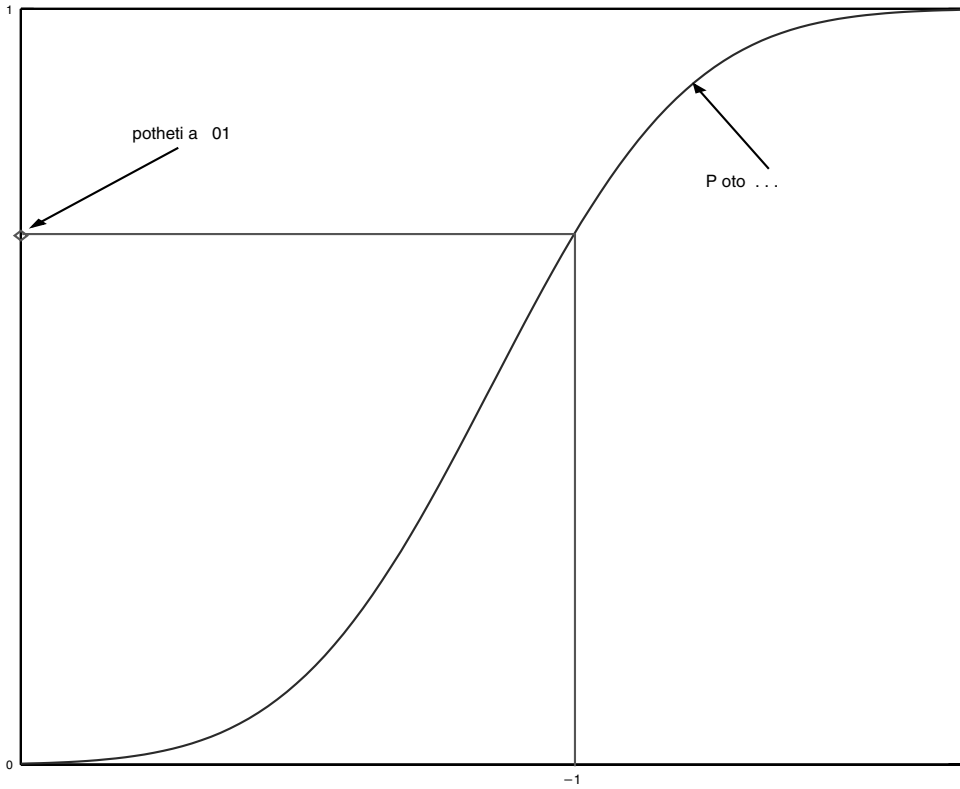


Figure 11.4 — Graphical depiction of inverse transform method.

Solution

We describe two different approaches for deriving the distribution of Y . First, we consider the c.d.f. method, and Figure 11.4 helps to illustrate this argument. Note that

$$\begin{aligned}\Pr[Y \leq a] &= \Pr[F^{-1}(U) \leq a] \\ &= \Pr[U \leq F(a)] \\ &= F(a),\end{aligned}$$

where the last line follows since $U \sim U(0, 1)$, and intuition behind the second line can be seen from Figure 11.4. This derivation shows that Y has distribution function F .

Another approach for establishing this result makes use of a change of variables. First note that $p(u) = I[0 \leq u \leq 1]$, with $I(\cdot)$ denoting an indicator function and $U = F(Y)$. Thus, $p(y) = f(y)I[0 \leq F(y) \leq 1] = f(y)$.

This result is extremely useful, because it provides a way to generate draws from f . If the c.d.f. F and its inverse F^{-1} are easily calculated, we can first draw $U \sim U(0, 1)$ and then calculate $Y = F^{-1}(U)$. It follows that Y is a draw from f . The next exercise provides two applications of this method.

Exercise 11.20 (Applications of the inverse transform: drawing exponential and truncated normal variates)

(a) Consider the exponential density (see Appendix Theorem 2) with density function

$$p(x|\theta) = \theta^{-1} \exp(-x/\theta), \quad x > 0.$$

Show how the inverse transform method can be used to generate draws from the exponential density.

(b) Let $x \sim TN_{[a,b]}(\mu, \sigma^2)$ denote that x is a *truncated normal* random variable. Specifically, this notation defines that x is generated from a normal density with mean μ and variance σ^2 , which is truncated to lie in the interval $[a, b]$. The density function for x in this case is given as

$$p(x) = \frac{\phi(x; \mu, \sigma^2)}{\Phi\left(\frac{b-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right)} I(a \leq x \leq b).$$

Show how the inverse transform method can be used to generate draws from the truncated normal density.

Solution

(a) Note that

$$\begin{aligned} F(x) &= \int_0^x \frac{1}{\theta} \exp\left(-\frac{t}{\theta}\right) dt \\ &= 1 - \exp\left(-\frac{x}{\theta}\right). \end{aligned}$$

The results of Exercise 11.19 imply that if we solve for x in the equation

$$u = 1 - \exp\left(-\frac{x}{\theta}\right),$$

with u denoting a realized draw from a $U(0, 1)$ distribution, then x has the desired exponential density. A little algebra provides

$$x = -\theta \ln(1 - u)$$

as the solution.

(b) For $a \leq x \leq b$, the c.d.f. of the truncated normal random variable is

$$\begin{aligned} F(x) &= \left[\Phi\left(\frac{b-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right) \right]^{-1} \int_a^x \phi(t; \mu, \sigma^2) dt \\ &= \frac{\Phi\left(\frac{x-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right)}{\Phi\left(\frac{b-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right)}. \end{aligned}$$

The results of Exercise 11.19 reveal that if x is a solution to the equation

$$u = \frac{\Phi\left(\frac{x-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right)}{\Phi\left(\frac{b-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right)},$$

where u is realized draw from a $U(0, 1)$ distribution, then $x \sim TN_{[a,b]}(\mu, \sigma^2)$. It follows that

$$x = \mu + \sigma \Phi^{-1} \left(\Phi\left(\frac{a-\mu}{\sigma}\right) + u \left[\Phi\left(\frac{b-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right) \right] \right).$$

Our ability to draw from the truncated normal density will be quite important for the estimation of many models considered in Chapter 14.

Exercise 11.21 (Acceptance/rejection sampling) Consider the following strategy for drawing from a density $f(x)$ defined over the compact support $a \leq x \leq b$:

1. Generate two independent uniform random variables U_1 and U_2 as follows:

$$U_i \stackrel{\text{iid}}{\sim} U(0, 1), \quad i = 1, 2.$$

2. Let

$$M \equiv \max_{a \leq x \leq b} f(x).$$

If

$$MU_2 > f(a + [b - a]U_1),$$

start over. That is, go back to the first step and generate new values for U_1 and U_2 , and again determine if $MU_2 > f(a + [b - a]U_1)$. When

$$MU_2 \leq f(a + [b - a]U_1)$$

set

$$x = a + (b - a)U_1$$

as a draw from $f(x)$.

- (a) What is the probability that any particular iteration in this algorithm will produce a draw that is accepted?
- (b) Sketch a proof as to why x , when it is accepted, has distribution function $F(x) = \int_a^x f(t) dt$.

Solution

(a) Note that

$$\begin{aligned}
 \Pr(\text{Acceptance}) &= \Pr[MU_2 \leq f(a + [b - a]U_1)] \\
 &= \Pr[U_2 \leq M^{-1}f(a + [b - a]U_1)] \\
 &= \int_0^1 \left[\int_0^{M^{-1}f(a + [b - a]U_1)} p(U_2) dU_2 \right] p(U_1) dU_1 \\
 &= \int_0^1 M^{-1}f(a + [b - a]U_1) p(U_1) dU_1 \\
 &= \int_a^b M^{-1}f(t)(b - a)^{-1} dt \\
 &= [M(b - a)]^{-1}.
 \end{aligned}$$

The third line uses the fact that U_1 and U_2 are independent, the fourth and fifth lines follow from the fact that $U_i \sim U(0, 1)$, $i = 1, 2$, and the fifth line also applies a change of variable, setting $t = a + (b - a)U_1$. Thus the probability of accepting a candidate draw in the algorithm is $[M(b - a)]^{-1}$. Note that, when using this method to sample from a uniform distribution on $[a, b]$, all candidates from the algorithm are accepted.

(b) Consider $\Pr(x \leq c | x \text{ is accepted})$. We seek to show that this probability equals the c.d.f. value $F(c) = \int_a^c f(t) dt$. We note

$$\begin{aligned}
 \Pr(x \leq c | x \text{ is accepted}) &= \frac{\Pr(x \leq c, x \text{ is accepted})}{\Pr(x \text{ is accepted})} \\
 &= M(b - a) \Pr[a + (b - a)U_1 \leq c, U_2 \leq M^{-1}f(a + [b - a]U_1)] \\
 &= M(b - a) \Pr\left[U_1 \leq \frac{c - a}{b - a}, U_2 \leq M^{-1}f(a + [b - a]U_1)\right] \\
 &= M(b - a) \int_0^{\frac{c - a}{b - a}} \int_0^{M^{-1}f(a + [b - a]U_1)} dU_2 dU_1 \\
 &= M(b - a) \int_0^{\frac{c - a}{b - a}} M^{-1}f(a + [b - a]U_1) dU_1 \\
 &= M(b - a) \int_a^c M^{-1}f(t)(b - a)^{-1} dt \\
 &= \int_a^c f(t) dt \\
 &= F(c).
 \end{aligned}$$

Therefore, a candidate draw that is accepted from the acceptance/rejection method has distribution function F , as desired.

Exercise 11.22 (Using acceptance/rejection sampling, 1)

(a) Consider the *triangular density* function, given as

$$p(x) = 1 - |x|, \quad x \in [-1, 1].$$

Use the acceptance/rejection sampling method of Exercise 11.21 to generate 25,000 draws from this distribution.

(b) Suppose it is desired to sample a random variable x from a $TN_{[0,4]}(1, 1)$ distribution, that is, a normal density with unit mean and variance, that has been truncated to the interval $[0, 4]$. Apply the acceptance/rejection method to generate 25,000 draws from this truncated normal distribution.

(c) Using the 25,000 draws obtained from (a) and (b), estimate the density functions of the accepted draws in each case. Evaluate the performances of the acceptance/rejection method by comparing the density estimates to the actual densities.

Solution

(a–c) The Matlab code used to simulate draws from these distributions is provided on the Web site associated with this book. For (a), note that $M = 1$ and $b - a = 2$, so the overall acceptance rate is one-half. (That is, we would expect that 50,000 pairs of independent uniform variates in the acceptance/rejection algorithm are needed to produce a final sample of 25,000 draws.) For (b), $M \approx .475$ and $b - a = 4$, so the overall acceptance rate is approximately .53.

Graphs of the actual densities (dotted) and kernel density estimates based on the 25,000 draws (solid) are provided in Figure 11.5. For the triangular case, the estimated density is nearly indistinguishable from the actual density. For the truncated normal case, the two graphs are again very similar. The slight discrepancy at the lower limit of zero seems related to the performance of the kernel density estimator at the boundary.

Exercise 11.23 (A generalized acceptance/rejection algorithm) Suppose it is of interest to generate draws from a density $f(\theta)$ (henceforth, referred to as the *target* density). Let Θ denote the support of $f(\theta)$, and suppose there exists some approximating density $s(\theta)$, called the *source density*, with support Θ^* , where $\Theta \subseteq \Theta^*$.

In many applications requiring posterior simulation, the normalizing constant of the target density is unknown, since the joint or conditional posteriors are only given up to proportionality. To this end, let us work with the kernels of both the source and target densities and write

$$\begin{aligned} f(\theta) &= c_f \tilde{f}(\theta), \\ s(\theta) &= c_s \tilde{s}(\theta), \end{aligned}$$

so that \tilde{f} and \tilde{s} denote the target and source kernels, respectively, and c_f and c_s denote the associated normalizing constants. Finally, let

$$\tilde{M} \equiv \sup_{\theta \in \Theta} \left(\frac{\tilde{f}(\theta)}{\tilde{s}(\theta)} \right).$$

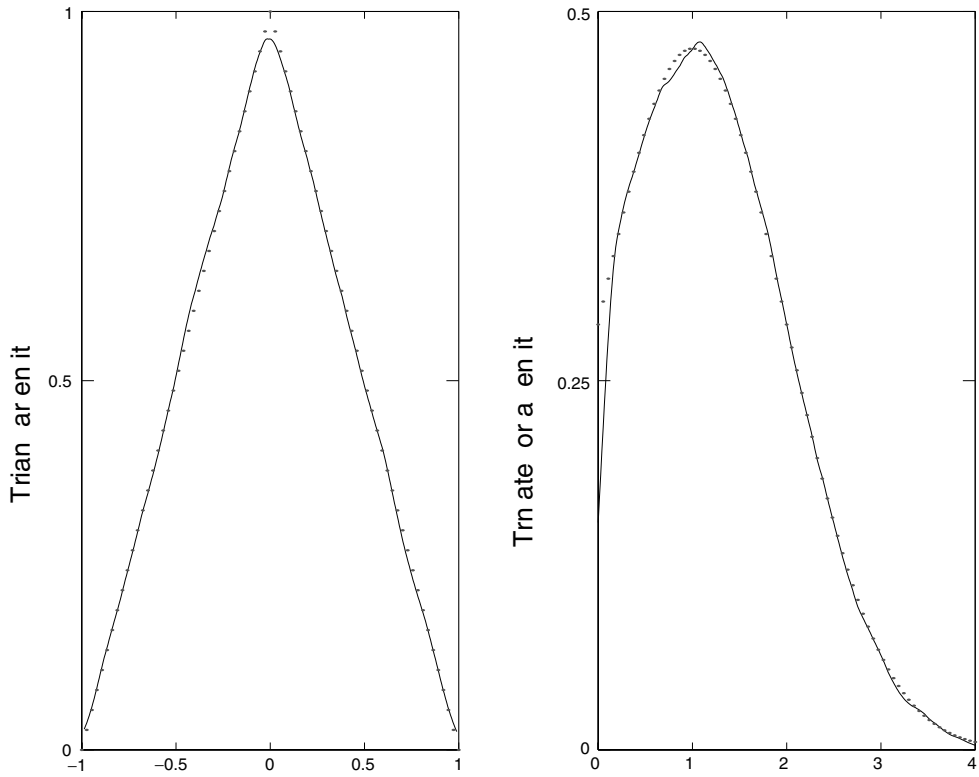


Figure 11.5 — Analytical densities (dotted) and kernel density estimates (solid) obtained using 25,000 draws from acceptance/rejection sampler: triangular (left) and truncated normal (right).

Consider the following algorithm:

1. Draw U uniformly on $[0, 1]$ [i.e., $U \sim U(0, 1)$].
2. Draw a candidate from the source density $s(\theta)$ [i.e., $\theta^{\text{cand}} \sim s(\theta)$].
3. If

$$U \leq \frac{\tilde{f}(\theta^{\text{cand}})}{\tilde{M}\tilde{s}(\theta^{\text{cand}})},$$

then set

$$\theta = \theta^{\text{cand}}$$

as a draw from $f(\theta)$. Otherwise, return to the first step and repeat the process until condition (3) is satisfied.

- (a) Show how this algorithm includes the one provided in Exercise 11.21 as a special case.
- (b) What is the overall acceptance rate in this algorithm?
- (c) Sketch a proof of why this algorithm provides a draw from $f(\theta)$.

Solution

(a) Consider using this algorithm to generate a draw from $f(\theta)$ with compact support $[a, b]$, as described in Exercise 11.21. In addition, employ a source density $s(\theta)$ that is uniform over the interval $[a, b]$.

In this case we can write

$$f(\theta) = c_f g(\theta) I(a \leq \theta \leq b) = c_f \tilde{f}(\theta),$$

where $\tilde{f}(\theta) = g(\theta) I(a \leq \theta \leq b)$, $\int_a^b g(\theta) = c_f^{-1}$, and

$$s(\theta) = [b - a]^{-1} I(a \leq \theta \leq b) = [b - a]^{-1} \tilde{s}(\theta).$$

It follows that

$$\tilde{M} = \max_{a \leq \theta \leq b} \left(\frac{\tilde{f}(\theta)}{\tilde{s}(\theta)} \right) = \max_{a \leq \theta \leq b} g(\theta) = c_f^{-1} \max_{a \leq \theta \leq b} f(\theta) = c_f^{-1} M,$$

where M is defined as the maximum of f in Exercise 11.21.

To implement the algorithm with the given uniform source density, we first generate $\theta^{\text{cand}} \sim U(a, b)$, which is equivalent to writing $\theta^{\text{cand}} = a + (b - a)U_1$, where $U_1 \sim U(0, 1)$. We then generate $U_2 \sim U(0, 1)$ and accept θ^{cand} provided

$$\begin{aligned} U_2 &\leq \frac{\tilde{f}(\theta^{\text{cand}})}{\tilde{M} \tilde{s}(\theta^{\text{cand}})} \\ &= \frac{c_f^{-1} f(a + [b - a]U_1)}{\tilde{M}} \\ &= \frac{c_f^{-1} f(a + [b - a]U_1)}{c_f^{-1} M} \\ &= \frac{f(a + [b - a]U_1)}{M}. \end{aligned}$$

This decision rule and the random variables U_1 and U_2 are identical to those described in Exercise 11.21. So the algorithm provided in this exercise reduces to that described in Exercise 11.21 when the target density has compact support and a source density that is uniform over that same support is employed.

(b) The overall acceptance rate is

$$\begin{aligned} \Pr\left(U \leq [\tilde{f}(\theta)/\tilde{M}\tilde{s}(\theta)]\right) &= \int_{\Theta} \left[\int_0^{\tilde{f}(\theta)/\tilde{M}\tilde{s}(\theta)} dU \right] s(\theta) d\theta \\ &= \int_{\Theta} \frac{\tilde{f}(\theta)}{\tilde{M}\tilde{s}(\theta)} s(\theta) d\theta \\ &= \frac{c_s}{\tilde{M}c_f}. \end{aligned}$$

(c) Following Geweke (2005, Section 4.2.1), we note that for any subset A of Θ

$$\begin{aligned}\Pr(\theta \text{ is accepted}, \theta \in A) &= \int_A \frac{\tilde{f}(\theta)}{\tilde{M}\tilde{s}(\theta)} s(\theta) d\theta \\ &= \frac{c_s}{\tilde{M}} \int_A \tilde{f}(\theta) d\theta.\end{aligned}$$

Since

$$\begin{aligned}\Pr(\theta \in A | \theta \text{ is accepted}) &= \frac{\Pr(\theta \text{ is accepted}, \theta \in A)}{\Pr(\theta \text{ is accepted})} \\ &= \frac{c_s \tilde{M}^{-1} \int_A \tilde{f}(\theta) d\theta}{c_s [\tilde{M} c_f]^{-1}} \\ &= c_f \int_A \tilde{f}(\theta) d\theta \\ &= \int_A f(\theta) d\theta,\end{aligned}$$

it follows that when θ is accepted from the algorithm, it is indeed a draw from $f(\theta)$.

Exercise 11.24 (Using acceptance/rejection sampling, 2)

(a) Using the algorithm of Exercise 11.23 and a $U(-1, 1)$ source density, generate 25,000 draws from the triangular distribution $p(\theta) = 1 - |\theta|$, $\theta \in [-1, 1]$. Comment on your results.

(b) Generate an equal number of draws from the triangular distribution using the same algorithm and an $N(0, \sigma^2)$ source density. First, consider a standard normal source with $\sigma^2 = 1$. Then, investigate the performance of the acceptance/rejection method with $\sigma^2 = 2$ and $\sigma^2 = 1/6$. Comment on your results.

Solution

(a–c). For the $U(-1, 1)$ source density, note that the solution to part (a) of the previous exercise shows that the algorithm of Exercise 11.23 reduces to the algorithm of Exercise 11.21. Thus, from part (a) of Exercise 11.21, it follows that the overall acceptance rate is .5. We regard this as a benchmark and seek to determine if an alternate choice of source density can lead to increased efficiency.

For the normal source density in part (b), when $\sigma^2 = 1$, the maximum of the target/source ratio occurs at $\theta = 0$, yielding a value of $\tilde{M} = 1$. Since the normalizing constant of the standard normal is $c_s = (2\pi)^{-1/2}$, it follows that the overall acceptance rate is

$(2\pi)^{-1/2} \approx .40$. When comparing this algorithm to one with $\sigma^2 = 2$, it is clear that the standard normal source will be preferred. The maximum of the target/source ratio with $\sigma^2 = 2$ again occurs at $\theta = 0$, yielding $\tilde{M} = 1$. However, the overall acceptance probability reduces to $1/\sqrt{4\pi} \approx .28$.

The final choice of $\sigma^2 = 1/6$ is reasonably tailored to fit this application. One can easily show that the mean of the target is zero and the variance is $1/6$, so the $N(0, 1/6)$ source is chosen to match these two moments of the triangular density. With a little algebra, one can show that the maximum of the target/source ratio occurs at

$$\theta = \frac{1 + \sqrt{1/3}}{2} \approx .789.$$

(Note that another maximum also occurs at $-.789$ since both the target and source are symmetric about zero.) With this result in hand, the maximized value of the target/source ratio is $\tilde{M} \approx 1.37$, yielding a theoretical acceptance rate of $1/(1.37\sqrt{2\pi[1/6]}) \approx .72$. Thus, the $N(0, 1/6)$ source density is the most efficient of the candidates considered here.

In Figure 11.6 we plot the alternate source densities, scaled up by \tilde{M} , against the triangular density. The efficiency of the algorithm increases with the proximity of the scaled

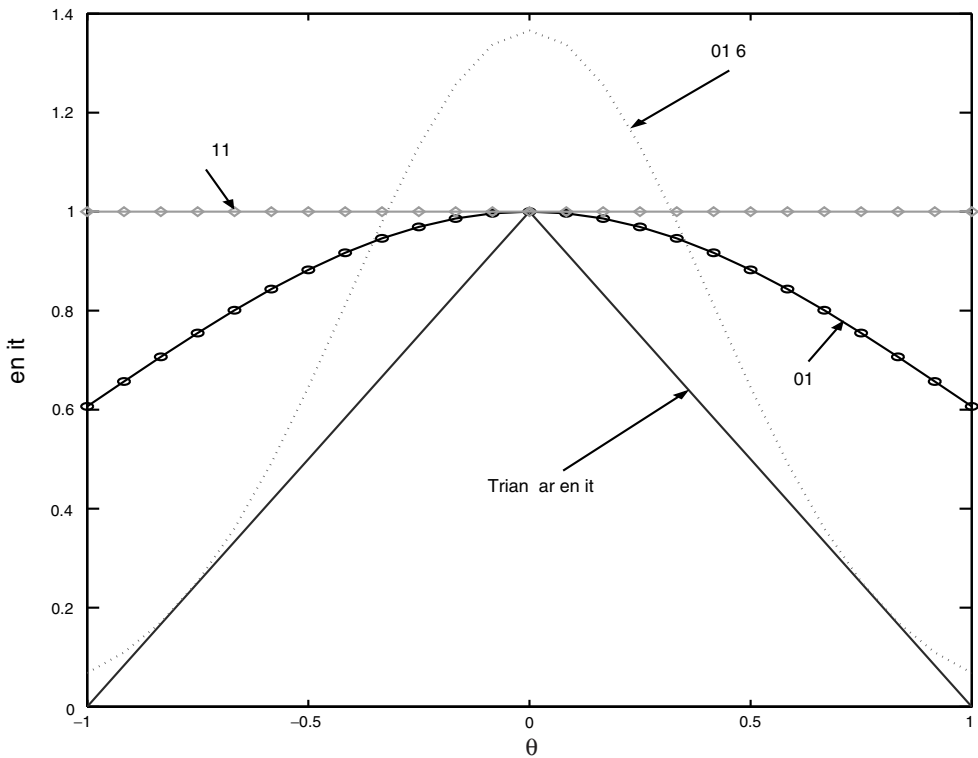


Figure 11.6 — Triangular density together with three different scaled source densities.

source density to the target. As the figure suggests, the $N(0, 1/6)$ source is most efficient, as its selection results in the fewest number of discarded draws. Although it might seem that the figure suggests a preference for the $N(0, 1)$ source over the $U(-1, 1)$ alternative, note that this figure is only plotted over the support of the target density. When a candidate outside the $[-1, 1]$ range is drawn from the $N(0, 1)$ source, it must be discarded, resulting in increased inefficiency of the $N(0, 1)$ choice relative to the $U(-1, 1)$ source.

Exercise 11.25 (The weighted bootstrap) The *weighted bootstrap* of Smith and Gelfand (1992) and the highly related sampling-importance resampling (SIR) algorithm of Rubin (1987, 1988) circumvents the need to calculate the “blanketing constant” \tilde{M} in acceptance sampling.

Consider the following procedure for obtaining a draw from a density of interest f :

1. Draw $\theta_1, \theta_2, \dots, \theta_n$ from some approximating source density $s(\theta)$.
2. Like Exercise 11.23, let us work with the kernels of the target and source densities and thus define

$$\begin{aligned} f(\theta) &= c_f \tilde{f}(\theta), \\ s(\theta) &= c_s \tilde{s}(\theta). \end{aligned}$$

Set

$$w_i = w_i(\theta_i) = \frac{\tilde{f}(\theta_i)}{\tilde{s}(\theta_i)}$$

and define the normalized weights

$$\tilde{w}_i = \frac{w_i}{\sum_{i=1}^n w_i}.$$

3. Draw θ^* from the discrete set $\{\theta_1, \theta_2, \dots, \theta_n\}$ with $\Pr(\theta^* = \theta_j) = \tilde{w}_j$, $j = 1, 2, \dots, n$.

Show that θ^* provides an *approximate* draw from $f(\theta)$, with the accuracy of the approach improving with n , the simulated sample size from the source density.

Solution

Note that

$$\begin{aligned}
 \Pr(\theta^* \leq c) &= \sum_{j=1}^n \tilde{w}_j I(\theta_j \leq c) \\
 &= \frac{\sum_{j=1}^n w_j I(\theta_j \leq c)}{\sum_{j=1}^n w_j} \\
 &= \frac{[1/n] \sum_{j=1}^n w_j I(\theta_j \leq c)}{[1/n] \sum_{j=1}^n w_j} \\
 &\rightarrow \frac{E[w(\theta) I(\theta \leq c)]}{E[w(\theta)]} \\
 &= \frac{\int_{-\infty}^{\infty} w(\theta) I(\theta \leq c) s(\theta) d\theta}{\int_{-\infty}^{\infty} w(\theta) s(\theta) d\theta} \\
 &= \frac{\int_{-\infty}^c [\tilde{f}(\theta)/\tilde{s}(\theta)] s(\theta) d\theta}{\int_{-\infty}^{\infty} [\tilde{f}(\theta)/\tilde{s}(\theta)] s(\theta) d\theta} \\
 &= \frac{\int_{-\infty}^c \tilde{f}(\theta) d\theta}{\int_{-\infty}^{\infty} \tilde{f}(\theta) d\theta} \\
 &= \int_{-\infty}^c f(\theta) d\theta \\
 &\equiv F(c).
 \end{aligned}$$

In the fourth line, we consider the limit as the simulated sample size n approaches infinity. To proceed from line 6 to 7, we note $s(\theta)/\tilde{s}(\theta) = c_s$, and similarly note that $f(\theta)/\tilde{f}(\theta) = c_f$ when moving from lines 7 to 8.

Again, this algorithm only provides an approximate draw from f , with its accuracy increasing with n . Of course, the performance of the algorithm depends on how accurately s approximates f . If, for example, the mass under s is concentrated in a region where f places little mass, then the algorithm will typically perform poorly.

Hierarchical models

In this chapter we take up issues relating to estimation in hierarchical models. Many models belong in this class, and indeed, several exercises appearing in later chapters (e.g., mixture models and panel probit models) involve specifications with hierarchical structures that could be included as exercises here. In this chapter, however, we confine our attention to linear specifications and pay particular attention to normal hierarchical linear regression models. In econometrics, these are commonly used with panel (also called longitudinal) data and several of the questions in this chapter use such data. With the exception of the first exercise, all questions involve posterior simulation using Gibbs samplers. Introductory exercises on Gibbs sampling were provided in Chapter 11. Matlab code for all the empirical work is also provided on the Web site associated with this book.

Exercise 12.1 (General results for normal hierarchical linear models) Let y be an N vector and θ_1 , θ_2 , and θ_3 be parameter vectors of length k_1 , k_2 , and k_3 , respectively. Let X , W , and Z be known $N \times k_1$, $k_1 \times k_2$, and $k_2 \times k_3$ matrices and C_1 , C_2 , and C_3 be $N \times N$, $k_1 \times k_1$, and $k_2 \times k_2$ known positive definite matrices. Assume

$$y|\theta_1 \sim N(X\theta_1, C_1), \quad (12.1)$$

$$\theta_1|\theta_2 \sim N(W\theta_2, C_2), \quad (12.2)$$

and

$$\theta_2|\theta_3 \sim N(Z\theta_3, C_3). \quad (12.3)$$

We make conditional independence assumptions such that $p(y|\theta_1, \theta_2) = p(y|\theta_1)$ and $p(\theta_1|\theta_2, \theta_3) = p(\theta_1|\theta_2)$.

(a) Show that

$$y|\theta_2 \sim N(XW\theta_2, C_1 + XC_2X') \quad (12.4)$$

and thus that this normal linear regression model with hierarchical prior can be written as a different normal linear regression model with nonhierarchical prior.

(b) Derive $p(\theta_2|y, \theta_3)$.

(c) Show that

$$\theta_1|y, \theta_3 \sim N(Dd, D),$$

where

$$D^{-1} = X'C_1^{-1}X + (C_2 + WC_3W')^{-1}$$

and

$$d = X'C_1^{-1}y + (C_2 + WC_3W')^{-1}WZ\theta_3.$$

Solution

(a) Equation (12.1) is equivalent to

$$y = X\theta_1 + \epsilon,$$

where $\epsilon \sim N(0, C_1)$. Similarly, (12.2) can be written as

$$\theta_1 = W\theta_2 + v,$$

where $v \sim N(0, C_2)$. Substituting, we obtain

$$y = XW\theta_2 + Xv + \epsilon.$$

Using the properties of the normal distribution leads immediately to (i) $y|\theta_2$ is normal, (ii) its mean is $XW\theta_2$, and (iii) its covariance is $E[(Xv + \epsilon)(Xv + \epsilon)'] = C_1 + XC_2X'$ (since the question assumptions imply ϵ and v are independent of one another). Hence, the required result is proven.

(b) By Bayes' theorem,

$$p(\theta_2|y, \theta_3) \propto p(\theta_2|\theta_3) p(y|\theta_2, \theta_3).$$

The first term on the right-hand side is simply the normal prior in (12.3) and the second term is the normal likelihood in (12.4). The derivation of the resulting posterior is standard (see Exercises 10.1 and 13.1 for related derivations). It can be verified that

$$\theta_2|y, \theta_3 \sim N(Hh, H),$$

where

$$H = \left(W'X'(C_1 + XC_2X')^{-1}XW + C_3^{-1} \right)^{-1}$$

and

$$h = W'X'(C_1 + XC_2X')^{-1}y + C_3^{-1}Z\theta_3.$$

(c) Again, Bayes' theorem implies $p(\theta_1|y, \theta_3) \propto p(\theta_1|\theta_3)p(y|\theta_1, \theta_3)$. The assumptions of the question imply $p(y|\theta_1, \theta_3)$ is given in (12.1). A proof the same as that for part (a), except that (12.2) and (12.3) are used, establishes that the prior for θ_1 with θ_2 integrated out takes the form

$$\theta_1|\theta_3 \sim N(WZ\theta_3, C_2 + WC_3W'). \quad (12.5)$$

Hence, just as in part (b), we have a normal likelihood (12.1) times a normal prior (12.5). The derivation of the resulting posterior is standard (see Exercises 2.3 and 10.1 for closely related derivations). It can be verified that the mean and covariance of the normal posterior, $p(\theta_1|y, \theta_3)$, are as given in the question.

Exercise 12.2 (Hierarchical modeling with longitudinal data) Consider the following longitudinal data model:

$$y_{it} = \alpha_i + \epsilon_{it}, \quad \epsilon_{it} \stackrel{\text{iid}}{\sim} N(0, \sigma_\epsilon^2), \quad (12.6)$$

$$\alpha_i \stackrel{\text{iid}}{\sim} N(\alpha, \sigma_\alpha^2), \quad (12.7)$$

where y_{it} refers to the outcomes for individual (or more generally, group) i at time t and α_i is a person-specific random effect. We assume $i = 1, 2, \dots, N$ and $t = 1, 2, \dots, T$ (i.e., a balanced panel).

(a) Comment on how the presence of the random effects accounts for correlation patterns within individuals over time.

(b) Derive the conditional posterior distribution $p(\alpha_i|\alpha, \sigma_\epsilon^2, \sigma_\alpha^2, y)$.

(c) Obtain the mean of the conditional posterior distribution in (b). Comment on its relationship to a shrinkage estimator [see, e.g., Poirier (1995), Chapter 6]. How does the mean change as T and $\sigma_\epsilon^2/\sigma_\alpha^2$ change?

Solution

(a) Conditional on the random effects $\{\alpha_i\}_{i=1}^N$, the y_{it} are independent. However, marginalized over the random effects, outcomes are correlated within individuals over time. To see this, note that we can write (12.6) equivalently as

$$y_{it} = \alpha + u_i + \epsilon_{it},$$

where we have rewritten (12.7) as

$$\alpha_i = \alpha + u_i, \quad u_i \stackrel{\text{iid}}{\sim} N(0, \sigma_\alpha^2),$$

and substituted this result into (12.6). Thus for $t \neq s$,

$$\text{Cov}(y_{it}, y_{is}|\alpha, \sigma_\epsilon^2, \sigma_\alpha^2) = \text{Cov}(u_i + \epsilon_{it}, u_i + \epsilon_{is}) = \text{Cov}(u_i, u_i) = \text{Var}(u_i) = \sigma_\alpha^2,$$

so that outcomes are correlated over time within individuals. However, the random effects have not permitted any degree of correlation between the outcomes of different individuals.

(b) From Exercise 12.1,

$$\alpha_i | \alpha, \sigma_\epsilon^2, \sigma_\alpha^2, y \sim N(Dd, D),$$

where

$$D = (\iota'_T \iota_T / \sigma_\epsilon^2 + 1 / \sigma_\alpha^2)^{-1} = \frac{\sigma_\alpha^2 \sigma_\epsilon^2}{T \sigma_\alpha^2 + \sigma_\epsilon^2}$$

and

$$d = \frac{\iota'_T y_i}{\sigma_\epsilon^2} + \frac{\alpha}{\sigma_\alpha^2} = \frac{\sigma_\alpha^2 \sum_t y_{it} + \alpha \sigma_\epsilon^2}{\sigma_\alpha^2 \sigma_\epsilon^2}.$$

In these expressions, we have let ι_T denote a $T \times 1$ vector of ones and $y_i = [y_{i1} \ y_{i2} \ \cdots \ y_{iT}]'$.

(c) The mean of this conditional posterior distribution is easily obtained from our solution in (b):

$$\begin{aligned} E(\alpha_i | \beta, \sigma_\epsilon^2, \sigma_\alpha^2, y) &= \frac{\sigma_\alpha^2 \sum_t y_{it} + \alpha \sigma_\epsilon^2}{T \sigma_\alpha^2 + \sigma_\epsilon^2} \\ &= \left(\frac{T \sigma_\alpha^2}{T \sigma_\alpha^2 + \sigma_\epsilon^2} \right) \bar{y}_i + \left(\frac{\sigma_\epsilon^2}{T \sigma_\alpha^2 + \sigma_\epsilon^2} \right) \alpha \\ &= \left(\frac{T}{T + (\sigma_\epsilon^2 / \sigma_\alpha^2)} \right) \bar{y}_i + \left(\frac{(\sigma_\epsilon^2 / \sigma_\alpha^2)}{T + (\sigma_\epsilon^2 / \sigma_\alpha^2)} \right) \alpha. \end{aligned}$$

Let

$$w = w(T, [\sigma_\epsilon^2 / \sigma_\alpha^2]) \equiv \frac{T}{T + (\sigma_\epsilon^2 / \sigma_\alpha^2)}.$$

Then we can write

$$E(\alpha_i | \beta, \sigma_\epsilon^2, \sigma_\alpha^2, y) = w \bar{y}_i + (1 - w) \alpha.$$

This is in the form of a shrinkage estimator, where the conditional posterior mean of α_i is a weighted average (with w serving as the weight) of the averaged outcomes for individual i , \bar{y}_i , and the common mean for all individuals, α . As $T \rightarrow \infty$ (and holding all else constant) we see that $w \rightarrow 1$, so the conditional posterior mean approaches the average of the individual outcomes. Intuitively, this makes sense, since as $T \rightarrow \infty$, we acquire more and more information on individual i , and thus data information from individual i dominates any information provided by the outcomes of other individuals in the sample. As $(\sigma_\epsilon^2 / \sigma_\alpha^2) \rightarrow \infty$, $w \rightarrow 0$, so our posterior mean collapses to the common mean α . In this case the error variability in the outcome equation continues to grow relative to the variability in the random effects equation, and thus our posterior mean reduces the common effect for all individuals, α .

Exercise 12.3 (A posterior simulator for a linear hierarchical model: Gelfand et al. [1990]) We illustrate Bayesian procedures for estimating hierarchical linear models using a data set that has become something of a “classic” in the MCMC literature. These data come from the study of Gelfand, Hills, Racine-Poon, and Smith (1990). In this exercise, we derive in full detail the complete posterior conditionals. In the following exercise we fit this model and provide some diagnostic checks for convergence. In subsequent examples complete derivations of the conditionals will not typically be provided, as these derivations of will often follow similarly to those described in this exercise.

In the rat growth model of Gelfand, et al. (1990), 30 different rats are weighed at five different points in time. We denote the weight of rat i at measurement j as y_{ij} and let x_{ij} denote the age of the i th rat at the j th measurement. Since each of the rats were weighed at exactly the same number of days since birth, we have

$$x_{i1} = 8, x_{i2} = 15, x_{i3} = 22, x_{i4} = 29, x_{i5} = 36 \quad \forall i.$$

The data used in the analysis are provided in Table 12.1.

Table 12.1: Rat growth data from Gelfand et al. (1990).

Rat i	Weight Measurements					Rat i	Weight Measurements				
	y_{i1}	y_{i2}	y_{i3}	y_{i4}	y_{i5}		y_{i1}	y_{i2}	y_{i3}	y_{i4}	y_{i5}
1	151	199	246	283	320	16	160	207	248	288	324
2	145	199	249	293	354	17	142	187	234	280	316
3	147	214	263	312	328	18	156	203	243	283	317
4	155	200	237	272	297	19	157	212	259	307	336
5	135	188	230	280	323	20	152	203	246	286	321
6	159	210	252	298	331	21	154	205	253	298	334
7	141	189	231	275	305	22	139	190	225	267	302
8	159	201	248	297	338	23	146	191	229	272	302
9	177	236	285	340	376	24	157	211	250	285	323
10	134	182	220	260	296	25	132	185	237	286	331
11	160	208	261	313	352	26	160	207	257	303	345
12	143	188	220	273	314	27	169	216	261	295	333
13	154	200	244	289	325	28	157	205	248	289	316
14	171	221	270	326	358	29	137	180	219	258	291
15	163	216	242	281	312	30	153	200	244	286	324

In our model, we want to permit unit-specific variation in initial birth weight and growth rates. This leads us to specify the following model:

$$y_{ij} | \alpha_i, \beta_i, \sigma^2, x_{ij} \stackrel{\text{ind}}{\sim} N(\alpha_i + \beta_i x_{ij}, \sigma^2), \quad i = 1, 2, \dots, 30, \quad j = 1, 2, \dots, 5,$$

so that each rat possesses its own intercept α_i and growth rate β_i .

We also assume that the rats share some degree of “commonality” in their weight at birth and rates of growth, and thus we assume that the intercept and slope parameters are drawn from the same normal population:

$$\theta_i = \begin{bmatrix} \alpha_i \\ \beta_i \end{bmatrix} \sim N(\theta_0, \Sigma),$$

where $\theta_0 = [\alpha_0 \ \beta_0]'$ and Σ is a 2×2 positive definite symmetric covariance matrix.

We complete our Bayesian analysis by specifying the following priors:

$$\begin{aligned} \sigma^2 | a, b &\sim IG(a, b), \\ \theta_0 | \eta, C &\sim N(\eta, C), \\ \Sigma^{-1} | \rho, R &\sim W([\rho R]^{-1}, \rho), \end{aligned}$$

with W denoting the Wishart distribution (see Appendix Definition 6 for details).

Derive the complete conditionals for this model and comment on any intuition behind the forms of these conditionals.

Solution

Given the assumed conditional independence across observations, the joint posterior distribution for all the parameters of this model can be written as

$$p(\Gamma | y) \propto \left[\prod_{i=1}^{30} p(y_i | x_i, \theta_i, \sigma^2) p(\theta_i | \theta_0, \Sigma^{-1}) \right] p(\theta_0 | \eta, C) p(\sigma^2 | a, b) p(\Sigma^{-1} | \rho, R),$$

where $\Gamma \equiv [\{\theta_i\}, \theta_0, \Sigma^{-1}, \sigma^2]$ denotes all the parameters of the model. We will also use the notation Γ_{-x} to denote all parameters other than x . We have stacked the observations over time for each individual rat so that

$$y_i = \begin{bmatrix} y_{i1} \\ y_{i2} \\ \vdots \\ y_{i5} \end{bmatrix}, \quad X_i = \begin{bmatrix} 1 & x_{i1} \\ 1 & x_{i2} \\ \vdots & \vdots \\ 1 & x_{i5} \end{bmatrix}.$$

Complete posterior conditional for θ_i

We first note that this complete conditional is proportional to the aforementioned joint posterior. Thus, all of the terms in the product that do not involve θ_i are absorbed into the normalizing constant of this conditional. Hence,

$$p(\theta_i | \Gamma_{-\theta_i}, y) \propto p(y_i | X_i, \theta_i, \sigma^2) p(\theta_i | \theta_0, \Sigma).$$

This fits directly into the framework of Exercise 12.1 [see also Lindley and Smith (1972)], and thus we find

$$\theta_i | \Gamma_{-\theta_i}, y \sim N(D_{\theta_i} d_{\theta_i}, D_{\theta_i}), \quad i = 1, 2, \dots, 30,$$

where

$$D_{\theta_i} = (X_i' X_i / \sigma^2 + \Sigma^{-1})^{-1}, \quad d_{\theta_i} = X_i' y_i / \sigma^2 + \Sigma^{-1} \theta_0.$$

Because of the conditional independence, we can draw each of the θ_i in turn by sampling from the corresponding complete conditional.

Complete posterior conditional for θ_0

For θ_0 , a similar argument shows

$$p(\theta_0 | \Gamma_{-\theta_0}, y) \propto \left(\prod_{i=1}^{30} p(\theta_i | \theta_0, \Sigma^{-1}) \right) p(\theta_0 | \eta, C).$$

Since the second stage of our model specifies $p(\theta_i | \theta_0, \Sigma^{-1})$ as iid, we can write

$$\begin{bmatrix} \theta_1 \\ \theta_2 \\ \vdots \\ \theta_{30} \end{bmatrix} = \begin{bmatrix} I_2 \\ I_2 \\ \vdots \\ I_2 \end{bmatrix} \theta_0 + \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_{30} \end{bmatrix},$$

or equivalently,

$$\tilde{\theta} = \tilde{I} \theta_0 + \tilde{u},$$

with $\tilde{\theta} = [\theta_1' \theta_2' \cdots \theta_{30}']'$, $\tilde{I} = [I_2 \ I_2 \cdots I_2]'$, $\tilde{u} = [u_1' \ u_2' \cdots u_{30}']'$, and $E(\tilde{u} \tilde{u}') = I_{30} \otimes \Sigma$.

In this form, the results of Exercise 12.1 again apply with

$$\theta_0 | \Gamma_{-\theta_0}, y \sim N(D_{\theta_0} d_{\theta_0}, D_{\theta_0}),$$

where

$$D_{\theta_0} = \left(\tilde{I}' (I_{30} \otimes \Sigma^{-1}) \tilde{I} + C^{-1} \right)^{-1} = (30 \Sigma^{-1} + C^{-1})^{-1}$$

and

$$d_{\theta_0} = (\tilde{I}' (I_{30} \otimes \Sigma^{-1}) \tilde{\theta} + C^{-1} \eta) = (30 \Sigma^{-1} \bar{\theta} + C^{-1} \eta),$$

and where $\bar{\theta} = (1/30) \sum_{i=1}^{30} \theta_i$.

Complete posterior conditional for σ^2

For σ^2 , we find

$$\begin{aligned}
 p(\sigma^2 | \Gamma_{-\sigma^2}, y) &\propto \left(\prod_{i=1}^{30} p(y_i | X_i \theta_i, \sigma^2) \right) p(\sigma^2 | a, b) \\
 &\propto \frac{1}{[\sigma^2]^{N/2}} \exp \left(-\frac{1}{2\sigma^2} \sum_{i=1}^{30} (y_i - X_i \theta_i)' (y_i - X_i \theta_i) \right) \\
 &\quad \times \frac{1}{[\sigma^2]^{a+1}} \exp \left(-\frac{1}{b\sigma^2} \right) \\
 &= \frac{1}{[\sigma^2]^{N/2+a+1}} \exp \left(-\frac{1}{\sigma^2} \left[1/2 \sum_{i=1}^{30} (y_i - X_i \theta_i)' (y_i - X_i \theta_i) + b^{-1} \right] \right).
 \end{aligned}$$

Thus,

$$\sigma^2 | \Gamma_{-\sigma^2}, y \sim IG \left(N/2 + a, \left[1/2 \sum_{i=1}^{30} (y_i - X_i \theta_i)' (y_i - X_i \theta_i) + b^{-1} \right]^{-1} \right),$$

where $N = 5(30) = 150$.

Complete posterior conditional for Σ^{-1}

We find

$$\begin{aligned}
 p(\Sigma^{-1} | \Gamma_{-\Sigma^{-1}}, y) &\propto \left[\prod_{i=1}^{30} p(\theta_i | \theta_0, \Sigma^{-1}) \right] p(\Sigma^{-1} | \rho, R) \\
 &\propto |\Sigma^{-1}|^{30/2} \exp \left(-1/2 \sum_{i=1}^{30} (\theta_i - \theta_0)' \Sigma^{-1} (\theta_i - \theta_0) \right) \\
 &\quad \times |\Sigma^{-1}|^{(\rho-3)/2} \exp \left(-[1/2] \text{tr}[\rho R \Sigma^{-1}] \right) \\
 &= |\Sigma^{-1}|^{(30+\rho-3)/2} \\
 &\quad \times \exp \left(-1/2 \left(\left[\sum_{i=1}^{30} \text{tr}(\theta_i - \theta_0)(\theta_i - \theta_0)' \Sigma^{-1} \right] + \text{tr}[\rho R \Sigma^{-1}] \right) \right) \\
 &= |\Sigma^{-1}|^{(30+\rho-3)/2} \\
 &\quad \exp \left[-(1/2) \text{tr} \left(\left[\sum_{i=1}^{30} (\theta_i - \theta_0)(\theta_i - \theta_0)' + \rho R \right] \Sigma^{-1} \right) \right].
 \end{aligned}$$

Therefore,

$$\Sigma^{-1}|\Gamma_{-\Sigma^{-1}}, y \sim W \left(\left[\sum_{i=1}^{30} (\theta_i - \theta_0)(\theta_i - \theta_0)' + \rho R \right]^{-1}, 30 + \rho \right).$$

We can gather some intuition regarding the forms of these conditionals. First, note from the complete conditional for θ_i that the conditional posterior mean can be regarded as a weighted average of the OLS estimate using data from that individual: $(X_i'X_i)^{-1}X_i'y_i$ and the common or “pooled” mean across all individuals θ_0 . In this sense, the conditional mean is similar to a “shrinkage” estimator as it combines information from the outcomes on the individual rat with cross-section information from the outcomes of all the rats. Further, as $\Sigma^{-1} \rightarrow 0$ and holding all else constant, the conditional posterior mean approaches the OLS estimate. This is seemingly reasonable since $\Sigma^{-1} \approx 0$ implies a large amount of second-stage variability, and thus most of the weight will be given to the OLS estimate from the outcomes of the individual rat.

For the complete conditional for θ_0 , note that for very flat priors (choosing C large and thus C^{-1} “small”), the conditional posterior mean reduces to the sample averages of all of the θ_i . Thus, in this case, the individual level parameter estimates are “shrunk” toward the average of all parameter estimates for all the rats.

Finally, for the complete conditional posterior mean of Σ^{-1} , note that for relatively vague priors corresponding to small values of ρ , the conditional posterior mean of Σ^{-1} is approximately $k[\sum_{i=1}^k (\theta_i - \theta_0)(\theta_i - \theta_0)']^{-1}$, so that $E(\Sigma) \approx 1/k \sum_{i=1}^k (\theta_i - \theta_0)(\theta_i - \theta_0)'$. Thus, when the prior information is not informative relative to the data information, the conditional posterior mean of the covariance matrix Σ is approximately equal to the sample variability of the individual-level parameters around the common mean θ_0 .

Exercise 12.4 (The linear hierarchical model, 2: estimation and convergence diagnostics) Use the data and algorithm provided in the previous exercise.

(a) Fit the rat growth model using the following priors:

$$\eta = \begin{bmatrix} 100 \\ 15 \end{bmatrix}, C = \begin{bmatrix} 40^2 & 0 \\ 0 & 10^2 \end{bmatrix}, \rho = 5, R = \begin{bmatrix} 10^2 & 0 \\ 0 & .5^2 \end{bmatrix}, a = 3, b = 1/(40).$$

(b) Perform some diagnostic checking to detect convergence of the sampler [see, e.g., Exercise 11.7 and, for a related discussion of MCMC convergence diagnostics, see, e.g., Cowles and Carlin (1996)].

(c) How do the Bayesian estimation results differ from OLS estimates? Can you explain the differences in the results?

Solution

(a) The priors employed specify that the expected weight at birth is 100 grams, and the growth per day is 15 grams. The prior covariance matrix C is chosen to be quite diffuse around these mean values. For the Wishart prior, we set $\rho = 5$ and choose the prior to reflect some degree of variability in the initial weights and growth rates across rats. Finally,

Table 12.2: Posterior quantities for a selection of parameters.

Parameter	Post. Mean	Post. Std.	10th Percentile	90th Percentile
α_0	106.6	2.34	103.7	109.5
β_0	6.18	.106	6.05	6.31
σ_α^2	124.5	42.41	77.03	179.52
σ_β^2	.275	.088	.179	.389
$\rho_{\alpha,\beta}$	-.120	.211	-.390	.161
α_{10}	93.76	5.24	87.09	100.59
α_{25}	86.91	5.81	79.51	94.45
β_{10}	5.70	.217	5.43	5.98
β_{25}	6.75	.243	6.44	7.05

for the variance parameter σ^2 , we have selected the hyperparameters so that the prior mean and prior standard deviation are equal to 20, a reasonably diffuse choice.

Estimation results for a selection of parameters are provided in Table 12.2. The results are obtained after running the Gibbs sampler for 10,000 iterations and discarding the first 500 of them as the burn-in period. Matlab code for this exercise is provided on the Web site associated with this book.

In Table 12.2, we present results for the common intercept α_0 , common slope β_0 , variance parameters of the second-stage covariance matrix Σ (denoted σ_α^2 and σ_β^2), the correlation between the intercept and slope, denoted $\rho_{\alpha,\beta}$, and the slope and intercept results for the 10th and 25th rat in the sample. For the covariance matrix, we report the variance parameters as well as the correlation between the intercept and slope.

Our results suggest that the birth weight of an “average” rat is about 106.6 grams, and the average weight gained per day is about 6.2 grams. For the covariance matrix results, we do see evidence of individual heterogeneity across rats as both σ_α^2 and σ_β^2 are reasonably large, and the correlation between the intercept and growth rate is mostly negative, though it is rather imprecisely estimated. This suggests that smaller rats at birth may tend to have higher growth rates, as they tend to “catch up” to the weights of the other rats.

(b) For our estimation results, we ran the Gibbs sampler for 10,000 iterations and discarded the first 500 of those draws. We now provide some suggestive evidence that these choices were adequate for convergence.

In Figure 12.1 we present trace plots for the common intercepts and slope (α_0 and β_0), the (1, 1) element of Σ (σ_α^2), and the intercept for the 20th rat in the sample (α_{20}). We obtain two different trace plots from two different chains that were run with different and over-dispersed starting values. As shown in the table, it appears as if the simulated draws from both chains appear to settle down and explore the same region remarkably quickly. In fact, after the first ten iterations or so, the progression of both chains seems very similar. This provides suggestive evidence that the use of 500 iterations as a burn-in period is more than sufficient for this application.

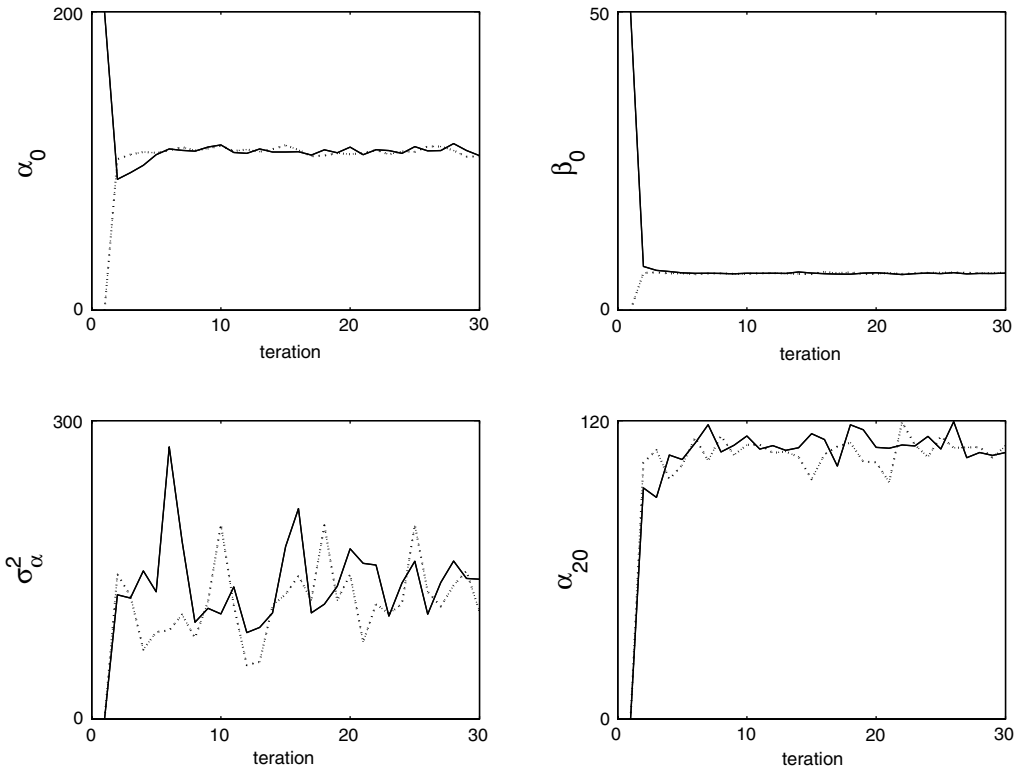


Figure 12.1 — Trace plots of selected parameters from two different chains.

We can also gauge the number of draws required by examining the degree of autocorrelation in our parameter chains. In Table 12.3 we present the within-parameter autocorrelations at different lags.

Table 12.3: Autocorrelations in parameter chains at various lag orders.

Parameter	Lag 1	Lag 5	Lag 10
α_0	.24	.010	.007
β_0	.22	−.004	−.009
σ_α^2	.36	.020	−.010
$\rho_{\alpha,\beta}$.37	.036	.003
α_{15}	.18	.025	.018

As shown in the table, even the first-order autocorrelations are not excessively large, and the lag 5 correlations are virtually zero. As such, we do not expect autocorrelation in our parameter chains to substantially slow our convergence, so the number of parameter values used here appears to be sufficient.

(c) One might think that there is a much easier way to obtain the same results. Specifically, one might be able to obtain the same results by taking the five observations corresponding

to each rat and using those five observations to obtain intercept and slope estimates via a linear regression model.

Although this OLS method is indeed computationally more attractive, this approach ignores the fact that the individual-level parameters are also involved in the second stage. That is, you can think about obtaining an estimate of α_i from the first-stage OLS regression as well as from the second-stage common mean α . In our hierarchical model, the conditional posterior distribution of the individual-level intercept and slope parameters balance information from the OLS regression using outcomes for the given rat with information obtained from the intercept and slope estimates from all other rats. This suggests that models like these should not be fit “sequentially,” where point estimates of parameters from a given stage of the hierarchy are then used as a dependent variable in subsequent stages of the hierarchy.

To show how these approaches can yield different results in our example, we provide in Figure 12.2 OLS estimates of the intercepts and slopes as well as posterior means of the same intercepts and slopes from our hierarchical model.

As shown in the figures, the hierarchical model estimates tend to be less variable and are “shrunk” toward the common mean α . The sequential OLS procedure, in general, exhibits a tendency to “overfit” the model.

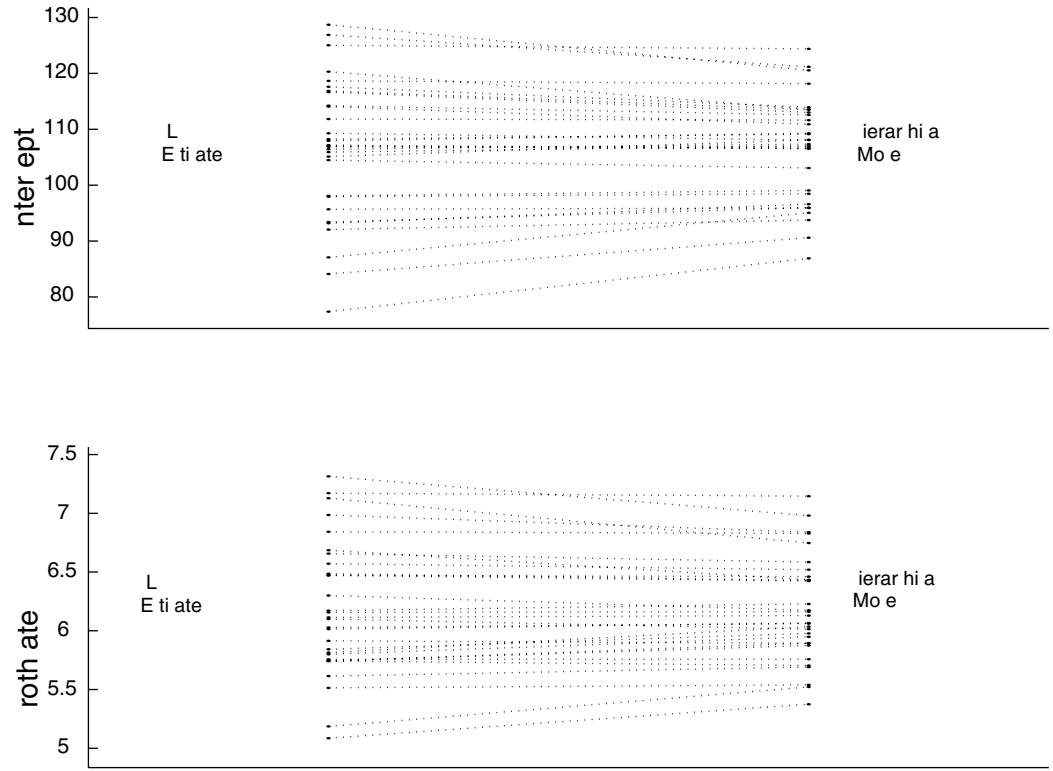


Figure 12.2 — OLS and hierarchical model point estimates of intercepts and growth rates.

Exercise 12.5 (Blocking steps in random effects models) In the rat growth model in the previous exercise, both the intercept and slope were permitted to vary across the units. Consider, instead, a model that combines some elements of the random coefficients model and some elements of the pooled model. In particular, we will permit the intercept to vary across individuals but restrict the other regression coefficients to be constant across individuals. That is, we consider the *mixed model*

$$y_{it} = \alpha_i + x_{it}\beta + \epsilon_{it}, \quad \epsilon_{it} \stackrel{\text{iid}}{\sim} N(0, \sigma_\epsilon^2),$$

$$\alpha_i \stackrel{\text{iid}}{\sim} N(\alpha, \sigma_\alpha^2).$$

For this model we employ independent priors of the form

$$\begin{aligned} \beta &\sim N(\beta_0, V_\beta), \\ \alpha &\sim N(\alpha_0, V_\alpha), \\ \sigma_\epsilon^2 &\sim IG(e_1, e_2), \\ \sigma_\alpha^2 &\sim IG(a_1, a_2). \end{aligned}$$

The model is sometimes termed a “mixed” model because the random effects α_i are permitted to vary across i , while we impose the restriction that the slope parameters β are constant across individuals. In the beginning of this exercise, we continue to assume a balanced panel ($i = 1, 2, \dots, N$, $t = 1, 2, \dots, T$), though we relax this condition in part (d).

(a) Derive the complete posterior conditionals

$$p(\alpha_i | \beta, \alpha, \sigma_\epsilon^2, \sigma_\alpha^2, y) \quad \text{and} \quad p(\beta | \{\alpha_i\}, \alpha, \sigma_\epsilon^2, \sigma_\alpha^2, y),$$

where $\{\alpha_i\}$ simply denotes the collection of the N individual random effects, $\{\alpha_i\}_{i=1}^N$.

(b) Describe how one could use a *blocking* or *grouping* step [e.g., Chib and Carlin (1999)] to obtain draws directly from the *joint posterior conditional* $p(\{\alpha_i\}, \beta | \alpha, \sigma_\epsilon^2, \sigma_\alpha^2, y)$.

(c) Describe how the Gibbs sampler can be used to fit the model, given your result in (b). Would you expect any improvements in this blocked algorithm relative to the standard Gibbs algorithm in (a)?

(d) How does your answer in (c) change for the case of an *unbalanced panel* where $T = T_i$?

Solution

(a) The complete posterior conditionals can be obtained in the same manner as in previous exercises in this chapter. Specifically, we obtain

$$p(\alpha_i | \beta, \alpha, \sigma_\epsilon^2, \sigma_\alpha^2, y) \sim N(Dd, D), \tag{12.8}$$

where

$$D = (T/\sigma_\epsilon^2 + 1/\sigma_\alpha^2)^{-1}, \quad d = \sum_{t=1}^T (y_{it} - x_{it}\beta)/\sigma_\epsilon^2 + \alpha/\sigma_\alpha^2.$$

The complete posterior conditional for β follows similarly:

$$\beta | \{\alpha_i\}, \alpha, \sigma_\epsilon^2, \sigma_\alpha^2, y \sim N(Hh, H),$$

where

$$H = \left(X'X/\sigma_\epsilon^2 + V_\beta^{-1} \right)^{-1}, \quad h = X'(y - \bar{\alpha})/\sigma_\epsilon^2 + V_\beta^{-1}\beta_0,$$

where X , y , and α are the $NT \times k$, $NT \times 1$, and $NT \times 1$ design matrix, outcome vector, and random effect vector, respectively, stacked appropriately. Specifically,

$$\bar{\alpha} = [(\iota_T \alpha_1)' (\iota_T \alpha_2)' \cdots (\iota_T \alpha_N)']'.$$

(b) Instead of the strategy described in (a), we seek to draw the random effects $\{\alpha_i\}$ and “fixed effects” β in a single block. (The notation “fixed effects” is somewhat common in statistics and is used to denote the set of parameters that do not vary across i .) This strategy of grouping together correlated parameters will generally facilitate the mixing of the chain and thereby reduce numerical standard errors associated with the Gibbs sampling estimates.

We break this joint posterior conditional into the following two pieces:

$$p(\{\alpha_i\}, \beta | \alpha, \sigma_\epsilon^2, \sigma_\alpha^2, y) = p(\{\alpha_i\} | \beta, \alpha, \sigma_\epsilon^2, \sigma_\alpha^2, y) p(\beta | \alpha, \sigma_\epsilon^2, \sigma_\alpha^2, y).$$

Because of the assumed structure of the problem, the random effects $\{\alpha_i\}$ are conditionally independent, so that

$$p(\{\alpha_i\}, \beta | \alpha, \sigma_\epsilon^2, \sigma_\alpha^2, y) = \left[\prod_{i=1}^N p(\alpha_i | \beta, \alpha, \sigma_\epsilon^2, \sigma_\alpha^2, y) \right] p(\beta | \alpha, \sigma_\epsilon^2, \sigma_\alpha^2, y).$$

This suggests that one can draw from this joint posterior conditional via the method of composition by first drawing from $p(\beta | \alpha, \sigma_\epsilon^2, \sigma_\alpha^2, y)$ and then drawing each α_i independently from its complete posterior conditional distribution in (12.8).

We now derive the conditional posterior distribution for β , marginalized over the random effects. Note that our model can be rewritten as follows:

$$y_{it} = \alpha + x_{it}\beta + u_i + \epsilon_{it},$$

$$\beta \sim N(\beta_0, V_\beta),$$

where $u_i \stackrel{\text{iid}}{\sim} N(0, \sigma_\alpha^2)$. Let $v_{it} = u_i + \epsilon_{it}$. If we stack this equation over t within i we obtain

$$y_i = \iota_T \alpha + x_i \beta + v_i,$$

where

$$E(v_i v_i') \equiv \Sigma = \sigma_\epsilon^2 I_T + \sigma_\alpha^2 \iota_T \iota_T'$$

and $y_i = [y_{i1} \ y_{i2} \ \cdots \ y_{iT}]'$, $x_i = [x'_{i1} \ x'_{i2} \ \cdots \ x'_{iT}]'$, and $v_i = [v_{i1} \ v_{i2} \ \cdots \ v_{iT}]'$. Stacking again over i we obtain

$$y = i_{NT} \alpha + X \beta + v,$$

where

$$E(v v') = I_N \otimes \Sigma$$

and $y = [y'_1 \ y'_2 \ \cdots \ y'_N]'$, $X = [x'_1 \ x'_2 \ \cdots \ x'_N]'$, and $v = [v'_1 \ v'_2 \ \cdots \ v'_N]'$. In this form, we can now appeal to our standard results for the regression model to obtain

$$\beta | \alpha, \sigma_\epsilon^2, \sigma_\alpha^2, y \sim N(Gg, G),$$

where

$$G = \left(X'(I_N \otimes \Sigma^{-1})X + V_\beta^{-1} \right)^{-1} = \left(\sum_{i=1}^N x'_i \Sigma^{-1} x_i + V_\beta^{-1} \right)^{-1}$$

and

$$g = X'(I_N \otimes \Sigma^{-1})(y - \iota_{NT}\alpha) + V_\beta^{-1}\beta_0 = \sum_{i=1}^N x'_i \Sigma^{-1}(y_i - \iota_T\alpha) + V_\beta^{-1}\beta_0.$$

Thus, to sample from the desired joint conditional, you first sample β from the distribution given here and then sample the random effects independently from their complete conditional posterior distributions.

Finally, it is also worth noting that β and α could be drawn together in the first step of this process. That is, one could draw from the joint posterior conditional

$$p(\{\alpha_i\}, \alpha, \beta | \sigma_\epsilon^2, \sigma_\alpha^2, y) = \left[\prod_{i=1}^N p(\{\alpha_i\} | \beta, \alpha, \sigma_\epsilon^2, \sigma_\alpha^2, y) \right] p(\beta, \alpha | \sigma_\epsilon^2, \sigma_\alpha^2, y)$$

in a similar way as described here.

(c) Given the result in (b), we now need to obtain the remaining complete conditionals. These are

$$\alpha | \{\alpha_i\}, \beta, \sigma_\epsilon^2, \sigma_\alpha^2, y \sim N(Rr, R),$$

where

$$R = (N/\sigma_\alpha^2 + V_\alpha^{-1})^{-1}, \quad r = \sum_{i=1}^N \alpha_i/\sigma_\alpha^2 + V_\alpha^{-1}\alpha_0,$$

$$\sigma_\alpha^2 | \{\alpha_i\}, \beta, \sigma_\epsilon^2, y \sim IG \left((N/2) + a_1, \left[a_2^{-1} + .5 \sum_{i=1}^N (\alpha_i - \alpha)^2 \right]^{-1} \right),$$

$$\sigma_\epsilon^2 | \{\alpha_i\}, \beta, \sigma_\alpha^2, y \sim IG \left((NT/2) + e_1, [e_2^{-1} + .5(y - \iota_{NT}\alpha - X\beta)'(y - \iota_{NT}\alpha - X\beta)]^{-1} \right).$$

(d) Only slight changes are required in the case of unbalanced panels. Let NT continue to denote the total number of observations (i.e., $NT \equiv \sum_{i=1}^N T_i$). In addition, let $\Sigma_i \equiv \sigma_\epsilon^2 I_{T_i} + \sigma_\alpha^2 \iota_{T_i} \iota'_{T_i}$. Then, for the unbalanced panel model, we will obtain the same complete conditionals as described in (c), upon replacing Σ^{-1} with Σ_i^{-1} , and T with T_i in the complete conditional for α_i and stacking x_i y_i , X , and y appropriately in the unbalanced model.

***Exercise 12.6 (Hierarchical recentering: Gelfand, Sahu, and Carlin [1995])** Consider the following hierarchical model:

$$y_i = \iota_T \eta_i + \epsilon_i, \quad i = 1, 2, \dots, n, \quad (12.9)$$

where $y_i = [y_{i1} \ y_{i2} \ \dots \ y_{iT}]'$, $\epsilon_i = [\epsilon_{i1} \ \epsilon_{i2} \ \dots \ \epsilon_{iT}]' \stackrel{\text{iid}}{\sim} N(0, \sigma_\epsilon^2 I_T)$, η_i is a scalar, and ι_T is a $T \times 1$ vector of ones. Suppose further that the following prior (sometimes referred to as a *population model*) for the η_i is employed:

$$\eta_i | \mu \stackrel{\text{iid}}{\sim} N(\mu, \sigma_\eta^2). \quad (12.10)$$

Finally, assume that the variance parameters σ_η^2 and σ_ϵ^2 are *known* and that a flat prior $p(\mu) \propto c$, for some constant c , is adopted.

(a) Find the marginal and conditional posterior distributions $\mu|y$, $\eta_i|\mu, y$ and $\eta_i|y$, where $y = [y'_1 \ y'_2 \ \dots \ y'_n]'$.

(b) Calculate $\text{Cov}(\eta_i, \mu|y)$ and $\text{Corr}(\eta_i, \mu|y)$.

(c) Consider an equivalent formulation of the model:

$$y_i = \iota_T \mu + \iota_T (\eta_i - \mu) + \epsilon_i, \quad (12.11)$$

$$\eta_i - \mu \stackrel{\text{iid}}{\sim} N(0, \sigma_\eta^2), \quad (12.12)$$

or, by letting $\alpha_i \equiv \eta_i - \mu$,

$$y_i = \iota_T \mu + \iota_T \alpha_i + \epsilon_i, \quad (12.13)$$

$$\alpha_i \stackrel{\text{iid}}{\sim} N(0, \sigma_\eta^2). \quad (12.14)$$

As in Gelfand, Sahu, and Carlin (1995), call the model in (μ, η) space [i.e., (12.9) and (12.10)] the *centered* parameterization, and call the model in the (μ, α) space [i.e., (12.13) and (12.14)] the *uncentered* parameterization.

Derive $\mu|y$ and $\alpha_i|y$ for the uncentered model.

(d) Use the results in (a) through (c) to calculate $\text{Cov}(\alpha_i, \mu|y)$ and $\text{Corr}(\alpha_i, \mu|y)$. Can you establish any ordering between the correlations in the (μ, α) space and the (μ, η) space? What might this imply for the performance of a posterior simulator under each parameterization?

(e) For the particular values $N = 10$, $T = 2$, $\sigma_\epsilon^2 = 1$ and $\sigma_\eta^2 = 10$, calculate $\text{Corr}(\eta_i, \mu|y)$ and $\text{Corr}(\alpha_i, \mu|y)$.

(f) Generate data consistent with the sample sizes and parameter values in (e). Using these data, fit both the centered and uncentered models using the Gibbs sampler. Calculate the first-order autocorrelation coefficients for a represented set of simulated parameters in each case and comment on your results.

Solution

(a) Given that the variance parameters are known, it follows using reasoning similar to that of Exercise 12.3 that the joint posteriors are multivariate normal under each

parameterization. In terms of the specific quantities required in part (a), we first note, upon marginalizing over η_i ,

$$y_i|\mu \sim N(\iota_T\mu, \sigma_\epsilon^2 I_T + \sigma_\eta^2 \iota_T \iota_T'). \quad (12.15)$$

Let $\Sigma_T \equiv \sigma_\epsilon^2 I_T + \sigma_\eta^2 \iota_T \iota_T'$. Under the flat prior for μ , it follows that

$$\mu|y \sim N(\hat{\mu}, V_\mu), \quad (12.16)$$

where

$$V_\mu = (n\iota_T' \Sigma_T^{-1} \iota_T)^{-1} \quad \text{and} \quad \hat{\mu} = V_\mu \left[\sum_{i=1}^n \iota_T' \Sigma_T^{-1} y_i \right].$$

Using derivations like those of Exercises 12.2 and 12.3, we find

$$\eta_i|u, y \sim N(Bb_i, B), \quad (12.17)$$

where

$$B = \frac{\sigma_\epsilon^2 \sigma_\eta^2}{T\sigma_\eta^2 + \sigma_\epsilon^2} \quad \text{and} \quad b_i = \left[\sum_{t=1}^T y_{it}/\sigma_\epsilon^2 \right] + (\mu/\sigma_\eta^2).$$

Note that the marginal posterior for η_i , will also be normal, and thus it remains to calculate its mean and variance. For the mean,

$$E(\eta_i|y) = E_{\mu|y}[E(\eta_i|\mu, y)] = BE_{\mu|y}(b_i) = B\hat{b}_i, \quad (12.18)$$

where $\hat{b}_i \equiv [\sum_t y_{it}/\sigma_\epsilon^2] + \hat{\mu}/\sigma_\eta^2$. To calculate the variance, we make use of a well-known identity:

$$\begin{aligned} \text{Var}(\eta_i|y) &= E_{\mu|y}[\text{Var}(\eta_i|\mu, y)] + \text{Var}_{\mu|y}[E(\eta_i|\mu, y)] \\ &= B + \text{Var}_{\mu|y}(Bb_i) \\ &= B + B^2 \text{Var}_{\mu|y}(b_i) \\ &= B + [B/\sigma_\eta^2]^2 V_\mu. \end{aligned} \quad (12.19)$$

Thus, the marginal posterior $\eta_i|y$ is normal with mean (12.18) and variance (12.19).

(b) Consider $\text{Cov}(\eta_i, \mu|y)$. The densities $\eta_i|\mu, y$ and $\mu|y$ derived in part (a) define the joint (normal) density $\eta_i, \mu|y$. Using well-known properties of the conditional normal distribution (see Appendix Theorem 3), we can obtain the desired covariance by carefully examining the conditional mean $E(\eta_i|\mu, y)$. Specifically, from (12.17), we can write

$$\begin{aligned} E(\eta_i|\mu, y) &= B \sum_t y_{it}/\sigma_\epsilon^2 + (B/\sigma_\eta^2)\mu \\ &= \left[B \sum_t y_{it}/\sigma_\epsilon^2 + (B/\sigma_\eta^2)\hat{\mu} \right] + (B/\sigma_\eta^2)(\mu - \hat{\mu}) \\ &= \left[B \sum_t y_{it}/\sigma_\epsilon^2 + (B/\sigma_\eta^2)\hat{\mu} \right] (BV_\mu/\sigma_\eta^2)(1/V_\mu)(\mu - \hat{\mu}). \end{aligned}$$

In this last form it becomes clear (again using properties of the conditional normal distribution) that

$$\text{Cov}(\eta_i, \mu|y) = BV_\mu/\sigma_\eta^2. \quad (12.20)$$

The correlation coefficient $\text{Corr}(\eta_i, \mu|y)$ simply takes the covariance in (12.20) and divides it by the posterior standard deviations for η_i and μ implied from (12.19) and (12.16), respectively.

(c) Upon marginalizing (12.13) over the prior for α_i in (12.14), a density $y_i|\mu$ is obtained that is identical to (12.15). Since we continue to employ a flat prior for μ , it follows that $\mu|y$ is identical to the posterior in (12.16).

The marginal posterior $\alpha_i|y$, like all of the posteriors here with known variance parameters, will also be normal. It thus remains to characterize its mean and variance. For its mean, we have

$$E(\alpha_i|y) = E(\eta_i - \mu|y) = E(Bb_i - \mu|y) = B\hat{b}_i - \hat{\mu}, \quad (12.21)$$

and for its variance, we have

$$\begin{aligned} \text{Var}(\alpha_i|y) &= \text{Var}(\eta_i - \mu|y) = \text{Var}(\eta_i|y) + \text{Var}(\mu|y) - 2\text{Cov}(\eta_i, \mu|y) \\ &= B + [B/\sigma_\eta^2]^2 V_\mu + V_\mu - 2[BV_\mu/\sigma_\eta^2], \end{aligned} \quad (12.22)$$

using previous results in parts (a) and (b). Thus, $\alpha_i|y$ is normal with mean (12.21) and variance (12.22).

(d) Using the results of the previous parts of this exercise, we obtain

$$\text{Cov}(\alpha_i, \mu|y) = \text{Cov}(\eta_i - \mu, \mu|y) = \text{Cov}(\eta_i, \mu|y) - \text{Var}(\mu|y) = [BV_\mu/\sigma_\eta^2] - V_\mu. \quad (12.23)$$

Gelfand, Sahu, and Carlin (1995) argue that the choice of parameterization depends critically on the value of B/σ_η^2 . Specifically, when $0 \leq B/\sigma_\eta^2 \leq 1$ is near zero, $\text{Corr}(\eta_i, \mu) < \text{Corr}(\alpha_i, \mu)$, and thus the centered parameterization is to be preferred. Conversely, when B/σ_η^2 is near one, $\text{Corr}(\eta_i, \mu) > \text{Corr}(\alpha_i, \mu)$, and thus the uncentered parameterization will be preferred. The idea behind this reasoning is that, all else being equal, it is better to have little correlation among the parameters of our posterior distribution, as highly correlated parameters will negatively affect the performance and convergence behavior of a posterior simulator (see, e.g., Exercise 11.7). In this particular context, if the researcher has strong prior information about the magnitude of the fraction $B/\sigma_\eta^2 = \sigma_\epsilon^2/(\sigma_\epsilon^2 + T\sigma_\eta^2)$, such information would suggest the best parameterization to employ. In the following two parts of this exercise, we quantify how the posterior correlation structure changes with the parameterization employed, and using generated data, we document how a posterior simulator can be affected by this change in correlation structure.

(e) By using the given parameter values and substituting into the expressions in (12.16), (12.17), (12.19), and (12.20), it follows that $V_\mu = 1.05$, $B = .476$, $\text{Var}(\eta_i|y) = .478$, $\text{Cov}(\eta_i, \mu|y) = .05$, and thus $\text{Corr}(\eta_i, \mu|y) = .0706$.

When performing the same calculations for the (μ, α) space, we obtain $\text{Var}(\alpha_i|y) = 1.43$, $\text{Cov}(\alpha_i, \mu|y) = -1$, and thus $\text{Corr}(\alpha_i, \mu|y) = -.816$.

In this particular example, $B/\sigma_\eta^2 = .0476$ is small, and thus the results of Gelfand, Sahu, and Carlin (1995) suggest working with the centered parameterization. Our calculations do indeed show that, with this experimental design, the centered parameterization yields a posterior with less correlation among these elements.

(f) On the Web site associated with this book, we provide MATLAB code for four posterior simulators and apply each of these simulators to data generated according to the design described in (e). For both the centered and uncentered parameterizations, we fit the model using both the Gibbs sampler and Monte Carlo integration. For the case of Monte Carlo integration, we first draw from the marginal posterior of μ in (12.16) and then draw independently from the posterior conditionals for η_i or α_i .

The results of these experiments echoed the conclusions of part (e) – the calculated correlations using both the Monte Carlo and Gibbs output were close to the analytical values derived in (e). For the case of the Gibbs algorithm, we also note a tremendous increase in the first-order autocorrelation coefficients associated with our simulated parameter draws when using the uncentered parameterization. The lag-1 correlation between the μ posterior simulations, for example, in the (μ, α) space exceeded .9, whereas the same correlation in the (μ, η) space was less than .05. This example suggests that parameterization issues can be quite important and can substantially impact the performance of a posterior simulator. Such a conclusion may seem surprising, since the difference between parameterizations in this simple example seems innocuous, yet leads to posterior simulators with substantially different performances.

Exercise 12.7 (A hierarchical/smoothness prior approach to nonparametric regression) Consider the nonparametric regression model

$$y_i = f(x_i) + \epsilon_i, \quad i = 1, 2, \dots, n, \quad (12.24)$$

where $\epsilon_i \stackrel{\text{iid}}{\sim} N(0, \sigma^2)$, and x_i is a scalar, and suppose that the data have been sorted so that $x_1 < x_2 < \dots < x_n$. (For simplicity, consider the case where x is continuous so that, within sample, $x_j \neq x_{j-1}$. When different observations share the same x values, the framework presented here can be modified in a straightforward fashion to accommodate this feature of the data.)

Let $\theta_i = f(x_i)$ and write $y_i = \theta_i + \epsilon_i$, so that every point on the nonparametric regression line is treated as a new parameter.

Define $\Delta_j = x_j - x_{j-1}$, $j = 2, 3, \dots, n$, and construct an $n \times n$ matrix H as follows:

$$H = \begin{bmatrix} 1 & 0 & 0 & 0 & \cdots & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & \cdots & 0 & 0 & 0 \\ \Delta_2^{-1} - (\Delta_2^{-1} + \Delta_3^{-1}) & \Delta_3^{-1} & 0 & \cdots & 0 & 0 & 0 \\ 0 & \Delta_3^{-1} & -(\Delta_3^{-1} + \Delta_4^{-1}) & \Delta_4^{-1} & \cdots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & \Delta_{n-1}^{-1} - (\Delta_{n-1}^{-1} + \Delta_n^{-1}) & \Delta_n^{-1} \end{bmatrix}.$$

Let $\theta = [\theta_1 \ \theta_2 \ \cdots \ \theta_n]'$, and suppose a prior of the following form is employed:

$$H\theta|\eta \sim N\left(\begin{bmatrix} \mu_1 \\ 0 \end{bmatrix}, \begin{bmatrix} V_1 & 0 \\ 0 & \eta I_{n-2} \end{bmatrix}\right) \equiv N(\mu, V), \quad (12.25)$$

where μ_1 is 2×1 and V_1 is 2×2 .

(a) Interpret this prior, paying particular attention to the role of η . Using a prior for σ^2 of the form $\sigma^2 \sim IG(a, b)$, reparameterize the model in terms of $\gamma = H\theta$ and derive a Gibbs sampling algorithm for fitting the reparameterized nonparametric regression model.

(b) Generate $n = 200$ observations from the following regression model:

$$y_i = .15x_i + .3 \exp[-4(x_i + 1)^2] + .7 \exp[-16(x_i - 1)^2] + \epsilon_i,$$

where $x_i \stackrel{\text{iid}}{\sim} U(-2, 2)$ (with U denoting a uniform distribution; see Appendix Definition 1) and $\epsilon_i \stackrel{\text{iid}}{\sim} N(0, .01)$. Using the algorithm in part (a), fit the model for three different values of η : .00001, .15, and 100. Plot the posterior mean of θ in each case and comment on the results.

Solution

(a) First, note that $H\theta$ serves to transform the coefficient vector into a vector of “initial conditions” (i.e., the first two points on the regression curve) and first differences in pointwise slopes of the curve. That is,

$$\gamma = H\theta = \begin{bmatrix} \theta_1 \\ \theta_2 \\ \frac{\theta_3 - \theta_2}{\Delta_3} - \frac{\theta_2 - \theta_1}{\Delta_2} \\ \frac{\theta_4 - \theta_3}{\Delta_4} - \frac{\theta_3 - \theta_2}{\Delta_3} \\ \vdots \\ \frac{\theta_n - \theta_{n-1}}{\Delta_n} - \frac{\theta_{n-1} - \theta_{n-2}}{\Delta_{n-1}} \end{bmatrix},$$

where terms of the form $(\theta_j - \theta_{j-1})/\Delta_j$ can be interpreted as pointwise derivatives of f at x_{j-1} .

The parameter η controls how “tightly” these first differences are centered around zero. When η is very small, the prior restricts these differences to be approximately equal. As a result, the prior will essentially restrict the regression function to be linear, with the initial conditions serving to define the intercept and slope of the line. When η is quite large, little prior information is imposed in the analysis; as a result, the posterior mean of $\theta = H^{-1}\gamma$ will be excessively “jumpy.” For intermediate η , the prior is designed to embed a degree of smoothness into the regression function.

For a posterior simulator, the prior for γ is given in (12.25). The nonparametric regression model can also be rewritten as follows:

$$\begin{aligned} y &= I_n \theta + \epsilon \\ &= H^{-1} \gamma + \epsilon, \end{aligned} \quad (12.26)$$

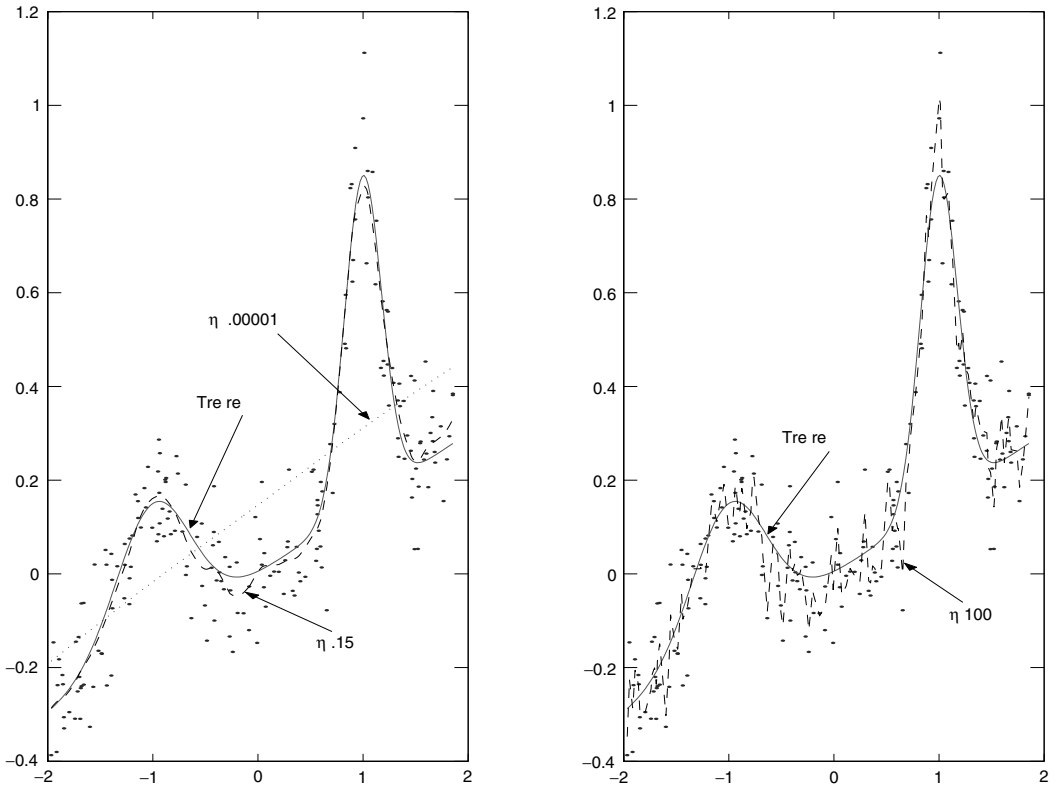


Figure 12.3 — True curve and posterior means of regression function under different values of η .

where $y = [y_1 \ y_2 \ \cdots \ y_n]'$ and $\epsilon = [\epsilon_1 \ \epsilon_2 \ \cdots \ \epsilon_n]'$, and thus we are in the framework of a normal linear regression model with normal and inverted gamma priors, as in Exercise 11.8.

A Gibbs sampler for the model in (12.26) under the given priors thus follows immediately:

$$\gamma | \sigma^2, y \sim N(D_\gamma d_\gamma, D_\gamma), \quad (12.27)$$

where

$$D_\gamma = (H^{-1'} H^{-1} / \sigma^2 + V^{-1})^{-1}, \quad d_\gamma = H^{-1'} y / \sigma^2 + V^{-1} \mu.$$

Similarly,

$$\sigma^2 | y \sim IG\left(\frac{n}{2} + a, [b^{-1} + (1/2)(y - H^{-1}\gamma)'(y - H^{-1}\gamma)]^{-1}\right). \quad (12.28)$$

The posterior simulator involves sampling from (12.27) and (12.28). To recover the original parameter vector θ , which represents $f(x)$ at each data point, simply calculate $\theta = H^{-1}\gamma$ at each iteration.

(b) Using data generated as in part (b) of this exercise, we run the Gibbs sampler for 250 iterations for each of the three values of η . The first 50 of these are discarded as the burn-in period. In Figure 12.3 we present posterior means of θ for this exercise. The leftmost graph presents the scatter plot, true curve, and posterior means when $\eta = .0001$ and $\eta = .15$. As you can see, when η is very small, the regression curve is constrained to be linear. When $\eta = .15$, the posterior mean matches the true curve reasonably well. The rightmost graph presents the posterior mean when $\eta = 100$. With weak prior information with this value of η , we obtain an undersmoothed, erratic point estimate of the regression function.

The linear regression model with general covariance matrix

Chapter 10 contained many questions relating to the normal linear regression model under classical assumptions. This model assumed the error covariance matrix to be $h^{-1}I$. Econometricians are often interested in relaxing this assumption to allow for factors such as heteroscedasticity, autocorrelation, etc. Other specifications such as the seemingly unrelated regressions model and some panel data models also require relaxation of this assumption. This chapter contains both theoretical and empirical exercises relating to such factors. To establish notation, all exercises in this chapter will involve the linear regression model defined as in Chapter 10:

$$y = X\beta + \epsilon, \quad (13.1)$$

except that we now assume ϵ to be $N(0_N, h^{-1}\Omega)$, where Ω is an $N \times N$ positive definite matrix. Some of the exercises involve posterior simulators and most typically impose additional structure on Ω to achieve identification. We remind the reader that Chapter 11 contains exercises relating to relevant MCMC algorithms such as the Gibbs sampler and M–H algorithm.

Two of the exercises in this chapter require empirical work using a multiple regression model. The reader can use any data set. However, the answers in the following are based on the data set HPRICE.TXT (available on the Web site associated with this book) containing the sales price of $N = 546$ houses sold in Windsor, Canada, in 1987. Interest centers on finding out which factors affect house prices and, hence, sales price is our dependent variable. We use four explanatory variables: the size of the lot the house is on, the number of bedrooms, the number of bathrooms, and the number of stories. Thus, we have the following:

- y_i = sales price of the i th house measured in Canadian dollars,
- x_{i2} = the lot size of the i th house measured in square feet,

x_{i3} = the number of bedrooms in the i th house,

x_{i4} = the number of bathrooms in the i th house,

x_{i5} = the number of storeys in the i th house.

Exercise 13.1 (Theoretical results for the general case)

Assume a prior of the form

$$p(\beta, h, \Omega) = p(\beta) p(h) p(\Omega), \quad (13.2)$$

where

$$p(\beta) = \phi(\beta; \underline{\beta}, \underline{V}), \quad (13.3)$$

$$p(h) = f_{\gamma}(\underline{s}^{-2}, \underline{\nu}), \quad (13.4)$$

and $p(\Omega)$ is left unspecified.

(a) Derive the full posterior conditionals for this model: $p(\beta|y, h, \Omega)$, $p(h|y, \beta, \Omega)$, and $p(\Omega|y, \beta, h)$.

(b) Using your results for part (a), discuss how a Gibbs sampler could be set up for the case where Ω is known.

Solution

(a) A standard theorem in matrix algebra [see, e.g., Koop (2003), p. 316] says that, since Ω is positive definite, an $N \times N$ matrix P exists with the property that $P\Omega P' = I_N$. If we multiply both sides of (13.1) by P , we obtain a transformed model

$$y^* = X^* \beta + \epsilon^*, \quad (13.5)$$

where $y^* = Py$, $X^* = PX$, $\epsilon^* = P\epsilon$, and ϵ^* is $N(0_N, h^{-1}I_N)$. If we work with this transformed model, the derivation of the posterior is very similar to Exercise 10.1. The only difference arises since Exercise 10.1 used a natural conjugate normal-gamma prior, whereas here we are using an independent normal-gamma prior. The following material provides details.

Using properties of the multivariate normal distribution gives us the likelihood function:

$$L(\beta, h, \Omega) = \frac{h^{\frac{N}{2}}}{(2\pi)^{\frac{N}{2}}} \left\{ \exp \left[-\frac{h}{2} (y^* - X^* \beta)' (y^* - X^* \beta) \right] \right\}. \quad (13.6)$$

Multiplying prior times the likelihood gives us the posterior:

$$\begin{aligned} p(\beta, h, \Omega|y) &\propto p(\Omega) \\ &\times \left\{ \exp \left[-\frac{1}{2} \left\{ h (y^* - X^* \beta)' (y^* - X^* \beta) + (\beta - \underline{\beta})' \underline{V}^{-1} (\beta - \underline{\beta}) \right\} \right] \right\} \\ &\times h^{\frac{N+\underline{\nu}-2}{2}} \exp \left[-\frac{h\underline{\nu}}{2\underline{s}^{-2}} \right]. \end{aligned} \quad (13.7)$$

It can be verified that this posterior is not of the form of any common density. However, we can collect terms involving β to figure out the conditional posterior $p(\beta|y, h, \Omega)$. If we do this, using the same derivations as in the solution to Exercise 10.1, we obtain

$$\beta|y, h, \Omega \sim N(\bar{\beta}, \bar{V}), \quad (13.8)$$

where

$$\bar{V} = (\underline{V}^{-1} + hX'\Omega^{-1}X)^{-1} \quad (13.9)$$

and

$$\bar{\beta} = \bar{V} \left(\underline{V}^{-1}\underline{\beta} + hX'\Omega^{-1}X\hat{\beta}(\Omega) \right), \quad (13.10)$$

where

$$\hat{\beta}(\Omega) = (X^*X^*)^{-1}X^*y^* = (X'\Omega^{-1}X)^{-1}X'\Omega^{-1}y. \quad (13.11)$$

Similarly, if we collect terms involving h to figure out the conditional posterior $p(h|y, \beta, \Omega)$, we obtain

$$h|y, \beta, \Omega \sim \gamma(\bar{s}^{-2}, \bar{\nu}), \quad (13.12)$$

where

$$\bar{\nu} = N + \underline{\nu} \quad (13.13)$$

and

$$\bar{s}^2 = \frac{(y - X\beta)' \Omega^{-1} (y - X\beta) + \underline{\nu} \underline{s}^2}{\bar{\nu}}. \quad (13.14)$$

The posterior for Ω conditional on β and h has a kernel of the form

$$p(\Omega|y, \beta, h) \propto p(\Omega) |\Omega|^{-\frac{1}{2}} \left\{ \exp \left[-\frac{h}{2} (y - X\beta)' \Omega^{-1} (y - X\beta) \right] \right\}. \quad (13.15)$$

(b) If we knew Ω , then we could simply plug its value into the posterior conditionals $p(\beta|y, h, \Omega)$ and $p(h|y, \beta, \Omega)$. Thus a Gibbs sampler involving (13.8) and (13.12) involves only the normal and gamma distributions. When Ω is unknown, as we shall see in the following, Gibbs samplers for a wide variety of models with general error covariance structure involve blocks of this form for $p(\beta|y, h, \Omega)$ and $p(h|y, \beta, \Omega)$, and that differ only in the form for $p(\Omega|y, \beta, h)$.

Exercise 13.2 (Heteroscedasticity of a known form) Heteroscedasticity occurs if

$$\Omega = \begin{bmatrix} \omega_1 & 0 & \cdots & 0 \\ 0 & \omega_2 & 0 & \cdots \\ \vdots & 0 & \cdots & \vdots \\ \vdots & \cdots & \cdots & 0 \\ 0 & \cdots & 0 & \omega_N \end{bmatrix}. \quad (13.16)$$

In this exercise, assume that

$$\omega_i = (1 + \alpha_1 z_{i1} + \alpha_2 z_{i2} + \cdots + \alpha_p z_{ip})^2, \quad (13.17)$$

where z_i is a p vector of data that may include some or all of the explanatory variables. Let $\alpha = [\alpha_1 \ \alpha_2 \ \cdots \ \alpha_p]'$.

(a) Assuming a prior of the form given in (13.2), (13.3), and (13.4), derive the full posterior conditionals for this model.

(b) Assuming a noninformative prior for the parameters that define Ω ,

$$p(\alpha) \propto 1,$$

describe how a Metropolis-within-Gibbs algorithm can be developed for carrying out Bayesian inference in this model.

(c) Program the algorithm described in part (b) and carry out a Bayesian investigation of heteroscedasticity using a suitable data set (e.g., the house price data set described at the beginning of this chapter) and choices for the prior hyperparameters $\underline{\beta}$, \underline{V} , $\underline{\nu}$, and \underline{s}^{-2} .

Solution

(a) The theoretical derivations are exactly the same as in Exercise 13.1. The conditional posterior for Ω has the form given in (13.15). There are a few ways of simplifying this expression a little bit (e.g., using $|\Omega| = \prod \omega_i$), but none are that instructive.

(b) The results for part (a) suggest that a Metropolis-within-Gibbs algorithm can be developed since $p(\beta|y, h, \alpha)$ is normal and $p(h|y, \beta, \alpha)$ is gamma and we require only a method for taking draws from $p(\alpha|y, \beta, h)$. The latter density does not have a convenient form, which means that a M–H step is required. In this solution, we will use a random walk chain M–H algorithm, although other algorithms could be used. Equation (13.15) evaluated at old and candidate draws is used to calculate the acceptance probability.

(c) The necessary Matlab code is provided on the Web site associated with this book. In this solution, we provide our empirical results using the house price data set. The variance of the proposal density, which we will call Σ , is chosen by first setting $\Sigma = cI$ and experimenting with different values of the scalar c until a value is found that yields reasonable acceptance probabilities. The posterior simulator is then run using this value to yield an estimate of the posterior variance of α , $\widehat{\text{Var}}(\alpha|y)$. We then set $\Sigma = c\widehat{\text{Var}}(\alpha|y)$ and experiment with different values of c until we find one that yields an average acceptance probability of roughly .50. Then a final long run of 30,000 replications, with 5,000 burn-in replications discarded, is taken. For the prior hyperparameters we choose $\underline{\nu} = 5$, $\underline{s}^{-2} = 4.0 \times 10^{-8}$,

$$\underline{\beta} = \begin{bmatrix} 0.0 \\ 10 \\ 5,000 \\ 10,000 \\ 10,000 \end{bmatrix},$$

Table 13.1: Posterior results for β , h , and α .

	Mean	Standard Deviation	95% HPDI
β_1	-5,455.66	2,941.55	$[-10281, -626]$
β_2	6.27	0.42	$[5.59, 6.97]$
β_3	3,034.73	999.93	$[1392, 4678]$
β_4	14,207.85	1,602.56	$[11597, 16901]$
β_5	7,814.81	922.64	$[6290, 9327]$
h	7.90×10^{-8}	2.55×10^{-8}	$[4 \times 10^{-8}, 1 \times 10^{-7}]$
α_1	5.02×10^{-4}	1.29×10^{-4}	$[3 \times 10^{-4}, 8 \times 10^{-4}]$
α_2	0.39	0.13	$[0.16, 0.56]$
α_3	0.32	0.07	$[0.20, 0.41]$
α_4	-0.22	0.08	$[-0.33, -0.09]$

and

$$\underline{V} = \begin{bmatrix} 10,000^2 & 0 & 0 & 0 & 0 \\ 0 & 5^2 & 0 & 0 & 0 \\ 0 & 0 & 2,500^2 & 0 & 0 \\ 0 & 0 & 0 & 5,000^2 & 0 \\ 0 & 0 & 0 & 0 & 5,000^2 \end{bmatrix}.$$

It can be confirmed that the resulting prior is located in a sensible region of the parameter space but is relatively noninformative.

Table 13.1 contains posterior means, standard deviations, and 95% HPDIs for all parameters in the model. Results indicate that heteroscedasticity does seem to exist for this data set. That is, the 95% HPD intervals do not include zero for α_1 , α_2 , α_3 , and α_4 .

Exercise 13.3 (Heteroscedasticity of an unknown form: student t errors) Now assume that you have heteroscedasticity as in (13.16), but you do not know its exact form (i.e., you do not know Equation 13.17) but are willing to assume all of the error variances are drawn from a common distribution, the gamma. If we work with error precisions rather than variances and, hence, define $\lambda \equiv [\lambda_1 \ \lambda_2 \ \cdots \ \lambda_N]'$ $\equiv [\omega_1^{-1} \ \omega_2^{-1} \ \cdots \ \omega_N^{-1}]'$ we assume the following hierarchical prior for λ :

$$p(\lambda|\nu_\lambda) = \prod_{i=1}^N f_\gamma(\lambda_i|1, \nu_\lambda). \quad (13.18)$$

This prior is hierarchical because λ depends on a parameter, ν_λ , which in turn has its own prior, $p(\nu_\lambda)$. We will assume the latter to be exponential:

$$p(\nu_\lambda) = f_\gamma(\nu_\lambda|\underline{\nu}_\lambda, 2). \quad (13.19)$$

As an aside, we note that this specification is a popular one since it can be shown (if we integrate out λ) that we obtain a linear regression model with Student t errors. That is,

$$p(\epsilon_i) = f_t(\epsilon_i | 0, h^{-1}, \nu_\lambda).$$

The interested reader is referred to Exercise 15.1 or Geweke (1993) for proof and further details. Geweke (1993) shows that if you use a common noninformative prior for β [i.e., $p(\beta) \propto 1$ on the interval $(-\infty, \infty)$], then the posterior mean does not exist, unless $p(\nu_\lambda)$ is zero on the interval $(0, 2]$. The posterior standard deviation does not exist unless $p(\nu_\lambda)$ is zero on the interval $(0, 4]$. Hence, the researcher who wants to use a noninformative prior for β should either use a prior that excludes small values for ν_λ or present posterior medians and interquartile ranges (which will exist for any valid p.d.f.). With an informative normal prior for β , the posterior mean and standard deviation of β will exist.

(a) Assume a prior given by (13.2), (13.3), (13.4), (13.18), and (13.19). Derive the full posterior conditionals for this model and, using these, discuss how posterior simulation can be done.

(b) Program the algorithm described in part (a) and carry out an empirical Bayesian analysis using a suitable data set (e.g., the house price data set described at the beginning of this chapter) and suitable choices for prior hyperparameters.

Solution

(a) The posterior conditionals $p(\beta | y, h, \lambda)$ and $p(h | y, \beta, \lambda)$ will be exactly as in (13.8) and (13.12) with the precise form for Ω given in (13.16) plugged in. Thus we need only derive $p(\lambda | y, \beta, h, \nu_\lambda)$ and $p(\nu_\lambda | y, \beta, h, \lambda)$. The former of these can be derived by plugging the prior given in (13.18) into the general form for the conditional posterior given in (13.15). An examination of the resulting density shows that the λ_i 's are independent of one another (conditional on the other parameters of the model) and each of the conditional posteriors for λ_i has the form of a gamma density. Thus, we have

$$p(\lambda | y, \beta, h, \nu_\lambda) = \prod_{i=1}^N p(\lambda_i | y, \beta, h, \nu_\lambda) \quad (13.20)$$

and

$$p(\lambda_i | y, \beta, h, \nu_\lambda) = f_\gamma \left(\lambda_i \mid \frac{\nu_\lambda + 1}{h\epsilon_i^2 + \nu_\lambda}, \nu_\lambda + 1 \right). \quad (13.21)$$

Note that, conditional on knowing β , ϵ_i can be calculated and, hence, the parameters of the gamma density in (13.21) can be calculated in the posterior simulator.

The density $p(\nu_\lambda | y, \beta, h, \lambda)$ is relatively easy to derive since ν_λ does not enter the likelihood and $p(\nu_\lambda | y, \beta, h, \lambda) = p(\nu_\lambda | \lambda)$. It follows from Bayes' theorem that

$$p(\nu_\lambda | \lambda) \propto p(\lambda | \nu_\lambda) p(\nu_\lambda),$$

and thus the kernel of the posterior conditional of ν_λ is (13.18) times (13.19). Thus, we obtain

$$p(\nu_\lambda | y, \beta, h, \lambda) \propto \left(\frac{\nu_\lambda}{2} \right)^{\frac{N\nu_\lambda}{2}} \Gamma \left(\frac{\nu_\lambda}{2} \right)^{-N} \exp(-\eta\nu_\lambda), \quad (13.22)$$

where

$$\eta = \frac{1}{\underline{\nu}_\lambda} + \frac{1}{2} \sum_{i=1}^N [\ln(\lambda_i^{-1}) + \lambda_i].$$

This density is a nonstandard one. Hence, a M–H step will be required in our posterior simulation algorithm.

Thus, a Metropolis-within-Gibbs algorithm that sequentially draws from (13.8), (13.12), (13.20), and (13.22) can be used for carry out Bayesian inference in this model.

(b) We will use the house price data set described at the beginning of this chapter. We will use the same prior for β and h as in Exercise 13.2. The prior for ν_λ depends on the hyperparameter $\underline{\nu}_\lambda$, its prior mean. We set $\underline{\nu}_\lambda = 25$, a value that allocates substantial prior weight both to very fat tailed error distributions (e.g., $\nu_\lambda < 10$) as well as error distributions that are roughly normal (e.g., $\nu_\lambda > 40$).

All posterior conditionals except for $p(\nu_\lambda|y, \beta, h, \lambda)$ are standard densities (i.e., normal or gamma). There are many ways of drawing from $p(\nu_\lambda|y, \beta, h, \lambda)$. For instance, Geweke (1993) uses acceptance sampling. The Matlab code on the Web site associated with this book uses a random walk chain M–H algorithm with a normal increment random variable. Equation (13.15), evaluated at old and candidate draws, is used to calculate the acceptance probability. Candidate draws of ν_λ that are less than or equal to zero have the acceptance probability set to zero. The variance of the proposal density, Σ , is chosen by first setting $\Sigma = c$ and experimenting with different values of the scalar c until a value is found that yields reasonable acceptance probabilities. The posterior simulator is then run using this value to yield an estimate of the posterior variance of ν_λ , $\widehat{\text{Var}}(\nu_\lambda|y)$. We then set $\Sigma = c\widehat{\text{Var}}(\nu_\lambda|y)$ and experiment with different values of c until we find one that yields an average acceptance probability of roughly .50. Then a final long run of 30,000 replications, with 5,000 burn-in replications discarded, is taken.

Table 13.2 contains posterior results for the key parameters. It can be seen that, although posteriors for the elements of β are qualitatively similar to those presented in Table 13.1, the posterior for ν_λ indicates the errors exhibit substantial deviations from normality. Because this crucial parameter is univariate, you may also want to plot a histogram approximation

Table 13.2: Posterior results for β and ν_λ .

	Mean	Standard Deviation	95% HPDI
β_1	−413.15	2,898.24	[−5153, 4329]
β_2	5.24	0.36	[4.65, 5.83]
β_3	2118.02	972.84	[501, 3709]
β_4	14,910.41	1,665.83	[12188, 17631]
β_5	8,108.53	955.74	[6706, 9516]
ν_λ	4.31	0.85	[3.18, 5.97]

of its posterior. For this data set, such a histogram will indicate that $p(\nu_\lambda|y)$ has a shape that is quite skewed and confirms the finding that virtually all of the posterior probability is allocated to small values for the degrees of freedom parameter. Note, however, that there is virtually no support for extremely small values, which would imply extremely fat tails. Remember that the Cauchy distribution is the Student t with $\nu_\lambda = 1$. It has such fat tails that its mean does not exist. There is no evidence for this sort of extreme behavior in the errors for the present data set.

Exercise 13.4 (Autocorrelated errors) In this exercise, assume that that you have time series data. We will use a subscript t to indicate time. That is, y_t for $t = 1, \dots, T$ indicates observations on the dependent variable from period 1 through T . We consider the regression model

$$y_t = x_t\beta + \epsilon_t$$

and assume the errors follow an autoregressive process of order 1 or AR(1) process:

$$\epsilon_t = \rho\epsilon_{t-1} + u_t. \quad (13.23)$$

We also assume $u_t \stackrel{\text{iid}}{\sim} N(0, h^{-1})$ and that the error process is stationary: $|\rho| < 1$.

It can be shown that the covariance matrix of ϵ is $h^{-1}\Omega$, where

$$\Omega = \frac{1}{1 - \rho^2} \begin{bmatrix} 1 & \rho & \rho^2 & \cdots & \rho^{T-1} \\ \rho & 1 & \rho & \cdots & \rho^{T-2} \\ \rho^2 & \rho & 1 & \cdots & \rho^{T-3} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho^{T-1} & \rho^{T-2} & \rho^{T-3} & \cdots & 1 \end{bmatrix}, \quad (13.24)$$

and thus the regression model with AR(1) errors falls in the class of models being studied in this chapter.

For the new parameter in this model assume a prior of the form

$$p(\rho) \propto \phi(\rho; \underline{\rho}, \underline{V}_\rho) I(|\rho| < 1), \quad (13.25)$$

where $I(|\rho| < 1)$ is the indicator function that equals one for the stationary region and zero otherwise.

(a) Given the priors in (13.2), (13.3), (13.4), and (13.25), derive the full posterior conditionals for this model and, using these, discuss how posterior simulation can be carried out.

(b) Show how the Savage–Dickey density ratio (see Chapter 6, Exercise 6.8) can be used to find the Bayes factor in favor of $\rho = 0$.

(c) Using the results from parts (a) and (b) carry out a Bayesian analysis using a suitable time series data set. You may use the baseball data set described in the following (available on the Web site associated with this book as YANKEES.TXT). Conduct a prior sensitivity analysis with respect to the prior variance of ρ by trying different values for \underline{V}_ρ between 0.01 and 100. Discuss the sensitivity of the Bayes factor of part (b) to changes in this prior.

Solution

(a) As noted in Exercise 13.1, a suitably transformed regression can be found where the errors satisfy the classical assumptions. First, take the regression model

$$y_t = x_t \beta + \epsilon_t, \quad (13.26)$$

where $x_t = [1 \ x_{t2} \ \cdots \ x_{tk}]$, and multiply both sides of (13.26) by $(1 - \rho L)$, where L is the lag operator. Define $y_t^* = (1 - \rho L) y_t$ and $x_t^* = (1 - \rho L) x_t$. We obtain

$$y_t^* = x_t^{*'} \beta + u_t. \quad (13.27)$$

We have assumed that u_t is iid $N(0, h^{-1})$, and thus the transformed model given in (13.27), which has $T - 1$ observations, is simply a normal linear regression model with iid errors. To simplify things, you may work with data from $T = 2, \dots, T$ and ignore the (minor) complications relating to the treatment of initial conditions. Thus, if we define $y^* = [y_2^* \ y_3^* \ \cdots \ y_T^*]'$ and X^* conformably, we can use the results from Exercise 13.1 to obtain

$$\beta | y, h, \rho \sim N(\bar{\beta}, \bar{V}), \quad (13.28)$$

where

$$\bar{V} = (\underline{V}^{-1} + h X^{*'} X^*)^{-1} \quad (13.29)$$

and

$$\bar{\beta} = \bar{V} (\underline{V}^{-1} \underline{\beta} + h X^{*'} y^*). \quad (13.30)$$

The posterior for h conditional on the other parameters in the model is gamma:

$$h | y, \beta, \rho \sim \gamma(\bar{s}^{-2}, \bar{\nu}), \quad (13.31)$$

where

$$\bar{\nu} = T - 1 + \underline{\nu} \quad (13.32)$$

and

$$\bar{s}^2 = \frac{(y^* - X^* \beta)' (y^* - X^* \beta) + \underline{\nu} s^2}{\bar{\nu}}. \quad (13.33)$$

Conditional on β and h , it follows that (13.23) defines a regression model for ϵ_t (with known error variance). By using standard results that we have derived in earlier exercises for the normal linear regression model (e.g., Chapter 10, Exercise 10.1, and Chapter 2, Exercise 2.3) it follows immediately that

$$p(\rho | y, \beta, h) \propto \phi(\rho; \bar{\rho}, \bar{V}_\rho) I(|\rho| < 1), \quad (13.34)$$

where

$$\bar{V}_\rho = (\underline{V}_\rho^{-1} + h \epsilon'_{-1} \epsilon_{-1})^{-1}, \quad (13.35)$$

$$\bar{\rho} = \bar{V}_\rho (\underline{V}_\rho^{-1} \underline{\rho} + h \epsilon'_{-1} \epsilon), \quad (13.36)$$

$\epsilon_{-1} = [\epsilon_1 \ \epsilon_2 \ \cdots \ \epsilon_{T-1}]'$, and $\epsilon = [\epsilon_2 \ \epsilon_3 \ \cdots \ \epsilon_T]'$.

Posterior simulation can be done by using an MCMC algorithm that sequentially draws from (13.28), (13.31), and (13.34), which are normal, gamma, and truncated normal, respectively. The fact that the last is truncated normal could be dealt with by drawing from the corresponding untruncated normal part, $\phi(\rho; \bar{\rho}, \bar{V}_\rho)$, discarding any draws outside the interval $|\rho| < 1$, and redrawing until a draw in the interval is obtained. This is equivalent to doing importance sampling using $\phi(\rho; \bar{\rho}, \bar{V}_\rho)$ as an importance function (i.e., importance sampling weights are zero for draws outside the interval $|\rho| < 1$, which is equivalent to simply discarding such draws). Alternatively, if the number of discarded draws proves to be quite high, one could draw from the truncated normal directly. (See the solution of Exercise 11.20 for an algorithm to draw from a univariate truncated normal).

(b) The Savage–Dickey density ratio produces the Bayes factor in favor of the hypothesis that $\rho = 0$ and is given by (see Chapter 6, Exercise 6.8)

$$B_{12} = \frac{p(\rho = 0|y)}{p(\rho = 0)},$$

where the numerator and denominator are both evaluated in the unrestricted model. For the numerator, we require the *marginal posterior ordinate* at $\rho = 0$. Given the simulated output, this could be calculated using a kernel density estimator, or, more efficiently, using “Rao-Blackwellization”:

$$\widehat{p(\rho = 0|y)} = \frac{\sum_{r=1}^R p(\rho = 0|y, \beta^{(r)}, h^{(r)})}{R},$$

an average of the *conditional posterior ordinates* at $\rho = 0$. This approaches $p(\rho = 0|y)$ as $R \rightarrow \infty$ by the law of large numbers. Calculation of the ordinate in this way is feasible in this example, since implementation of the Gibbs sampler assumes that the conditional distribution $\rho|\beta, h, y$ is known, at least up to proportionality. However, the Savage–Dickey density ratio calculation requires you to know the complete density and not just its posterior kernel. Specifically, to implement this procedure, we need to know the full density

$$p(\rho|y, \beta, h) = \frac{\phi(\rho|\bar{\rho}, \bar{V}_\rho) 1(\rho \in \Phi)}{\int_{-1}^1 \phi(\rho|\bar{\rho}, \bar{V}_\rho) d\rho},$$

which requires evaluation of the integrating constant in the denominator. This, however, can be easily calculated since $p(\rho|y, \beta, h)$ is a univariate truncated normal and the properties of this density are easily evaluated in any relevant software program. To calculate the denominator in our expression for the Bayes factor, one needs to obtain the *prior ordinate* at zero, which, again, can be easily calculated from the univariate truncated normal density.

(c) To illustrate Bayesian inference in the normal regression model with autocorrelated errors, we use a data set pertaining to baseball. (You can use this data set, YANKEES.TXT, or one of your own or artificially generate data.) The dependent variable is the winning

percentage of the New York Yankees baseball team every year between 1903 and 1999. Interest centers on explaining the Yankees performance using various measures of team offensive and defensive performance. Thus

y_t = winning percentage (PCT) in year t = wins/(wins + losses),

x_{t2} = team on-base percentage (OBP) in year t ,

x_{t3} = team slugging average (SLG) in year t , and

x_{t4} = team earned run average (ERA) in year t .

We will use a noninformative prior for β and set $\underline{V}^{-1} = 0_{k \times k}$. We will also use a noninformative prior for the error precision and set $\underline{\nu} = 0$. With these choices, the values of $\underline{\beta}$ and \underline{s}^{-2} are irrelevant. The informative prior for ρ we use sets $\underline{\rho} = 0$ and $\underline{V}_\rho = cI_p$.

The posterior simulator is described in part (a) of this exercise and Matlab code is provided on the Web site associated with this book. All results are based on 30,000 replications, with 5,000 burn-in replications discarded and 25,000 replications retained.

Table 13.3 presents posterior results for β with $c = 0.09$, a reasonably small value reflecting a prior belief that autocorrelation in the errors is fairly small (i.e., the prior standard deviation of ρ is 0.3). It can be seen that OBP and SLG are positively associated and ERA is negatively associated with winning.

Table 13.4 contains results from the prior sensitivity analysis where various values of c are used. This table reveals that prior information has little affect on the posterior, unless prior information is extremely strong as in the $c = 0.01$ case. This can be seen by noting that posterior means, standard deviations, and HPDIs are almost the same for all values of

Table 13.3: Posterior results for β .

	Mean	Standard Deviation	95% HPDI
β_1	0.01	0.07	$[-0.11, 0.12]$
β_2	1.09	0.35	$[0.52, 1.66]$
β_3	1.54	0.18	$[1.24, 1.83]$
β_4	-0.12	0.01	$[-0.13, -0.10]$

Table 13.4: Posterior results for ρ .

	Mean	Standard Deviation	95% HPDI	Bayes Factor for $\rho_1 = 0$
$c = 0.01$	0.10	0.07	$[-0.02, 0.23]$	0.49
$c = 0.09$	0.20	0.10	$[0.03, 0.36]$	0.43
$c = 0.25$	0.21	0.11	$[0.04, 0.39]$	0.56
$c = 1.0$	0.22	0.11	$[0.05, 0.40]$	0.74
$c = 100$	0.22	0.11	$[0.05, 0.40]$	0.84

c between 0.09 and 100. The latter is a very large value, which, to all intents and purposes, implies that the prior is flat and noninformative over the stationary region. The Bayes factors are also fairly robust to changes in the prior. It is worth noting that this robustness of the Bayes factor in part has to do with the fact that the prior is truncated to a bounded interval – the stationary region. We remind the reader of problems associated with Bayes factors when noninformative improper priors are used on parameters whose support is unbounded (see, e.g., Chapter 10, Exercise 10.4).

Latent variable models

Many microeconomic applications (including binary, discrete choice, tobit, and generalized tobit analyses) involve the use of latent data. These latent data are unobserved by the econometrician, but the observed choices economic agents make typically impose some type of truncation or ordering among the latent variables.

In this chapter we show how the Gibbs sampler can be used to fit a variety of common microeconomic models involving the use of latent data. In particular, we review how *data augmentation* [see, e.g., Tanner and Wong (1987), Chib (1992), and Albert and Chib (1993)] can be used to simplify the computations in these models. Importantly, we recognize that many popular models in econometrics are essentially linear regression models on suitably defined latent data. Thus, conditioned on the latent data's values, we can apply all of the known techniques discussed in previous chapters (especially Chapter 10) for inference in the linear regression model. The idea behind data augmentation is to add (or augment) the joint posterior distribution with the latent data itself. In the Gibbs sampler, then, we can typically apply standard techniques to draw from the model parameters given the latent data, and then add an additional step to draw values of the latent data given the model parameters.

To review the idea behind data augmentation in general terms, suppose that we are primarily interested in characterizing the joint posterior $p(\theta|y)$ of a k -dimensional parameter vector θ . An alternative to analyzing this posterior directly involves working with a seemingly more completed posterior $p(\theta, z|y)$ for some *latent data* z . The desired marginal posterior for θ is then obtained by integrating out z .

Although the addition of the variables z might seem to needlessly complicate the estimation exercise, it often proves to be computationally convenient. To implement the Gibbs sampler here, we need to obtain the posterior conditionals $p(\theta|z, y)$ and $p(z|\theta, y)$. Often, the posterior conditional $p(\theta|z, y)$ will take a convenient form, thus making it possible to fit the model using the Gibbs sampler. Upon convergence, the set of draws for θ can be used

to characterize its marginal posterior. We begin illustrating these methods with perhaps the most celebrated latent variable model: the probit. We remind the reader that Matlab code for carrying out the empirical parts of the questions are provided on the Web site associated with this book. In addition to our previously used notation of y/X for the vector/matrix of observations on the dependent/explanatory variables, in this chapter we will also let $z = [z_1 \cdots z_n]'$, $z^* = [z_1^* \cdots z_n^*]'$, and $x_{-\theta}$ denote all quantities in x other than θ .

Exercise 14.1 (Posterior simulation in the probit model) Consider the following latent variable representation of the probit model:

$$z_i = x_i\beta + \epsilon_i, \quad \epsilon_i \stackrel{\text{iid}}{\sim} N(0, 1), \quad (14.1)$$

$$y_i = \begin{cases} 1 & \text{if } z_i > 0, \\ 0 & \text{if } z_i \leq 0. \end{cases}$$

The value of the binary variable y_i is observed, as are the values of the explanatory variables x_i . The latent data z_i , however, are unobserved.

(a) Derive the likelihood function for this model.

(b) Using a prior for β of the form $\beta \sim N(\mu_\beta, V_\beta)$, derive the *augmented* joint posterior $p(\beta, z|y)$.

(c) Verify that, marginalized over z , the joint posterior for the parameters β is exactly the same as the posterior you would obtain without introducing any latent variables to the model. For an interesting comparison, see Zellner and Rossi (1984) for a discussion of Bayesian analysis of binary choice models prior to the use of MCMC methods.

(d) Given your result in (b), derive the complete posterior conditionals and discuss how the Gibbs sampler can be employed to fit the model.

Solution

(a) Note that

$$\Pr(y_i = 1|x_i, \beta) = \Pr(x_i\beta + \epsilon_i > 0) = \Pr(\epsilon_i > -x_i\beta) = 1 - \Phi(-x_i\beta) = \Phi(x_i\beta).$$

Similarly,

$$\Pr(y_i = 0|x_i, \beta) = 1 - \Phi(x_i\beta).$$

Because of the assumed independence across observations, the likelihood function is obtained as

$$L(\beta) = \prod_{i=1}^n \Phi(x_i\beta)^{y_i} [1 - \Phi(x_i\beta)]^{1-y_i}.$$

(b) To derive the *augmented* joint posterior, note that

$$p(\beta, z|y) = \frac{p(y, z|\beta)p(\beta)}{p(y)},$$

implying that

$$p(\beta, z|y) \propto p(y, z|\beta)p(\beta).$$

The term $p(\beta)$ is simply our prior, whereas $p(y, z|\beta)$ represents the *complete* or *augmented* data density. To characterize this density in more detail, note that

$$p(y, z|\beta) = p(y|z, \beta)p(z|\beta).$$

Immediately, from our latent variable representation, we know

$$p(z|\beta) = \prod_{i=1}^n \phi(z_i; x_i\beta, 1). \quad (14.2)$$

For the conditional for y given z and β , note that when $z_i > 0$ then y_i must equal one, whereas when $z_i \leq 0$, then y_i must equal zero. In other words, the sign of z_i perfectly predicts the value of y . Hence, we can write

$$p(y|z, \beta) = \prod_{i=1}^n [I(z_i > 0)I(y_i = 1) + I(z_i \leq 0)I(y_i = 0)], \quad (14.3)$$

with I denoting the indicator function, which takes on the value one if the statement in the parentheses is true and is otherwise zero. This expression simply states that when z_i is positive, then y_i is one with probability one, and conversely, when z_i is negative, then y_i is zero with probability one.

Putting the pieces in (14.2) and (14.3) together, we obtain the augmented data density $p(y, z|\beta)$. We combine this with our prior to obtain

$$p(\beta, z|y) \propto p(\beta) \prod_{i=1}^n [I(z_i > 0)I(y_i = 1) + I(z_i \leq 0)I(y_i = 0)] \phi(z_i, x_i\beta, 1). \quad (14.4)$$

(c) To show the equality of these quantities, we integrate (14.4) over z to obtain

$$\begin{aligned} p(\beta|y) &= \int_z p(\beta, z|y) dz \\ &\propto p(\beta) \int_z \prod_{i=1}^n [I(z_i > 0)I(y_i = 1) + I(z_i \leq 0)I(y_i = 0)] \phi(z_i, x_i\beta, 1) dz_1 \cdots dz_n \\ &= p(\beta) \prod_{i=1}^n \left[\int_{-\infty}^{\infty} [I(z_i > 0)I(y_i = 1) + I(z_i \leq 0)I(y_i = 0)] \phi(z_i, x_i\beta, 1) dz_i \right] \\ &= p(\beta) \prod_{i=1}^n \left[\int_{-\infty}^0 I(y_i = 0) \phi(z_i; x_i\beta, 1) dz_i + \int_0^{\infty} I(y_i = 1) \phi(z_i; x_i\beta, 1) dz_i \right] \\ &= p(\beta) \prod_{i=1}^n [I(y_i = 0)[1 - \Phi(x_i\beta)] + I(y_i = 1)\Phi(x_i\beta)] \\ &= p(\beta) \prod_{i=1}^n \Phi(x_i\beta)^{y_i} [1 - \Phi(x_i\beta)]^{1-y_i}. \end{aligned}$$

Note that this is exactly the prior times the likelihood obtained from part (a), which can be obtained without explicitly introducing any latent variables. What is important to note is that the posterior of β is unchanged by the addition of the latent variables. So, augmenting the posterior with z will not change any inference regarding β , though it does make the problem computationally easier, as described in the solution to the next question.

(d) With the addition of the latent data, the complete conditionals are easily obtained. In particular, the complete conditional for β given z and the data y follows directly from (14.4) and standard results from the linear regression model (details of which are provided in Chapter 12, Exercise 12.1):

$$\beta|z, y \sim N(D_\beta d_\beta, D_\beta), \quad (14.5)$$

where

$$D_\beta = (X'X + V_\beta^{-1})^{-1}, \quad d_\beta = X'z + V_\beta^{-1}\mu_\beta.$$

For the complete conditional for z , first note that the conditional independence across observations implies that each z_i can be drawn independently. We also note from (14.4) that

$$\begin{aligned} z_i|\beta, y &\propto I(z_i > 0)\phi(z_i; x_i\beta, 1) \quad \text{if } y_i = 1, \\ z_i|\beta, y &\propto I(z_i \leq 0)\phi(z_i; x_i\beta, 1) \quad \text{if } y_i = 0. \end{aligned}$$

Thus,

$$z_i|\beta, y \stackrel{\text{ind}}{\sim} \begin{cases} \text{TN}_{(-\infty, 0]}(x_i\beta, 1) & \text{if } y_i = 0, \\ \text{TN}_{(0, \infty)}(x_i\beta, 1) & \text{if } y_i = 1, \end{cases} \quad (14.6)$$

where the notation $\text{TN}_{[a, b]}(\mu, \sigma^2)$ denotes a normal distribution with mean μ and variance σ^2 truncated to the interval $[a, b]$.

Implementation of the Gibbs sampler involves iteratively sampling from (14.5) and (14.6). To generate draws from the truncated normal in (14.6), one can use the inverse transform method as described in the solution to Exercise 11.20.

Exercise 14.2 (Parameter identification in the probit model) Generate data from a probit model, as described in Exercise 14.1, and consider what happens when you relax the normalization that the variance parameter is unity. Instead, specify $\epsilon_i \stackrel{\text{iid}}{\sim} N(0, \sigma^2)$. Derive the complete conditionals for this model. Implement the Gibbs sampler and plot the posterior simulations of β , σ , and β/σ obtained from your generated-data experiment.

Solution

For this exercise, we employed priors of the form

$$\beta \sim N(\mu_\beta, V_\beta)$$

and

$$\sigma^2 \sim IG(a, b).$$

Let us first derive the complete posterior conditionals when σ^2 is not normalized to unity:

$$\beta|z, \sigma^2, y \sim N(D_\beta d_\beta, D_\beta), \quad (14.7)$$

where

$$D_\beta = (X'X/\sigma^2 + V_\beta^{-1})^{-1}, \quad d_\beta = X'z/\sigma^2 + V_\beta^{-1}\mu_\beta,$$

$$z_i|\sigma^2, \beta, y_i \stackrel{\text{ind}}{\sim} \begin{cases} \text{TN}_{(-\infty, 0]}(x_i\beta, \sigma^2) & \text{if } y_i = 0, \\ \text{TN}_{(0, \infty)}(x_i\beta, \sigma^2) & \text{if } y_i = 1, \end{cases} \quad (14.8)$$

and

$$\sigma^2|z, \beta, y \sim IG\left[\frac{n}{2} + a, \left(b^{-1} + (1/2) \sum_i (z_i - x_i\beta)^2\right)^{-1}\right]. \quad (14.9)$$

It is important to recognize that the complete conditionals in (14.7)–(14.9) do not necessarily alert us of the problem that the regression parameters are only identified up to scale. The densities in this algorithm are well defined, easily sampled, and offer no direct evidence of an identification problem. We will see, however, that the output from the posterior simulator will suggest a problem and reveal that the regression coefficients and scale parameter are not jointly identifiable.

We generate 2,000 observations from a probit model containing an intercept β_0 and one slope coefficient β_1 . Specifically, we generate $x_i \stackrel{\text{iid}}{\sim} N(0, 1)$, $\beta = [\beta_0 \ \beta_1]' = [.2 \ .6]'$ and fix $\sigma^2 = 1$. For our hyperparameters, we set $\mu_\beta = 0$, $V_\beta = 100I_2$, $a = 3$, and $b = .5$. We run the Gibbs sampler by cycling through (14.7), (14.8), and (14.9). The first 500 simulations are discarded as the burn-in and the final 2,000 are retained and used for our posterior analysis. Since the β parameters are only identified up to scale, we expect that the actual draws of β_0 , σ , and β_1 will not converge to anything, whereas the quantities β_0/σ and β_1/σ should converge to their generated-data values of .2 and .6, respectively.

Presented in Figure 14.1 are plots of the β_1 , σ , β_0/σ , and β_1/σ draws from our posterior simulator. As seen from the figure, the draws from the marginal posteriors of β_1 and σ wander throughout the parameter space and do not settle down around their generated-data values. However, they do move together, and as shown in the bottom row of the figure, the identified quantities β_0/σ and β_1/σ do approximately converge to .2 and .6, the values used to generate these data.

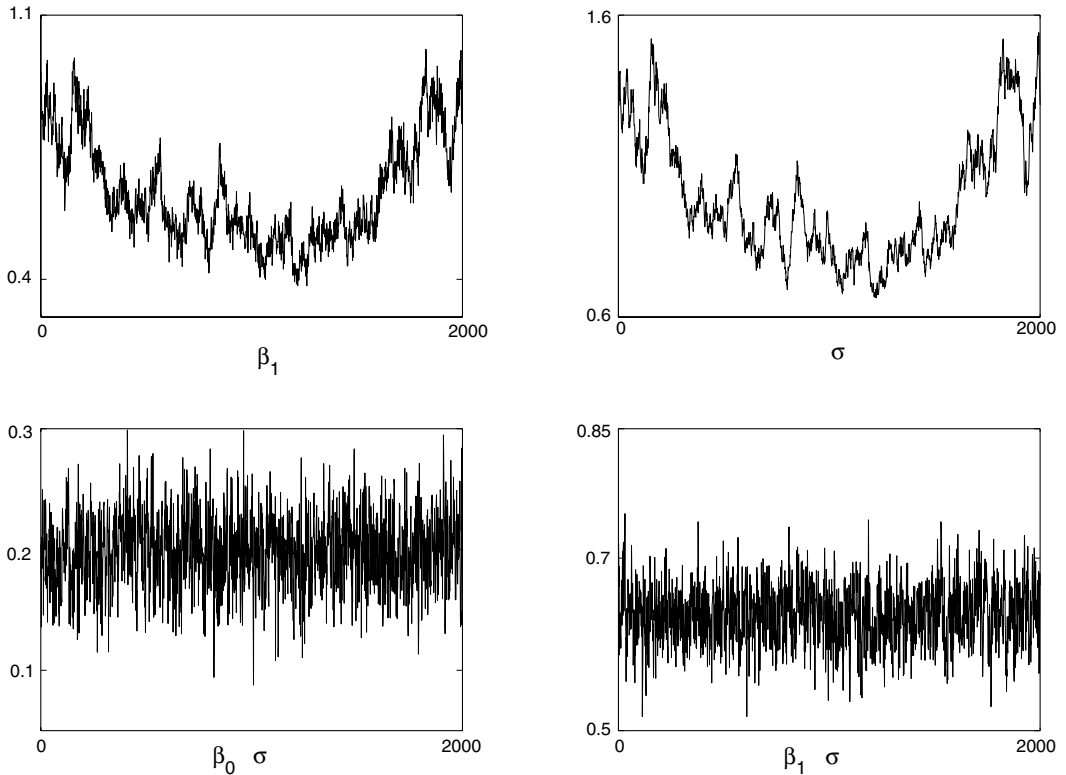


Figure 14.1 — Plots of parameter values for 2,000 Gibbs iterations using generated probit data.

Exercise 14.3 (Probit model: application using Fair’s [1978] data) In an interesting paper that has been used to illustrate the binary choice model in econometrics texts [e.g., Gujarati (2003), pp. 618–620], Fair (1978) analyzed the decision to have an extramarital affair. We have made a version of these data available on our Web site. Take these data and perform a binary choice analysis. Fit the model using the Gibbs sampler and interpret your results.

Solution

Fair (1978) used the results of surveys conducted by *Psychology Today* and *Redbook* magazines to model extramarital affairs. In this application we take 601 responses from the *Psychology Today* survey, noting that the variables constructed and used here are slightly different than the ones used in his original study. All observations are taken from individuals currently married and married for the first time.

We specify a probit model where the decision to have an extramarital affair depends on an intercept (CONS), a male dummy (MALE), number of years married (YS-MARRIED), a dummy variable if the respondent has children from the marriage (KIDS), a dummy for classifying one’s self as “religious” (RELIGIOUS), years of schooling completed (ED),

Table 14.1: Coefficient and marginal effect posterior means and standard deviations from the probit model using Fair's (1978) data.

	Coefficient		Marginal Effect	
	Mean	Std. Dev.	Mean	Std. Dev.
CONS	-.726	(.417)	—	—
MALE	.154	(.131)	.047	(.040)
YS-MARRIED	.029	(.013)	.009	(.004)
KIDS	.256	(.159)	.073	(.045)
RELIGIOUS	-.514	(.124)	-.150	(.034)
ED	.005	(.026)	.001	(.008)
HAPPY	-.514	(.125)	-.167	(.042)

and a final dummy variable denoting whether the person views the marriage as happier than an average marriage (HAPPY).

For our prior for the 7×1 parameter vector β , we specify

$$\beta \sim N(0, 10^2 I_7),$$

so that the prior is quite noninformative and centered at zero. We run the Gibbs sampler for 2,000 iterations and discard the first 500 as the burn-in.

Shown in Table 14.1 are coefficient posterior means and standard deviations from the binary choice model. To illustrate how quantities of interest other than regression coefficients can also be calculated, we report posterior means and standard deviations associated with *marginal effects* from the probit model. For a continuous explanatory variable we note that

$$\frac{\partial E(y|x, \beta)}{\partial x_j} = \beta_j \phi(x\beta).$$

For given x , this is a function of the regression parameters β . Thus, the *posterior distribution of the marginal effect* can be obtained by calculating and collecting the values $\{\beta_j^{(i)} \phi(x\beta^{(i)})\}_{i=1}^{1500}$, where $\beta^{(i)}$ represents the i th postconvergence draw obtained from the Gibbs sampler. When x_j is binary, we calculate the marginal effect as the difference between the normal c.d.f.s when the binary indicator is set to 1 and then 0. In all cases, we choose to evaluate the marginal effects at the sample means of the x 's. (If desired, one can instead set the binary variables to specific integer values rather than sample means, though we do not pursue this in our calculations.)

From the table, we see that people reporting themselves as religious are 15 percent less likely to have affairs (holding the x 's at mean values), and those reporting to be in happy marriages are 17 percent less likely. The posterior standard deviations for these quantities are small relative to their means, indicating that virtually all of the posterior masses associated with these marginal effects are placed over negative values.

Exercise 14.4 (Posterior simulation in the panel probit model) Consider a panel probit model of the form

$$z_{it} = \alpha_i + x_{it}\beta + \epsilon_{it}, \quad \epsilon_{it} \stackrel{\text{iid}}{\sim} N(0, 1), \quad i = 1, 2, \dots, n, \quad t = 1, 2, \dots, T. \quad (14.10)$$

The observed binary responses y_{it} are generated according to

$$y_{it} = \begin{cases} 1 & \text{if } z_{it} > 0, \\ 0 & \text{if } z_{it} \leq 0, \end{cases}$$

and the random effects $\{\alpha_i\}$ are drawn from a common distribution:

$$\alpha_i \stackrel{\text{iid}}{\sim} N(\alpha, \sigma_\alpha^2).$$

Suppose that you employ priors of the following forms:

$$\beta \sim N(\mu_\beta, V_\beta),$$

$$\alpha \sim N(\mu_\alpha, V_\alpha),$$

$$\sigma_\alpha^2 \sim IG(a, b).$$

Show how a Gibbs sampler can be employed to fit the panel probit model.

Solution

Just like the probit without a panel structure (e.g., Exercise 14.1), we will use data augmentation to fit the model. Specifically, we will work with an augmented posterior distribution of the form $p(z, \{\alpha_i\}, \alpha, \sigma_\alpha^2, \beta | y)$. Derivation of the complete posterior conditionals follows similarly to those derived in Exercise 14.1 for the probit and Exercises 12.2 and 12.5 for linear hierarchical models. Specifically, we obtain

$$z_{it} | \alpha, \{\alpha_i\}, \sigma_\alpha^2, \beta, y \stackrel{\text{ind}}{\sim} \begin{cases} TN_{(-\infty, 0]}(\alpha_i + x_{it}\beta, 1) & \text{if } y_{it} = 0, \\ TN_{(0, \infty)}(\alpha_i + x_{it}\beta, 1) & \text{if } y_{it} = 1, \end{cases} \quad (14.11)$$

$$\alpha_i | \alpha, \sigma_\alpha^2, \beta, z, y \stackrel{\text{ind}}{\sim} N(D_{\alpha_i} d_{\alpha_i}, D_{\alpha_i}), \quad i = 1, 2, \dots, n, \quad (14.12)$$

where

$$D_{\alpha_i} = (T + \sigma_\alpha^{-2})^{-1}, \quad d_{\alpha_i} = \sum_t (z_{it} - x_{it}\beta) + \sigma_\alpha^{-2}\alpha,$$

$$\beta | \alpha, \{\alpha_i\}, \sigma_\alpha^2, z, y \sim N(D_\beta d_\beta, D_\beta), \quad (14.13)$$

where

$$D_\beta = (X'X + V_\beta^{-1})^{-1}, \quad d_\beta = X'(z - \bar{\alpha}) + V_\beta^{-1}\mu_\beta,$$

X and z have been stacked appropriately, and $\bar{\alpha} \equiv [\alpha_1 \iota_T' \cdots \alpha_n \iota_T']'$, with ι_T denoting a $T \times 1$ vector of ones, and

$$\alpha | \beta, \{\alpha_i\}, \sigma_\alpha^2, z, y \sim N(D_\alpha d_\alpha, D_\alpha), \quad (14.14)$$

where

$$D_\alpha = (n/\sigma_\alpha^2 + V_\alpha^{-1})^{-1}, \quad d_\alpha = \sum_i \alpha_i/\sigma_\alpha^2 + V_\alpha^{-1}\mu_\alpha.$$

Finally,

$$\sigma_\alpha^2 | \alpha, \{\alpha_i\}, \beta, z, y \sim IG \left(\frac{n}{2} + a, [b^{-1} + .5 \sum_i (\alpha_i - \alpha)^2]^{-1} \right). \quad (14.15)$$

Fitting the model involves cycling through (14.11) through (14.15). Of course, blocking steps can and should be used to improve the mixing of the posterior simulator (see, e.g., Exercise 12.5).

Exercise 14.5 (Posterior simulation in the ordered probit, 1) Consider the latent variable representation of the *ordered probit model*:

$$z_i = x_i \beta + \epsilon_i, \quad \epsilon_i \stackrel{\text{iid}}{\sim} N(0, 1), \quad (14.16)$$

and

$$y_i = \begin{cases} 1 & \text{if } \alpha_0 < z_i \leq \alpha_1, \\ 2 & \text{if } \alpha_1 < z_i \leq \alpha_2, \\ \vdots & \vdots \\ M & \text{if } \alpha_{M-1} < z_i \leq \alpha_M. \end{cases}$$

We do not observe the latent variable z , but instead, we observe the ordered response $y_i \in \{1, 2, \dots, M\}$, where y is generated in terms of z . For identification purposes, we set $\alpha_0 = -\infty$, $\alpha_1 = 0$, and $\alpha_M = \infty$. The α_i are often called *cutpoints*.

- Write down the likelihood function for this model.
- How would you calculate marginal effects in the ordered probit?
- Show how the Gibbs sampler can be used to fit this model [e.g., Albert and Chib (1993)] under flat priors for α and β [i.e., $p(\alpha, \beta) \propto c$].
- Consider an ordered probit model where $y_i \in \{1, 2, 3\}$. Generate 500 observations from this ordered probit where $z_i = .5 + .3x_i + \epsilon_i$, $x_i \stackrel{\text{iid}}{\sim} N(0, 1)$ and $\alpha_2 = 1$. Present a table of posterior means and standard deviations of the model parameters, and also present a plot of the lagged autocorrelations for the cutpoint α_2 .

Solution

- For each i , the contribution to the likelihood function is

$$\begin{aligned} \Pr(y_i = k | \beta, \{\alpha_m\}) &= \Pr(\alpha_{k-1} < z_i \leq \alpha_k) \\ &= \Pr(\alpha_{k-1} - x_i \beta < z_i - x_i \beta \leq \alpha_k - x_i \beta) \\ &= \Phi(\alpha_k - x_i \beta) - \Phi(\alpha_{k-1} - x_i \beta). \end{aligned}$$

Thus, the likelihood function can be written as

$$L(\beta, \{\alpha_m\}) = \prod_{i=1}^n \Phi(\alpha_{y_i} - x_i \beta) - \Phi(\alpha_{y_i-1} - x_i \beta). \quad (14.17)$$

(b) For this model we have

$$\Pr(y = 1|x, \beta, \{\alpha_m\}) = 1 - \Phi(x\beta),$$

$$\Pr(y = k|x, \beta, \{\alpha_m\}) = \Phi(\alpha_k - x\beta) - \Phi(\alpha_{k-1} - x\beta), \quad 1 < k < M,$$

$$\Pr(y = M|x, \beta, \{\alpha_m\}) = 1 - \Phi(\alpha_{M-1} - x\beta).$$

Thus, we obtain the following marginal effects:

$$\frac{\partial \Pr(y = 1|x, \beta, \{\alpha_m\})}{\partial x_j} = -\phi(x\beta)\beta_j,$$

$$\frac{\partial \Pr(y = k|x, \beta, \{\alpha_m\})}{\partial x_j} = [\phi(\alpha_{k-1} - x\beta) - \phi(\alpha_k - x\beta)]\beta_j, \quad 1 < k < M,$$

$$\frac{\partial \Pr(y = M|x, \beta, \{\alpha_m\})}{\partial x_j} = \phi(\alpha_{M-1} - x\beta)\beta_j.$$

Hence, we can calculate how a change in a covariate changes the probabilities of y equaling each ordered value. Interestingly, note that the signs of the marginal effects for the “interior” values of y cannot be determined just given the sign of β_j . This was not the case in the linear regression and probit models.

(c) We again pursue a data augmentation approach in our Gibbs sampling algorithm. Let $\alpha = [\alpha_2 \ \alpha_3 \ \cdots \ \alpha_{M-1}]$ denote the vector of cutpoints and note that

$$p(\beta, \alpha, z|y) \propto p(y, z|\beta, \alpha)p(\beta, \alpha), \quad (14.18)$$

where

$$p(y, z|\beta, \alpha) = p(y|z, \beta, \alpha)p(z|\beta, \alpha)$$

is the complete or augmented likelihood function. Conditioned on z and α , the value of y is known with certainty. Specifically, we can write

$$\Pr(y_i = k|z_i, \alpha, \beta) = I(\alpha_{k-1} < z_i \leq \alpha_k), \quad (14.19)$$

or equivalently,

$$p(y_i|z_i, \alpha, \beta) = I(\alpha_{y_i-1} < z_i \leq \alpha_{y_i}), \quad y_i \in \{1, 2, \dots, M\}. \quad (14.20)$$

Under flat priors for α and β , it follows that the joint posterior distribution can be written as

$$p(\alpha, \beta, z|y) \propto \prod_{i=1}^n \phi(z_i; x_i\beta, 1) I(\alpha_{y_i-1} < z_i \leq \alpha_{y_i}). \quad (14.21)$$

Given this joint posterior distribution, we obtain the following complete posterior conditionals:

$$\beta|\alpha, z, y \sim N((X'X)^{-1}X'z, (X'X)^{-1}), \quad (14.22)$$

with X and z stacked appropriately and

$$z_i|\alpha, \beta, z_{-i}, y \stackrel{\text{ind}}{\sim} TN_{(\alpha_{y_i-1}, \alpha_{y_i})}(x_i\beta, 1), \quad i = 1, 2, \dots, n. \quad (14.23)$$

For the complete posterior conditional for the cutpoints, note that

$$\alpha_k | \alpha_{-k}, \beta, z, y \propto \prod_{i: y_i = k} I(\alpha_{k-1} < z_i \leq \alpha_k) \prod_{i: y_i = k+1} I(\alpha_k < z_i \leq \alpha_{k+1}).$$

It follows that

$$\alpha_k | \alpha_{-k}, \beta, z, y \stackrel{\text{ind}}{\sim} U[a_k, b_k], \quad (14.24)$$

where

$$a_k = \max \left\{ \alpha_{k-1}, \max_{i: y_i = k} z_i \right\}$$

and

$$b_k = \min \left\{ \alpha_{k+1}, \min_{i: y_i = k+1} z_i \right\}.$$

Implementing the Gibbs sampler involves successively sampling from (14.22), (14.23), and (14.24).

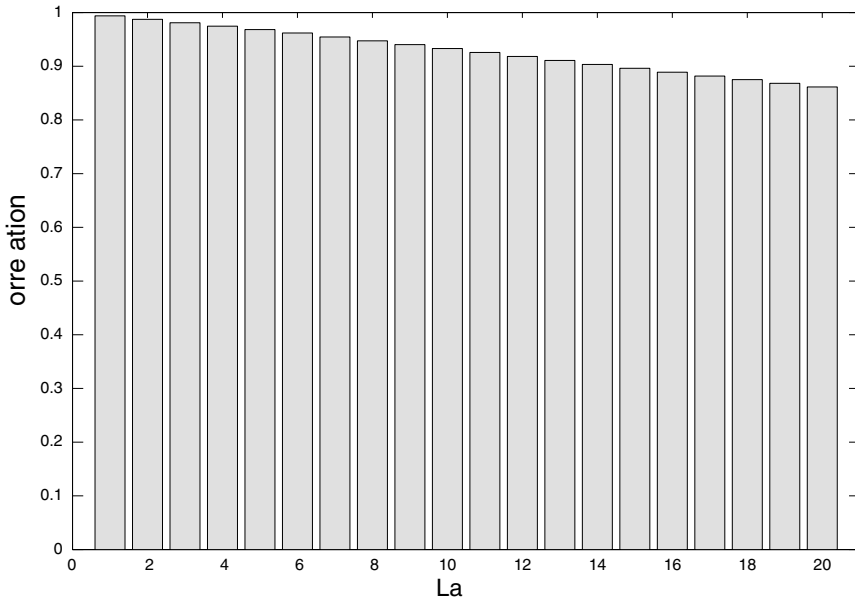
(d) Starting with initial conditions $\alpha_2 = .5$ and $\beta = [\beta_0 \ \beta_1]' = [0 \ 0]'$, we cycle through these complete posterior conditionals (starting with the latent data). We run the sampler for 2,500 iterations and discard the first 500 as the burn-in period. Posterior means and standard deviations are presented in Table 14.2.

Table 14.2: Posterior means and standard deviations from the ordered probit model.

Parameter	True Value	Post. Mean	Post. Std.
β_0	.5	.479	.057
β_1	.3	.569	.051
α_2	1	.972	.053

We see that our posterior means with 500 observations are close to the values used to generate the data. To get a feeling of the mixing of the chain, we plot in Figure 14.2 the lagged correlations up to order 20 for the cutpoint α_2 .

As one can see from Figure 14.2, the autocorrelations among the Gibbs draws are quite high and diminish very slowly. The correlation between simulated α_2 values that are 20 iterations away from one another is almost .9. This suggests very slow mixing of the chain, and, as a result, high numerical standard errors (see, e.g., Exercise 11.14). The following exercise suggests a reparameterization that may help facilitate mixing in the ordered probit model.

Figure 14.2 — Plots of lagged autocorrelations for α_2 .

Exercise 14.6 (Posterior simulation in the ordered probit, 2: Nandram and Chen [1996]) Consider the ordered probit model with $M = 3$ discussed in Exercise 14.5 and make the following reparameterization as suggested in Nandram and Chen (1996):

$$\delta = 1/\alpha_2, \quad \alpha^* = \delta\alpha, \quad \beta^* = \delta\beta, \quad z^* = \delta z.$$

- Derive the joint posterior distribution $p(\delta, \beta^*, z^*|y)$. Note that, for the case of $M = 3$ with this parameterization, there are no unknown cutpoints, since $\alpha_0^* = -\infty$, $\alpha_1^* = 0$, $\alpha_2^* = 1$, and $\alpha_3^* = \infty$.
- Describe how the Gibbs sampler can be used to fit this reparameterized model.
- Using the same data that were employed in part (d) of Exercise 14.5, estimate the reparameterized model and present a similar plot of the lagged autocorrelations for α_2 .

Solution

- For the case of $M = 3$, under flat priors, we showed in the previous exercise that

$$p(\alpha_2, \beta, z|y) \propto \prod_{i=1}^n \phi(z_i; x_i\beta, 1) I(\alpha_{y_i-1} < z_i \leq \alpha_{y_i}). \quad (14.25)$$

We seek to determine the joint posterior distribution $p(\delta, \beta^*, z^*|y)$.

The joint posterior $p(\delta, \beta^*, z^*|y)$ follows from a change of variables from $p(\alpha_2, \beta, z|y)$. With a bit of work, one can show that the Jacobian of this transformation is $\delta^{-[n+r+2]}$, where β is an $r \times 1$ vector. Thus, we have

$$\begin{aligned} p(\delta, \beta^*, z^*|y) &\propto \delta^{-[n+r+2]} \prod_{i=1}^n \exp\left(-\frac{1}{2} \left[\frac{z_i^*}{\delta} - \frac{1}{\delta} x_i \beta^*\right]^2\right) I\left(\frac{\alpha_{y_i-1}^*}{\delta} < \frac{z_i^*}{\delta} \leq \frac{\alpha_{y_i}^*}{\delta}\right) \\ &= \delta^{-[n+r+2]} \prod_{i=1}^n \exp\left(-\frac{1}{2\delta^2} (z_i^* - x_i \beta^*)^2\right) I(\alpha_{y_i-1}^* < z_i^* \leq \alpha_{y_i}^*), \end{aligned}$$

where, again, all $\alpha_{y_i}^*$ are known.

(b) For this parameterization, we obtain the following complete conditional posterior distributions:

$$\beta^*|z^*, \delta, y \sim N((X'X)^{-1}X'z^*, \delta^2(X'X)^{-1}), \quad (14.26)$$

$$z_i^*|\beta^*, \delta, y \stackrel{\text{ind}}{\sim} TN_{(\alpha_{y_i-1}^*, \alpha_{y_i}^*]}(x_i \beta^*, \delta^2), \quad (14.27)$$

where the α_i^* are known. Finally,

$$\delta^2|\beta^*, z^*, y \sim IG\left(\frac{n+r}{2}, \left[\frac{1}{2}(z^* - X\beta^*)'(z^* - X\beta^*)\right]^{-1}\right). \quad (14.28)$$

We fit the model by cycling through the complete conditionals in (14.26) through (14.28). To recover the original coefficient vector β and cutpoint α_2 , we simply use the inverse transformations $\beta = (1/\delta)\beta^*$ and $\alpha_2 = (1/\delta)$.

(c) Table 14.3 presents posterior means and standard deviations, using the same data that were employed in Exercise 14.5(d).

Parameter	True Value	Post. Mean	Post. Std.
β_0	.5	.476	.057
β_1	.3	.324	.053
α_2	1	.957	.061

Figure 14.3 presents the lagged autocorrelations for α_2 using the reparameterized model. As you can see, the mixing of the chain in this reparameterized model improves dramatically relative to that reported in the previous exercise. This rescaling transformation can also be used in a similar manner (also adding α^* to the sampler) when $M > 3$.

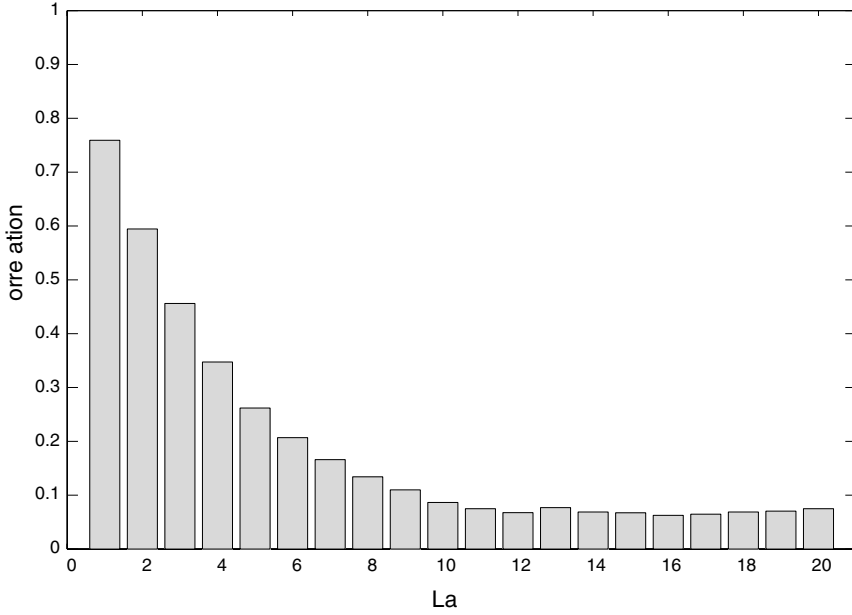


Figure 14.3 — Plots of lagged autocorrelations for α_2 using the reparameterized model.

Exercise 14.7 (Posterior simulation in the multinomial probit) Suppose that an agent has a choice among J alternatives, with no natural ordering among the alternatives. Let y_i denote the observed choice of the agent, with $y_i \in \{0, 1, \dots, J-1\}$.

Let U_{ij} represent the latent utility received by agent i from making choice j . We assume

$$U_{ij} = x_{ij}\beta_j + \epsilon_{ij}, \quad i = 1, 2, \dots, n, \quad j = 0, 1, \dots, J-1.$$

We treat choice 0 as the base alternative and choose to work with utility differences between the choices:

$$U_{ij} - U_{i0} = (x_{ij} - x_{i0})\beta_j + \epsilon_{ij} - \epsilon_{i0},$$

or equivalently,

$$U_{ij}^* = z_{ij}\beta_j + \epsilon_{ij}^*, \quad i = 1, 2, \dots, n, \quad j = 1, 2, \dots, J-1, \quad (14.29)$$

with $U_{ij}^* \equiv U_{ij} - U_{i0}$, $z_{ij} \equiv x_{ij} - x_{i0}$, and $\epsilon_{ij}^* \equiv \epsilon_{ij} - \epsilon_{i0}$. We assume that

$$\epsilon_i^* = \begin{bmatrix} \epsilon_{i1}^* \\ \epsilon_{i2}^* \\ \vdots \\ \epsilon_{iJ-1}^* \end{bmatrix} \stackrel{\text{iid}}{\sim} N(0, \Sigma), \quad i = 1, 2, \dots, n,$$

where Σ is a $(J-1) \times (J-1)$ positive definite matrix with off-diagonal element (i, j) denoted σ_{ij} and j th diagonal element denoted σ_j^2 . Since we only observe the choice yielding maximal utility, there is an indeterminacy problem; the likelihood $L(\beta, \Sigma)$ is identical to $L(c\beta, c^2\Sigma)$. There are a variety of strategies one can use to impose identification, the most

common of which is to restrict one of the diagonal elements of Σ to unity. We follow this identification strategy in the discussion that follows.

The observed choice made by the agent, y_i , is related to the latent U_{ij}^* as follows:

$$y_i = \begin{cases} 0 & \text{if } \max\{U_{ij}^*\}_{j=1}^{J-1} \leq 0, \\ k & \text{if } \max\left\{0, \{U_{ij}^*\}_{j=1}^{J-1}\right\} = U_{ik}^*. \end{cases} \quad (14.30)$$

That is, the agent chooses the base alternative if all utility differences are negative and chooses k if k yields more utility than the base choice and all other alternatives.

Finally, suppose priors of the following forms are specified:

$$\begin{aligned} \beta &\sim N(\mu_\beta, V_\beta), \\ \Sigma^{-1} &\sim W([\rho R]^{-1}, \rho) I(\sigma_{J-1}^2 = 1), \end{aligned}$$

where the indicator function on the Wishart prior serves to enforce the identification restriction that the $(J-1, J-1)$ element of Σ is unity.

(a) Using data augmentation, describe how the Gibbs sampler can be employed to fit this model.

(b) Perform a generated-data experiment by simulating data from a multinomial probit model with three alternatives and one covariate. Specifically, generate 500 observations from the following model:

$$U_{ij}^* = \beta_0 + \beta_1 z_{ij} + \epsilon_{ij}^*, \quad j = 1, 2,$$

where $\beta_0 = -.5$, $\beta_1 = 1$, $z_{ij} \stackrel{\text{iid}}{\sim} N(0, 1)$, and $\epsilon_i^* = [\epsilon_{i1}^* \ \epsilon_{i2}^*]' \stackrel{\text{iid}}{\sim} N(0, \Sigma)$, where

$$\Sigma = \begin{bmatrix} \sigma_1^2 & \rho_{12}\sigma_1 \\ \rho_{12}\sigma_1 & 1 \end{bmatrix}.$$

Note that we have restricted the regression coefficients to be constant across alternatives for simplicity (i.e., $\beta_j = \beta$), though such a restriction need not be imposed in the analysis [see part (a) of this exercise]. Finally, set $\sigma_1^2 = .5$ and $\rho_{12} = .4$.

Fit the model using the generated data and algorithm presented in part (a). Present posterior means and compare them to the parameter values used to generate the data.

Solution

(a) First, note that we can stack the observations over j for each i as follows:

$$\begin{bmatrix} U_{i1}^* \\ U_{i2}^* \\ \vdots \\ U_{iJ-1}^* \end{bmatrix} = \begin{bmatrix} z_{i1} & 0 & \cdots & 0 \\ 0 & z_{i2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & z_{iJ-1} \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_{J-1} \end{bmatrix} + \begin{bmatrix} \epsilon_{i1}^* \\ \epsilon_{i2}^* \\ \vdots \\ \epsilon_{iJ-1}^* \end{bmatrix},$$

or equivalently,

$$U_i^* = Z_i \beta + \epsilon_i^*, \quad \epsilon_i^* \stackrel{\text{iid}}{\sim} N(0, \Sigma), \quad i = 1, 2, \dots, n, \quad (14.31)$$

where

$$U_i^* \equiv \begin{bmatrix} U_{i1}^* \\ U_{i2}^* \\ \vdots \\ U_{iJ-1}^* \end{bmatrix}, \quad Z_i \equiv \begin{bmatrix} z_{i1} & 0 & \cdots & 0 \\ 0 & z_{i2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & z_{iJ-1} \end{bmatrix}, \quad \beta \equiv \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_{J-1} \end{bmatrix},$$

and ϵ_i^* was defined previously.

We work with the augmented joint posterior of the form $p(U^*, \beta, \Sigma^{-1}|y)$, where

$$U^* = \begin{bmatrix} U_1^* \\ U_2^* \\ \vdots \\ U_n^* \end{bmatrix}.$$

This posterior distribution can be expressed as follows:

$$p(U^*, \beta, \Sigma^{-1}|y) \propto p(\beta)p(\Sigma^{-1}) \left[\prod_{i=1}^n \phi(U_i^*; Z_i\beta, \Sigma) \left(I(y_i = 0)I(\max\{U_{ij}^*\} \leq 0) \right. \right. \\ \left. \left. + \sum_{k=1}^{J-1} I(y_i = k)I[U_{ik}^* > \max\{0, U_{i-k}^*\}] \right) \right],$$

with U_{i-k}^* denoting the collection of utility differences for agent i other than the k th difference.

In this form, derivation of the complete posterior conditionals is relatively straightforward. In particular one can show that

$$\beta|\Sigma^{-1}, U^*, y \sim N(D_\beta d_\beta, D_\beta), \quad (14.32)$$

where

$$D_\beta = \left(\sum_i Z_i' \Sigma^{-1} Z_i + V_\beta^{-1} \right)^{-1}, \quad d_\beta = \sum_i Z_i' \Sigma^{-1} U_i^* + V_\beta^{-1} \mu_\beta, \\ \Sigma^{-1}|U^*, \beta, y \sim W \left(\left[\rho R + \sum_i \epsilon_i^* \epsilon_i^{*'} \right]^{-1}, n + \rho \right) I(\sigma_{J-1}^2 = 1). \quad (14.33)$$

The algorithm of Nobile (2000) can be employed to generate draws from this Wishart, conditioned on the $(J-1, J-1)$ element. For more on this issue (and multinomial probit models generally), see, for example, Keane (1992), McCulloch and Rossi (1994), Geweke, Keane, and Runkle (1997), and McCulloch, Polson, and Rossi (2000).

Finally,

$$U_i^*|\beta, \Sigma^{-1}, y \stackrel{\text{ind}}{\sim} MTN_{R_i(y_i)}(Z_i'\beta, \Sigma), \quad i = 1, 2, \dots, n, \quad (14.34)$$

where $MTN_R(\mu, \Omega)$ denotes a multivariate normal distribution with mean μ and covariance matrix Ω that is truncated to the region R . In this case, if $y_i = 0$, R_i consists of the region

where each component of U_i^* is negative. Similarly, when $y_i = k$, R_i restricts U_{ik}^* to be positive and greater than all other U_{ij}^* . Procedures for drawing from the multivariate truncated normal (and Student t) are provided in Geweke (1991), and Matlab code is provided on the Web site associated with this book.

(b) We fit the model by sampling from (14.32) through (14.34), discarding the first 200 of 1,000 simulations as the burn-in period. For our priors, we specify $\beta \sim N(0, 400I_2)$, and for the hyperparameters in the inverse Wishart prior, we select $\rho = 3$ and $R = I_2$. For the regression parameters β_0 and β_1 , we obtain posterior means (and standard deviations) of $-.522 (.070)$ and $1.05 (.079)$, respectively, which are close to the intercept and slope values used to generate the data. For the variance parameter σ_2^2 and correlation parameter ρ_{12} , we obtain $.549 (.102)$ and $.418 (.154)$, respectively. These posterior means are, again, close to the values used to generate the data.

Exercise 14.8 (Posterior simulation in the tobit model) The tobit model [e.g., Tobin (1958)] specifies a mixed discrete–continuous distribution for a censored outcome variable y . In most applications of tobit models, values of y are assumed to follow a normal distribution provided y is positive, whereas we simultaneously see a clustering of y values at zero. Formally, we can write the tobit specification in terms of a latent variable model:

$$z_i = x_i\beta + \epsilon_i, \quad \epsilon_i \stackrel{\text{iid}}{\sim} N(0, \sigma^2), \quad i = 1, 2, \dots, n, \quad (14.35)$$

and

$$y_i = \max\{0, z_i\}. \quad (14.36)$$

For every observation, we observe the (y_i, x_i) pair.

- (a) Write down the likelihood function for the tobit model.
- (b) What is the expression for the marginal effect $\partial E(y|x)/\partial x_j$?
- (c) Describe how data augmentation can be used in conjunction with the Gibbs sampler to carry out a Bayesian analysis of this model.

Solution

(a) The likelihood function breaks into two parts. For the set of observations censored at zero, the contribution to the likelihood is:

$$\Pr(y_i = 0|x_i, \beta, \sigma^2) = \Pr(\epsilon_i \leq -x_i\beta) = 1 - \Phi(x_i\beta/\sigma).$$

Similarly, when $y_i > 0$, the likelihood is identical to that from a standard linear regression model, and thus the contribution of the i th uncensored observation to the likelihood is

$$\phi(y_i; x_i\beta, \sigma^2) = \frac{1}{\sigma} \phi\left[\left(\frac{y_i - x_i\beta}{\sigma}\right)\right],$$

with $\phi(\cdot)$ denoting the *standard* normal density function. Hence, the likelihood function can be expressed as

$$L(\beta, \sigma^2) = \prod_{i:y_i=0} \left[1 - \Phi\left(\frac{x_i\beta}{\sigma}\right)\right] \prod_{i:y_i>0} \frac{1}{\sigma} \phi\left[\left(\frac{y_i - x_i\beta}{\sigma}\right)\right]. \quad (14.37)$$

(b) Unlike the linear regression model, the marginal effect for the tobit or censored regression model is not simply the regression coefficient β_j . This was pointed out in Poirier and Melino (1978) and McDonald and Moffitt (1980), and we review the derivations here.

Note that

$$\begin{aligned} E(y|x, \beta, \sigma^2) &= E(y|x, \beta, \sigma^2, z > 0) \Pr(z > 0|x, \beta, \sigma^2) \\ &\quad + E(y|x, \beta, \sigma^2, z \leq 0) \Pr(z \leq 0|x, \beta, \sigma^2) \\ &= E(y|x, \beta, \sigma^2, z > 0) \Pr(z > 0|x, \beta, \sigma^2). \end{aligned}$$

For convenience in notation, simply let $\Phi \equiv \Phi(x\beta/\sigma)$ and $\phi \equiv \phi(x\beta/\sigma)$. For the two pieces in our expression for $E(y|x, \beta, \sigma^2)$, we know that $\Pr(z > 0|x, \beta, \sigma^2) = \Phi$ and that

$$\begin{aligned} E(y|x, \beta, \sigma^2, z > 0) &= x\beta + E(\epsilon|\epsilon > -x\beta) \\ &= x\beta + \sigma \frac{\phi}{\Phi}, \end{aligned}$$

using known properties of the truncated normal distribution. Since

$$\begin{aligned} \frac{\partial E(y|x, \beta, \sigma^2)}{\partial x_j} &= \frac{\partial E(y|x, \beta, \sigma^2, z > 0)}{\partial x_j} \Pr(z > 0|x, \beta, \sigma^2) \\ &\quad + E(y|x, \beta, \sigma^2, z > 0) \frac{\partial \Pr(z > 0|x, \beta, \sigma^2)}{\partial x_j}, \end{aligned}$$

we obtain

$$\begin{aligned} \frac{\partial E(y|x, \beta, \sigma^2)}{\partial x_j} &= \left[\beta_j + \sigma \left(\frac{-\phi\Phi(x\beta/\sigma)(\beta_j/\sigma) - \phi^2(\beta_j/\sigma)}{\Phi^2} \right) \right] \Phi \\ &\quad + \left(x\beta + \sigma \frac{\phi}{\Phi} \right) \phi(\beta_j/\sigma). \end{aligned}$$

Canceling terms gives

$$\frac{\partial E(y|x, \beta, \sigma^2)}{\partial x_j} = \Phi \beta_j. \quad (14.38)$$

Thus, the marginal effect in the tobit model does not equal β_j in general and only equals β_j as $\Phi \rightarrow 1$. We might interpret this extreme case as the linear regression model in the limit, since $\Phi \rightarrow 1$ implies that no probability is placed over the discrete mass point at zero.

(c) Chib (1992) was the original Bayesian tobit paper. Before discussing how the model can be fit using the Gibbs sampler, we note that the following prior specifications are employed:

$$\begin{aligned} \beta &\sim N(\mu_\beta, V_\beta), \\ \sigma^2 &\sim IG(a, b). \end{aligned}$$

Following Chib (1992), we augment the joint posterior with the latent data z_i and thus will work with $p(z, \beta, \sigma^2|y)$. It is also important to recognize that when $y_i > 0$, z_i is observed (and equals y_i), whereas when $y_i = 0$, we know that z_i is truncated from above at zero.

To this end, let D_i be a binary variable that equals one if the observation is censored ($y_i = 0$) and equals zero otherwise. We define

$$y_z = Dz + (1 - D)y,$$

so that y_z just takes the value y for the uncensored observations, and for the censored observations, y_z will take the value of the latent data z . Conditioned on z , this makes it easy to draw from the β conditional. We thus obtain the complete posterior conditionals:

$$\beta|z, \sigma^2, y \sim N(D_\beta d_\beta, D_\beta), \quad (14.39)$$

where

$$D_\beta = \left(X'X/\sigma^2 + V_\beta^{-1} \right)^{-1}, \quad d_\beta = X'y_z/\sigma^2 + V_\beta^{-1}\mu_\beta.$$

$$\sigma^2|\beta, z, y \sim IG\left(\frac{n}{2} + a, \left(b^{-1} + (1/2) \sum_{i=1}^n (y_{z_i} - x_i\beta)^2\right)^{-1}\right), \quad (14.40)$$

and

$$z_i|\beta, \sigma^2, y \stackrel{\text{ind}}{\sim} TN_{(-\infty, 0]}(x_i\beta, \sigma^2), \quad \forall i : y_i = 0. \quad (14.41)$$

A Gibbs sampler involves drawing from (14.39) through (14.41).

Exercise 14.9 (A tobit model with two-sided censoring: a female labor supply application) On our Web site we have provided data describing the number of weeks worked in 1990 by a sample of 2,477 married women. The data are taken from the National Longitudinal Survey of Youth (NLSY).

Write down a tobit model appropriate for these data. Importantly, note that number of weeks worked in this sample is clustered both at 0 and at 52 (full-time employment). Account for this two-sided censoring in your model specification.

- Describe an algorithm for fitting this model.
- How can you calculate marginal effects given the two-sided censoring problem?
- Using the NLSY data, estimate the model using all the covariates provided and comment on your results.

Solution

- We use the same notation as used in Exercise 14.8. For our prior hyperparameters we set

$$\mu_\beta = 0, \quad V_\beta = 10^2 I_5, \quad a = 3, \quad b = (1/40).$$

As described in this exercise, we have a pronounced two-sided censoring problem, where there are discrete masses placed at 0 weeks worked (out of the labor force) and 52 weeks worked (full-time employment). Approximately 45 percent of the sample reports working 52 hours per week and 17 percent of the sample reports working 0 weeks per year. To account for this feature of our data, we specify a slightly generalized form of the tobit model:

$$z_i = x_i\beta + \epsilon_i, \quad \epsilon_i \stackrel{\text{iid}}{\sim} N(0, \sigma^2), \quad (14.42)$$

and

$$y_i = \begin{cases} 0 & \text{if } z_i \leq 0, \\ z_i & \text{if } 0 < z_i \leq 52, \\ 52 & \text{if } z_i > 52. \end{cases} \quad (14.43)$$

Equivalently,

$$y_i = \min \{52, \max \{0, z_i\}\}. \quad (14.44)$$

In terms of the Gibbs sampler, the only difference between the algorithm required for this model and the one described in Exercise 14.8(c) is that the value of z_i must be truncated from below at 52 for all of those observations with $y_i = 52$.

One possible programming expedient (though certainly not computationally efficient) is to sample a latent variable truncated from above at 0 (say z_i^1) and one truncated from below at 52 (say z_i^2) for every observation in the sample. Let D_i^1 be a dummy variable if $y_i = 0$ and D_i^2 be a dummy if $y_i = 52$. We can then construct our complete data vector y_z as

$$y_z = D^1 z^1 + D^2 z^2 + (1 - D^1 - D^2)y. \quad (14.45)$$

With this construction, the Gibbs sampler then proceeds as in Exercise 14.8(c).

(b and c) We use the data provided to model the number of hours worked by a sample of 2,477 married women. We seek to determine how a measure of cognitive ability (denoted AFQT for Armed Forces Qualifying Test, which was administered to the NLSY participants), years of schooling completed (ED), husband's income in the previous year (1989) (denoted SPOUSE-INC), and an indicator for having children (KIDS) affect the decision regarding number of weeks worked.

We report in the following posterior means and standard deviations of parameters and marginal effects based on the final 5,000 of 5,500 total iterations. For the marginal effects with two-sided censoring, one can use a similar derivation to that used to derive the marginal effect in Exercise 14.8(b) for one-sided censoring. For this case we can show that

$$\frac{\partial E(y|x, \beta)}{\partial x_j} = \beta_j \left[\Phi \left(\frac{52 - x\beta}{\sigma} \right) - \Phi \left(\frac{-x\beta}{\sigma} \right) \right].$$

[A general result is provided in Greene (1999).] Table 14.4 reveals that the parameters generally possess the signs that we expect *a priori*. Married women with low test scores, few

Table 14.4: Coefficient and marginal effect posterior means and standard deviations from tobit model of weeks worked by married women.

	Coefficient		Marginal Effect	
	Mean	Std. Dev.	Mean	Std. Dev.
CONST	29.65	(6.06)	—	—
AFQT	.107	(.047)	.043	(.019)
SPOUSE-INC	-.245	(.053)	-.098	(.021)
KIDS	-26.65	(2.44)	-10.61	.971
ED	2.94	(.513)	1.17	(.202)
σ	44.35	(1.20)	—	—

years of education, and children in the home are likely to work fewer weeks during the year. Apparently, the largest factor governing female labor supply behavior is whether or not children are present in the home. The presence of children in the home reduces the expected number of weeks of work by approximately 11 weeks, and the posterior standard deviation associated with this effect is quite small relative to the mean. The acquisition of one more year of education also increases the expected number of weeks of work by approximately 1 week.

Exercise 14.10 (Posterior simulation in a treatment–effect model) In economics and social sciences generally, it is often of interest to estimate the *causal* impact of a binary treatment on a continuous outcome. For example, one might wish to quantify the impact of a job training program on earnings, the effect of smoking on health outcomes, or the impact of dropping out of high school on wages. The primary problem when trying to extract such a “causal” effect is one of *unobserved confounding*, whereby unobservable characteristics of the agent not captured in the model are correlated with both the treatment decision (denoted D) and the outcome of interest (denoted y). A useful model to investigate such problems is given by the following two-equation system:

$$y_i = \alpha_0 + \alpha_1 D_i + \epsilon_i, \quad (14.46)$$

$$D_i^* = z_i \theta + u_i, \quad (14.47)$$

where

$$\begin{pmatrix} \epsilon_i \\ u_i \end{pmatrix} \stackrel{\text{iid}}{\sim} N \left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_\epsilon^2 & \sigma_{u\epsilon} \\ \sigma_{u\epsilon} & 1 \end{pmatrix} \right] \equiv N(0, \Sigma). \quad (14.48)$$

The presence of the covariance term $\sigma_{u\epsilon}$ in (14.48) is added to account for the potential of confounding on unobservables. The binary decision D is included in the mean function of y in (14.46) (and α_1 might be referred to as the “causal” impact of the treatment on y). The second equation of the system in (14.47) is a latent variable equation that generates D [i.e., $D = I(D^* > 0)$], like the probit model discussed in Exercise 14.1.

Describe a Gibbs sampling algorithm for fitting this two-equation system and the priors employed in your analysis.

Solution

First, note that we can stack the model into the form

$$\tilde{y}_i = X_i \beta + \tilde{u}_i, \quad (14.49)$$

where

$$\tilde{y}_i = \begin{bmatrix} y_i \\ D_i^* \end{bmatrix}, \quad X_i = \begin{bmatrix} 1 & D_i & 0 \\ 0 & 0 & z_i \end{bmatrix}, \quad \beta = \begin{bmatrix} \alpha_0 \\ \alpha_1 \\ \theta \end{bmatrix}, \quad \tilde{u}_i = \begin{bmatrix} \epsilon_i \\ u_i \end{bmatrix}.$$

To simplify drawing parameters of the covariance matrix Σ , which is complicated by the restriction that $\text{Var}(u_i) = 1$, it is desirable to work with the population expectation of ϵ given u :

$$\epsilon_i = \sigma_{u\epsilon} u_i + v_i, \quad (14.50)$$

where $v_i \sim N(0, \sigma_v^2)$ and $\sigma_v^2 \equiv \sigma_\epsilon^2 - \sigma_{u\epsilon}^2$. This representation of the error structure suggests that we can choose to work with an *equivalent* representation of the model instead:

$$y_i = \alpha_0 + \alpha_1 D_i + \sigma_{u\epsilon} u_i + v_i, \quad (14.51)$$

$$D_i^* = z_i \theta + u_i, \quad (14.52)$$

where u and v are independently distributed, as implied by (14.50). With this definition of σ_v^2 , we choose to parameterize the model in terms of σ_v^2 and $\sigma_{u\epsilon}$ (instead of σ_ϵ^2 and $\sigma_{u\epsilon}$). The covariance matrix Σ in (14.48) then takes the form

$$\Sigma = \begin{bmatrix} \sigma_v^2 + \sigma_{u\epsilon}^2 & \sigma_{u\epsilon} \\ \sigma_{u\epsilon} & 1 \end{bmatrix}.$$

Rewriting the model in this way, and parameterizing in terms of σ_v^2 and $\sigma_{u\epsilon}$, will help us avoid computational difficulties that arise when diagonal restrictions are imposed on Σ , as in (14.48). We choose priors of the following forms:

$$\beta \sim N(\mu_\beta, V_\beta),$$

$$\sigma_{u\epsilon} \sim N(\mu_0, V_0),$$

$$\sigma_v^2 \sim IG(a, b),$$

and we note that Σ is positive definite for $\sigma_v^2 > 0$, which is enforced through our prior.

We work with an augmented posterior distribution of the form $p(\beta, D^*, \sigma_v^2, \sigma_{u\epsilon} | y, D)$. One can show that the complete posterior conditionals are as follows:

$$\beta | D^*, \Sigma, y, D \sim N(D_\beta d_\beta, D_\beta) \quad (14.53)$$

where

$$D_\beta = \left(\sum_i X_i' \Sigma^{-1} X_i + V_\beta^{-1} \right)^{-1}, \quad d_\beta = \sum_i X_i' \Sigma^{-1} \tilde{y}_i + V_\beta^{-1} \mu_\beta,$$

$$D_i^* | \beta, \Sigma, y, D \stackrel{\text{ind}}{\sim} \begin{cases} TN_{(0, \infty)}[z_i \theta + [\sigma_{u\epsilon}/(\sigma_v^2 + \sigma_{u\epsilon}^2)](y_i - \alpha_0 - \alpha_1 D_i), (1 - \rho_{u\epsilon}^2)] & \text{if } D_i = 1, \\ TN_{(-\infty, 0]}[z_i \theta + [\sigma_{u\epsilon}/(\sigma_v^2 + \sigma_{u\epsilon}^2)](y_i - \alpha_0), (1 - \rho_{u\epsilon}^2)] & \text{if } D_i = 0, \end{cases} \quad (14.54)$$

where $\rho_{u\epsilon}^2 = \sigma_{u\epsilon}^2 / (\sigma_v^2 + \sigma_{u\epsilon}^2)$ and

$$\sigma_{u\epsilon} | \beta, D^*, \sigma_v^2, y, D \sim N(Dd, D), \quad (14.55)$$

where

$$D = (u'u/\sigma_v^2 + V_0^{-1})^{-1}, \quad d = u'(y - \alpha_0 - \alpha_1 D)/\sigma_v^2 + V_0^{-1} \mu_0.$$

(In (14.55), note that the error vector $u = D^* - z_i \theta$ is “known” given D^* and θ .) Finally,

$$\sigma_v^2 | \sigma_{u\epsilon}, \beta, D^*, y, D \sim IG \left(\frac{n}{2} + a, [b^{-1} + .5 \sum_i (y_i - \alpha_0 - \alpha_1 D_i - \sigma_{u\epsilon} u_i)^2]^{-1} \right). \quad (14.56)$$

The sampler proceeds by drawing from (14.53) through (14.56). Given $\sigma_{u\epsilon}$ and σ_v^2 from the final two steps of the sampler, one can construct Σ . See Li (1998) for a discussion of a related model.

Exercise 14.11 (Posterior simulation in a generalized tobit: a model of potential outcomes, 1) There have been many generalizations of the tobit model described in Exercise 14.8. For example, Amemiya (1985) outlines five different types of tobit models, discusses estimation via two-step methods [e.g., Heckman (1976)] and maximum likelihood, and provides numerous references to applications of these different models. In these generalized tobit models, there is often the problem of *incidental truncation*, whereas the value of an outcome variable y is only observed in magnitude depending on the sign of some other variable, say z (and not y itself).

In this question, we take up Bayesian estimation of the most general type 5 tobit model (using Amemiya's enumeration system), often referred to as a model of *potential outcomes*. Inference in the remaining generalized models follows similarly. This particular model, much like the model discussed in Exercise 14.10, has seen considerable use in the "program evaluation" literature, where the goal is to determine the effectiveness of a program or treatment when individuals are not randomly assigned into the treated or untreated states. Formally, the model we consider here can be written as follows

$$D_i^* = z_i\theta + U_{Di}, \quad (14.57)$$

$$y_{1i} = x_i\beta_1 + U_{1i}, \quad (14.58)$$

$$y_{0i} = x_i\beta_0 + U_{0i}. \quad (14.59)$$

In these expressions, y_1 and y_0 are continuous *outcome* variables, with y_1 denoting the *treated* outcome and y_0 denoting the *untreated* outcome. The variable D^* is a latent variable that generates an *observed binary treatment decision* D according to

$$D_i = I(D_i^* > 0) = I(z_i\theta + U_{Di}) > 0. \quad (14.60)$$

We also assume the availability of some *exclusion restriction* or instrument in z that does not appear in x .

One interesting feature of this model is that only one outcome is observed for each observation in the sample. That is, if we let y denote the observed vector of outcomes in the data, we can write

$$y = Dy_1 + (1 - D)y_0. \quad (14.61)$$

In other words, if an individual takes the treatment ($D_i = 1$), then we only observe his or her treated outcome y_{1i} , and conversely, if $D_i = 0$, we only observe y_{0i} .

We make the following joint normality assumption:

$$U_i \stackrel{\text{iid}}{\sim} N(0, \Sigma), \quad (14.62)$$

where

$$U_i = [U_{Di} \ U_{1i} \ U_{0i}]' \quad \text{and} \quad \Sigma \equiv \begin{bmatrix} 1 & \rho_{1D}\sigma_1 & \rho_{0D}\sigma_0 \\ \rho_{1D}\sigma_1 & \sigma_1^2 & \rho_{10}\sigma_1\sigma_0 \\ \rho_{0D}\sigma_0 & \rho_{10}\sigma_1\sigma_0 & \sigma_0^2 \end{bmatrix}.$$

The variance parameter in the latent variable equation for the binary treatment indicator D has been normalized to unity for identification purposes. In the following, we write this restriction in an alternate, though equivalent, form as $\sigma_{D^*}^2 = 1$ and impose this condition through our prior.

(a) Write down the likelihood function for this model. Comment on the role of the correlation parameter ρ_{10} in the likelihood function.

(b) Derive the complete posterior conditionals for this model using priors of the form

$$\beta \equiv \begin{bmatrix} \theta \\ \beta_1 \\ \beta_0 \end{bmatrix} \sim N(\mu_\beta, V_\beta)$$

and

$$\Sigma^{-1} \sim W([\rho R]^{-1}, \rho) I(\sigma_{D^*}^2 = 1).$$

Note, again, that the indicator function is added to the standard Wishart prior so that the variance parameter in the latent data D^* equation is normalized to unity.

Solution

(a) When $D_i = 1$, we observe y_{1i} and the event that $D_i = 1$. Conversely, when $D_i = 0$, we observe y_{0i} and the event that $D_i = 0$. We thus obtain

$$\begin{aligned} L(\beta, \Sigma) &= \prod_{\{i:D_i=1\}} p(y_{1i}, D_i^* > 0) \prod_{\{i:D_i=0\}} p(y_{0i}, D_i^* \leq 0) \\ &= \prod_{\{i:D_i=1\}} \int_0^\infty p(y_{1i}, D_i^*) dD_i^* \prod_{\{i:D_i=0\}} \int_{-\infty}^0 p(y_{0i}, D_i^*) dD_i^* \\ &= \prod_{\{i:D_i=1\}} \int_0^\infty p(D_i^* | y_{1i}) p(y_{1i}) dD_i^* \prod_{\{i:D_i=0\}} \int_{-\infty}^0 p(D_i^* | y_{0i}) p(y_{0i}) dD_i^*. \end{aligned}$$

The conditional and marginal densities in these expressions can be determined from the assumed joint normality of the error vector. Performing the required calculations (which involves applying Appendix Theorem 3 to derive the conditional normal distributions, and then performing the necessary integration), we obtain

$$\begin{aligned} L(\beta, \Sigma) &= \prod_{\{i:D_i=1\}} \Phi \left(\frac{\tilde{U}_{Di} + \rho_{1D}\tilde{U}_{1i}}{(1 - \rho_{1D}^2)^{-1/2}} \right) \frac{1}{\sigma_1} \phi(\tilde{U}_{1i}) \\ &\quad \times \prod_{\{i:D_i=0\}} \left[1 - \Phi \left(\frac{\tilde{U}_{Di} + \rho_{0D}\tilde{U}_{0i}}{(1 - \rho_{0D}^2)^{-1/2}} \right) \right] \frac{1}{\sigma_0} \phi(\tilde{U}_{0i}), \end{aligned} \tag{14.63}$$

where

$$\begin{aligned}\tilde{U}_{Di} &= z_i\theta, \\ \tilde{U}_{1i} &= (y_{1i} - x_i\beta_1)/\sigma_1, \\ \tilde{U}_{0i} &= (y_{0i} - x_i\beta_0)/\sigma_0,\end{aligned}$$

and $\phi(\cdot)$ denotes the standard normal density.

Note that the correlation parameter ρ_{10} does not enter this likelihood function, and thus it is *not identified*. The reason is that the pair of outcomes (y_1, y_0) are never observed for any individual, and thus the correlation between these two outcomes will not enter the density function for the observed data.

(b) To implement the Gibbs sampler, we will augment the posterior with the latent desire for the receipt of treatment (D^*) and the missing potential outcome, which we will denote y^{Miss} . We therefore work with the augmented joint posterior $p(y^{\text{Miss}}, D^*, \beta, \Sigma^{-1} | y, D)$. Conditional on the unobserved latent quantities D^* and y^{Miss} , the regression parameters and inverse covariance matrix can be sampled in a straightforward manner. We now turn to a discussion of the complete posterior conditionals for each of these quantities:

Missing outcome data y^{Miss}

Since the Jacobian of the transformation from the error vector to the outcome quantities is unity, the trivariate normal distribution for $[D_i^* \ y_{1i} \ y_{0i}]$ is easily derived. The missing outcome data are then sampled by drawing from the conditional posterior:

$$y_i^{\text{Miss}} | \Gamma_{-y_i^{\text{Miss}}}, y, D \stackrel{\text{ind}}{\sim} N((1 - D_i)\mu_{1i} + (D_i)\mu_{0i}, (1 - D_i)\omega_{1i} + (D_i)\omega_{0i}), \quad (14.64)$$

where

$$\begin{aligned}\mu_{1i} &= x_i\beta_1 + (D_i^* - z_i\theta) \left[\frac{\sigma_0^2\sigma_{1D} - \sigma_{10}\sigma_{0D}}{\sigma_0^2 - \sigma_{0D}^2} \right] + (y_i - x_i\beta_0) \left[\frac{\sigma_{10} - \sigma_{0D}\sigma_{1D}}{\sigma_0^2 - \sigma_{0D}^2} \right], \\ \mu_{0i} &= x_i\beta_0 + (D_i^* - z_i\theta) \left[\frac{\sigma_1^2\sigma_{0D} - \sigma_{10}\sigma_{1D}}{\sigma_1^2 - \sigma_{1D}^2} \right] + (y_i - x_i\beta_1) \left[\frac{\sigma_{10} - \sigma_{0D}\sigma_{1D}}{\sigma_1^2 - \sigma_{1D}^2} \right], \\ \omega_{1i} &= \sigma_1^2 - \frac{\sigma_{1D}^2\sigma_0^2 - 2\sigma_{10}\sigma_{0D}\sigma_{1D} + \sigma_{10}^2}{\sigma_0^2 - \sigma_{0D}^2}, \\ \omega_{0i} &= \sigma_0^2 - \frac{\sigma_{0D}^2\sigma_1^2 - 2\sigma_{10}\sigma_{0D}\sigma_{1D} + \sigma_{10}^2}{\sigma_1^2 - \sigma_{1D}^2},\end{aligned}$$

and σ_{10} , σ_{1D} , and σ_{0D} refer to the covariance parameters between the treated and untreated outcome errors (i.e., $\sigma_{10} \equiv \rho_{10}\sigma_1\sigma_0$) and outcome and selection equation errors (i.e., $\sigma_{1D} \equiv \rho_{1D}\sigma_1$ and $\sigma_{0D} \equiv \rho_{0D}\sigma_0$). Note that this construction automatically samples from the posterior conditional for y_1 if $D = 0$ and samples from the posterior conditional for y_0 if $D = 1$.

*Missing latent data D^**

The latent data D_i^* are also drawn from a conditional normal, which is truncated by the observed value of D_i :

$$D_i^* | \Gamma_{-D_i^*}, y, D \stackrel{\text{ind}}{\sim} \begin{cases} TN_{(0,\infty)}(\mu_{Di}, \omega_{Di}) & \text{if } D_i = 1, \\ TN_{(-\infty,0]}(\mu_{Di}, \omega_{Di}) & \text{if } D_i = 0, \quad i = 1, 2, \dots, n, \end{cases} \quad (14.65)$$

where

$$\begin{aligned} \mu_{Di} &= z_i \theta + (D_i y_i + (1 - D_i) y_i^{\text{Miss}} - x_i \beta_1) \left[\frac{\sigma_0^2 \sigma_{1D} - \sigma_{10} \sigma_{0D}}{\sigma_1^2 \sigma_0^2 - \sigma_{10}^2} \right] \\ &\quad + (D_i y_i^{\text{Miss}} + (1 - D_i) y_i - x_i \beta_0) \left[\frac{\sigma_1^2 \sigma_{0D} - \sigma_{10} \sigma_{1D}}{\sigma_1^2 \sigma_0^2 - \sigma_{10}^2} \right], \\ \omega_{Di} &= 1 - \frac{\sigma_{1D}^2 \sigma_0^2 - 2\sigma_{10} \sigma_{0D} \sigma_{1D} + \sigma_1^2 \sigma_{0D}^2}{\sigma_1^2 \sigma_0^2 - \sigma_{10}^2}, \end{aligned}$$

and $TN_{(a,b)}(\mu, \sigma^2)$ denotes a univariate normal density with mean μ and variance σ^2 , truncated to the interval (a, b) .

Given these drawn quantities, we then compute the *complete data* vector

$$r_i^* = \begin{bmatrix} D_i^* \\ D_i y_i + (1 - D_i) y_i^{\text{Miss}} \\ D_i y_i^{\text{Miss}} + (1 - D_i) y_i \end{bmatrix}.$$

Regression parameters β

Conditioned on the values of the missing and latent data, it is straightforward to sample from the complete conditional for the $k \times 1$ vector $\beta \equiv [\theta' \ \beta_1' \ \beta_0']'$:

$$\beta | \Gamma_{-\beta}, y, D \sim N(\mu_\beta, \omega_\beta), \quad (14.66)$$

where

$$\begin{aligned} \mu_\beta &= [W'(\Sigma^{-1} \otimes I_n)W + V_\beta^{-1}]^{-1} [W'(\Sigma^{-1} \otimes I_n)\bar{y} + V_\beta^{-1}\mu_\beta] \\ \omega_\beta &= [W'(\Sigma^{-1} \otimes I_n)W + V_\beta^{-1}]^{-1}, \end{aligned}$$

where

$$W_{3n \times k} \equiv \begin{bmatrix} Z & 0 & 0 \\ 0 & X & 0 \\ 0 & 0 & X \end{bmatrix} \quad \text{and} \quad \bar{y}_{3n \times 1} \equiv \begin{bmatrix} D^* \\ Dy + (1 - D)y^{\text{Miss}} \\ Dy^{\text{Miss}} + (1 - D)y \end{bmatrix}.$$

Inverse covariance matrix Σ^{-1}

For the inverse covariance matrix, Σ^{-1} , a slight complication is again introduced as the complete conditional is no longer Wishart, given that the (1,1) element must be normalized

to unity. However, as described in Exercise 14.7, Nobile (2000) provides an algorithm for drawing from such a Wishart, conditional on a diagonal restriction. (Matlab code for this algorithm is available on the Web site associated with this book.) We express this conditional as

$$\Sigma^{-1}|\Gamma_{-\Sigma}, y, D \sim W \left(\left[\sum_{i=1}^n (r_i^* - \tilde{W}_i \beta)(r_i^* - \tilde{W}_i \beta)' + \rho R \right]^{-1}, n + \rho \right) I(\sigma_{D^*}^2 = 1), \quad (14.67)$$

where

$$\tilde{W}_i \equiv \begin{bmatrix} z_i & 0 & 0 \\ 0 & x_i & 0 \\ 0 & 0 & x_i \end{bmatrix}.$$

A Gibbs sampler involves cycling through the conditionals in (14.64) through (14.67). Finally, this exercise has assumed that the x variables are the same across the treated and untreated states. This assumption is not necessary, though it is arguably appropriate for most empirical work.

Exercise 14.12 (A model of potential outcomes, 2)

(a) Using the notation in Exercise 14.11, generate 5,000 observations from the following selection model:

$$D_i^* = \theta_0 + z_i \theta_1 + U_{Di},$$

$$y_{1i} = \beta_1 + U_{1i},$$

$$y_{0i} = \beta_0 + U_{0i},$$

with

$$\theta_0 = 0, \quad \theta_1 = .5, \quad \beta_1 = 2, \quad \beta_0 = 1, \quad z_i \stackrel{\text{iid}}{\sim} N(0, 1)$$

and

$$\sigma_1^2 = \sigma_0^2 = 1, \quad \rho_{1D} = .9, \quad \rho_{0D} = .7, \quad \rho_{10} = .6.$$

(b) Using the data generated in part (a), estimate β_1 and β_0 from single-equation analyses using OLS. How do the OLS estimates perform?

(c) Fit the model via the Gibbs sampler using the data generated in part (a). (See Exercise 14.11 for the Gibbs algorithm.) How do your estimation results compare to those obtained from OLS in (b)?

(d) Compare and contrast the prior and posterior for the correlation parameter ρ_{10} . Comment on this result.

Solution

(a and b) We generate 5,000 observations according to the model described in part (a) of this exercise. Running separate OLS regressions for the set of observations with $D_i = 1$

and $D_i = 0$, we found that $\hat{\beta}_1 = 2.63$ and $\hat{\beta}_0 = .501$ with these data. The OLS point estimates do not perform well because of the strong correlation between unobservables in the selection and outcome equations. Specifically,

$$E(U_{1i}|D_i = 1) = E(U_{1i}|U_{Di} < -z_i\theta) = \rho_{1D}\sigma_1 \left[\frac{\phi(z_i\theta)}{\Phi(z_i\theta)} \right] \neq 0.$$

Thus, mean independence between the regressor and error is violated unless $\rho_{1D} = 0$, as a result, OLS estimates will be biased and inconsistent in general.

(c) The Gibbs sampler is run for 5,500 iterations, and the first 500 simulations are discarded as the burn-in. For our prior hyperparameters, we specified

$$\mu_\beta = 0, \quad V_\beta = 10^2 I_4, \quad \rho = 4, \quad R = I_3.$$

Posterior means and standard deviations of the parameters are presented in Table 14.5.

Table 14.5: Posterior means and standard deviations from a generalized tobit-generated data experiment.

Parameter	True Value	Post. Mean	Post. Standard Deviation
β_1	2	2.07	.04
β_0	1	1.02	.04
θ_0	0	-.001	.03
θ_1	.5	.535	.02
σ_1	1	.950	.03
σ_0	1	1.02	.03
ρ_{1D}	.9	.837	.02
ρ_{0D}	.7	.722	.03
ρ_{10}	.6	.675	.07

From the table, we see that the posterior means are quite close to the true values of the parameters used to generate the data.

(d) We simulate 5,500 draws from our (restricted) Wishart prior and use these draws to plot the prior density for the nonidentified correlation parameter ρ_{10} . We also use the 5,000 postconvergence draws to plot the posterior density for this parameter. These plots are provided in Figure 14.4 and are obtained by nonparametrically smoothing the simulated prior and posterior draws [see, e.g., Silverman (1986) and DiNardo and Tobias (2001) for an introduction to nonparametric density estimation].

As seen in the figure, the priors and posteriors differ dramatically. But, how is this possible since ρ_{10} does not enter the likelihood function? How can the data update our beliefs about the values of this parameter?

As shown in Vijverberg (1993), Koop and Poirier (1997), Poirier and Tobias (2003), and Li, Poirier, and Tobias (2003), the positive definiteness of the 3×3 covariance matrix serves to place bounds on ρ_{10} . Specifically, these bounds imply that, conditionally,

$$\underline{\rho}_{10} \leq \rho_{10} \leq \bar{\rho}_{10},$$

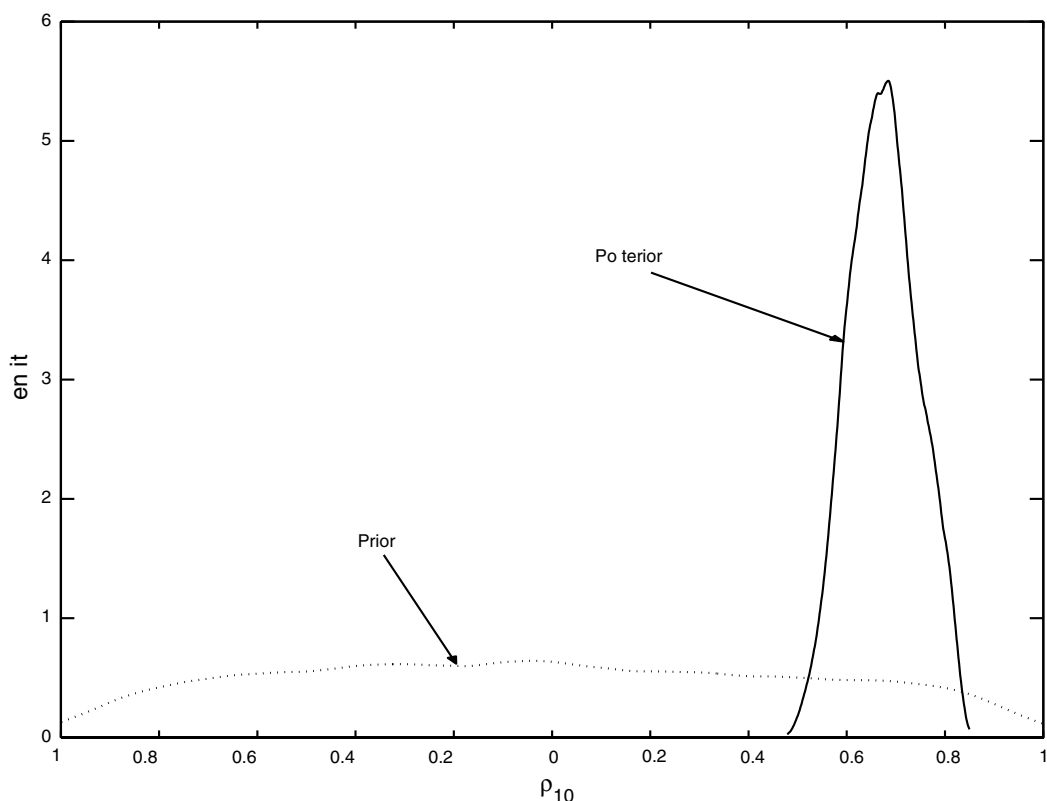


Figure 14.4 — Plots of priors and posteriors of ρ_{10} using generated data from a generalized tobit model.

where

$$\underline{\rho}_{10} = \rho_{1D}\rho_{0D} - [(1 - \rho_{1D}^2)(1 - \rho_{0D}^2)]^{1/2} \quad (14.68)$$

and

$$\bar{\rho}_{10} = \rho_{1D}\rho_{0D} + [(1 - \rho_{1D}^2)(1 - \rho_{0D}^2)]^{1/2}. \quad (14.69)$$

These bounds are only functions of the *identified* correlation parameters ρ_{1D} and ρ_{0D} . The values of the identified correlation parameters used to generate these data imply a lower bound of $\underline{\rho}_{10} \approx .32$ and an upper bound of $\bar{\rho}_{10} \approx .94$.

As seen from Table 14.5, the posteriors for the identified correlations ρ_{1D} and ρ_{0D} are quite tight around their generated-data values. As a result, the data provide substantial information about the lower and upper conditional support bounds $\underline{\rho}_{10}$ and $\bar{\rho}_{10}$. Hence, as the figure suggests, virtually all of the mass of the posterior for ρ_{10} is placed within the interval $[.3 \ .93]$. In short, information learned about ρ_{1D} and ρ_{0D} “spills over” and serves to update our beliefs about the nonidentified correlation parameter ρ_{10} via the enforced positive definiteness of Σ . See Poirier (1998) for more examples of such learning.

Exercise 14.13 (Posterior simulation in an ordered-treatment, ordered-outcome model)

Like Exercise 14.10, consider the general problem of estimating the effect of treatment on an outcome of interest. In this exercise, however, suppose that y , the outcome of interest, and r , the treatment, can take only three distinct values: $y_i \in \{1, 2, 3\}$ and $r_i \in \{1, 2, 3\}$, $i = 1, 2, \dots, n$. We additionally assume that each variable is *ordered* in nature so that one could reasonably adopt, say, an ordered probit (see, e.g., Exercise 14.5) to model each random variable.

To formally account for the discrete, ordered nature of our variables y and r within the treatment–response framework of Exercise 14.10, we consider a latent variable model

$$z_{yi} = x_{yi}\beta_y + d_{ri}\theta + \epsilon_{yi}, \quad (14.70)$$

$$z_{ri} = x_{ri}\beta_r + \epsilon_{ri}, \quad (14.71)$$

where an intercept has been excluded from x_y and d_{ri} is the 1×3 dummy variable vector for r_i that contains a one in the r_i th column and zeros elsewhere. We interpret the parameter vector θ as quantifying the *treatment effect* of levels of r on y . In our triangular system, we model the endogeneity of r_i by permitting correlation between ϵ_y and ϵ_r . Throughout this exercise, we therefore assume

$$\begin{bmatrix} \epsilon_{yi} \\ \epsilon_{ri} \end{bmatrix} \stackrel{\text{iid}}{\sim} N \left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \sigma_{yr} \\ \sigma_{yr} & 1 \end{pmatrix} \right], \quad (14.72)$$

where the error variances in each of the latent variable equations have been normalized to unity for identification purposes.

Like the single-equation ordered probit of Exercise 14.5, we relate the observed ordered y and r variables to the latent z_y in (14.70) and z_r in (14.71) through the restrictions

$$y_i = j \text{ iff } \gamma_j < z_{yi} \leq \gamma_{j+1}, \quad j = 1, 2, 3, \quad (14.73)$$

and

$$r_i = k \text{ iff } \tilde{\gamma}_k < z_{ri} \leq \tilde{\gamma}_{k+1}, \quad k = 1, 2, 3. \quad (14.74)$$

For identification purposes, we impose standard restrictions on certain values of the *cut-points* γ_j and $\tilde{\gamma}_k$, namely $\gamma_1 = \tilde{\gamma}_1 = -\infty$, $\gamma_2 = \tilde{\gamma}_2 = 0$ and $\gamma_4 = \tilde{\gamma}_4 = \infty$. Thus, with three possible values for each random variable y and r , γ_3 and $\tilde{\gamma}_3$ are the only cutpoints whose values are unknown.

(a) Describe a posterior simulator for fitting this treatment effect model. When doing so, consider, as in Li and Tobias (2005), making use of a reparameterization of the following form:

$$\sigma_y = 1/[\gamma_3^2], \quad \beta_y^* = \sqrt{\sigma_y}\beta_y, \quad z_y^* = \sqrt{\sigma_y}z_y, \quad (14.75)$$

and

$$\sigma_r = 1/[\tilde{\gamma}_3^2], \quad \beta_r^* = \sqrt{\sigma_r}\beta_r, \quad z_r^* = \sqrt{\sigma_r}z_r, \quad \tilde{\sigma}_{yr} = \sqrt{\sigma_y}\sqrt{\sigma_r}\sigma_{yr}. \quad (14.76)$$

How does this reparameterization facilitate posterior computation in the ordered-treatment, ordered-response model?

(b) The data provided on the Web site associated with this book can be used to investigate the impact of maternal alcohol consumption during pregnancy on subsequent infant health. For our purposes here, we create alcohol consumption and health status variables that are ordered and take three possible values. Specifically, infant health is constructed from a variable describing the number of doctor visits made by the infant during his or her first year of life for reasons related to illness or injury. Our constructed measure of infant health takes the value 1 if the child had no such doctor visits, 2 if the child had one or two such visits, and 3 if the child had at least three visits. Maternal alcohol consumption while pregnant takes the value of 1 if the mother never consumed alcohol while pregnant, 2 if the mother drank about once per month (or less), and 3 if the mother drank at least once a month.

Using these data and the algorithm described in (a), fit the ordered-treatment, ordered-outcome model and interpret your results.

Solution

(a) To motivate the suggested reparameterization, consider taking Equations (14.70) and (14.71) and multiplying them by $\sqrt{\sigma_y}$ and $\sqrt{\sigma_r}$, respectively. With the notation of Equations (14.75) and (14.76), such a multiplication produces a model of the form

$$z_{yi}^* = \bar{x}_{yi}\beta_y^* + \epsilon_{yi}^*, \quad (14.77)$$

$$z_{ri}^* = x_{ri}\beta_r^* + \epsilon_{ri}^*, \quad (14.78)$$

where

$$\begin{bmatrix} \epsilon_{yi}^* \\ \epsilon_{ri}^* \end{bmatrix} \stackrel{\text{iid}}{\sim} N \left[\begin{pmatrix} 0 \\ 0 \end{pmatrix} \begin{pmatrix} \sigma_y & \tilde{\sigma}_{yr} \\ \tilde{\sigma}_{yr} & \sigma_r \end{pmatrix} \right] \equiv N(0, \Sigma^*), \quad (14.79)$$

and $\tilde{\sigma}_{yr} = \sqrt{\sigma_y}\sqrt{\sigma_r}\sigma_{yr}$. To simplify notation, we have included d_{ri} as an element of \bar{x}_{yi} (i.e., $\bar{x}_{yi} = [x_{ri} \ d_{ri}]$), and similarly, we have absorbed $\theta^* = \sqrt{\sigma_y}\theta$ in β_y^* .

In terms of the latent data z_y^* and z_r^* , note that the restrictions in (14.73) and (14.74) imply

$$y_i = j \text{ iff } \gamma_j^* < z_{yi}^* \leq \gamma_{j+1}^*, \quad j = 1, 2, 3, \quad (14.80)$$

and

$$r_i = k \text{ iff } \tilde{\gamma}_k^* < z_{ri}^* \leq \tilde{\gamma}_{k+1}^*, \quad k = 1, 2, 3, \quad (14.81)$$

where $\gamma^* \equiv \sqrt{\sigma_y}\gamma$ and $\tilde{\gamma}^* \equiv \sqrt{\sigma_y}\tilde{\gamma}$. Note that all elements of γ^* and $\tilde{\gamma}^*$ are *known*, since our reparameterization implies γ_3^* and $\tilde{\gamma}_3^* = 1$.

The primary benefits afforded by this parameterization are twofold. First, as already noted, there are no unknown cutpoints in this reparameterized model, and thus the researcher does not need to address the issue of how to best simulate the cutpoint values. Next, note that our reparameterization essentially eliminates the diagonal restrictions on the covariance matrix in (14.72) and thus “restores” our ability to apply standard simulation-based posterior analysis, where the requisite posterior conditionals will have convenient forms (as shown in the following).

To this end, we employ the following priors:

$$\beta_y^* \sim N(0, V_y), \quad (14.82)$$

$$\beta_r^* \sim N(0, V_r), \quad (14.83)$$

$$\Sigma^{*-1} \sim W(H, a). \quad (14.84)$$

Let $\beta^* = [\beta_y^{*'} \beta_r^{*'}]'$. These priors combined with the augmented likelihood for the observed data yield the following posterior:

$$p(\beta^*, \Sigma^{*-1}, z_y^*, z_r^* | y, r) \propto p(\beta^*) p(\Sigma^{*-1}) \left\{ \prod_{i=1}^n \phi_2 \left[\begin{pmatrix} z_{yi}^* \\ z_{ri}^* \end{pmatrix}; \begin{pmatrix} \bar{x}_{yi} \beta_y^* \\ x_{ri} \beta_r^* \end{pmatrix}, \begin{pmatrix} \sigma_y & \tilde{\sigma}_{yr} \\ \tilde{\sigma}_{yr} & \sigma_r \end{pmatrix} \right] \right. \\ \left. \times I(\gamma_{yi}^* < z_{yi}^* \leq \gamma_{yi+1}^*) I(\tilde{\gamma}_{ri}^* < z_{ri}^* \leq \tilde{\gamma}_{ri+1}^*) \right\}, \quad (14.85)$$

where $\beta^* \sim N(0, V_\beta)$ and V_β is block diagonal with V_y and V_r along the upper and lower blocks, respectively.

The posterior simulator for this model involves sampling from the posterior conditionals for β^* , the latent variables, and inverse covariance matrix Σ^{*-1} . For the first of these,

$$\beta^* | \Xi_{-\beta^*}, y, r \sim N(D_\beta d_\beta, D_\beta), \quad (14.86)$$

where $\Xi_{-\theta}$ denotes all quantities in the posterior other than θ ,

$$D_\beta = \left[X'(\Sigma^{*-1} \otimes I_n)X + V_\beta^{-1} \right]^{-1}, \quad d_\beta = X'(\Sigma^{*-1} \otimes I_n)z^*,$$

$$X = \begin{pmatrix} \bar{X}_y & 0_{n \times k_r} \\ 0_{n \times k_y} & X_r \end{pmatrix}, \quad V_\beta = \begin{pmatrix} V_y & 0_{k_y \times k_r} \\ 0_{k_r \times k_y} & V_r \end{pmatrix},$$

$$z^* = \begin{bmatrix} z_{y1}^* \\ \vdots \\ z_{yn}^* \\ z_{r1}^* \\ \vdots \\ z_{rn}^* \end{bmatrix}, \quad \bar{X}_y = \begin{bmatrix} \bar{x}_{y1} \\ \vdots \\ \bar{x}_{yn} \end{bmatrix}, \quad \text{and} \quad X_r = \begin{bmatrix} x_{r1} \\ \vdots \\ x_{rn} \end{bmatrix}.$$

Here, k_y and k_r denote the dimensions of β_y^* and β_r^* , respectively.

The latent data in the outcome equation, z_{yi}^* , are sampled independently from conditionals of the form

$$z_{yi}^* | \Xi_{-z_{yi}^*}, y, r \stackrel{\text{ind}}{\sim} TN_{(\gamma_{yi}^*, \gamma_{yi+1}^*]}(\mu_{i,y|r}, \sigma_{y|r}), \quad i = 1, 2, \dots, n, \quad (14.87)$$

where $\mu_{i,y|r} \equiv \bar{x}_{yi} \beta_y^* + \tilde{\sigma}_{yr} \sigma_r^{-1} (z_{ri}^* - x_{ri} \beta_r^*)$, $\sigma_{y|r} \equiv \sigma_y - \tilde{\sigma}_{yr}^2 \sigma_r^{-1}$, and $TN_{(a,b]}(\mu, \sigma)$ denotes a Normal distribution with mean μ and variance σ truncated to the interval between a and b .

The latent data in the treatment equation, z_{ri}^* , are sampled similarly:

$$z_{ri}^* | \Xi_{-z_{ri}^*}, y, r \stackrel{\text{ind}}{\sim} TN_{(\tilde{\gamma}_{ri}^*, \tilde{\gamma}_{ri+1}^*)}(\mu_{i,r|y}, \sigma_{r|y}), \quad i = 1, 2, \dots, n, \quad (14.88)$$

where $\mu_{i,r|y} \equiv x_{ri}\beta_r^* + \tilde{\sigma}_{yr}\sigma_y^{-1}(z_{yi}^* - \bar{x}_{yi}\beta_y^*)$ and $\sigma_{r|y} \equiv \sigma_r - \tilde{\sigma}_{yr}^2\sigma_y^{-1}$.

Finally, the inverse covariance matrix Σ^{*-1} is sampled from its posterior conditional:

$$\Sigma^{*-1} | \Xi_{-\Sigma^{*-1}}, y, r \sim W \left(\left(H^{-1} + \sum_{i=1}^n [\epsilon_{yi}^* \epsilon_{ri}^*]' [\epsilon_{yi}^* \epsilon_{ri}^*] \right)^{-1}, n + a \right), \quad (14.89)$$

where ϵ_y^* and ϵ_r^* are effectively known given z_y^* , z_r^* , and β^* . The posterior simulator involves successively sampling from (14.86) through (14.89). To recover the original parameters of interest in (14.70) through (14.73), simply invert the identities given in (14.75) and (14.76) at each iteration of the sampler. In cases where y and r take on more than three values, an additional step must be added to the posterior simulator to sample the added cutpoint values. For such cases, Nandram and Chen (1996) and Li and Tobias (2005) recommend sampling *differences* between cutpoint values using a Dirichlet proposal density.

(b) For this analysis, we employ priors with hyperparameters given as follows: $a = 2$, $H = I_2$, $V_y = 10I_{k_y}$, and $V_r = 10I_{k_r}$. The Gibbs sampler is run for 500 iterations, the first 100 of which are discarded as the burn-in.

In our analysis we include controls for education and family income in both the drinking and doctor visit equations. In the drinking equation, we additionally include an indicator denoting whether the mother had a parent who had a drinking problem or was an alcoholic (Parent Drink). In the doctor visit equation, we also include an indicator denoting whether the mother has had children before and an indicator denoting whether the mother has a health problem that limits her ability to work (Health Problem). Table 14.6 presents posterior means and standard deviations associated with parameters of this model.

From an examination of the posterior means in Table 14.6, we see that higher levels of alcohol consumption while pregnant generally resulted in the infant experiencing more doctor visits during his or her first year of life. Note, of course, that coefficients in this case are not directly interpretable as marginal effects of interest, though such effects can be calculated using the simulated draws from the posterior. We have not done any formal model comparison tests here, but an informal examination of posterior standard deviations suggests that versions of this model with some coefficients restricted to zero may be preferred. Interestingly, the posterior mean of the covariance parameter σ_{yr} (which in this case is also the correlation) is negative, and most of the mass of this posterior is placed over negative values. A potential explanation for this result is that “indifferent” mothers – who are more likely to drink while pregnant – may also be less likely to take their children to the doctor at the onset of illness, leading to fewer observed visits, on average.

Table 14.6: Coefficient posterior means and standard deviations from an ordered-treatment, ordered-outcome model.

Variable/Parameter	Drinking Frequency Equation	
	Post. Mean	Post. Std. Dev.
Constant	−1.10	.233
Education	.050	.020
Family Income (\$10,000)	.024	.029
Parent Drink	.255	.083
$\tilde{\gamma}_3$.852	.045

Variable / Parameter	Doctor Visit Equation	
	Post. Mean	Post. Std. Dev
Never Drink	−.286	.213
Drink once/month or less	.286	.305
Drink more than once/month	.483	.406
Education	.022	.020
Havekids Before	.002	.063
Family Income (\$10,000)	.023	.029
Health Problem	.130	.066
γ_3	1.16	.064
σ_{yr}	−.328	.149

Exercise 14.14 (Posterior simulation in the stochastic frontier model) Consider a model of production where output of firm i at time t , Y_{it} , is produced using a vector of inputs, X_{it}^* , ($i = 1, \dots, N, t = 1, \dots, T$). Firms have access to a common best-practice technology for turning inputs into output characterized by a production frontier $f(X_{it}^*; \beta)$. Since actual output can fall short of the maximum possible, we write

$$Y_{it} = f(X_{it}^*; \beta) \tau_i, \quad (14.90)$$

where $0 < \tau_i \leq 1$ is a measure of firm-specific efficiency and $\tau_i = 1$ indicates firm i is fully efficient. Following standard econometric practice, we allow for a random error to the model, ζ_{it} , to capture measurement (or specification) error, resulting in

$$Y_{it} = f(X_{it}^*; \beta) \tau_i \zeta_{it}. \quad (14.91)$$

The inclusion of measurement error makes the frontier stochastic and hence the model is named the *stochastic frontier* model. If the production frontier $f(\cdot)$ is log-linear we can take logs and obtain

$$y_{it} = X_{it} \beta + \epsilon_{it} - z_i, \quad (14.92)$$

where $\beta = [\beta_1 \ \beta_2 \ \dots \ \beta_k]'$, $y_{it} = \ln(Y_{it})$, $\epsilon_{it} = \ln(\zeta_{it})$, $z_i = -\ln(\tau_i)$, and X_{it} is the counterpart of X_{it}^* with the inputs transformed to logarithms. The variable z_i is referred

to as *inefficiency* and, since $0 < \tau_i \leq 1$, it is a nonnegative random variable. We assume ϵ_{it} to be iid $N(0, h^{-1})$ and independent of z_i . This is a latent variable model since the inefficiencies z_i are unobserved.

Equation (14.92) can be written

$$y_i = X_i \beta + \epsilon_i - z_i \iota_T, \quad (14.93)$$

if we stack all variables into matrices in the obvious fashion.

In this exercise, assume $p(\beta, h) = p(\beta) p(h)$, with

$$\beta \sim N(\underline{\beta}, \underline{V}) \quad (14.94)$$

and

$$h \sim \gamma(\underline{s}^{-2}, \underline{\nu}). \quad (14.95)$$

For the inefficiencies assume an exponential hierarchical prior, with z_i and z_j being *a priori* independent for $i \neq j$. Thus, for all i ,

$$z_i \sim \gamma(\mu_z, 2). \quad (14.96)$$

For μ_z , use a prior of the form

$$\mu_z^{-1} \sim \gamma(\underline{\mu}_z^{-1}, \underline{\nu}_z). \quad (14.97)$$

(a) Derive the full posterior conditionals for this model.

(b) Using the results from part (a) carry out an empirical Bayesian analysis using a suitable data set. Note that the answer in the following uses an artificial data set that was generated from

$$y_{it} = 1.0 + 0.75x_{2,it} + 0.25x_{3,it} - z_i + \epsilon_{it}$$

for $i = 1, \dots, 100$ and $t = 1, \dots, 5$. We assume $\epsilon_{it} \sim N(0, 0.04)$, $z_i \sim \gamma(-\ln[.85], 2)$, $x_{2,it} \sim U(0, 1)$, and $x_{3,it} \sim U(0, 1)$, where all the random variables are independent of one another and independent over all i and t .

Solution

(a) From the properties of the normal distribution, we have

$$p(y|\beta, h, z) = \prod_{i=1}^N \frac{h^{\frac{T}{2}}}{(2\pi)^{\frac{T}{2}}} \left\{ \exp \left[-\frac{h}{2} (y_i - X_i \beta + z_i \iota_T)' (y_i - X_i \beta + z_i \iota_T) \right] \right\}. \quad (14.98)$$

However, conditional on z , this is simply a standard normal likelihood function. So derivations very similar to those in previous exercises can be used (e.g., Chapter 10, Exercise 10.1, or Chapter 13, Exercise 13.1). In particular, if we multiply (14.94) by (14.98) it can be seen that the resulting posterior conditional is

$$\beta|y, h, z, \mu_z \sim N(\bar{\beta}, \bar{V}), \quad (14.99)$$

where

$$\bar{V} = \left(\underline{V}^{-1} + h \sum_{i=1}^N X_i' X_i \right)^{-1}$$

and

$$\bar{\beta} = \bar{V} \left(\underline{V}^{-1} \underline{\beta} + h \sum_{i=1}^N X_i' [y_i + z_i \iota_T] \right).$$

Multiplying (14.95) by (14.98), we obtain

$$h|y, \beta, z, \mu_z \sim \gamma(\bar{s}^{-2}, \bar{\nu}), \quad (14.100)$$

$$\bar{\nu} = TN + \underline{\nu},$$

and

$$\bar{s}^2 = \frac{\sum_{i=1}^N (y_i + z_i \iota_T - X_i \beta)' (y_i + z_i \iota_T - X_i \beta) + \underline{\nu} s^2}{\bar{\nu}}.$$

Using Bayes' theorem, we have

$$p(z|y, \beta, h, \mu_z) \propto p(y|z, \beta, h, \mu_z) p(z|\beta, h, \mu_z),$$

which is simply (14.96) times (14.98). Doing this multiplication and noting that the posterior conditionals for the inefficiencies are independent of one another (i.e., z_i and z_j are independent for $i \neq j$), it follows immediately that

$$p(z_i|y_i, X_i, \beta, h, \mu_z) \propto \phi(z_i|\bar{X}_i \beta - \bar{y}_i - (Th\mu_z)^{-1}, (Th)^{-1}) I(z_i \geq 0), \quad (14.101)$$

where $\bar{y}_i = (\sum_{t=1}^T y_{it})/T$ and \bar{X}_i is a $(1 \times k)$ matrix containing the average value of each explanatory variable for individual i . Recall that $I(z_i \geq 0)$ is the indicator function, which equals one if $z_i \geq 0$ and is otherwise equal to zero.

An examination of (14.97) times (14.98) yields the posterior conditional for μ_z^{-1} , which can be seen to be

$$\mu_z^{-1}|y, \beta, h, z \sim \gamma(\bar{\mu}_z, \bar{\nu}_z), \quad (14.102)$$

$$\bar{\nu} = 2N + \underline{\nu}_z,$$

and

$$\bar{\mu}_z = \frac{2N + \underline{\nu}_z}{2 \sum_{i=1}^N z_i + \underline{\nu}_z \mu_z}.$$

(b) A Gibbs sampler with data augmentation involves sequentially drawing from (14.99), (14.100), (14.101), and (14.102). All of these posterior conditionals involve standard distributions. (See Exercise 14.1 for a discussion of drawing from the truncated normal.) Matlab code that does this is available on the Web site associated with this book.

With regards to prior hyperparameters we choose

$$\underline{\beta} = \begin{bmatrix} 0.0 \\ 0.5 \\ 0.5 \end{bmatrix}$$

and

$$\underline{V} = \begin{bmatrix} 100.0 & 0.0 & 0.0 \\ 0.0 & 0.25^2 & 0.0 \\ 0.0 & 0.0 & 0.25^2 \end{bmatrix}.$$

These values are chosen to be relatively noninformative about the intercept, but they are slightly more informative about the slope coefficients. In particular, they reflect a belief that large deviations from constant returns to scale are unlikely. For the error precision, we make the noninformative choice of $\underline{\nu} = 0$, which makes the choice of \underline{s}^2 irrelevant. For μ_z , we make the relatively noninformative choice of $\underline{\nu}_z = 2$ with $\underline{\mu}_z = -\ln(0.85)$, which implies that the median of the prior efficiency distribution is 0.85. Posterior results are based on 30,000 replications, with 5,000 burn-in replications discarded and 25,000 replications retained.

Table 14.7 contains posterior means and standard deviations for the parameters of the stochastic frontier model. With stochastic frontier models, interest often centers on the firm-specific efficiencies, τ_i for $i = 1, \dots, N$. Since $\tau_i = \exp(-z_i)$, and the Gibbs sampler yields draws of z_i , we can simply transform them and average to obtain estimates of $E(\tau_i|y)$. For the sake of brevity, we do not present results for all $N = 100$ efficiencies. Rather, we select the firms that have the minimum, median, and maximum values for $E(\tau_i|y)$. These are labeled τ_{\min} , τ_{med} and τ_{\max} in Table 14.7.

It can be seen that the posterior means of all parameters are quite close to the true values used to generate the data set [note that $-\ln(0.85) = 0.16$] and they are all accurately estimated. The posterior means of the efficiencies are reasonable, although their posterior standard deviations are relatively large.

Table 14.7: Posterior results for an artificial data set from the stochastic frontier model.

	Mean	Standard Deviation
β_1	0.98	0.03
β_2	0.74	0.03
β_3	0.27	0.03
h	26.69	1.86
μ_z	0.15	0.02
τ_{\min}	0.56	0.05
τ_{med}	0.89	0.06
τ_{\max}	0.97	0.03

Exercise 14.15 (Posterior simulation in the two-part [hurdle] model) A variant of the tobit model (see Exercise 14.8) for analyzing a censored outcome is the *two-part model* [e.g., Cragg (1971)]. Like the tobit, the two-part specification provides a framework for modeling a mixed discrete–continuous variable that is characterized by a discrete mass over zero and a continuous distribution over strictly positive values. As the name suggests, the two-part model is composed of two equations: one equation determining whether the outcome is positive or zero (i.e., a “decision to use” equation) and a second equation specifying a continuous distribution over strictly positive values (i.e., a “level-of-use” equation). In contrast to the tobit model, which introduces only one equation, the two-part model can allow different covariates to enter the decision to use and level-of-use equations and does not impose restrictions on parameters across equations. This added flexibility is often argued to be beneficial for capturing key features of the censored variable, particularly in situations where the model is characterized by a large fraction of zero responses.

The two-part model has seen considerable use in empirical work and was championed by researchers at RAND for use in health economics and the modeling of medical expenditure data [e.g., Manning, Morris, and Newhouse (1981) and Manning, Duan, and Rogers (1987)]. These works predate the MCMC revolution in econometrics, though recent Bayesian treatments of the model can be found in van Hasselt (2005) and Deb, Munkin, and Trivedi (2006).

In what follows, we consider the following representation of the two-part model:

$$D_i^* = x_{1i}\beta_1 + \epsilon_{1i}, \quad (14.103)$$

$$\ln y_i | D_i^* > 0 = x_{2i}\beta_2 + \epsilon_{2i}. \quad (14.104)$$

Equation (14.103) for the latent variable D_i^* denotes the “first” part of the model and describes whether the outcome y_i is positive or zero. Specifically,

$$D_i \equiv I(D_i^* > 0) = I(y_i > 0).$$

The quantities $(D_i, y_i, x_{1i}, x_{2i})$ denote the data observed for i . Finally, in most applications, the level-of-use equation is specified in logarithmic form, which explains the appearance of $\ln y$ on the left-hand side of (14.104).

The two-part model typically assumes $\epsilon_{1i} \stackrel{\text{iid}}{\sim} N(0, 1)$ and $\epsilon_{2i} \stackrel{\text{iid}}{\sim} N(0, \sigma_\epsilon^2)$. Thus, the model assumes that the level of the outcome is independent of the decision to use, conditioned on the event that some level of use has taken place. That is, the errors of (14.103) and (14.104) are assumed independent, conditioned on $\epsilon_1 > -x_1\beta_1$.

(a) Using priors of the form

$$\beta_j \sim N(\mu_{\beta_j}, V_{\beta_j}), \quad j = 1, 2, \quad (14.105)$$

$$\sigma_\epsilon^2 \sim IG(a, b), \quad (14.106)$$

describe a posterior simulator for fitting the two-part model.

(b) Consider the two-part specification with only an intercept in equations (14.103) and (14.104). Generate 1,000 observations from this two-part model by setting $\beta_1 = .5, \beta_2 = 1$,

and $\sigma_\epsilon^2 = 4$. For the prior hyperparameters, set $\mu_j = 0$, $V_{\beta_j} = 10$ ($j = 1, 2$), $a = 3$, and $b = 1$. Write a program to implement the algorithm derived in part (a), and apply this program to your generated data.

(c) On the Web site associated with this book, we provide data on the volumes of bilateral trade (in thousands of U.S. dollars) among 79 pairs of countries. Specifically, the data set provided quantifies the amount of trade occurring between 79 different countries in a type of products called differentiated goods. Each country is observed to trade (or not trade) with all other countries, producing a grand total of 3,081 observations. Using only an intercept in Equations (14.103) and (14.104), fit the two-part model to these data and comment on your results.

Solution

(a) To fit this model, we make use of data augmentation and work with a joint posterior of the form $p(D^*, \beta_1, \beta_2, \sigma_\epsilon^2 | D, y)$, where

$$D^* = \begin{bmatrix} D_1^* \\ D_2^* \\ \vdots \\ D_n^* \end{bmatrix}, \quad D = \begin{bmatrix} D_1 \\ D_2 \\ \vdots \\ D_n \end{bmatrix}, \quad \text{and} \quad y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}.$$

Though the conditioning on D in the posterior is essentially redundant given y , we maintain this notation and conditioning for the sake of clarity.

Note that

$$\begin{aligned} p(D^*, \beta_1, \beta_2, \sigma_\epsilon^2 | D, y) &\propto p(\beta_1)p(\beta_2)p(\sigma_\epsilon^2)p(D^*, y | \beta_1, \beta_2, \sigma_\epsilon^2) \\ &\quad \times p(D | D^*, y, \beta_1, \beta_2, \sigma_\epsilon^2) \\ &= p(\beta_1)p(\beta_2)p(\sigma_\epsilon^2)[p(D | D^*, y, \beta_1, \beta_2, \sigma_\epsilon^2) \\ &\quad \times p(y | D^*, \beta_1, \beta_2, \sigma_\epsilon^2)p(D^* | \beta_1, \beta_2, \sigma_\epsilon^2)] \\ &= p(\beta_1)p(\beta_2)p(\sigma_\epsilon^2) \\ &\quad \times \prod_{i=1}^n \left([I(D_i = 1)I(D_i^* > 0) + I(D_i = 0)I(D_i^* \leq 0)] \right. \\ &\quad \times \left[\frac{1}{y_i} \phi(\ln y_i; x_{2i}\beta_2, \sigma_\epsilon^2)I(D_i^* > 0) + I(y_i = 0)I(D_i^* \leq 0) \right] \\ &\quad \left. \times \phi(D_i^*; x_{1i}\beta_1, 1) \right). \end{aligned} \tag{14.107}$$

A Gibbs sampler involves cycling through conditionals of the form $D^* | \cdot, y, D$, $\beta_1 | \cdot, y, D$, $\beta_2 | \cdot, y, D$, and $\sigma_\epsilon^2 | \cdot, y, D$ where “ \cdot ” is simplifying notation serving to denote the remaining parameters of the model. When sampling β_2 and σ_ϵ^2 from their posterior conditionals, the augmented posterior in (14.107) reveals that only the x_2 and y data arising from the

subsample for which $y_i > 0$ are used. Let X_2^+ and y^+ denote this subsample of data (arranged conformably) and also define

$$X_1 = \begin{bmatrix} x_{11} \\ x_{12} \\ \vdots \\ x_{1n} \end{bmatrix}.$$

With this notation in hand, we obtain posterior conditionals of the following forms:

$$D_i^* | \beta_1, \beta_2, \sigma_\epsilon^2, y, D \stackrel{\text{ind}}{\sim} \begin{cases} TN_{(-\infty, 0]}(x_{1i}\beta_1, 1) & \text{if } D_i = 0, \\ TN_{(0, \infty)}(x_{1i}\beta_1, 1) & \text{if } D_i = 1, \end{cases} \quad (14.108)$$

and

$$\beta_1 | D^*, \beta_2, \sigma_\epsilon^2, y, D \sim N(D_{\beta_1} d_{\beta_1}, D_{\beta_1}), \quad (14.109)$$

where

$$D_{\beta_1} = \left(X_1' X_1 + V_{\beta_1}^{-1} \right)^{-1}, \quad d_{\beta_1} = X_1' D^* + V_{\beta_1}^{-1} \mu_{\beta_1}.$$

For the remaining two posterior conditionals, we obtain

$$\beta_2 | D^*, \beta_1, \sigma_\epsilon^2, y, D \sim N(D_{\beta_2} d_{\beta_2}, D_{\beta_2}), \quad (14.110)$$

where

$$D_{\beta_2} = \left(X_2^{+'} X_2^+ / \sigma_\epsilon^2 + V_{\beta_2}^{-1} \right)^{-1} \quad d_{\beta_2} = X_2^{+'} \ln y^+ / \sigma_\epsilon^2 + V_{\beta_2}^{-1} \mu_{\beta_2},$$

and

$$\sigma_\epsilon^2 | D^*, \beta_1, \beta_2, y, D \sim IG \left(\frac{n^+}{2} + a, [b^{-1} + .5(\ln y^+ - X_2^+ \beta_2)'(\ln y^+ - X_2^+ \beta)]^{-1} \right), \quad (14.111)$$

where n^+ is defined as the number of elements in $\ln y^+$. A posterior simulator proceeds by sampling from (14.108) through (14.111).

(b) MAT LAB code for generating the data and fitting the two-part model are provided on the Web site associated with this book. In our code, the Gibbs sampler is run for 3,000 iterations, and the first 1,000 of these are discarded as the burn-in. In Table 14.8 we provide posterior means and the actual parameters used to generate the data. The results show that the algorithm fares well in recovering the parameters of the data-generating process.

Table 14.8: Posterior means
and true values of the parameters.

	β_1	β_2	σ_ϵ^2
Post. Mean	.489	.976	3.94
True Value	.500	1.00	4.00

(c) We apply the same code used in part (b) to fit the trade data, setting the prior variances of β_1 and β_2 to 1,000. Posterior means of β_1 , β_2 , and σ^2 were equal to .790, .935, and 11.63, respectively. These results imply the probability that no trade occurs is equal to $\Pr(\epsilon_i \leq -\beta_1) = 1 - \Phi(\beta_1) \approx .215$, which matches the fraction of observations with zero trade in the sample. The posterior means of β_2 and σ^2 also match the sample mean and sample variance of log trade values for those observations associated with positive trade. Finally, given our lognormality assumption, the mean of trade, given that trade occurs, is $\exp(\beta_2 + \sigma_\epsilon^2/2)$. Evaluated at posterior means, this quantity equals \$3.8 billion (where we recall that trade has been measured in thousands of dollars).

Exercise 14.16 (Posterior simulation in the sample-selection model) Closely related to the two-part model of Exercise 14.15 is the *sample-selection model*. Unlike the two-part model, the sample-selection model allows for correlation between unobservables in the decision to use and level-of-use equations and also provides an explicit role for the *counterfactual* or *potential outcome* – the amount of expenditure that could have been observed but was not. The merits of and assumptions made by the sample-selection and two-part models have received considerable attention in the literature [e.g., Manning et al. (1987), Hay and Olsen (1984), and Leung and Yu (1996)]. We do not review these issues here; instead, we seek to introduce a Bayesian posterior simulator for the sample-selection model, along the lines of the analysis of van Hasselt (2005).

We consider the following representation of the sample-selection model:

$$D_i^* = x_{1i}\beta_1 + \epsilon_{1i}, \quad (14.112)$$

$$y_i^* = x_{2i}\beta_2 + \epsilon_{2i}, \quad (14.113)$$

where

$$\begin{aligned} \begin{bmatrix} \epsilon_{1i} \\ \epsilon_{2i} \end{bmatrix} &\stackrel{\text{iid}}{\sim} N \left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{pmatrix} \right], \quad i = 1, 2, \dots, n \\ &\equiv N(0, \Sigma). \end{aligned}$$

Like Equation (14.103), (14.112) is the latent variable decision-to-use equation. If $D_i \equiv I(D_i^* > 0)$ is zero, then observed expenditure is zero, whereas if $D_i = 1$, then a positive amount of expenditure is observed. The latent variable y_i^* in (14.113), like (14.104), describes the *logarithm* of observed or *potential* expenditure. Formally, observed expenditure y_i is linked to the latent D_i^* and y_i^* as follows:

$$y_i = \begin{cases} 0 & \text{if } D_i^* \leq 0, \\ \exp(y_i^*) & \text{if } D_i^* > 0. \end{cases} \quad (14.114)$$

The observed data for individual i consist of y_i , D_i , x_{1i} , and x_{2i} . Derive a posterior simulator for fitting the sample-selection model and the priors employed in your analysis.

Solution

Let

$$D^* = \begin{bmatrix} D_1^* \\ \vdots \\ D_n^* \end{bmatrix}, \quad y^* = \begin{bmatrix} y_1^* \\ \vdots \\ y_n^* \end{bmatrix}, \quad D = \begin{bmatrix} D_1 \\ \vdots \\ D_n \end{bmatrix}, \quad \text{and} \quad y = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}.$$

To facilitate posterior computations, we will work with an augmented posterior distribution of the form $p(D^*, y^*, \beta, \Sigma | y, D)$. Note that, under independent priors for β and Σ , denoted $p(\beta)$ and $p(\Sigma)$, we obtain

$$\begin{aligned} p(D^*, y^*, \beta, \Sigma | y, D) &\propto p(D^*, y^*, D, y | \beta, \Sigma) p(\beta) p(\Sigma) \\ &= p(y, D | y^*, D^*, \beta, \Sigma) p(y^*, D^* | \beta, \Sigma) p(\beta) p(\Sigma) \\ &= p(y | D, y^*) p(D | D^*) p(y^*, D^* | \beta, \Sigma) p(\beta) p(\Sigma) \\ &= p(\beta) p(\Sigma) \prod_{i=1}^n p(y_i | D_i, y_i^*) p(D_i | D_i^*) \phi \left(\begin{bmatrix} D_i^* \\ y_i^* \end{bmatrix}; \begin{bmatrix} x_{1i}\beta_1 \\ x_{2i}\beta_2 \end{bmatrix}, \Sigma \right). \end{aligned}$$

In the third line of this expression, we write the joint distribution of y and D given the parameters of the augmented posterior as the conditional $y|D$ times the marginal for D , and we have dropped extraneous variables from the conditioning set. These two densities can be shown to equal

$$p(D_i | D_i^*) = I(D_i = 1)I(D_i^* > 0) + I(D_i = 0)I(D_i^* \leq 0)$$

and

$$p(y_i | D_i, y_i^*) = D_i I[y_i = \exp(y_i^*)] + (1 - D_i) I(y_i = 0).$$

Before describing the posterior conditionals of this model, let

$$r_i \equiv \begin{bmatrix} D_i^* \\ y_i^* \end{bmatrix}, \quad X_i \equiv \begin{bmatrix} x_{1i} & 0 \\ 0 & x_{2i} \end{bmatrix}, \quad \text{and} \quad \beta = \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix}.$$

Then, under a prior for β of the form $\beta \sim N(\mu_\beta, V_\beta)$,

$$\beta | D^*, y^*, \Sigma, y, D \sim N(D_\beta d_\beta, D_\beta), \quad (14.115)$$

where

$$D_\beta = \left(\sum_{i=1}^n X_i' \Sigma^{-1} X_i + V_\beta^{-1} \right)^{-1}, \quad d_\beta = \sum_{i=1}^n X_i' \Sigma^{-1} r_i + V_\beta^{-1} \mu_\beta.$$

To sample from the joint posterior conditional distribution of D^* and y^* , we need to consider two distinct cases. When $D_i = 1$, the conditional posterior distribution of y_i^* is degenerate [i.e., $y_i^* = \ln(y_i)$], and

$$D_i^* | \beta, \Sigma, y^*, y, D \stackrel{\text{ind}}{\sim} TN_{(0,\infty)} \left[x_{1i}\beta_1 + \frac{\sigma_{12}}{\sigma_2^2} [y_i^* - x_{2i}\beta_2], \left(1 - \frac{\sigma_{12}^2}{\sigma_2^2} \right) \right]. \quad (14.116)$$

When $D_i = 0$ we obtain

$$p(D_i^*, y_i^* | \beta, \Sigma, y, D) \propto I(D_i^* \leq 0) \phi \left(\begin{bmatrix} D_i^* \\ y_i^* \end{bmatrix}, \begin{bmatrix} x_{1i}\beta_1 \\ x_{2i}\beta_2 \end{bmatrix}, \Sigma \right).$$

We can employ a blocking step to sample D_i^*, y_i^* jointly from this conditional. Specifically, marginalized over y_i^* , it follows that

$$D_i^* | \beta, \Sigma, y, D \stackrel{\text{ind}}{\sim} TN_{(-\infty, 0]}(x_{1i}\beta, 1), \quad (14.117)$$

and the complete conditional for y_i^* is given as

$$y_i^* | D_i^*, \beta, \Sigma, y, D \stackrel{\text{ind}}{\sim} N \left(x_{2i}\beta_2 + \sigma_{12}(D_i^* - x_{1i}\beta_1), \sigma_2^2 \left[1 - \frac{\sigma_{12}^2}{\sigma_2^2} \right] \right). \quad (14.118)$$

To sample the parameters of the covariance matrix Σ , we follow the reparameterization strategy described in Exercise 14.10, equations (14.50)–(14.52). Specifically, we write $\epsilon_{2i} = \sigma_{12}\epsilon_{1i} + \nu_i$, $\nu_i \sim N(0, \sigma_\nu^2)$ with $\sigma_\nu^2 = \sigma_2^2 - \sigma_{12}^2$. Under this parameterization, we can write Σ as

$$\Sigma = \begin{bmatrix} 1 & \sigma_{12} \\ \sigma_{12} & \sigma_\nu^2 + \sigma_{12}^2 \end{bmatrix}$$

and employ priors of the form $\sigma_\nu^2 \sim IG(a, b)$ and $\sigma_{12} \sim N(\mu_0, V_0)$. Our two-equation system in (14.112) and (14.113) can then be rewritten as

$$\begin{aligned} D_i^* &= x_{1i}\beta_1 + \epsilon_{1i}, \\ y_i^* &= x_{2i}\beta_2 + \sigma_{12}\epsilon_{1i} + \nu_i, \end{aligned}$$

whence it follows

$$\sigma_{12} | \sigma_\nu^2, \beta, D^*, y^*, y, D \sim N(D_\sigma d_\sigma, D_\sigma), \quad (14.119)$$

where

$$D_\sigma = (\epsilon_1' \epsilon_1 / \sigma_\nu^2 + V_0^{-1})^{-1}, \quad d_\sigma = \epsilon_1' (y^* - X_2 \beta_2) / \sigma_\nu^2 + V_0^{-1} \mu_0,$$

and

$$\sigma_\nu^2 | \sigma_{12}, \beta, D^*, y^*, y, D \sim IG \left(\frac{n}{2} + a, \left[b^{-1} + \frac{1}{2} \nu' \nu \right]^{-1} \right). \quad (14.120)$$

In (14.119), note that $\epsilon_1 = [\epsilon_{1i} \ \epsilon_{2i} \ \cdots \ \epsilon_{ni}]'$ is known given D^* and β , in (14.120) note that $\nu \equiv [\nu_1 \ \cdots \ \nu_n]'$ is known given y^*, D^*, β , and σ_{12} , and

$$X_2 = \begin{bmatrix} x_{21} \\ x_{22} \\ \vdots \\ x_{2n} \end{bmatrix}.$$

The Gibbs sampler involves drawing from (14.115) through (14.120). Note that when drawing from the posterior conditional for the latent data D^* and y^* , either (14.116) or (14.117)–(14.118) is applied for a given i , depending on the observed value of D_i .

Exercise 14.17 (Missing data, 1) Consider the regression model

$$y_i^* = \beta_0 + \beta_1 x_i + \epsilon_i, \quad \epsilon_i \stackrel{\text{iid}}{\sim} N(0, \sigma^2).$$

Assume that x_i is observed for all $i = 1, 2, \dots, n$, and that y_i^* is observed (and equal to y_i) for $i = 1, 2, \dots, m, m < n$. For $i = m+1, m+2, \dots, n$, the dependent variable is missing, and is presumed to be generated by the equation given here. Note that in this question and those that follow, do not deal with issues related to data being missing completely at random, at random, or not at random. Here, the “missing” indicator is presumed to be generated by a process unrelated to the regression parameters and y^* (i.e., the data are missing completely at random). [See, e.g., Gelman, Carlin, Stern, and Rubin (2004, Chapter 21) for further discussion and references.]

(a) Let $\beta = [\beta_0 \ \beta_1]'$ and employ priors of the form $\beta \sim N(\mu_\beta, V_\beta)$ and $\sigma^2 \sim IG(a, b)$. Describe a data augmentation approach for “filling in” the missing data and fitting the regression model.

(b) Is posterior inference regarding β_0, β_1 , and σ^2 sharpened by including the latent missing data into the joint posterior?

Solution

(a) First, stack the regression equation into vector and matrix form as follows:

$$\begin{bmatrix} y_1 \\ \vdots \\ y_m \\ y_{m+1}^* \\ \vdots \\ y_n^* \end{bmatrix} = \begin{bmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_m \\ 1 & x_{m+1} \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \vdots \\ \epsilon_m \\ \epsilon_{m+1} \\ \vdots \\ \epsilon_n \end{bmatrix},$$

or compactly as

$$y^* = X\beta + \epsilon,$$

with y^*, X, β , and ϵ defined in the obvious ways.

Let y_{obs} denote the first m entries of y^* and let $y_{\text{miss}} = [y_{m+1}^* \cdots y_n^*]'$. Including y_{miss} as an element of the joint posterior, we can write

$$\begin{aligned} p(y_{\text{miss}}, \beta, \sigma^2 | y_{\text{obs}}) &\propto p(y_{\text{obs}}, y_{\text{miss}} | \beta, \sigma^2) p(\beta) p(\sigma^2) \\ &= \phi(y^*; X\beta, \sigma^2 I_n) p(\beta) p(\sigma^2). \end{aligned}$$

Given this result, it follows that the posterior conditionals $\beta | y_{\text{miss}}, \sigma^2, y_{\text{obs}}$ and $\sigma^2 | y_{\text{miss}}, \beta, y_{\text{obs}}$ are of the following forms:

$$\beta | y_{\text{miss}}, \sigma^2, y_{\text{obs}} \sim N(D_\beta d_\beta, D_\beta), \quad (14.121)$$

where

$$D_\beta = (X'X/\sigma^2 + V_\beta^{-1})^{-1}, \quad d_\beta = X'y^*/\sigma^2 + V_\beta^{-1}\mu_\beta,$$

and

$$\sigma^2 | y_{\text{miss}}, \beta, y_{\text{obs}} \sim IG \left\{ \frac{n}{2} + a, [b^{-1} + (1/2)(y^* - X\beta)'(y^* - X\beta)]^{-1} \right\}. \quad (14.122)$$

Similarly, since

$$p(y_{\text{miss}}, \beta, \sigma^2 | y_{\text{obs}}) \propto p(y_{\text{obs}} | \beta, \sigma^2) p(y_{\text{miss}} | \beta, \sigma^2) p(\beta) p(\sigma^2), \quad (14.123)$$

it follows that the missing data y_{miss} can be drawn independently as follows:

$$y_i^* | \beta, \sigma^2, y_{\text{obs}} \stackrel{\text{ind}}{\sim} N(\beta_0 + \beta_1 x_i, \sigma^2), \quad i = m+1, m+2, \dots, n. \quad (14.124)$$

A posterior simulator proceeds by sampling from (14.121), (14.122), and (14.124).

(b) Though somewhat obvious, we note that

$$p(\beta, \sigma^2 | y_{\text{obs}}) = \int_{R_{y_{\text{miss}}}} p(y_{\text{miss}}, \beta, \sigma^2 | y_{\text{obs}}) dy_{\text{miss}} \propto p(y_{\text{obs}} | \beta, \sigma^2) p(\beta) p(\sigma^2).$$

Thus, the addition of the latent, missing data does nothing to “sharpen” inference regarding β and σ^2 . We obtain the same marginal posterior for β and σ^2 that we would have obtained had we worked directly with the density for y_{obs} and ignored the contribution of all observations with missing y data. Since the addition of the missing variables will likely slow down the mixing of the posterior simulator, it really does not pay to work with the augmented likelihood in this case.

Exercise 14.18 (Missing data, 2) Similar to the SUR model of Exercise 11.10, suppose

$$\begin{bmatrix} y_{1i} \\ y_{2i} \end{bmatrix} \stackrel{\text{ind}}{\sim} N \left(\begin{bmatrix} x_{1i}\beta_1 \\ x_{2i}\beta_2 \end{bmatrix}, \begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{bmatrix} \right), \quad i = 1, 2, \dots, n, \quad (14.125)$$

or equivalently,

$$\tilde{y}_i \stackrel{\text{ind}}{\sim} N(\tilde{X}_i\beta, \Sigma), \quad i = 1, 2, \dots, n,$$

where

$$\tilde{y}_i \equiv \begin{bmatrix} y_{1i} \\ y_{2i} \end{bmatrix}, \quad \tilde{X}_i \equiv \begin{bmatrix} x_{1i} & 0 \\ 0 & x_{2i} \end{bmatrix}, \quad \text{and} \quad \beta = \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix}.$$

In addition to this structure, suppose that $[y_{2m+1} \ y_{2m+2} \ \dots \ y_n]'$ are missing, but data on all other variables (i.e., y_1 , x_1 , and x_2) are available. Denote the first m observed elements of y_2 as $y_{2,\text{obs}}$ and the final $n - m$ elements as $y_{2,\text{miss}}$ so that $y_2 = [y'_{2,\text{obs}} \ y'_{2,\text{miss}}]'$.

(a) Employing priors of the forms

$$\beta \sim N(\mu_\beta, V_\beta), \quad \Sigma^{-1} \sim W(\Omega, \nu),$$

augment the posterior with $y_{2,\text{miss}}$ and describe a posterior simulator for fitting the model.

(b) Comment on how the incorporation of the missing data into the posterior simulator facilitates computation in this model.

Solution

Let $y_{\text{obs}} = [y_1' \ y_{2,\text{obs}}']'$ denote the observed outcome data, where y_1 denotes the $n \times 1$ vector of y_{1i} outcomes stacked over i , and y_2 is defined similarly. The augmented posterior distribution can be expressed as

$$\begin{aligned} p(y_{2,\text{miss}}, \beta, \Sigma^{-1} | y_{\text{obs}}) &\propto p(y_1, y_2 | \beta, \Sigma) p(\beta) p(\Sigma^{-1}) \\ &= p(\beta) p(\Sigma^{-1}) \prod_{i=1}^n \phi(\tilde{y}_i; \tilde{X}_i \beta, \Sigma). \end{aligned}$$

A Gibbs sampler applied to this model involves drawing from the complete posterior conditionals for β , Σ^{-1} , and $y_{2,\text{miss}}$. Since, conditioned on the missing outcome data, this model has the same structure as the SUR model of Exercise 11.10, the posterior conditionals for β and Σ^{-1} are the same as those reported in that exercise and thus are not repeated here.

For the posterior conditional for $y_{2,\text{miss}}$, note that the assumptions of our model imply

$$y_{2i} | \beta, \Sigma^{-1}, y_{\text{obs}} \stackrel{\text{ind}}{\sim} N \left(x_{2i} \beta_2 + \frac{\sigma_{12}}{\sigma_1^2} (y_{1i} - x_{1i} \beta_1), \sigma_2^2 (1 - \rho_{12}^2) \right), \quad i = m+1, \dots, n, \quad (14.126)$$

where $\rho_{12}^2 \equiv \sigma_{12}^2 / \sigma_1^2 \sigma_2^2$. The Gibbs sampler involves drawing from the posterior conditionals for β and Σ^{-1} , as described in the SUR analysis of Exercise 11.10, and then drawing independently from (14.126) for $i = m+1, \dots, n$.

(b) Note that when marginalizing the joint posterior in (a) over the latent data, we obtain

$$p(\beta, \Sigma^{-1} | y_{\text{obs}}) \propto p(\beta) p(\Sigma^{-1}) \prod_{i=1}^m \phi(\tilde{y}_i; \tilde{X}_i \beta, \Sigma) \prod_{i=m+1}^n \phi(y_{1i}; x_{1i} \beta_1, \sigma_1^2). \quad (14.127)$$

This is a quite natural and expected result – when the y_2 data are missing, the appropriate contribution to the likelihood is simply the marginal density for the observed y_{1i} .

When working with the observed data only, posterior computation becomes somewhat more cumbersome, given the form of the posterior in (14.127). The appeal of the augmentation approach presented in (a) is that, by conditioning on the latent data, we essentially restore the problem to the standard SUR analysis of Exercise 11.10. This contrasts with Exercise 14.17, where the addition of the missing data to the posterior simulator offered no real computational benefit and needlessly complicated the estimation exercise.

Exercise 14.19 (Missing data, 3) Consider a two-equation model of the form

$$y_i = \beta_0 + \beta_1 x_i + u_i, \quad (14.128)$$

$$x_i = \theta_0 + \theta_1 z_i + \epsilon_i. \quad (14.129)$$

We assume $u_i \stackrel{\text{iid}}{\sim} N(0, \sigma_u^2)$, $\epsilon_i \stackrel{\text{iid}}{\sim} N(0, \sigma_\epsilon^2)$, u and ϵ are independent, and $z_i \stackrel{\text{iid}}{\sim} N(0, 1)$. In addition to this structure, suppose that $x_i = [x_1 \ \cdots \ x_m \ x_{m+1} \ \cdots \ x_n]'$, where $x_{\text{obs}} \equiv [x_1 \ \cdots \ x_m]'$ are observed in the data set, whereas $x_{\text{miss}} \equiv [x_{m+1} \ \cdots \ x_n]'$ are not observed. Data on y and z are available for every observation.

Finally, let

$$\beta = [\beta_0 \ \beta_1]', \quad \theta = [\theta_0 \ \theta_1]', \quad \pi = [\beta' \ \theta' \ \sigma_u^2 \ \sigma_\epsilon^2]'$$

and employ priors of the form $\beta \sim N(\mu_\beta, V_\beta)$, $\theta \sim N(\mu_\theta, V_\theta)$, $\sigma_u^2 \sim IG(a_1, b_1)$, and $\sigma_\epsilon^2 \sim IG(a_2, b_2)$.

(a) Augment the posterior with the missing data x_{miss} and describe a posterior simulator for fitting this model.

(b) Derive an expression for the marginal posterior $p(\pi|y, x_{\text{obs}})$. Is the posterior for β_0 and β_1 affected only by those observations with observed x values?

(c) Generate 2,000 observations from the model in (a) with $\beta_0 = 2$, $\beta_1 = -.7$, $\theta_0 = 1$, $\theta_1 = .35$, $\sigma_u^2 = .4$, and $\sigma_\epsilon^2 = .01$. Once the complete data set has been generated, treat the first 500 x values as observed, and discard the remaining 1,500 as missing.

Consider two different approaches in this situation. For the first case, consider only the equation $y_i = \beta_0 + \beta_1 x_i + u_i$ for the first 500 observations with nonmissing data. Fit the model and comment on your results. Next, estimate the complete model, as given in (14.128) and (14.129) using the algorithm described in (a). Comment on your results. For the priors, center β and θ around a prior mean of zero with $V_\beta = V_\theta = 10I_2$, and choose $a_1 = a_2 = 3$, and $b_1 = b_2 = 1/(2 \times .25)$ (so that the prior means and standard deviations of σ_u^2 and σ_ϵ^2 are all .25).

Solution

(a) Given the assumptions of our model, it follows that

$$p(x_{\text{miss}}, \pi|y, x_{\text{obs}}) \propto p(\pi) \prod_{i=1}^n p(y_i, x_i|\pi).$$

Given the distributional assumptions made in this problem, we obtain

$$y_i|x_i, \pi \stackrel{\text{ind}}{\sim} N(\beta_0 + \beta_1 x_i, \sigma_u^2), \quad (14.130)$$

$$y_i|\pi \stackrel{\text{ind}}{\sim} N(\beta_0 + \beta_1[\theta_0 + \theta_1 z_i], \sigma_u^2 + \beta_1^2 \sigma_\epsilon^2), \quad (14.131)$$

$$x_i|\pi \stackrel{\text{ind}}{\sim} N(\theta_0 + \theta_1 z_i, \sigma_\epsilon^2), \quad (14.132)$$

$$x_i|y_i, \pi \stackrel{\text{ind}}{\sim} N\left(\theta_0 + \theta_1 z_i + \frac{\beta_1 \sigma_\epsilon^2}{\sigma_u^2 + \beta_1^2 \sigma_\epsilon^2} [y_i - \beta_0 - \beta_1(\theta_0 + \theta_1 z_i)], \right. \\ \left. \sigma_\epsilon^2 \left[1 - \frac{\beta_1^2 \sigma_\epsilon^4}{\sigma_u^2 + \beta_1^2 \sigma_\epsilon^2}\right]\right). \quad (14.133)$$

A posterior simulator can be implemented by first sampling x_{miss} from $x_i|y_i, \pi$ in (14.133), for $i = m+1, m+2, \dots, n$. The remaining posterior conditionals are as follows:

$$\theta|x_{\text{miss}}, \pi_{-\theta}, y, x_{\text{obs}} \sim N(D_\theta d_\theta, D_\theta), \quad (14.134)$$

where

$$D_\theta = (Z'Z/\sigma_\epsilon^2 + V_\theta^{-1})^{-1}, \quad d_\theta = Z'x/\sigma_\epsilon^2 + V_\theta^{-1}\mu_\theta,$$

$$Z \equiv \begin{bmatrix} 1 & z_1 \\ 1 & z_2 \\ \vdots & \vdots \\ 1 & z_n \end{bmatrix},$$

and $x = [x'_{\text{obs}} \ x'_{\text{miss}}]'$ has been updated with the most recent values of x_{miss} from the previous step,

$$\sigma_\epsilon^2 | x_{\text{miss}}, \pi_{-\sigma_\epsilon^2}, y, x_{\text{obs}} \sim IG\left(\frac{n}{2} + a_2, [b_2^{-1} + (1/2)(x - Z\theta)'(x - Z\theta)]^{-1}\right), \quad (14.135)$$

$$\beta | x_{\text{miss}}, \pi_{-\beta}, y, x_{\text{obs}} \sim N(D_\beta d_\beta, D_\beta), \quad (14.136)$$

where

$$D_\beta = (X'X/\sigma_u^2 + V_\beta^{-1})^{-1}, \quad d_\beta = X'y/\sigma_u^2 + V_\beta^{-1}\mu_\beta,$$

and

$$X \equiv \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix}.$$

(Alternatively, θ and β could be sampled jointly.) Finally,

$$\sigma_u^2 | x_{\text{miss}}, \pi_{-\sigma_u^2}, y, x_{\text{obs}} \sim IG\left(\frac{n}{2} + a_1, [b_1^{-1} + (1/2)(y - X\beta)'(y - X\beta)]^{-1}\right). \quad (14.137)$$

A posterior simulator proceeds by sampling from (14.133) (for the missing x data) through (14.137).

(b) From our general expression for the joint posterior in (a), it follows that

$$p(\pi | y, x_{\text{obs}}) \propto p(\pi) \prod_{i=1}^m p(y_i | x_i, \pi) p(x_i | \pi) \prod_{i=m+1}^n p(y_i | \pi).$$

Each of the densities on the right-hand side of this expression are represented in (14.130)–(14.133). From the expression for $p(y_i | \pi)$ in (14.131), it is clear that even the observations with missing x values play a role in the posterior for β_0, β_1 , and σ_u^2 . The observed x data aid us in learning about θ_0, θ_1 , and σ_ϵ^2 , which, in turn, help us to “impute” x when it is missing. In the limiting case where $\sigma_\epsilon^2 \approx 0$ and θ_0 and θ_1 are known, the missing data feature of the problem essentially vanishes, as the imputation equation in (14.129) perfectly predicts x .

(c) MATLAB code for implementing this exercise is provided on the Web site associated with this book. For each of the two procedures, we ran the sampler for 5,000 iterations, discarding the first 1,000 of those as the burn-in. Generally speaking (and as one should

expect), both procedures provided reasonably accurate point estimates (e.g., posterior means) of β and σ_u^2 . The posterior standard deviations, however, revealed a different story. For the approach using only the first 500 observations, the posterior standard deviations of β_0 and β_1 were .083 and .078, respectively. When estimating the full model, the standard deviations were nearly cut in half, equaling .043 and .041, respectively. This suggests the potential for some efficiency gains, particularly when an accurate imputation equation is available.

Mixture models

The reader has undoubtedly noted that the majority of exercises in previous chapters have included normality assumptions. These assumptions should not be taken as necessarily “correct,” and indeed, such assumptions will not be appropriate in all situations (see, e.g., Exercise 11.13 for a question related to error-term diagnostic checking). In this chapter, we review a variety of computational strategies for extending analyses beyond the textbook normality assumption. The strategies we describe, perhaps ironically, begin with the normal model and proceed to augment it in some way. In the first section, we describe *scale mixtures of normals* models, which enable the researcher to generalize error distributions to the Student t and double exponential classes (among others). Exercises 15.5 and 15.6 then describe *finite normal mixture models*, which can accommodate, among other features, skewness and multimodality in error distributions. Importantly, these models are *conditionally normal* (given values of certain mixing or component indicator variables), and thus standard computational techniques for the normal linear regression model can be directly applied (see, e.g., Exercises 10.1 and 11.8). The techniques described here are thus computationally tractable, flexible, and generalizable. That is, the basic methods provided in this chapter can often be adapted in a straightforward manner to add flexibility to many of the models introduced in previous and future exercises. Some exercises related to mixture models beyond those involving conditional normal sampling distributions are also provided within this chapter.

For the case of finite mixtures, we do not discuss issues of identifiability of the model’s parameters up to relabeling of the mixture components, and we invite the reader to consult, for example, Richardson and Green (1997), McLachlan and Peel (2000), Celeux, Hurn, and Robert (2000), and Geweke (2006) for further discussion of this and related issues. Instead, we focus our attention on algorithms for fitting mixture models as well as providing generated-data experiments and illustrative applications. We remind the reader that Matlab

code for carrying out the empirical parts of the questions are on the Web site associated with this book.

15.1 Some scale mixture of normals models

Exercise 15.1 (The student t -distribution as a scale mixture of normals) Suppose you specify the density

$$y|\beta, \lambda, \sigma^2 \sim N(x\beta, \lambda\sigma^2) \quad (15.1)$$

and specify the following prior for λ (treating ν as given):

$$\lambda|\nu \sim IG\left(\frac{\nu}{2}, \frac{2}{\nu}\right) \Rightarrow p(\lambda) = \left[\Gamma\left(\frac{\nu}{2}\right) \left(\frac{2}{\nu}\right)^{\nu/2}\right]^{-1} \lambda^{-(\nu/2)+1} \exp\left(-\frac{\nu}{2\lambda}\right). \quad (15.2)$$

(a) Derive the density

$$p(y|\beta, \sigma^2) = \int_0^\infty p(y|\beta, \lambda, \sigma^2) p(\lambda) d\lambda.$$

(b) Given the result in (a), comment on how the addition of λ to the error variance can be a useful computational device for an applied Bayesian researcher.

Solution

(a) It follows that

$$\begin{aligned} p(y|\beta, \sigma^2) &= \int_0^\infty \left([\sqrt{2\pi}\sigma]^{-1} \lambda^{-(1/2)} \exp\left[-\frac{1}{2\lambda} \left(\frac{y-x\beta}{\sigma}\right)^2\right] \right) \\ &\quad \times \left(\left[\Gamma\left(\frac{\nu}{2}\right) \left(\frac{2}{\nu}\right)^{\nu/2} \right]^{-1} \lambda^{-(\nu/2)+1} \exp\left(-\frac{\nu}{2\lambda}\right) \right) d\lambda \\ &= [\sqrt{2\pi}\sigma]^{-1} \left[\Gamma\left(\frac{\nu}{2}\right) \left(\frac{2}{\nu}\right)^{\nu/2} \right]^{-1} \\ &\quad \times \int_0^\infty \lambda^{-(\nu+3)/2} \exp\left[-\frac{1}{\lambda} \left(\frac{1}{2} \left[\frac{y-x\beta}{\sigma}\right]^2 + \frac{\nu}{2}\right)\right] d\lambda. \end{aligned}$$

The expression within this integral is the kernel of an

$$IG\left(\frac{\nu+1}{2}, \left(\frac{1}{2} \left[\left(\frac{y-x\beta}{\sigma}\right)^2 + \nu\right]\right)^{-1}\right)$$

density. Thus,

$$p(y|\beta, \sigma^2) = \left[\sqrt{2\pi}\sigma \right]^{-1} \left[\Gamma\left(\frac{\nu}{2}\right) \left(\frac{2}{\nu}\right)^{\nu/2} \right]^{-1} \\ \times \Gamma\left(\frac{\nu+1}{2}\right) \left(\frac{1}{2} \left[\left(\frac{y-x\beta}{\sigma}\right)^2 + \nu \right] \right)^{-(\nu+1)/2}.$$

Rearranging and canceling terms, we obtain

$$p(y|\beta, \sigma^2) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi}\sigma\Gamma\left(\frac{\nu}{2}\right)} \left[1 + \frac{1}{\nu} \left(\frac{y-x\beta}{\sigma}\right)^2 \right]^{-(\nu+1)/2}, \quad (15.3)$$

which is in the form of a Student t -density [i.e., $y|\beta, \sigma^2 \sim t(x\beta, \sigma, \nu)$].

(b) This result is useful to the applied Bayesian researcher as it, in conjunction with the Gibbs sampler, allows the estimation of models with Student t errors, thus relaxing the normality assumption. To see this, regard the conditional density in (a) as one observation's contribution to the likelihood function and note (adding i subscripts to denote the individual observations) that

$$p(\beta, \sigma^2, \{\lambda_i\}|y) \propto \left[\prod_{i=1}^n \phi(y_i; x_i\beta, \lambda_i\sigma^2) p(\lambda_i) \right] p(\beta, \sigma^2), \quad (15.4)$$

which implies that

$$p(\beta, \sigma^2|y) = \int_0^\infty \cdots \int_0^\infty p(\beta, \sigma^2, \{\lambda_i\}|y) d\lambda_1 \cdots d\lambda_n \quad (15.5)$$

$$\propto \left[\prod_{i=1}^n \int_0^\infty \phi(y_i; x_i\beta, \lambda_i\sigma^2) p(\lambda_i) d\lambda_i \right] p(\beta, \sigma^2) \quad (15.6)$$

$$= \prod_{i=1}^n f_t(y_i|x_i\beta, \sigma, \nu) p(\beta, \sigma^2). \quad (15.7)$$

Thus working with the seemingly more complicated joint posterior in (15.4), which contains the inverted-gamma mixing variables λ_i , yields the same inference for β and σ^2 that would be obtained by directly working with a regression model with Student t errors in (15.7). By working with the joint posterior in (15.4), we can work in terms of conditional normal distributions, and thus we can exploit well-known results associated with those distributions. Indeed this feature will provide a computationally more attractive approach to modeling Student t errors, as revealed in the next exercise.

Exercise 15.2 (Student t errors in regression and binary choice models) Consider the regression model

$$y_i = x_i\beta + \epsilon_i, \quad \epsilon_i | \lambda_i, \sigma^2 \stackrel{\text{ind}}{\sim} N(0, \lambda_i \sigma^2), \quad (15.8)$$

$$\lambda_i \stackrel{\text{iid}}{\sim} IG\left(\frac{\nu}{2}, \frac{2}{\nu}\right), \quad i = 1, 2, \dots, n, \quad (15.9)$$

$$\beta \sim N(\mu_\beta, V_\beta), \quad (15.10)$$

$$\sigma^2 \sim IG(a, b). \quad (15.11)$$

- (a) Show how the Gibbs sampler can be used to fit this model.
 (b) Describe how the degrees-of-freedom parameter ν can also be estimated within the model.
 (c) Describe how the addition of the inverted gamma mixing variables λ_i can be used to generalize the probit model discussed in Exercise 14.1.

Solution

(a) Exercise 13.3 describes a closely related solution within the framework of the linear regression model. We review many of those derivations here to facilitate answering parts (b) and (c) of this exercise.

To implement the Gibbs sampler we need to obtain the complete posterior conditionals for the parameters β , σ^2 , and $\{\lambda_i\}$. The joint posterior distribution is given as

$$p(\beta, \{\lambda_i\}, \sigma^2 | y) \propto \left[\prod_{i=1}^n \phi(y_i; x_i\beta, \lambda_i \sigma^2) p(\lambda_i) \right] p(\beta) p(\sigma^2).$$

From this joint posterior, the following complete conditional posterior distributions are obtained:

$$\beta | \{\lambda_i\}, \sigma^2 | y \sim N \left[\left(X' \Lambda^{-1} X / \sigma^2 + V_\beta^{-1} \right)^{-1} \left(X' \Lambda^{-1} y + V_\beta^{-1} \mu_\beta \right), \right. \\ \left. \left(X' \Lambda^{-1} X / \sigma^2 + V_\beta^{-1} \right)^{-1} \right], \quad (15.12)$$

where $\Lambda \equiv \text{diag}\{\lambda_i\}$ and thus $\Lambda^{-1} = \text{diag}\{\lambda_i^{-1}\}$, and X and y are stacked appropriately,

$$\sigma^2 | \beta, \{\lambda_i\}, y \sim IG \left[\frac{n}{2} + a, \left(b^{-1} + \frac{1}{2} (y - X\beta)' \Lambda^{-1} (y - X\beta) \right)^{-1} \right], \quad (15.13)$$

and

$$\lambda_i | \beta, \sigma^2, y \stackrel{\text{ind}}{\sim} IG \left[\frac{\nu + 1}{2}, \left(\frac{1}{2} \left[\frac{y_i - x_i\beta}{\sigma} \right]^2 + \frac{\nu}{2} \right)^{-1} \right], \quad i = 1, 2, \dots, n. \quad (15.14)$$

A Gibbs sampler involves drawing from (15.12), (15.13), and (15.14).

(b) To allow ν to be determined within the model [e.g., Albert and Chib (1993)], one can specify the following hierarchical prior:

$$\lambda_i | \nu \stackrel{\text{iid}}{\sim} IG\left(\frac{\nu}{2}, \frac{2}{\nu}\right),$$

$$\nu \sim \pi,$$

with π denoting a prior over the degrees-of-freedom parameter.

With this formulation, the complete posterior conditionals for the parameters β , σ^2 , and $\{\lambda_i\}$ remain the same. (Of course, noting that ν is now a parameter of our model, one must condition on its most recent values when sampling from these conditionals.) The complete posterior conditional for ν is obtained as

$$\nu | \sigma^2, \beta, \{\lambda_i\} \propto \pi(\nu) \left[\Gamma\left(\frac{\nu}{2}\right) \left(\frac{2}{\nu}\right)^{(\nu/2)} \right]^{-n} \prod_{i=1}^n \lambda_i^{-[(\nu/2)+1]} \exp\left(-\frac{\nu}{2\lambda_i}\right). \quad (15.15)$$

This distribution is not of standard form. One approach to simulation is to discretize this distribution and draw from the discrete approximation [e.g., Albert and Chib (1993)]. Another approach is to use an M–H step (sometimes called a Metropolis-within-Gibbs step) to sample from this conditional.

(c) Consider the probit model as described in Exercise 14.1. To generalize the assumption of normal errors to the Student t errors, we would write

$$z_i = x_i \beta + \epsilon_i, \quad \epsilon_i \stackrel{\text{ind}}{\sim} N(0, \lambda_i), \quad (15.16)$$

where

$$y_i = \begin{cases} 1 & \text{if } z_i > 0, \\ 0 & \text{if } z_i \leq 0, \end{cases} \quad (15.17)$$

and $\{y_i, x_i\}$ constitutes the observed data. As was the case in parts (a) and (b) of this exercise, we specify the priors

$$\lambda_i \stackrel{\text{ind}}{\sim} IG\left(\frac{\nu}{2}, \frac{2}{\nu}\right), \quad i = 1, 2, \dots, n, \quad (15.18)$$

so that marginalized over the mixing variables λ_i , $\epsilon_i \sim t(0, 1, \nu)$. We also specify the prior

$$\beta \sim N(\mu_\beta, V_\beta).$$

With this in hand, we obtain the *augmented* joint posterior distribution

$$\begin{aligned} p(\beta, \lambda, z | y) &\propto p(y, z | \lambda, \beta) p(\lambda) p(\beta) \\ &= \left[\prod_{i=1}^n \phi(z_i; x_i \beta, \lambda_i) [I(y_i = 1)I(z_i > 0) + I(y_i = 0)I(z_i \leq 0)] p(\lambda_i) \right] p(\beta), \end{aligned}$$

where $z = [z_1 \ z_2 \ \cdots \ z_n]'$, $y = [y_1 \ y_2 \ \cdots \ y_n]'$, and $\lambda = [\lambda_1 \ \lambda_2 \ \cdots \ \lambda_n]'$. Using standard arguments, it follows that the complete posterior conditionals are given by

$$\beta | \{\lambda_i\}, \{z_i\}, y \sim N \left[\left(X' \Lambda^{-1} X + V_\beta^{-1} \right)^{-1} \left(X' \Lambda^{-1} z + V_\beta^{-1} \mu_\beta \right), \right. \\ \left. \left(X' \Lambda^{-1} X + V_\beta^{-1} \right)^{-1} \right], \quad (15.19)$$

$$\lambda_i | \beta, \{z_i\}, y \stackrel{\text{ind}}{\sim} IG \left[\frac{\nu + 1}{2}, \left(\frac{1}{2} [z_i - x_i \beta]^2 + \frac{\nu}{2} \right)^{-1} \right], \quad i = 1, 2, \dots, n \quad (15.20)$$

and

$$z_i \stackrel{\text{ind}}{\sim} \begin{cases} TN_{(0, \infty)}(x_i \beta, \lambda_i) & \text{if } y_i = 1, \\ TN_{(-\infty, 0]}(x_i \beta, \lambda_i) & \text{if } y_i = 0. \end{cases} \quad (15.21)$$

A Gibbs sampler involves drawing from (15.19) through (15.21).

Exercise 15.3 (Posterior simulation in a regression model with double exponential errors) Consider the regression model

$$y_i = x_i \beta + \epsilon_i, \quad \epsilon_i | \lambda_i, \sigma^2 \stackrel{\text{ind}}{\sim} N(0, \lambda_i \sigma^2).$$

Suppose that the following hierarchical prior for the scale-mixing variables λ_i is employed:

$$\lambda_i \stackrel{\text{iid}}{\sim} \text{Exp}(2),$$

an exponential density with mean 2 (see Appendix Definition 2 and Theorem 2).

(a) Derive $p(\epsilon_i | \sigma^2)$.

(b) Describe a posterior simulator for fitting this model, using priors of the form $\beta \sim N(\mu_\beta, V_\beta)$ and $\sigma^2 \sim IG(a, b)$.

Solution

(a) We note that

$$\begin{aligned} p(\epsilon_i | \sigma^2) &= \int_0^\infty p(\epsilon_i | \lambda_i, \sigma^2) p(\lambda_i) d\lambda_i \\ &= \int_0^\infty (2\pi \lambda_i \sigma^2)^{-1/2} \exp \left[-\epsilon_i^2 / (2\lambda_i \sigma^2) \right] (1/2) \exp \left(-\lambda_i / 2 \right) d\lambda_i \\ &= (1/2) (2\pi \sigma^2)^{-1/2} \int_0^\infty \lambda_i^{-1/2} \exp \left[- (1/2) (\lambda_i + [\epsilon_i / \sigma]^2 \lambda_i^{-1}) \right] d\lambda_i. \end{aligned}$$

Now, make a change of variable and let $\psi_i = \lambda_i^{1/2}$. We can then express this integral as

$$p(\epsilon_i | \sigma^2) = (2\pi \sigma^2)^{-1/2} \int_0^\infty \exp \left(- [1/2] (\psi_i^2 + [\epsilon_i / \sigma]^2 \psi_i^{-2}) \right) d\psi_i. \quad (15.22)$$

The integral in (15.22) can be evaluated analytically. For example, Andrews and Mallows (1974) observe that

$$\int_0^\infty \exp \left[- (1/2)(a^2 u^2 + b^2 u^{-2}) \right] du = \left(\frac{\pi}{2a^2} \right)^{1/2} \exp(-|ab|).$$

Applying this to Equation (15.22), we obtain

$$\begin{aligned} p(\epsilon_i | \sigma^2) &= (2\pi\sigma^2)^{-1/2} \left(\frac{\pi}{2} \right)^{1/2} \exp \left(-\frac{|\epsilon_i|}{\sigma} \right) \\ &= \frac{1}{2\sigma} \exp \left(-\frac{|\epsilon_i|}{\sigma} \right), \end{aligned}$$

a *double exponential* or *Laplace* distribution with scale parameter σ [see, e.g., Poirier (1995, p. 110)]. So, the addition of the exponential mixing variables to the disturbance variance allows for the possibility of double exponential errors in our regression model.

(b) A posterior simulator via the Gibbs sampler will draw from the complete conditionals for β , σ^2 , and $\lambda = [\lambda_1 \lambda_2 \cdots \lambda_n]'$. Apart from the distribution associated with the mixing variables λ_i , this exercise is identical to Exercise 15.2, and as a consequence, the posterior conditionals for β and σ^2 are the same as those in Equations (15.12) and (15.13), respectively.

The kernel of the λ_i posterior conditional is given as

$$\lambda_i | \beta, \sigma^2, y \propto \lambda_i^{-1/2} \exp \left[-\frac{1}{2} \left(\lambda_i + \left(\frac{y_i - x_i \beta}{\sigma} \right)^2 \lambda_i^{-1} \right) \right]. \quad (15.23)$$

This distribution is of the *generalized inverse Gaussian* (GIG) form. Following Shuster (1968), Michael, Schucany, and Hass (1976), and Carlin and Polson (1991), we outline a strategy for obtaining a draw from this GIG density. To this end we first consider a cousin of the GIG density, the *inverse Gaussian density* [see, e.g., Poirier (1995, p. 100)]. In the discussion that follows, we say that x follows an inverse Gaussian distribution [which we denote as $x \sim \text{invGauss}(\psi, \mu)$] if

$$p(x | \psi, \mu) \propto x^{-3/2} \exp \left(-\frac{\psi(x - \mu)^2}{2x\mu^2} \right), \quad x > 0. \quad (15.24)$$

Now, let $z = x^{-1}$. It follows by a change of variables that

$$\begin{aligned} p(z | \psi, \mu) &\propto z^{-2} z^{3/2} \exp \left(-\frac{\psi(z^{-1} - \mu)^2}{2z^{-1}\mu^2} \right) \\ &\propto z^{-1/2} \exp \left(-\frac{\psi}{2} [z + \mu^{-2} z^{-1}] \right). \end{aligned}$$

In terms of the posterior conditional for λ_i , it follows that the reciprocal of an

$$\text{invGauss}(1, |\sigma/(y_i - x_i \beta)|)$$

random variable has the density in (15.23).

It remains to discuss how to draw from the inverse Gaussian distribution. Shuster (1968) notes that if x has the inverse Gaussian density in (15.24), then $\psi(x - \mu)^2/x\mu^2 \sim \chi^2(1)$, a chi-square distribution with one degree of freedom. Let ν_0 denote a draw from the $\chi^2(1)$ density. The roots of $\nu_0 = \psi(x - \mu)^2/\mu^2 x$ are given as x_1 and x_2 , where

$$x_1 = \mu + \frac{\mu^2 \nu_0}{2\psi} - \frac{\mu}{2\psi} \sqrt{4\mu\psi\nu_0 + \mu^2 \nu_0^2},$$

$$x_2 = \mu^2/x_1.$$

Michael et al. (1976) use this idea to show that one can obtain a draw from the inverse Gaussian(ψ, μ) density by first drawing $\nu_0 \sim \chi^2(1)$, calculating the roots x_1 and x_2 from the preceding equations, and then setting x equal to the smaller root x_1 with probability $\mu/(\mu + x_1)$ and equal to the larger root x_2 with probability $x_1/(\mu + x_1)$. In terms of sampling from (15.23), we first use this strategy to draw from the inverse Gaussian distribution with $\psi = 1$ and $\mu = |\sigma/(y_i - x_i\beta)|$. This draw is then inverted, producing a draw from (15.23). This process is repeated (independently) for $i = 1, 2, \dots, n$ in order to sample each scale-mixing variable λ_i from its posterior conditional.

15.2 Other continuous and finite-mixture models

Exercise 15.4 (The negative binomial as a continuous mixture of the Poisson) We say a (count) random variable $x = 0, 1, 2, \dots$ has a *negative binomial* distribution, denoted $NB(a, p)$ with $a > 0, 0 < p < 1$, if

$$p(x|a, p) = \frac{\Gamma(a+x)}{x!\Gamma(a)} p^a (1-p)^x. \quad (15.25)$$

Now, consider the *Poisson* density for a count variable y_i :

$$p(y_i|\lambda_i) = \frac{\exp(-\lambda_i)\lambda_i^{y_i}}{y_i!}, \quad y_i = 0, 1, 2, \dots, \quad \lambda_i > 0. \quad (15.26)$$

It is common practice in empirical work to incorporate covariates in the analysis by specifying the mean function λ_i as follows:

$$\ln \lambda_i = x_i\beta.$$

Suppose, instead, that you seek to capture a greater degree of heterogeneity across individuals i and thus include a random error in the specification of $\ln \lambda_i$:

$$\begin{aligned} \ln \lambda_i &= x_i\beta + \epsilon_i \\ &= \ln \mu_i + \ln v_i, \end{aligned} \quad (15.27)$$

with $\mu_i = \exp(x_i\beta)$ and $v_i = \exp(\epsilon_i)$. For the heterogeneity distribution, you assume

$$v_i \sim G(\theta, \theta^{-1}), \quad (15.28)$$

with G denoting the gamma density (see Appendix Definition 2).

Under these assumptions, derive $p(y_i|\mu_i, \theta)$.

Solution

First, note that

$$p(y_i|\mu_i, \theta) = \int_0^\infty p(y_i|\mu_i, v_i, \theta)p(v_i|\theta) dv_i.$$

This follows since v_i is assumed independent of μ_i . Substituting the Poisson density in (15.26) [with λ_i defined in (15.27)] and the gamma density for v_i in (15.28), we obtain

$$\begin{aligned} p(y_i|\mu_i, \theta) &= \int_0^\infty \left(\frac{\exp[-\mu_i v_i](\mu_i v_i)^{y_i}}{(y_i!)} \right) \frac{\theta^\theta v_i^{\theta-1} \exp[-\theta v_i]}{\Gamma(\theta)} dv_i \\ &= \frac{\theta^\theta \mu_i^{y_i}}{(y_i!)\Gamma(\theta)} \int_0^\infty \exp[-(\theta + \mu_i)v_i] v_i^{(\theta+y_i-1)} dv_i. \end{aligned}$$

Let us now make a change of variables. Let $\tilde{v}_i = (\theta + \mu_i)v_i$. Then

$$\begin{aligned} \int_0^\infty \exp[-(\theta + \mu_i)v_i] v_i^{(\theta+y_i-1)} dv_i &= \int_0^\infty \left(\exp[-\tilde{v}_i] \tilde{v}_i^{(\theta+y_i-1)} \right. \\ &\quad \left. \times (\theta + \mu_i)^{-(\theta+y_i-1)} (\theta + \mu_i)^{-1} d\tilde{v}_i \right) \\ &= (\theta + \mu_i)^{-(\theta+y_i)} \int_0^\infty \exp[-\tilde{v}_i] \tilde{v}_i^{(\theta+y_i-1)} d\tilde{v}_i \\ &= (\theta + \mu_i)^{-(\theta+y_i)} \Gamma(\theta + y_i). \end{aligned}$$

Therefore,

$$p(y_i|\mu_i, \theta) = \frac{\theta^\theta \mu_i^{y_i} \Gamma(\theta + y_i) (\theta + \mu_i)^{-(\theta+y_i)}}{(y_i!)\Gamma(\theta)}. \quad (15.29)$$

Let $p_i = \theta/[\mu_i + \theta]$. This implies

$$\mu_i^{y_i} = [(1 - p_i)(\mu_i + \theta)]^{y_i} = [(1 - p_i)(\theta/p_i)]^{y_i}.$$

Similarly,

$$(\theta + \mu_i)^{-(\theta+y_i)} = [\theta/p_i]^{-(\theta+y_i)}.$$

Substituting these results back into (15.29) gives

$$\begin{aligned} p(y_i|\mu_i, \theta) &= \frac{\Gamma(\theta + y_i)}{(y_i!)\Gamma(\theta)} (\theta^\theta) [(1 - p_i)(\theta/p_i)]^{y_i} [\theta/p_i]^{-(\theta+y_i)} \\ &= \frac{\Gamma(\theta + y_i)}{(y_i!)\Gamma(\theta)} p_i^\theta (1 - p_i)^{y_i}, \end{aligned}$$

which is indeed an $NB(\theta, p_i)$ density.

Exercise 15.5 (Experiments with two-component normal mixtures) Consider the two-component normal mixture model

$$y_i | P, \mu_1, \mu_2, \sigma_1^2, \sigma_2^2 \stackrel{\text{iid}}{\sim} PN(\mu_1, \sigma_1^2) + (1 - P)N(\mu_2, \sigma_2^2), \quad 0 \leq P \leq 1. \quad (15.30)$$

To generate values y from this model, note that one can first draw from a two-point distribution with probabilities P and $1 - P$. Given a draw from this two-point distribution, one can then draw from the associated *component* of the mixture [either $N(\mu_1, \sigma_1^2)$ or $N(\mu_2, \sigma_2^2)$] to obtain a draw y . Repeating this process a variety of times will produce a set of draws from the desired mixture density.

(a) Using this intuition, *augment* the mixture model with a set of *component indicator variables*, say $\{\tau_i\}_{i=1}^n$, where τ_i is either zero or one, and $\tau_i = 1$ implies that the i th observation is drawn from the first component of the mixture. (When $\tau_i = 0$, the implication is that the i th observation is drawn from the second component.) Assign a hierarchical prior to τ_i so that the probability associated with the first component is P , and then place a beta prior on P . Use this augmented structure to describe how a Gibbs sampler can be employed to fit the mixture model.

(b–d) Now, illustrate the flexibility of the two-component mixture model by first generating data from some nonnormal distributions. Write a program to implement the Gibbs sampler for this two-component model, given your answer to part (a), and compare your estimated results to the actual densities. To fix ideas, generate the following:

(b) 2,000 observations from a lognormal distribution with parameters $\mu = \ln 10$ and $\sigma^2 = .04$. In this case, the lognormal random variable has density function

$$p(x) = \frac{1}{x\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}[\ln x - \mu]^2\right), \quad x > 0.$$

(c) 5,000 observations from a chi-square distribution with 10 degrees of freedom.

(d) 3,000 observations from a two-component mixture model with $P = .4$, $\mu_1 = 0$, $\mu_2 = 2$, $\sigma_1^2 = 1$, and $\sigma_2^2 = .5$. For this last case, compare the parameter estimates with the actual values used to generate the data.

Solution

(a) Before describing the augmented representation, let θ denote all the model's parameters and θ_{-x} denote all parameters other than x . The model can be written as

$$p(y|\theta, \{\tau_i\}) = \prod_{i=1}^n [\phi(y_i; \mu_1, \sigma_1^2)]^{\tau_i} [\phi(y_i; \mu_2, \sigma_2^2)]^{1-\tau_i}, \quad (15.31)$$

$$\tau_i \stackrel{\text{iid}}{\sim} B(1, P), \quad i = 1, 2, \dots, n, \quad (15.32)$$

$$P \sim \beta(\underline{p}_1, \underline{p}_2), \quad (15.33)$$

$$\mu_i \stackrel{\text{ind}}{\sim} N(\underline{\mu}_i, \underline{v}_i), \quad i = 1, 2, \quad (15.34)$$

$$\sigma_i^2 \stackrel{\text{ind}}{\sim} IG(\underline{a}_i, \underline{b}_i), \quad i = 1, 2. \quad (15.35)$$

Here, $B(1, P)$ denotes a binomial density on one trial with “success” probability P , or equivalently, a Bernoulli density with success probability P . Similarly, $\beta(a, b)$ denotes the beta density with parameters a and b ; properties of these densities and their respective parameterizations are described in Definitions 7 and 10 of the Appendix.

Note that when marginalizing the conditional likelihood $p(y|\theta, \{\tau_i\})$ over τ_i , we are left with the two-component mixture model described at the outset of this section. To see this, note that the assumed conditional independence across observations, together with the fact that τ_i is binary, implies

$$\begin{aligned} p(y|\theta) &= \prod_{i=1}^n \sum_{j=0}^1 p(y_i|\theta, \tau_i = j) \Pr(\tau_i = j|\theta) \\ &= \prod_{i=1}^n [P\phi(y_i; \mu_1, \sigma_1^2) + (1 - P)\phi(y_i; \mu_2, \sigma_2^2)]. \end{aligned}$$

Thus, the component indicators serve the practical purpose of facilitating computation, but their presence does not affect the joint posterior distribution of our parameters of interest, θ . Our complete posterior distribution consists of $p(\theta, \{\tau_i\}|y) = p(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, P, \{\tau_i\}|y)$.

The following complete posterior conditionals are obtained:

$$\mu_1 | \theta_{-\mu_1}, \tau, y \sim N(D\mu_1 d\mu_1, D\mu_1) \quad (15.36)$$

where

$$D\mu_1 = (n_1/\sigma_1^2 + \underline{v}_1^{-1})^{-1}, \quad d\mu_1 = \sum_i \tau_i y_i + \underline{v}_1^{-1} \underline{\mu}_1,$$

$n_1 \equiv \sum_i \tau_i$ denotes the number of observations “in” the first component of the mixture, and n_2 will be defined as $n_2 \equiv \sum_i (1 - \tau_i) = n - n_1$. The complete conditional for μ_2 follows similarly. The conditional posterior distribution for the variance parameter σ_2^2 is

$$\sigma_2^2 | \theta_{-\sigma_2^2}, \{\tau\}, y \sim IG \left(\frac{n_2}{2} + \underline{a}_2, \left[\underline{b}_2^{-1} + .5 \sum_{i=1}^n (1 - \tau_i)(y_i - \mu_2)^2 \right]^{-1} \right), \quad (15.37)$$

and the complete conditional for σ_1^2 follows similarly. Finally, for the component indicator variables $\{\tau_i\}$ and component probability P ,

$$\tau_i | \theta, \tau_{-i}, y \stackrel{\text{ind}}{\sim} B \left[1, \frac{P\phi(y_i; \mu_1, \sigma_1^2)}{P\phi(y_i; \mu_1, \sigma_1^2) + (1 - P)\phi(y_i; \mu_2, \sigma_2^2)} \right], \quad (15.38)$$

$$P | \theta_{-P}, \{\tau_i\}, y \sim \beta(n_1 + \underline{p}_1, n_2 + \underline{p}_2). \quad (15.39)$$

With these conditionals in hand, we can implement the Gibbs sampler by drawing from (15.36) through (15.39), noting, of course, that similar conditionals need to be obtained for μ_2 and σ_1^2 .

(b–d) The preceding algorithm is coded in Matlab (and made available on the Web site associated with this book), and the Gibbs sampler is run in cases where the data are generated from lognormal, Chi-square, and mixture densities. Hyperparameters for all cases

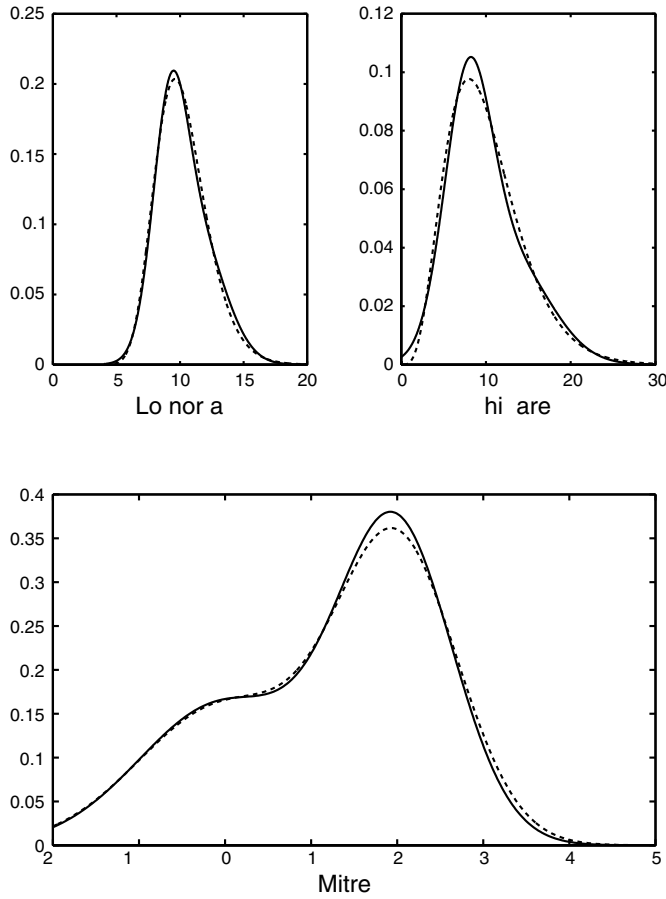


Figure 15.1 — Actual densities and two-component normal mixture estimates.

were chosen so that the priors were noninformative relative to the data. The results of these three experiments are presented in Figure 15.1.

As shown in the figure, the mixture estimates tend to do a good job at recovering the shape of the “true” density for all the examples. In the first two cases, we see that the mixtures can accommodate the right skew associated with the lognormal and chi-square distributions. In the last example – an example of correct specification – the mixture estimate performs quite well. Table 15.1 shows the actual parameter values used, as well as the posterior means obtained from our simulations.

Table 15.1: Results of generated-data experiment.

	P	μ_1	σ_1^2	μ_2	σ_2^2
Actual	.400	.000	1.00	2.00	.500
Post. Mean	.400	-.007	.961	1.98	.445

Exercise 15.6 (Using finite normal mixtures in a regression model) Consider the general setup for a regression model using G Normal mixture components:

$$y_i | x_i, \{\beta_g, \sigma_g^2, \pi_g\}_{g=1}^G \stackrel{\text{ind}}{\sim} \sum_{g=1}^G \pi_g N(x_i \beta_g, \sigma_g^2), \quad \sum_{g=1}^G \pi_g = 1. \quad (15.40)$$

In this model we allow each mixture component to possess its own variance parameter, σ_g , and set of regression parameters, β_g . This level of generality is not required – if desired, we could restrict some of these parameters to be constant across the mixture components.

(a) Consider, for the purposes of computation, augmenting this model with a set of *component label vectors*, $\{z_i\}_{i=1}^n$, where

$$z_i = [z_{1i} \ z_{2i} \ \cdots \ z_{Gi}], \quad (15.41)$$

and $z_{gi} = 1$ implies that the i th individual is “drawn from” the g th component of the mixture.

To complete the augmentation step, add a multinomial prior (see Appendix Definition 9) for the component label vector z_i that depends on a vector of component probabilities π , and then specify a Dirichlet prior (see Appendix Definition 10) for π . Given this setup, write down the augmented likelihood and the above priors.

(b) Add the following prior specifications to complete the model:

$$\begin{aligned} \beta_g &\stackrel{\text{ind}}{\sim} N(\beta_{0g}, V_{\beta_g}), \quad g = 1, 2, \dots, G, \\ \sigma_g^2 &\stackrel{\text{ind}}{\sim} IG(a_g, b_g), \quad g = 1, 2, \dots, G, \end{aligned}$$

and derive the complete posterior conditionals for use with the Gibbs sampler.

Solution

(a) Let θ denote all parameters and component indicator variables in the model, and let θ_{-x} denote all quantities other than x . If we condition on the values of the component indicator variables, the conditional likelihood function can be expressed as

$$L(\theta) = \prod_{i=1}^n [\phi(y_i; x_i \beta_1, \sigma_1^2)]^{z_{1i}} [\phi(y_i; x_i \beta_2, \sigma_2^2)]^{z_{2i}} \cdots [\phi(y_i; x_i \beta_G, \sigma_G^2)]^{z_{Gi}}. \quad (15.42)$$

As stated, we add the following priors for the component indicators and component probabilities:

$$z_i | \pi \stackrel{\text{iid}}{\sim} M(1, \pi) \Rightarrow p(z_i | \pi) = \prod_{g=1}^G \pi_g^{z_{gi}}, \quad \pi = [\pi_1 \ \pi_2 \ \cdots \ \pi_G]', \quad (15.43)$$

$$\pi \sim D(\alpha_1, \alpha_2, \dots, \alpha_G) \Rightarrow p(\pi) \propto \pi_1^{\alpha_1-1} \pi_2^{\alpha_2-1} \cdots \pi_G^{\alpha_G-1}. \quad (15.44)$$

The multinomial prior (M) is a multivariate generalization of the binomial, and the Dirichlet prior (D) is a multivariate generalization of the beta. Note that, analogous to the derivation of the previous exercise, if we take the likelihood function conditional on the component indicators in (15.42) and integrate out the component indicators using (15.43), we obtain an unconditional likelihood that is equivalent to the one implied by (15.40).

(b) The *augmented posterior density* $p(\{\beta_g, \sigma_g^2, \pi_g\}_{g=1}^G, \{z_i\}_{i=1}^n | y)$ is proportional to the product of the augmented likelihood, the multinomial and beta priors, and the given priors for the regression and variance parameters.

The following complete posterior conditionals can be obtained:

$$\beta_g | \theta_{-\beta_g}, y \stackrel{\text{ind}}{\sim} N(D_{\beta_g} d_{\beta_g}, D_{\beta_g}), \quad g = 1, 2, \dots, G, \quad (15.45)$$

where, again, θ_{-x} denotes all parameters in our posterior other than x and

$$D_{\beta_g} = \left[\left(\sum_i z_{gi} x_i x_i' \right) / \sigma_g^2 + V_{\beta_g}^{-1} \right]^{-1}, \quad d_{\beta_g} = \left(\sum_i z_{gi} x_i' y_i \right) / \sigma_g^2 + V_{\beta_g}^{-1} \beta_{0g},$$

$$\sigma_g^2 | \theta_{-\sigma_g^2}, y \stackrel{\text{ind}}{\sim} IG \left(\frac{n_g}{2} + a_g, \left[b_g^{-1} + (1/2) \sum_i z_{gi} (y_i - x_i \beta_g)^2 \right]^{-1} \right) \quad g = 1, 2, \dots, G, \quad (15.46)$$

where $n_g \equiv \sum_{i=1}^N z_{gi}$ denotes the number of observations in the g^{th} component of the mixture.

$$z_i | \theta_{-z_i}, y \stackrel{\text{ind}}{\sim} M \left(1, \left[\frac{\pi_1 \phi(y_i; x_i \beta_1, \sigma_1^2)}{\sum_{g=1}^G \pi_g \phi(y_i; x_i \beta_g, \sigma_g^2)} \frac{\pi_2 \phi(y_i; x_i \beta_2, \sigma_2^2)}{\sum_{g=1}^G \pi_g \phi(y_i; x_i \beta_g, \sigma_g^2)} \right. \right. \\ \left. \left. \dots \frac{\pi_G \phi(y_i; x_i \beta_G, \sigma_G^2)}{\sum_{g=1}^G \pi_g \phi(y_i; x_i \beta_g, \sigma_g^2)} \right] \right), \quad (15.47)$$

and

$$\pi | \theta_{-\pi}, y \sim D(n_1 + \alpha_1, n_2 + \alpha_2, \dots, n_G + \alpha_G). \quad (15.48)$$

Algorithms for drawing from the multinomial and Dirichlet distributions are provided on the Web site associated with this book and, for example, in Gelman, Carlin, Stern, and Rubin (1995, pp. 481–483). The Gibbs sampler is implemented by drawing from (15.45) through (15.48).

Exercise 15.7 (A normal mixture with covariates affecting the mixture assignment: response delays) This question uses data from an experiment in psychology that measures the (log) reaction times of 17 different subjects. Each of the 17 subjects had his or her reaction times measured at 30 different trials, and we focus our attention on modeling the average of those 30 response times. The data and application presented here is a simplified version of the one discussed in Gelman et al. (1995, Section 16.4). The data are provided in Table 15.2 (or are available on the Web site associated with this book).

What is interesting to note is that the last 6 subjects presented in the final three rows of Table 15.2 were diagnosed schizophrenics, whereas the first 11 were not schizophrenics. Psychological theory suggests that, for a variety of reasons, schizophrenic individuals are likely to suffer from response delays relative to those who are not schizophrenic. For this reason, we adopt a model that supposes that the data in the table arise from two separate component populations. We also recognize that the observed schizophrenia indicator should

Table 15.2: Average natural log of response times for 17 subjects

5.73	5.89	5.71
5.71	5.57	5.80
5.87	5.58	5.54
5.78	5.72	5.71
6.03	6.27	6.32
6.34	6.14	

play a strong role in allocating individuals to the delayed or nondelayed component of the mixture.

In Figure 15.2 we take a quick pass at the data and smooth these 17 observations non-parametrically. We do indeed see that the data appear bimodal, and given the entries from Table 15.2, schizophrenics appear to be far more likely to be drawn from the delayed component of the mixture.

Write down a normal mixture model for analyzing these data. In this model, determine the additional delay received by the high response-time component, and introduce the schizophrenia indicator variable to help allocate individuals to the mixture components.

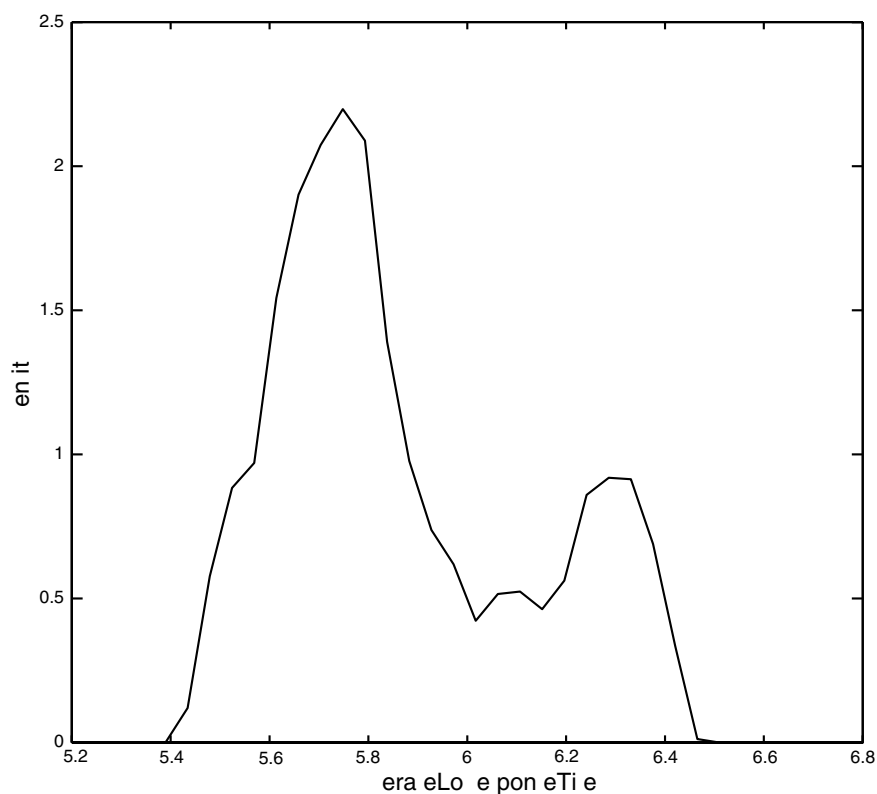


Figure 15.2 — Nonparametric estimate of response-time density.

Solution

Let T_i denote the average response time of individual i , $T = [T_1 \ T_2 \ \cdots \ T_{17}]'$ denote the vector of all average response times, and S_i , $i = 1, 2, \dots, 17$, be the observed schizophrenia indicator, taking the value of 1 if individual i is schizophrenic.

We assume independence across observations and specify our response-time density as follows:

$$p(T_i|S_i, \theta) = [1 - \Phi(\beta_0 + \beta_1 S_i)]\phi(T_i; \mu_1, \sigma_1^2) + \Phi(\beta_0 + \beta_1 S_i)\phi(T_i; \mu_1 + \alpha, \sigma_2^2). \quad (15.49)$$

We wish to restrict $\alpha > 0$ so that the second component of the mixture corresponds to our delayed component. The probability of being in the first component of the mixture is $1 - \Phi(\beta_0 + \beta_1 S_i)$. We expect that schizophrenics are more likely to be in the delayed component, and thus expect $\beta_1 > 0$.

As with the other mixture models, it will prove to be computationally advantageous to augment our model with component indicator variables. Thus, we consider an equivalent model:

$$p(T_i|S_i, \theta, z_i^*) = \phi(T_i; \mu_1, \sigma_1^2)^{I(z_i^* \leq 0)} \phi(T_i; \mu_1 + \alpha, \sigma_2^2)^{I(z_i^* > 0)}, \quad (15.50)$$

$$z_i^* = \beta_0 + \beta_1 S_i + \epsilon_i, \quad \epsilon_i \stackrel{\text{iid}}{\sim} N(0, 1). \quad (15.51)$$

Note, again, that when marginalized over the component indicators, we are back to our original specification of the data density. We add the following priors to our model:

$$\mu_1 \sim N(\underline{\mu}, V_\mu),$$

$$\alpha \sim TN_{(0, \infty)}(\mu_\alpha, V_\alpha),$$

$$\sigma_i^2 \sim IG(a_i, b_i), \quad i = 1, 2,$$

$$[\beta_0 \ \beta_1]' \sim N(\underline{\mu}_\beta, V_\beta),$$

where a truncated normal prior for α has been employed for component identification purposes so that the second component will represent the delayed component. Thus, our joint posterior $p(\mu_1, \alpha, \sigma_1^2, \sigma_2^2, \beta_0, \beta_1, \{z_i^*\}|y)$ is proportional to

$$\left(\prod_{i=1}^N \left[\phi(T_i; \mu_1, \sigma_1^2)^{I(z_i^* \leq 0)} \phi(T_i; \mu_1 + \alpha, \sigma_2^2)^{I(z_i^* > 0)} \right] \phi(z_i^*; \beta_0 + \beta_1 S_i, 1) \right) \\ \times p(\beta_0, \beta_1) p(\mu_1) p(\alpha) p(\sigma_1^2) p(\sigma_2^2),$$

where the priors for the parameters listed in the second row have been specified explicitly above.

Fitting the model

The following complete posterior conditionals are obtained:

$$\mu_1 | \theta_{-\mu_1}, y \sim N(D_{\mu_1} d_{\mu_1}, D_{\mu_1}), \quad (15.52)$$

where

$$D_{\mu_1} = (\iota' \Sigma^{-1} \iota + V_\mu^{-1})^{-1}, \quad d_{\mu_1} = \iota' \Sigma^{-1} (T - z\alpha) + V_\mu^{-1} \underline{\mu},$$

ι is a 17×1 vector of ones, $z_i \equiv I(z_i^* > 0)$, $z = [z_1 z_2 \cdots z_n]$, $\Sigma \equiv \text{Diag}\{(1 - z_i)\sigma_1^2 + z_i\sigma_2^2\}$,
 $\alpha|\theta_{-\alpha}, y \sim TN_{(0,\infty)}(D_\alpha d_\alpha, D_\alpha)$, (15.53)

where

$$D_\alpha = (n_2/\sigma_2^2 + V_\alpha^{-1})^{-1}, \quad d_\alpha = \sum_{i=1}^{17} z_i(T_i - \mu_1)/\sigma_2^2 + V_\alpha^{-1}\mu_\alpha,$$

and $n_2 \equiv \sum_i z_i$ [and, as before, we will also define $n_1 = 17 - n_2 = \sum_i (1 - z_i)$],

$$\beta_0, \beta_1|\theta_{-\beta_0, \beta_1}, y \sim N(D_\beta d_\beta, D_\beta), \quad (15.54)$$

where

$$D_\beta = (X'X + V_\beta^{-1})^{-1}, \quad d_\beta = X'z^* + V_\beta^{-1}\mu_\beta,$$

where $z^* = [z_1^* \ z_2^* \ \cdots \ z_{17}^*]'$, $S = [S_1 \ S_2 \ \cdots \ S_{17}]'$, and $X = [\iota \ S]$,

$$\sigma_1^2|\theta_{-\sigma_1^2}, y \sim IG\left(\frac{n_1}{2} + a_1, \left[b_1^{-1} + .5 \sum_{i=1}^{17} (1 - z_i)(T_i - \mu_1)^2\right]^{-1}\right), \quad (15.55)$$

and

$$\sigma_2^2|\theta_{-\sigma_2^2}, y \sim IG\left(\frac{n_2}{2} + a_2, \left[b_2^{-1} + .5 \sum_{i=1}^{17} z_i(T_i - \mu_1 - \alpha)^2\right]^{-1}\right). \quad (15.56)$$

For the complete posterior conditional for z^* , note that each z_i^* can be drawn independently and that

$$p(z_i^*|\theta_{-z_i^*}, y) \propto \left[\phi(T_i; \mu_1, \sigma_1^2)^{I(z_i^* \leq 0)} \phi(T_i; \mu_1 + \alpha, \sigma_2^2)^{I(z_i^* > 0)} \right] \phi(z_i^*; \beta_0 + \beta_1 S_i, 1). \quad (15.57)$$

Though seemingly nonstandard, one can obtain draws from this distribution by first drawing a binary U from a two-point distribution $U \in \{0, 1\}$, where

$$\Pr(U = 1) = \frac{[1 - \Phi(\beta_0 + \beta_1 S_i)]\phi(T_i; \mu_1, \sigma_1^2)}{[1 - \Phi(\beta_0 + \beta_1 S_i)]\phi(T_i; \mu_1, \sigma_1^2) + \Phi(\beta_0 + \beta_1 S_i)\phi(T_i; \mu_1 + \alpha, \sigma_2^2)}.$$

We then draw z_i^* from its posterior conditional as follows:

$$z_i^*|U, \theta_{-z_i^*}, y \begin{cases} TN_{(-\infty, 0]}(\beta_0 + \beta_1 S_i, 1) & \text{if } U = 1, \\ TN_{(0, \infty)}(\beta_0 + \beta_1 S_i, 1) & \text{if } U = 0. \end{cases}$$

The conditionals in (15.52) through (15.57) define the Gibbs sampler.

Estimation results

We ran the Gibbs sampler for 50,000 iterations, discarding the first 1,000 simulations as the burn-in period. To obtain the results in Tables 15.3 and 15.4 we employed relatively informative priors and selected the following hyperparameter values: $\underline{\mu} = 5.5$, $\underline{v}_\mu = .2^2$, $\underline{\alpha} = .4$, $\underline{v}_\alpha = .2^2$, $a_1 = a_2 = 2$, $b_1 = b_2 = (.01)^{-1}$, $\underline{\beta} = [0 \ 0]'$, and $\underline{V}_\beta = 4I_2$.

Table 15.3: Posterior means of quantities of interest.

μ_1	σ_1	α	σ_2	Prob. Nondelay		Prob. Delay	
				$S = 1$	$S = 0$	$S = 1$	$S = 0$
5.72	.11	.50	.12	.26	.94	.74	.06

From the results in Table 15.3, we see that the first component is centered around a log response time of 5.7, and the second component is centered at about 6.2, given our estimated value of α . These results are consistent with the descriptive nonparametric density estimate in Figure 15.2.

The last four entries of the table describe probabilities of assignment to the delayed or nondelayed component of the model according to the value of the schizophrenia indicator. For example, the probability of belonging to the nondelayed component of the mixture given that a person is not a schizophrenic is $1 - \Phi(\beta_0)$, and the posterior mean of this quantity was found to be .94. To summarize these results, it is very unlikely that non-schizophrenics will suffer any delays, and schizophrenics are most likely to belong to the delayed component. However, there is a modest probability that schizophrenics will not suffer any additional delays, as the posterior mean of the probability that schizophrenics will come from the nondelayed component of the mixture is .26.

Additionally, we can look at the component assignment mechanism at the individual-level. At each iteration in our sampler we obtain a value of the component indicator variable z_i for every individual. We can thus compute individual-level probabilities of belonging to the delayed component of the mixture. These results are presented in Table 15.4.

With the exception of one of the schizophrenic subjects, the mapping of individuals to components can be completely divided along values of the schizophrenia indicator. That is, all of the 11 nonschizophrenic individuals are allocated to the nondelayed component with very high probability, whereas 5 of the 6 schizophrenic individuals are allocated to the delayed component with very high probability. The schizophrenic individual with an average response time of 5.71 does not apparently suffer any response delays and is assigned to the nondelayed component with high probability.

Exercise 15.8 (Are my students randomly guessing?) Those of us who have had the opportunity to teach a large introductory class in statistics or econometrics have probably held the suspicion that a subset of the class has simply randomly guessed at the answers on a multiple-choice exam. We create a similar scenario in this exercise using artificial data and illustrate how mixture analysis can be used in nonnormal models. Specifically, let us suppose that you administer a multiple choice exam with $Q = 25$ questions with 5 possible answers to each question. All questions are assumed to have an equal level of difficulty. You obtain the (hypothetical) data set giving the test scores of 200 students, as shown in Table 15.5.

Table 15.4:
Probabilities of belonging to delayed component,
observed average natural log of response time (T_i),
and schizophrenia indicator (S_i) for 17 subjects.

$\Pr(z_i = 1 y)$	T_i	S_i
.00	5.73	0
.02	5.90	0
.00	5.71	0
.00	5.71	0
.00	5.57	0
.00	5.80	0
.01	5.87	0
.00	5.58	0
.00	5.54	0
.00	5.78	0
.00	5.72	0
.03	5.71	1
.87	6.03	1
.99	6.27	1
.99	6.32	1
.99	6.34	1
.98	6.14	1

You assume that scores are independent across individuals.

- Consistent with your prior suspicions about the nature of the students in the class, write down a two-component mixture model appropriate for analyzing these test score data.
- Discuss how the Gibbs sampler can be employed to fit this model and then fit it using the data given in Table 15.5. Interpret your results.
- Now, suppose that three students miss the exam, and thus they retake the same exam at a later date. After grading the three exams, you find that the scores are 8, 12, and 22. Based on the estimation results you obtained for the 200 students in the class, what would be your (predictive) probability that each of these three students were randomizing; specifically, what are the probabilities that they belong to the low-scoring component of the mixture?

Solution

- Let y_i denote the number of correct answers on test i , $i = 1, 2, \dots, 200$, and suppose the test scores are drawn from two distinct subpopulations, which we might think of conceptually as randomizers and nonrandomizers. Let p_1 denote the probability of success (correctly answering each question) associated with the first component, p_2 denote the probability of success associated with the second component, and π denote the probability that people are drawn from the first component. To fix ideas, we will regard the first component as the

Table 15.5:
Hypothetical test scores for 200 students.

22	20	21	20	23	21	20	17	21	20
21	19	21	21	20	20	20	17	19	19
19	21	19	22	24	19	6	17	21	24
20	20	21	22	19	20	19	20	19	18
23	21	19	19	21	19	21	17	19	20
23	21	21	8	19	21	21	20	18	16
18	21	17	22	21	22	19	19	19	22
23	23	18	7	21	21	19	18	19	5
21	20	17	20	20	19	20	19	24	18
20	5	5	19	18	19	20	20	22	22
21	19	22	23	21	16	21	20	21	20
23	4	18	16	23	20	21	22	22	21
22	17	21	6	23	20	18	19	20	23
22	3	21	23	19	21	21	19	23	18
19	21	21	19	17	20	21	24	14	17
20	17	19	21	20	22	21	18	22	4
19	16	18	17	24	21	22	3	16	21
19	22	22	20	20	20	20	16	18	19
20	21	22	18	21	19	16	22	18	18
23	18	20	24	19	16	18	5	23	23

component of “randomizers.” Within each component, the distribution of test scores is assumed to follow a binomial distribution, where the number of questions (trials) is denoted by Q (here $Q = 25$), the number of successes on test i is denoted as y_i , and the probability of success in each component is p_g , $g = 1, 2$.

With this notation in hand our model can be written as

$$p(y|p_1, p_2, \{\tau_i\}) = \prod_{i=1}^n \left[\binom{Q}{y_i} p_1^{y_i} (1 - p_1)^{Q-y_i} \right]^{\tau_i} \left[\binom{Q}{y_i} p_2^{y_i} (1 - p_2)^{Q-y_i} \right]^{1-\tau_i}, \quad (15.58)$$

$$\tau_i | \pi \stackrel{\text{ind}}{\sim} B(1, \pi),$$

$$\pi \sim \beta(a_1, b_1),$$

$$p_1 \sim \beta(a_2, b_2),$$

$$p_2 \sim \beta(a_3, b_3),$$

where B and β denote the binomial and beta densities, respectively (see Appendix Definitions 7 and 10).

(b) The complete posterior conditionals for this model are as follows:

$$p_1 | \theta_{-p_1}, y \sim \beta \left[\sum_i \tau_i y_i + a_2, \sum_i \tau_i (Q - y_i) + b_2 \right], \quad (15.59)$$

$$p_2 | \theta_{-p_2}, y \sim \beta \left[\sum_i (1 - \tau_i) y_i + a_3, \sum_i (1 - \tau_i) (Q - y_i) + b_3 \right], \quad (15.60)$$

$$\pi | \theta_{-\pi}, y \sim \beta \left[\sum_i \tau_i + a_1, \sum_i (1 - \tau_i) + b_1 \right], \quad (15.61)$$

$$\tau_i | \theta_{-\tau_i}, y \stackrel{\text{ind}}{\sim} B \left[1, \frac{\pi (p_1^{y_i} (1 - p_1)^{Q - y_i})}{\pi (p_1^{y_i} (1 - p_1)^{Q - y_i}) + (1 - \pi) (p_2^{y_i} (1 - p_2)^{Q - y_i})} \right]. \quad (15.62)$$

Estimation

Using the 200 observations presented in Table 15.5, we implement the Gibbs sampler, iteratively drawing from (15.59) through (15.62). We run the sampler for 1,000 iterations, discarding the first 200 as the burn-in. For the prior hyperparameters, we set $a_i = b_i = 2$, $i = 1, 2, 3$.

In Table 15.6 we present posterior means of parameters in this model. In addition, we also present the actual values of the parameters that were used to generate the 200 observations in Table 15.5. Specifically, we generated these data so that the randomized component had a probability of success equal to $(1/5) = .2$, the nonrandomizing component had a probability of success equal to .8, and 5 percent of the class pursued the randomization strategy.

Table 15.6:
Posterior means and
actual parameter values

	p_1	p_2	π
Posterior Mean	.21	.80	.07
Actual Values	.20	.80	.05

As you can see from Table 15.6, the first component is associated with the group of randomizers, where the estimated probability of success, .21, is quite close to the actual value used. In the second component, the average probability of success for each question is significantly higher, and the estimated value of .80 equals the actual value to the first two decimal places. Based on our estimation results, approximately 7 percent of the class pursues the strategy of randomization, and the data were actually generated so that 5 percent of the sample was randomly guessing. Thus, we see that our estimates tend to approximate the actual values used to generate the data very well. Also note that, as a byproduct of the

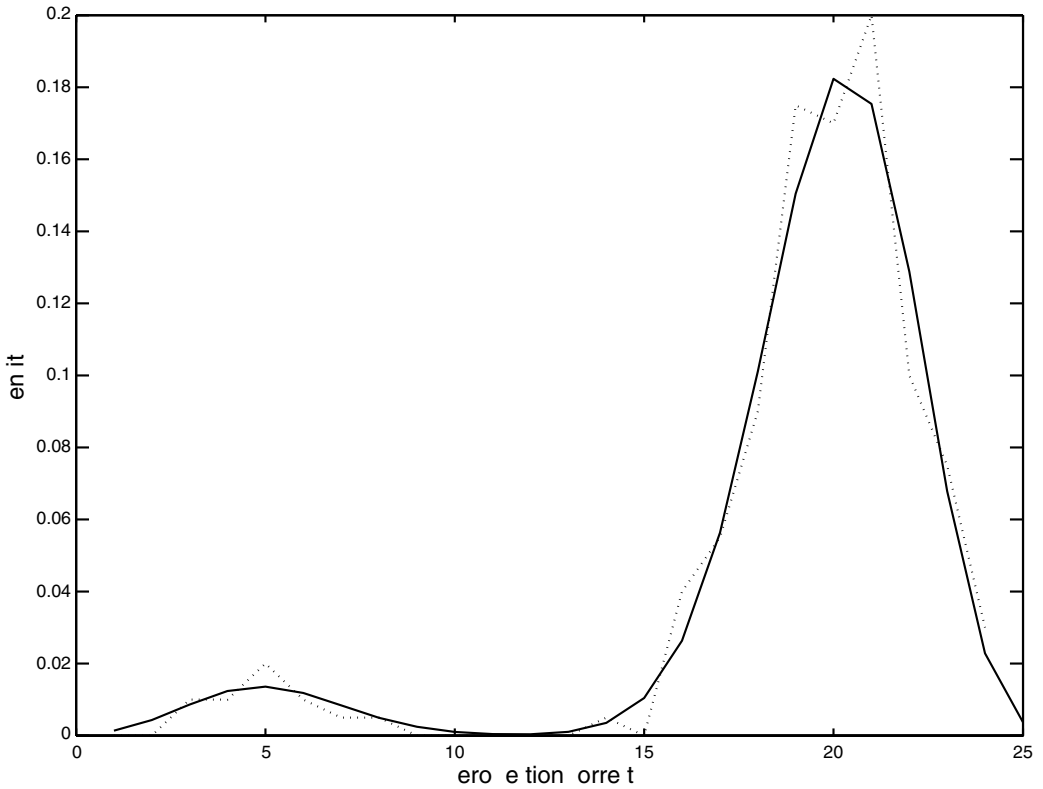


Figure 15.3 — Two-component mixture of binomials estimate (solid) and observed test score probabilities (dashed).

algorithm, individual-level component indicators are produced, and thus it is possible to compute probabilities that each individual is randomizing.

To illustrate the performance of our mixture model in a slightly different way, Figure 15.3 presents the estimated distribution of test scores from our mixture model (Figure 15.3) as well as the probability of each score based on our sample of $n = 200$ outcomes. As seen from the figure, the mixture model fits these data quite well.

(c) To address this question, let τ_f denote the as-yet unobserved component indicator for each of the three “future” students and let y_f denote the future test score outcome. We seek to obtain

$$\Pr(\tau_f = 1|y_f, y), \quad (15.63)$$

the predicted probability of being in the randomizing (or low-scoring) component given the future test score y_f and past data. Note that

$$\Pr(\tau_f = 1|y_f, y) = \int_{\Theta} \Pr(\tau_f = 1|y_f, \theta, y) p(\theta|y_f, y) d\theta, \quad (15.64)$$

where θ denotes all the parameters in the model. Taking each of these terms in the integral of (15.64) separately, we note that

$$\begin{aligned}\Pr(\tau_f = 1|y_f, \theta, y) &= \frac{p(y_f|\tau_f = 1, \theta, y)\Pr(\tau_f = 1, \theta, y)}{p(y_f, \theta, y)} \\ &= \frac{p(y_f|\tau_f = 1, \theta, y)\Pr(\tau_f = 1|\theta, y)}{p(y_f|\theta, y)}.\end{aligned}\quad (15.65)$$

The second term in the integral of (15.64) is

$$p(\theta|y_f, y) = \frac{p(y_f|\theta, y)p(\theta|y)p(y)}{p(y_f, y)} = \frac{p(y_f|\theta, y)p(\theta|y)}{p(y_f|y)}.\quad (15.66)$$

In terms of the components of (15.66), $p(y_f|\theta, y)$ is seen as a mixture of binomials, upon marginalizing over τ_f :

$$p(y_f|\theta, y) = \pi \binom{Q}{y_f} p_1^{y_f} (1 - p_1)^{Q - y_f} + (1 - \pi) \binom{Q}{y_f} p_2^{y_f} (1 - p_2)^{Q - y_f}$$

and

$$\begin{aligned}p(y_f|y) &= \int_{\Theta} p(y_f|\theta, y)p(\theta|y)d\theta \\ &\approx \frac{1}{R} \left[\sum_{i=1}^R \pi(i) \binom{Q}{y_f} p_1(i)^{y_f} [1 - p_1(i)]^{Q - y_f} \right. \\ &\quad \left. + [1 - \pi(i)] \binom{Q}{y_f} p_2(i)^{y_f} [1 - p_2(i)]^{Q - y_f} \right],\end{aligned}\quad (15.67)$$

with $x(i)$ denoting the i th postconvergence draw for the parameter x and R denoting the number of postconvergence Gibbs sampler replications.

Substituting (15.65) and (15.66) into (15.64), canceling terms, and noting that $\Pr(\tau_f = 1|\theta, y) = \pi$, we obtain

$$\Pr(\tau_f = 1|y_f, y) = \frac{1}{p(y_f|y)} \int_{\Theta} \pi p(y_f|\tau_f = 1, \theta, y)p(\theta|y)d\theta.\quad (15.68)$$

Substituting for $p(y_f|y)$ in (15.67) and noting that $y_f|\tau_f = 1, \theta, y$ is just the density of the first mixture component, we obtain

$$\begin{aligned}\Pr(\tau_f = 1|y_f, y) &\approx \frac{\frac{1}{R} \sum_{i=1}^R \pi(i) p_1(i)^{y_f} [1 - p_1(i)]^{Q - y_f}}{\frac{1}{R} \sum_{i=1}^R [\pi(i) p_1(i)^{y_f} [1 - p_1(i)]^{Q - y_f} + [1 - \pi(i)] p_2(i)^{y_f} [1 - p_2(i)]^{Q - y_f}]}.\end{aligned}$$

Using all our postconvergence draws, we apply this estimator and evaluate the probabilities of coming from the first component given that the test scores were 8, 12, or 22. Our point predictions for each of these quantities are 1.00, .37 and .00 for the test scores 8, 12, and

22, respectively. Thus, we see that a person with a score of 8 is very likely to arise from the group of randomizers (or, more specifically, the low-scoring component), whereas a person scoring 22 is almost certainly a nonrandomizer. The person scoring 12 is somewhat ambiguous: Either he or she is a fortunate guesser or an unfortunate nonrandomizer.

Exercise 15.9 (A skewed regression model) Consider the following regression model:

$$y_i = x_i\beta + \delta z_i + \epsilon_i, \quad \epsilon_i \stackrel{\text{iid}}{\sim} N(0, \sigma^2), \quad z_i \stackrel{\text{iid}}{\sim} TN_{(0, \infty)}(0, 1), \quad i = 1, 2, \dots, n, \quad (15.69)$$

with $TN_{(a, b)}(\mu, \sigma^2)$ denoting a normal distribution with mean μ and variance σ^2 truncated to the interval (a, b) . Thus, z has a *half-normal* distribution. We observe (x_i, y_i) , yet z_i is unobserved.

(a) For y , a scalar generated from the model in (15.69), derive the mixture density $p(y|x, \beta, \delta, \sigma^2)$. Comment on the role of δ in this conditional distribution.

(b) Let $\beta^* = [\beta' \delta']'$. Employing priors of the form $\sigma^2 \sim IG(a, b)$ and $\beta^* \sim N(0, V_\beta)$, derive a posterior simulator for fitting this regression model.

(c) Using the wage data available on the Web site associated with this book, fit the regression model in (15.69) and comment on your results. In this application, $x_i = 1 \forall i$; that is, the model only contains an intercept.

Solution

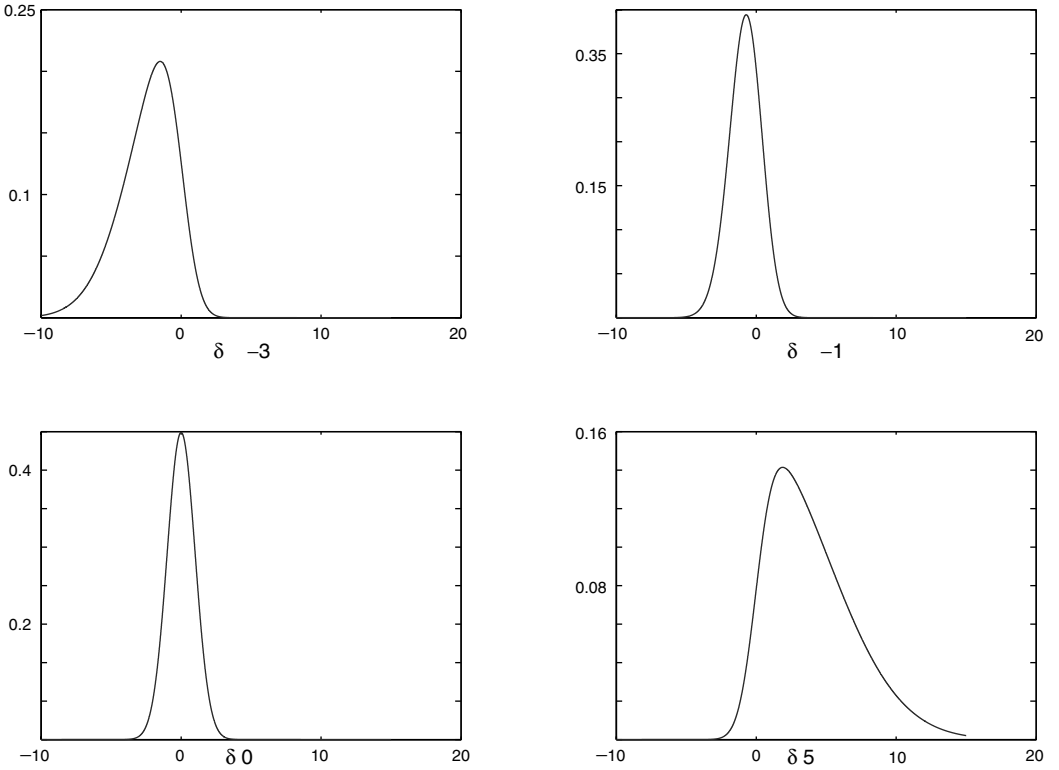
(a) For ease of exposition, let us drop the conditioning on β, σ^2 , and δ in our notation and leave this implicit. That is, simply write $p(y|x, \beta, \delta, \sigma^2)$ as $p(y|x)$. We note that

$$\begin{aligned} p(y|x) &= \int_0^\infty p(y|x, z)p(z) dz \\ &= \int_0^\infty \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(y - x\beta - \delta z)^2\right) \frac{2}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}z^2\right) dz \\ &= \frac{1}{\pi\sqrt{\sigma^2}} \exp\left(-\frac{(y - x\beta)^2}{2(\sigma^2 + \delta^2)}\right) \int_0^\infty \exp\left(-\frac{\sigma^2 + \delta^2}{2\sigma^2} \left[z - \frac{\delta}{\sigma^2 + \delta^2}(y - x\beta)\right]^2\right) dz \\ &= 2\phi(y; x\beta, \sigma^2 + \delta^2) \Phi\left(\frac{\delta(y - x\beta)}{\sigma\sqrt{\sigma^2 + \delta^2}}\right). \end{aligned}$$

This density is known as a *skew-normal* distribution [e.g., Azzalini and Dalla Valle (1996) and Branco and Dey (2002)] and is sometimes written as $y \sim SN(x\beta, \sigma^2 + \delta^2, \delta/\sigma)$.

The parameter δ acts as a skewness parameter; specifically, the density is right-skewed when $\delta > 0$ and left-skewed when $\delta < 0$. When $\delta = 0$, the density is symmetric and we obtain $y \sim N(x\beta, \sigma^2)$. We provide in Figure 15.4 plots of the skew-normal density across different values of δ when $\sigma^2 = 1$ and $x\beta = 0$.

(b) For our posterior simulator, we make use of data augmentation and include $z = [z_1 \ z_2 \ \dots \ z_n]'$ in the posterior distribution. When doing so, sampling from the posterior conditional for β^* and σ^2 is straightforward and will proceed similarly to Exercise 11.8.

Figure 15.4 — Skew-normal densities for various values of δ .

Before describing how one can sample from the posterior conditional for z , we first note that

$$\begin{aligned}
 p(z, \beta^*, \sigma^2 | y) &\propto p(\beta^*) p(\sigma^2) p(y, z | \beta^*, \sigma^2) \\
 &\propto p(\beta^*) p(\sigma^2) \prod_{i=1}^n \phi(y_i; x_i \beta + z_i \delta, \sigma^2) \exp\left(-\frac{1}{2} z_i^2\right) I(z_i > 0).
 \end{aligned}$$

Let θ denote all the quantities in our joint posterior, and let θ_{-x} denote all quantities other than x . It follows that

$$z_i | \theta_{-z_i}, y \propto \exp\left(-\frac{1}{2\sigma^2} (y_i - x_i \beta - z_i \delta)^2\right) \exp\left(-\frac{1}{2} z_i^2\right) I(z_i > 0).$$

Completing the square on z_i , and noting that z_i is truncated at zero, we obtain

$$z_i | \theta_{-z_i}, y \stackrel{\text{ind}}{\sim} TN_{(0, \infty)}\left(\frac{\delta(y_i - x_i \beta)}{\sigma^2 + \delta^2}, \frac{\sigma^2}{\sigma^2 + \delta^2}\right), \quad i = 1, 2, \dots, n. \quad (15.70)$$

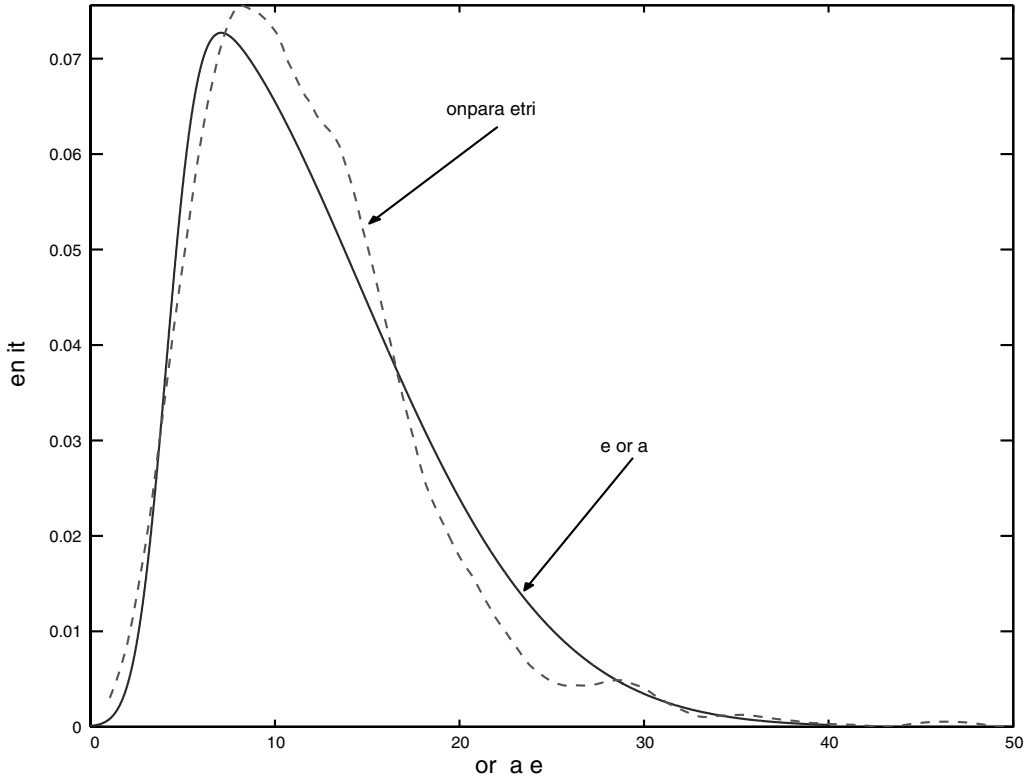


Figure 15.5 — Nonparametric and skew-normal wage densities.

Let

$$X = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}$$

and $W = [X \ z]$. With this notation, the posterior conditional for β^* is of the form

$$\beta^* | \theta_{-\beta^*}, y \sim N(D_\beta d_\beta, D_\beta), \quad (15.71)$$

where

$$D_\beta = (W'W/\sigma^2 + V_\beta^{-1})^{-1}, \quad d_\beta = W'y/\sigma^2.$$

Finally,

$$\sigma^2 | \theta_{-\sigma^2}, y \sim IG \left(\frac{n}{2} + a, \left[b^{-1} + \left(\frac{1}{2} \right) (y - W\beta^*)'(y - W\beta^*) \right]^{-1} \right). \quad (15.72)$$

The posterior simulator proceeds by sampling independently from (15.70), for $i = 1, 2, \dots, n$, and then from (15.71) and (15.72).

(c) For our prior hyperparameters, we set $a = 3$, $b = 1/2$, and $V_\beta = 10I_2$. The Gibbs sampler is run for 5,000 iterations, and the first 2,000 are discarded as the burn-in. Based on the final 3,000 simulations, we calculate posterior means of β , δ , and σ^2 equal to 4.12, 10.3, and 1.80, respectively. These results suggest a strong right skew associated with the wage density, which is well documented in the empirical literature.

To investigate the skew-normal density's ability to capture key features of these wage data, we first calculate a nonparametric density estimate and then compare it to the density in (a) evaluated at posterior mean values of β , σ^2 , and δ (which, admittedly, is a crude comparison). We present these results in Figure 15.5. As you can see, the densities are rather similar and both have clearly accounted for the right skew of the wage data.

Bayesian model averaging and selection

In Chapter 6 on hypothesis testing, exercises relating to model selection were provided. The approaches introduced there, such as marginal likelihoods, Bayes factors, and posterior odds ratios, can be used to select a model. In the later exercises of this chapter, we review some existing computational strategies for model selection via the calculation of marginal likelihoods.

As an alternative to model *selection*, many Bayesians prefer to *average across models* rather than select a single model. The motivation for Bayesian model averaging (BMA), follows immediately from the laws of probability. That is, if we let M_r for $r = 1, \dots, R$ denote the R different models under consideration and ϕ be a vector of parameters that has a common interpretation in all models, then the rules of probability say

$$p(\phi|y) = \sum_{r=1}^R p(\phi|y, M_r) p(M_r|y). \quad (16.1)$$

Alternatively, if $g(\phi)$ is a function of ϕ , the rules of conditional expectation imply that

$$E[g(\phi)|y] = \sum_{r=1}^R E[g(\phi)|y, M_r] p(M_r|y). \quad (16.2)$$

In words, the logic of Bayesian inference says that one should obtain results for every model under consideration and average them, with weights given by the posterior model probabilities, $p(M_r|y)$.

Given that hypothesis testing was covered in Chapter 6 and the quantities in (16.1) and (16.2) can be calculated using familiar methods, the reader may be wondering why we devote a separate chapter to exercises on Bayesian model selection and averaging. One reason is that marginal likelihoods in some models can be difficult to calculate and we provide, in Exercises 16.6 through 16.8, some exercises on general methods of marginal likelihood calculation. Another reason is that, in many empirical examples, the number of

models under consideration is simply too large for evaluation of, e.g., a marginal likelihood for every model. For instance, in the regression model (which we will focus on in most of this chapter), one has K potential explanatory variables. If models are defined by the inclusion/exclusion of every potential explanatory variable then $R = 2^K$. If K is moderate or large (e.g., $K > 20$), then it is computationally difficult or impossible to evaluate every model. Accordingly, a large Bayesian literature has emerged for developing computational methods (closely related to posterior simulation methods) for doing BMA and Bayesian model selection in the regression model when the number of potential explanatory variables is large. For further reading on BMA see Raftery, Madigan, and Hoeting (1997), Hoeting, Madigan, Raftery, and Volinsky (1999), and the Bayesian Model Averaging Web site (<http://www.research.att.com/~volinsky/bma.html>). George and McCulloch (1993, 1997) are important references for Bayesian model selection.

Exercises 16.1 through 16.3 in this chapter will use the following basic setup. All the potential explanatory variables are stacked in an $N \times K$ matrix X but the first column of this matrix is not an intercept. Our set of models is defined by

$$y = \alpha \iota_N + X_r \beta_r + \varepsilon, \quad (16.3)$$

where ι_N is an $N \times 1$ vector of ones and X_r is an $N \times k_r$ matrix containing some (or all) columns of X . The N vector of errors, ε , is assumed to be $N(0_N, h^{-1}I_N)$. Since there are 2^K possible choices for X_r , this is the total number of models under consideration.

16.1 Bayesian model averaging

Exercise 16.1 (Basic g-prior results commonly used with BMA) When doing BMA, it is common to use reference, or *benchmark*, priors that yield analytical results for quantities such as $p(M_r|y)$ and $E(\beta_r|y)$. Such priors ease the job of prior elicitation for huge numbers of models and lessen the computational burden. A common benchmark prior [see, e.g., Fernandez, Ley, and Steel (2001a,b)] sets

$$p(h) \propto \frac{1}{h}, \quad (16.4)$$

and for the intercept

$$p(\alpha) \propto 1. \quad (16.5)$$

To make absolutely certain that the noninformative prior for the intercept has the same implications for every model, Fernandez et al. (2001b) recommend standardizing all the explanatory variables by subtracting off their means. This will have no effect on the slope coefficients β_r , but it ensures that the intercept can be interpreted in the same manner in every model as measuring the mean of y .

For the regression coefficients the prior is of the form

$$\beta_r|h \sim N(\underline{\beta}_r, h^{-1}\underline{V}_r), \quad (16.6)$$

with

$$\underline{\beta}_r = 0_{k_r}$$

and

$$\underline{V}_r = [g_r X_r' X_r]^{-1}. \quad (16.7)$$

With this choice of \underline{V}_r , the prior in (16.6) is termed a *g-prior*, as introduced in Zellner (1986a).

Derive $p(\beta_r|y, M_r)$ and $p(M_r|y)$.

Solution

The *g-prior* is simply a special case of the natural conjugate prior. Thus, the derivations of Exercise 10.1 of Chapter 10 apply directly. Using these results and properties of the normal-gamma distribution, it follows immediately that the posterior for β_r follows a multivariate *t* distribution with mean

$$E(\beta_r|y, M_r) \equiv \bar{\beta}_r = \bar{V}_r X_r' y, \quad (16.8)$$

covariance matrix

$$\text{Var}(\beta_r|y, M_r) = \frac{\bar{\nu} \bar{s}_r^2}{\bar{\nu} - 2} \bar{V}_r, \quad (16.9)$$

and $\bar{\nu} = N$ degrees of freedom. Plugging in (16.7) into the general form for \bar{V}_r , we obtain

$$\bar{V}_r = [(1 + g_r) X_r' X_r]^{-1} \quad (16.10)$$

and

$$\bar{s}_r^2 = \frac{\frac{1}{g_r+1} y' P_{X_r} y + \frac{g_r}{g_r+1} (y - \bar{y} \iota_N)' (y - \bar{y} \iota_N)}{\bar{\nu}}, \quad (16.11)$$

where

$$P_{X_r} = I_N - X_r (X_r' X_r)^{-1} X_r'.$$

Plugging the *g-prior* form for \bar{V}_r into the formula for the marginal likelihood for the normal linear regression model with natural conjugate prior yields

$$p(y|M_r) \propto \left(\frac{g_r}{g_r + 1} \right)^{\frac{k_r}{2}} \left[\frac{1}{g_r + 1} y' P_{X_r} y + \frac{g_r}{g_r + 1} (y - \bar{y} \iota_T)' (y - \bar{y} \iota_T) \right]^{-\frac{N-1}{2}}. \quad (16.12)$$

The posterior model probabilities can be calculated using

$$p(M_r|y) = c p(y|M_r) p(M_r), \quad (16.13)$$

where c is a constant, which is the same for all models. This constant will cancel out in all relevant formulas in the following and can be ignored. Alternatively, the fact that $\sum_{r=1}^R p(M_r|y) = 1$ can be used to determine c .

Exercise 16.2 (Markov chain Monte Carlo model composition [MC³]) The exercise asks you to derive the MC³ algorithm of Madigan and York (1995). The model space used in this chapter, defined by M_r for $r = 1, \dots, R$, can be expressed in terms of a $K \times 1$ vector

$\gamma = (\gamma_1, \dots, \gamma_K)'$, where all elements are either 0 or 1. Models defined by $\gamma_j = 1$ indicate that the j th explanatory variable enters the model (or else $\gamma_j = 0$). There are 2^K possible configurations of γ and the space of this parameter is equivalent to the model space. Construct an M–H algorithm for γ and discuss how this can be used to carry out BMA using the setup of Exercise 16.1.

Solution

In the case of continuous random variables, the random walk chain M–H algorithm employs a candidate density in which a candidate draw equals the current draw plus an increment. We can modify this intuition involving the discrete random vector γ . That is, a candidate value, γ^* , is drawn with equal probability from the set containing the current draw, $\gamma^{(s-1)}$, and all configurations that change *one* element of $\gamma^{(s-1)}$. In terms of model space, this means a candidate model, M^* , is proposed that is drawn randomly (with equal probability) from the set of models including (i) the current model, $M^{(s-1)}$, (ii) all models that delete one explanatory variable from $M^{(s-1)}$, and (iii) all models that add one explanatory variable to $M^{(s-1)}$. If candidate models are generated in this way, we can use the standard acceptance probability formula for the random walk chain M–H algorithm [see Chapter 11, Exercise 11.18, or Koop (2003), p. 97]:

$$\alpha \left(M^{(s-1)}, M^* \right) = \min \left[\frac{p(y|\gamma^*)p(\gamma^*)}{p(y|\gamma^{(s-1)})p(\gamma^{(s-1)})}, 1 \right]. \quad (16.14)$$

The quantities $p(y|\gamma^{(s-1)})$ and $p(y|\gamma^*)$ can be calculated using (16.12). In the common case where equal prior weight is allocated to each model, $p(\gamma^*) = p(\gamma^{(s-1)})$, and these terms cancel out of (16.14).

Posterior results based on the sequence of models generated from the MC³ algorithm can be calculated by averaging over draws in the standard MCMC manner. So, for instance, (16.2) can be approximated by \hat{g}_{S_1} , where

$$\hat{g}_{S_1} = \frac{1}{S_1} \sum_{s=S_0+1}^S E \left[g(\phi) | y, \gamma^{(s)} \right]. \quad (16.15)$$

As with other MCMC algorithms, \hat{g}_{S_1} converges to $E[g(\phi)|y]$ as S_1 goes to infinity [where $S_1 = S - (S_0 + 1)$]. Note that a starting value for the chain, $\gamma^{(0)}$, must be chosen and hence S_0 burn-in replications should be discarded to eliminate the effects of this choice. Note also that the frequencies with which models are drawn can be used to calculate Bayes factors. For instance, if the MC³ algorithm draws the model M_r A times and the model M_s B times, then the ratio A/B will converge to the Bayes factor comparing M_r to M_s .

Exercise 16.3 (BMA: an application in economic growth) Using the solutions to Exercises 16.1 and 16.2, carry out a BMA exercise with a data set from the economic growth literature and the g-prior value $g = 1/N$. This choice of g is recommended by Fernandez et al. (2001b). The data set used in this exercise is taken from Fernandez et al. (2001a) and is

available from the *Journal of Applied Econometrics* data archive (www.econ.queensu.ca/jae). [We also make this data set available on the Web site associated with this book (labeled GROWTH.DAT).] This data set covers $N = 72$ countries and contains $K = 41$ potential explanatory variables. The dependent variable is average per capita GDP growth for the period 1960–1992. For the sake of brevity, we will not list all of the explanatory variables here (see the original paper for a detailed description of all the data). Furthermore, Table 16.1 in the solution provides a list of short-form names for all explanatory variables, which should be enough to provide a rough idea of what each explanatory variable is measuring.

Solution

The computer code implementing the MC^3 algorithm described in Exercise 16.2 is available on the Web site associated with this book. The following results are based on taking 1, 100, 000 draws and discarding the first 100, 000 as burn-in replications (i.e., $S_0 = 100, 000$ and $S_1 = 1, 000, 000$).

Elements in the column of Table 16.1 labeled “BMA Post. Prob.” can be interpreted as the probability that the corresponding explanatory variable should be included. It is calculated as the proportion of models drawn by the MC^3 algorithm that contain the corresponding explanatory variable. Informally, this is a useful diagnostic for deciding whether an individual explanatory variable has an important role in explaining economic growth. It can be seen that several variables (i.e., Life expectancy, GDP level in 1960, Equipment Investment, and Fraction Confucian) do have an important role in explaining economic growth. Regardless of which other explanatory variables are included, these variables almost always exhibit strong explanatory power. However, for the remainder of the explanatory variables there is some uncertainty as to whether they have important roles to play in explaining economic growth. And, for many of the explanatory variables, there is strong evidence that they should not be included.

The other two columns of Table 16.1 contain posterior means and standard deviations for each regression coefficient, averaged across models. Remember that models where a particular explanatory variable is excluded are interpreted as implying a zero value for its coefficient. Hence, the average in (16.15) involves some terms where $E[g(\phi) | y, M^{(s)}]$ is calculated and others where the value of zero is used. With the exception of the few variables with high BMA posterior probability, most of the posterior means are small relative to their standard deviations. Thus, BMA is indicating a high degree of uncertainty about which factors explain economic growth. Undoubtedly, this is an honest representation of the fact that the data are not informative enough to send a clear story.

The MC^3 algorithm allows for the calculation of posterior model probabilities by simply counting the proportion of draws taken from each model. For the top ten models, the column of Table 16.2 labeled “ $p(M_r | y)$ MC^3 Estimate” contains posterior model probabilities calculated in this way. The column labeled “ $p(M_r | y)$ Analytical” contains the exact values calculated using (16.12) and (16.13). It can be seen that the posterior model probability is widely scattered across models with no single model dominating. In fact, the top ten models account for only a little more than 4% of the total posterior model probability.

Table 16.1: Bayesian model averaging results.

Explanatory Variable	BMA Post. Prob.	Posterior Mean	Posterior St. Dev.
Primary School Enrollment	0.207	0.004	0.010
Life expectancy	0.935	0.001	3.4×10^{-4}
GDP level in 1960	0.998	-0.016	0.003
Fraction GDP in Mining	0.460	0.019	0.023
Degree of Capitalism	0.452	0.001	0.001
No. Years Open Economy	0.515	0.007	0.008
% of Pop. Speaking English	0.068	-4.3×10^{-4}	0.002
% of Pop. Speaking Foreign Lang.	0.067	2.9×10^{-4}	0.001
Exchange Rate Distortions	0.081	-4.0×10^{-6}	1.7×10^{-5}
Equipment Investment	0.927	0.161	0.068
Nonequipment Investment	0.427	0.024	0.032
St. Dev. of Black Market Premium	0.049	-6.3×10^{-7}	3.9×10^{-6}
Outward Orientation	0.039	-7.1×10^{-5}	5.9×10^{-4}
Black Market Premium	0.181	-0.001	0.003
Area	0.031	-5.0×10^{-9}	1.1×10^{-7}
Latin America	0.207	-0.002	0.004
Sub-Saharan Africa	0.736	-0.011	0.008
Higher Education Enrollment	0.043	-0.001	0.010
Public Education Share	0.032	0.001	0.025
Revolutions and Coups	0.030	-3.7×10^{-6}	0.001
War	0.076	-2.8×10^{-4}	0.001
Political Rights	0.094	-1.5×10^{-4}	0.001
Civil Liberties	0.127	-2.9×10^{-4}	0.001
Latitude	0.041	9.1×10^{-7}	3.1×10^{-5}
Age	0.083	-3.9×10^{-6}	1.6×10^{-5}
British Colony	0.037	-6.6×10^{-5}	0.001
Fraction Buddhist	0.201	0.003	0.006
Fraction Catholic	0.126	-2.9×10^{-4}	0.003
Fraction Confucian	0.989	0.056	0.014
Ethnolinguistic Fractionalization	0.056	3.2×10^{-4}	0.002
French Colony	0.050	2.0×10^{-4}	0.001
Fraction Hindu	0.120	-0.003	0.011
Fraction Jewish	0.035	-2.3×10^{-4}	0.003
Fraction Muslim	0.651	0.009	0.008
Primary Exports	0.098	-9.6×10^{-4}	0.004
Fraction Protestant	0.451	-0.006	0.007
Rule of Law	0.489	0.007	0.008
Spanish Colony	0.057	2.2×10^{-4}	1.5×10^{-3}
Population Growth	0.036	0.005	0.046
Ratio Workers to Population	0.046	-3.0×10^{-4}	0.002
Size of Labor Force	0.072	6.7×10^{-9}	3.7×10^{-8}

Table 16.2: Posterior model probabilities for top ten models.

	$p(M_r y)$ Analytical	$p(M_r y)$ MC ³ Estimate
1	0.0089	0.0088
2	0.0078	0.0080
3	0.0052	0.0052
4	0.0035	0.0035
5	0.0032	0.0035
6	0.0029	0.0029
7	0.0028	0.0028
8	0.0028	0.0025
9	0.0028	0.0025
10	0.0024	0.0023

The numbers in tables such as this (perhaps with many more models) allow for an assessment of the convergence of the MC³ algorithm. The numbers in the last three rows of Table 16.2 are slightly different from one another. These differences are small enough for present purposes and we can be confident that the numbers in Table 16.1 are approximately correct. However, for a research paper a higher degree of accuracy is typically called for and the number of replications should be increased. Here, we have used 1,000,000 replications, which sounds like a large number. However, the number of possible models is 2^{41} and, with 1,000,000 replications, we have visited only a tiny fraction of them.

16.2 Bayesian variable selection and marginal likelihood calculation

Exercise 16.4 (Stochastic search variable selection) Another popular method that addresses issues similar to Bayesian model averaging, described in George and McCulloch (1993), is known as *stochastic search variable selection* (SSVS). This approach involves using the normal linear regression model with an independent normal-gamma prior (see Exercise 11.8) with one minor alteration. This alteration is in the prior for the regression coefficients and is useful for the case where a large number of explanatory variables exist, but the researcher does not know which ones are likely to be important. To capture this, the prior for each regression coefficient is specified as a mixture of two normals, both with mean zero. One of the terms in the mixture has very small variance (i.e., it says the coefficient is virtually zero and thus the variable can be effectively excluded from the model) and the other has a large variance (i.e., it is most likely different from zero and thus the variable should be retained in the model). To be precise, for each coefficient β_j for $j = 1, \dots, K$, the prior is

$$\beta_j | \gamma_j \sim (1 - \gamma_j) N(0, \tau_j^2) + \gamma_j N(0, c_j^2 \tau_j^2), \quad (16.16)$$

where c_j and τ_j are known prior hyperparameters with τ_j^2 being small and $c_j^2\tau_j^2$ large. In addition, γ_j are binary variables that equal 1 with prior probability $\Pr(\gamma_j = 1) = p_j$, $0 \leq p_j \leq 1$. If $\gamma_j = 0$, the coefficient β_j “belongs to” the first component of the mixture, which is tightly centered around zero. We interpret this as evidence that variable x_j (the j th column of the stacked regressor matrix) can be excluded from the model. Conversely, if $\gamma_j = 1$, we interpret this as evidence that x_j should be retained in the model.

(a) Suppose you place a point prior on $p_j: p_j = p \ \forall j$, for some value of p , and thus the prior probability that each variable is included in the model is simply p [i.e., $\Pr(\gamma_j = 1) = p \ \forall j$]. Derive a Gibbs sampler for fitting a linear regression model with the SSVS hierarchical prior described here.

(b) Using the results and prior specification from part (a), perform a generated-data experiment to illustrate the performance of the SSVS approach. Specifically, generate

$$\begin{bmatrix} x_{1i} \\ x_{2i} \\ \vdots \\ x_{5i} \end{bmatrix} \stackrel{\text{iid}}{\sim} N \left[\begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & .4 & .4 & 0 & .6 \\ .4 & 1 & .7 & 0 & .3 \\ .4 & .7 & 1 & 0 & .3 \\ 0 & 0 & 0 & 1 & 0 \\ .6 & .3 & .3 & 0 & 1 \end{pmatrix} \right], \quad i = 1, 2, \dots, 1,000. \quad (16.17)$$

Then, set

$$y_i = 2 + .25x_{1i} - .4x_{2i} + .6x_{3i} + .2\epsilon_i, \quad (16.18)$$

where

$$\epsilon_i \stackrel{\text{iid}}{\sim} N(0, 1).$$

In this design, note that we are generating five potential explanatory variables (some of which are correlated with one another), and y is generated using only the first three of these. Apply the SSVS algorithm using all five explanatory variables, and determine the posterior probabilities that each variable should be retained in the model. For the intercept, simply employ a normal prior, and fix attention on the issue of excluding or retaining the x 's from the model.

Solution

(a) First, let us write the linear regression model as

$$y_i = \beta_0 + \sum_{j=1}^K x_{ji}\beta_j + \epsilon_i, \quad \epsilon_i \stackrel{\text{iid}}{\sim} N(0, \sigma^2). \quad (16.19)$$

Place an inverted gamma prior on σ^2 of the form

$$\sigma^2 \sim IG(a, b),$$

use an independent normal prior for the intercept β_0 of the form

$$\beta_0 \sim N(0, \underline{V}_0),$$

and specify conditionally independent priors for the slope coefficients using (16.16). The augmented joint posterior (which contains the component indicators γ) is $p(\beta, \sigma^2, \gamma|y)$, where $\beta = [\beta_0 \ \beta_1 \ \cdots \ \beta_K]'$ and $\gamma = [\gamma_1 \ \gamma_2 \ \cdots \ \gamma_K]'$. We obtain the following posterior conditionals for this model:

$$\beta|\gamma, \sigma^2, y \sim N(D_\beta d_\beta, D_\beta), \quad (16.20)$$

where

$$D_\beta = \left(X'X/\sigma^2 + V_\beta^{-1}(\gamma) \right)^{-1}, \quad d_\beta = X'y/\sigma^2.$$

In these definitions of D_β and d_β , X and y have been stacked over i and j in the obvious way and $V_\beta(\gamma)$ is a $(K+1) \times (K+1)$ diagonal matrix with \underline{V}_0 as the (1,1) element and $\gamma_i c_i^2 \tau_i^2 + (1 - \gamma_i) \tau_i^2$ as the (i, i) element, $i = 2, 3, \dots, K+1$. We also have

$$\sigma^2|\beta, \gamma, y \sim IG\left(\frac{N}{2} + a, [b^{-1} + .5(y - X\beta)'(y - X\beta)]^{-1}\right), \quad (16.21)$$

and finally,

$$\alpha_j|\alpha_{-j}, \beta, \sigma^2, y \stackrel{\text{ind}}{\sim} B\left(1, \frac{p\phi(\beta_j; 0, c_j^2 \tau_j^2)}{p\phi(\beta_j; 0, c_j^2 \tau_j^2) + (1-p)\phi(\beta_j; 0, \tau_j^2)}\right), \quad j = 1, 2, \dots, K, \quad (16.22)$$

with B denoting the Binomial distribution (see Appendix Definition 7).

(b) We generate 1,000 observations from the stated model and choose prior hyperparameters of the form $\tau_j^2 = .0000001$, $c_j^2 = 9/\tau_j^2$, and $p = 1/2$. With these hyperparameters, our prior probability that each variable should be included in the model is 1/2, the prior standard deviation for the “slab” component (for which the variable is to be retained in the model) is 3, and the prior standard deviation for the “spike” component (for which the variable is to be excluded from the model) is .0003. The sampler is run for 10,000 iterations, drawing from (16.20) through (16.22), and the first 200 simulations are discarded as the burn-in.

Posterior means for the β parameters were found to be 2.01, .258, $-.395$, .595, $-.0001$, and .002, which are close to the values actually used to generate the data. The values of γ_1 , γ_2 , and γ_3 at each iteration of the sampler were equal to unity, providing clear evidence that these variables belong in the model. This, of course, is consistent with the true data-generating process. Finally, the posterior means of γ_4 and γ_5 were .003 and .008, respectively, indicating rather strongly that these variables can be excluded from the model.

***Exercise 16.5 (Nonparametric regression via variable selection: Smith and Kohn [1996])** Consider a nonparametric regression model of the form

$$y_i = f(x_i) + \epsilon_i, \quad \epsilon_i \stackrel{\text{iid}}{\sim} N(0, \sigma^2), \quad (16.23)$$

$i = 1, 2, \dots, N$. Suppose that $f(x)$ can be written as a cubic regression spline [e.g., Poirier (1973)],

$$f(x) = \delta_0 + \delta_1 x + \delta_2 x^2 + \sum_{j=3}^J \delta_j (x - \tau_{j-2})_+^3 = X\delta, \quad (16.24)$$

where $z_+ \equiv \max\{0, z\}$ and $\{\tau_j\}_{j=1}^{J-2}$ denote a collection of known potential *knot* points placed in the interior of x with $\tau_j < \tau_{j+1}$. [Note that in this exercise, we do not consider the case where the knot points themselves are unknown (either in number or location), but instead, we suppose that f can be approximated reasonably accurately by a large number of potential knot points placed throughout the support of x . See Chib (1998) and Koop and Potter (2005) for related discussion and Denison, Holmes, Mallick, and Smith (2002) for a more general treatment of the knot points. Alternate approaches to nonparametric regression in various models can be found in, e.g., Koop (2003), Koop and Poirier (2004), Chib and Jeliazkov (2006), and Koop and Tobias (2006).]

Smith and Kohn (1996) consider the problem of estimation of f as a problem of variable selection and proceed in a manner similar to the SSVS approach of George and McCulloch (1993) described in Exercise 16.4. To this end, let $\theta = [\theta_1 \theta_2 \cdots \theta_{J+1}]$ be an *indicator vector* denoting the columns of X that are to be retained in the model (i.e., $\theta_k = 1$ denotes that the k th column is to be kept and $\theta_k = 0$ denotes that the k th column is to be excluded), and define X_θ as the collection of those columns of X for which the corresponding elements of θ equal one. Let δ_θ denote the set of slope parameters associated with X_θ .

Specify a g-prior for the slope parameters of the form

$$\delta_\theta | g, \theta, \sigma^2 \sim N[0, g^{-1} \sigma^2 (X'_\theta X_\theta)^{-1}], \quad (16.25)$$

and also specify the following priors:

$$p(\sigma^2) \propto 1/\sigma^2, \quad (16.26)$$

$$\theta | \pi = \prod_{j=1}^{J+1} \pi_j^{\theta_j} (1 - \pi_j)^{1-\theta_j}. \quad (16.27)$$

For simplicity throughout the remainder of this exercise, assume that $\pi_j = 1/2$ for all j , so that all models are equally likely *a priori*.

- Describe a posterior simulator for fitting this nonparametric regression model.
- Generate 1,500 observations from the following regression model:

$$y_i = 2 + .1x_i - .15x_i^2 + \sin(x_i) + \epsilon_i, \quad (16.28)$$

where x_i are independently generated from a Uniform distribution on $[-2, 2]$ and $\epsilon_i \stackrel{\text{iid}}{\sim} N(0, .15)$.

Apply the algorithm derived in part (a) to fit the model. In our solution to this portion of the exercise, we set $g = 1/N = (1,500)^{-1}$ and $J+1 = 12$ (so that there are 12 “optional” variables in total and 9 knot points placed throughout the support of x). We also chose to spread these knots evenly over the interval $[-1.75, 1.75]$. Compare your estimated results to the function used to generate the data.

Solution

- To fit this model we use a *direct Monte Carlo integration* algorithm, where we sample from the conditionals $\theta | y, \sigma^2 | \theta, y$ and finally $\delta_\theta | \theta, \sigma^2, y$. These conditional distributions are enough to specify the joint posterior distribution.

Given θ , the regression model takes the form

$$y = X_\theta \delta_\theta + \epsilon.$$

First, note that $y|\theta, \sigma^2$ can be obtained by marginalizing the conditional density $y|\theta, \delta_\theta, \sigma^2$ implied here over the conditional prior for δ_θ . Doing this we obtain

$$y|\theta, \sigma^2 \sim N \left[0, \sigma^2 \left(I_N + g^{-1} X_\theta (X'_\theta X_\theta)^{-1} X'_\theta \right) \right]. \quad (16.29)$$

Using properties of the inverse of a matrix sum of the form $(A + BCD)^{-1}$ [e.g., Poirier (1995, p. 627)] and an analogous property for a determinant of a similar form, one can show that

$$\begin{aligned} (I_N + g^{-1} X_\theta (X'_\theta X_\theta)^{-1} X'_\theta)^{-1} &= I_N - g^{-1} X_\theta \left[g^{-1} X'_\theta X_\theta + X'_\theta X_\theta \right]^{-1} X'_\theta \\ &= I_N - [1/(1+g)] X_\theta (X'_\theta X_\theta)^{-1} X'_\theta \end{aligned}$$

and

$$\begin{aligned} |I_N + g^{-1} X_\theta (X'_\theta X_\theta)^{-1} X'_\theta| &= |I_N| |(X'_\theta X_\theta)^{-1}| |X'_\theta X_\theta + g^{-1} X'_\theta X_\theta| \\ &= (1 + g^{-1})^{q_\theta}, \end{aligned}$$

where q_θ denotes the number of columns in X_θ . Putting these together, we can write (16.29) as

$$y|\theta, \sigma^2 \propto (1 + g^{-1})^{-q_\theta/2} \exp \left(-[1/(2\sigma^2)] y' [I_N - [1/(1+g)] X_\theta (X'_\theta X_\theta)^{-1} X'_\theta] y \right). \quad (16.30)$$

To obtain the density $y|\theta$, we must marginalize the conditional density in (16.30) over the prior for σ^2 in (16.26):

$$\begin{aligned} p(y|\theta) &= \int_0^\infty p(y|\theta, \sigma^2) (1/\sigma^2) d\sigma^2 \\ &= (1 + g^{-1})^{-q_\theta/2} \\ &\quad \times \int_0^\infty (\sigma^2)^{-(N/2+1)} \exp \left(-[1/(2\sigma^2)] y' [I_N - [1/(1+g)] X_\theta (X'_\theta X_\theta)^{-1} X'_\theta] y \right) d\sigma^2. \end{aligned}$$

We recognize this integral as the kernel of an

$$IG[N/2, 2(y'[I_N - [1/(1+g)] X_\theta (X'_\theta X_\theta)^{-1} X'_\theta] y)^{-1}]$$

density. Using the known form of the normalizing constant of the inverted gamma, and noting that $\pi_j = 1/2$ implies $p(\theta|y) \propto p(y|\theta)$, we obtain

$$p(\theta|y) \propto (1 + g^{-1})^{-q_\theta/2} [y' y - [1/(1+g)] y' X_\theta (X'_\theta X_\theta)^{-1} X'_\theta y]^{-N/2}. \quad (16.31)$$

Though this density is not of a standard form, θ is discrete valued. Thus we can calculate the (unnormalized) ordinates for every possible value of θ and then normalize the density to make it proper. This defines the first conditional needed for our posterior simulator.

For the second conditional, we have

$$\begin{aligned} p(\sigma^2|\theta, y) &\propto p(y|\theta, \sigma^2) p(\sigma^2) \\ &= (\sigma^2)^{-(N/2+1)} \exp \left(-[1/(2\sigma^2)] y' [I_N - [1/(1+g)] X_\theta (X'_\theta X_\theta)^{-1} X'_\theta] y \right), \end{aligned}$$

where the last line follows from our preceding derivation of $y|\sigma^2, \theta$. We recognize this as the following inverted gamma density:

$$\sigma^2|\theta, y \sim IG\left(\frac{N}{2}, 2[y'(I_N - [1/(1+g)]X_\theta(X'_\theta X_\theta)^{-1}X'_\theta)y]^{-1}\right). \quad (16.32)$$

Finally, the complete conditional for δ_θ is as follows:

$$\delta_\theta|\theta, \sigma^2, y \sim N\left([1/(1+g)](X'_\theta X_\theta)^{-1}X'_\theta y, \sigma^2[1/(1+g)](X'_\theta X_\theta)^{-1}\right). \quad (16.33)$$

The posterior simulator is implemented by sampling from $\theta|y$, $\sigma^2|\theta, y$ and $\delta_\theta|\sigma^2, \theta, y$ in (16.31), (16.32) and (16.33), respectively, conditioning on the most recent values of the parameters drawn from the sampler.

(b) We generate 1,500 observations as described in part (b) of this exercise. We run the algorithm described in (a) for 100 iterations, noting that no burn-in period is required. For each iteration, we calculate our model's prediction of $f(x)$ over a discrete grid of points throughout the $[-2, 2]$ support. Our final point estimate of f is pieced together as the posterior means of the 100 collected values at each grid point.

In Figure 16.1, we present this posterior mean together with the true value of f that was used to generate the data. As one can see, the algorithm performs quite well at recovering the shape of the regression function.

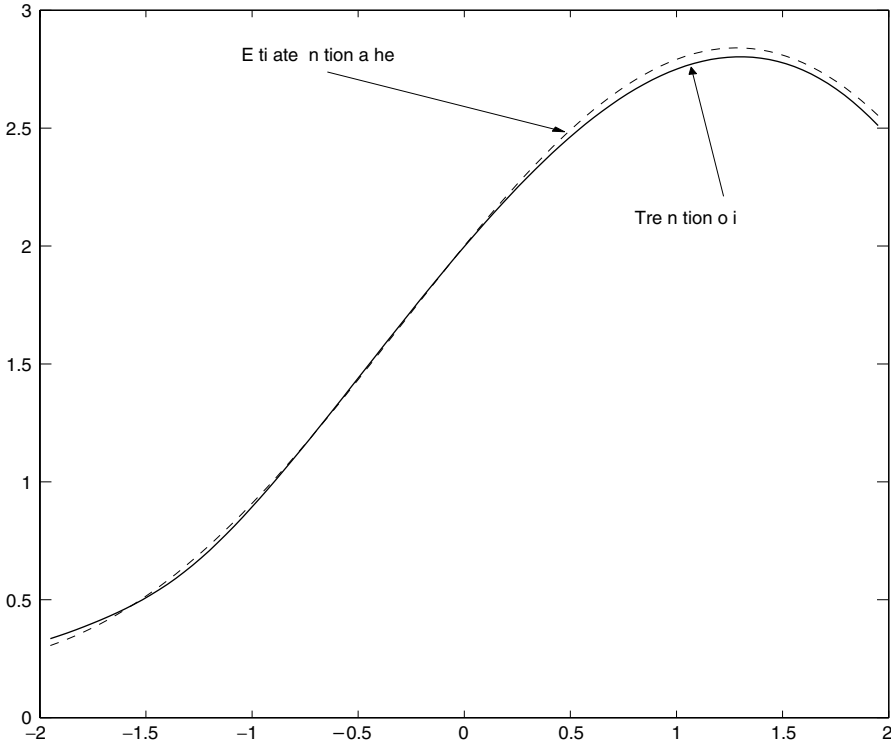


Figure 16.1 — “True” and estimated regression functions.

Exercise 16.6 (Calculating the marginal likelihood with simulated output: Gelfand–Dey [1994])

(a) Show that, for any proper density function $f(\theta)$,

$$E_{\theta|y} \left[\frac{f(\theta)}{p(\theta)p(y|\theta)} \right] = p(y)^{-1}.$$

(b) How is this result useful when trying to calculate the marginal likelihood for a particular model?

Solution

(a) Note that

$$\begin{aligned} E_{\theta|y} \left[\frac{f(\theta)}{p(\theta)p(y|\theta)} \right] &= \int_{\Theta} \frac{f(\theta)}{p(\theta)p(y|\theta)} p(\theta|y) d\theta \\ &= \int_{\Theta} \frac{f(\theta)}{p(\theta)p(y|\theta)} \frac{p(\theta)p(y|\theta)}{p(y)} d\theta \\ &= p(y)^{-1} \int_{\Theta} f(\theta) d\theta \\ &= p(y)^{-1}. \end{aligned}$$

(b) This is a very useful result as it shows how one can use *posterior* simulations from a sampler to calculate the marginal likelihood $p(y)$. The “direct” way to calculate the marginal likelihood $p(y) = \int_{\Theta} p(y|\theta)p(\theta) d\theta$ by Monte Carlo integration would require draws from the *prior* $p(\theta)$. Since the prior is typically flat relative to the likelihood, the resulting estimate of the marginal likelihood is quite inaccurate and unstable under this approach. The Gelfand–Dey algorithm instead makes use of posterior simulations to calculate the marginal likelihood; one simply calculates $f(\theta)/[p(\theta)p(y|\theta)]$ for each θ value simulated from the posterior and averages this collection of values to estimate the (inverse) marginal likelihood.

Geweke (1999) points out that for the asymptotics of the Gelfand–Dey method to go through, the expression $f(\theta)/[p(\theta)p(y|\theta)]$ must be bounded from above. (One might worry that, in the tails, the denominator will be quite small, potentially leading this expression to explode.) Geweke’s suggestion is to alleviate this problem by choosing $f(\theta)$ to be a normal approximation to the posterior $p(\theta|y)$ with a tail truncation. See Exercise 16.8, Geweke (1999), and Koop (2003, pp. 105–106) for more on this issue.

Exercise 16.7 (Calculating the marginal likelihood with simulated output: Chib [1995]) Chib (1995) notes that, for any model M , data y , and parameter vector θ ,

$$p(y, \theta|M) = p(y|\theta, M)p(\theta|M) = p(\theta|y, M)p(y|M),$$

and thus

$$p(y|M) = \frac{p(y|\theta, M)p(\theta|M)}{p(\theta|y, M)}.$$

Taking logs of both sides, we obtain

$$\ln p(y|M) = \ln p(y|\theta, M) + \ln p(\theta|M) - \ln p(\theta|y, M).$$

This expression provides an identity in θ : The log marginal likelihood on the left hand side does not involve θ , and so the equation must hold for any value of θ . Thus, for purposes of computational reliability, let us choose a $\hat{\theta}$ of high posterior density (e.g., a posterior mean or mode) and note that

$$\ln p(y|M) = \ln p(y|\hat{\theta}, M) + \ln p(\hat{\theta}|M) - \ln p(\hat{\theta}|y, M).$$

Some quick accounting suggests that for most problems, the first two terms on the right hand side of the equality are readily available. That is, $\ln p(y|\hat{\theta}, M)$ is the log likelihood evaluated at $\hat{\theta}$ and $\ln p(\hat{\theta}|M)$ is the log prior ordinate at $\hat{\theta}$. In general, these two quantities can be computed. The last term, $\ln p(\hat{\theta}|y, M)$, requires more care, since the normalizing constant of the posterior is typically unknown.

(a) Suppose that the parameter vector θ can be broken into two blocks (i.e., $\theta = [\theta'_1 \theta'_2]'$) and that a Gibbs sampler can be applied to fit this model, cycling through the two posterior conditionals $\theta_1|\theta_2, y, M$ and $\theta_2|\theta_1, y, M$. Show how the Gibbs output can be used to estimate $p(\hat{\theta}|y, M)$, the posterior ordinate needed to calculate the log marginal likelihood. You may assume that the normalizing constants of the complete conditional distributions are known.

(b) Repeat the exercise in (a) when the Gibbs sampler cycles through three blocks ($\theta = [\theta'_1 \theta'_2 \theta'_3]'$).

Solution

(a) First, write

$$p(\hat{\theta}|y, M) = p(\hat{\theta}_1, \hat{\theta}_2|y, M) = p(\hat{\theta}_1|\hat{\theta}_2, y, M)p(\hat{\theta}_2|y, M). \quad (16.34)$$

Note that the first term following the last equality, $p(\hat{\theta}_1|\hat{\theta}_2, y, M)$, can be evaluated as it is an ordinate of the complete conditional $\theta_1|\theta_2, y, M$, which is assumed to be known. For the term $p(\hat{\theta}_2|y, M)$, note that

$$p(\theta_2|y, M) = \int_{\Theta_1} p(\theta_2|\theta_1, y, M)p(\theta_1|y, M) d\theta_1.$$

This representation suggests that one can estimate the desired ordinate via “Rao–Blackwellization.” That is,

$$p(\hat{\theta}_2|y, M) \approx \frac{1}{R} \sum_{r=1}^R p(\hat{\theta}_2|\theta_1^{(r)}, y, M), \quad (16.35)$$

and $\theta_1^{(r)} \sim \theta_1|y, M$, which are provided from the postconvergence output from the Gibbs sampler. Again, note that the conditional distribution $\theta_2|\theta_1, y, M$ (including its normalizing constant) is assumed known. Equation (16.35) plus the conditional ordinate $p(\hat{\theta}_1|\hat{\theta}_2, y, M)$ provide what are needed to calculate (16.34) and thus the log marginal likelihood.

(b) Much like (a), to calculate the marginal likelihood, we need to obtain

$$p(\hat{\theta}_1, \hat{\theta}_2, \hat{\theta}_3 | y, M) = p(\hat{\theta}_1 | \hat{\theta}_2, \hat{\theta}_3, y, M) p(\hat{\theta}_2 | \hat{\theta}_3, y, M) p(\hat{\theta}_3 | y, M). \quad (16.36)$$

The first and last of these three pieces can be calculated in a manner similar to that described in the solution to part (a). That is, $p(\hat{\theta}_1 | \hat{\theta}_2, \hat{\theta}_3, y, M)$ is known, as it is simply an ordinate of the complete conditional for θ_1 . Similarly,

$$p(\hat{\theta}_3 | y, M) \approx \frac{1}{R} \sum_{r=1}^R p(\hat{\theta}_3 | \theta_1^{(r)}, \theta_2^{(r)}, y, M)$$

where $(\theta_1^{(r)}, \theta_2^{(r)})$ denote the r th set of (θ_1, θ_2) values simulated from the posterior.

The middle term of (16.36) introduces a new complication. However, one can estimate this ordinate using some additional *reduced runs* of our posterior simulator. That is, suppose that we ran the Gibbs sampler again, and specifically, suppose we iterate through the posterior conditionals $\theta_2 | \hat{\theta}_3, \theta_1, y, M$ and $\theta_1 | \hat{\theta}_3, \theta_2, y, M$. Then, after a suitable pre-convergence period, we would expect these draws to converge to $\theta_1, \theta_2 | \hat{\theta}_3, y, M$ (and thus, individually, the θ_1 and θ_2 draws would converge to their corresponding marginal distributions). Since

$$p(\theta_2 | \theta_3, y, M) = \int_{\Theta_1} p(\theta_2 | \theta_1, \theta_3, y, M) p(\theta_1 | \theta_3, y, M) d\theta_1,$$

it follows that

$$p(\hat{\theta}_2 | \hat{\theta}_3, y, M) \approx \frac{1}{R} \sum_{r=1}^R p(\hat{\theta}_2 | \theta_1^{(r)}, \hat{\theta}_3, y, M),$$

where

$$\theta_1^{(r)} \sim \theta_1 | \hat{\theta}_3, y, M,$$

and these draws are obtained from the reduced runs by cycling through the conditionals $\theta_2 | \hat{\theta}_3, \theta_1, y, M$ and $\theta_1 | \hat{\theta}_3, \theta_2, y, M$. With this ordinate in hand, we have the three pieces needed to calculate the posterior ordinate in (16.36) and thus the log marginal likelihood. A similar procedure can be applied to calculate the marginal likelihood in cases of more than three blocks.

Exercise 16.8 (Calculating marginal likelihoods with generated data) Generate 100 observations from the simple regression model

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \quad \epsilon_i \stackrel{\text{iid}}{\sim} N(0, \sigma^2),$$

with $\beta_0 = 1$, $\beta_1 = .5$, $\sigma^2 = .2$, and $x_i \stackrel{\text{iid}}{\sim} N(0, 1)$. Employ a normal-gamma prior of the form

$$\beta | \sigma^2 \sim N(0, 4\sigma^2 I_2),$$

$$\sigma^2 \sim IG(3, 2.5).$$

Analytically determine the log marginal likelihood (see Exercise 10.4 for a related discussion) and calculate it given your generated data. Next, use the Gibbs sampler to fit the linear regression model (see, e.g., Exercise 11.8). Given the Gibbs draws, calculate the log marginal likelihood using the methods of Gelfand and Dey (1994) (see Exercise 16.6) and Chib (1995) (see Exercise 16.7) and compare these results to those you obtained analytically.

Solution

Matlab code for this exercise is provided on the Web site associated with this book. The analytical expression of the marginal likelihood was covered in Exercise 10.4, and we do not repeat those derivations here. For the Gelfand–Dey (1994) and Chib (1995) approaches, we fit the model using the Gibbs sampler, running it for 20,000 iterations and discarding the first 100 as the burn-in. In the Gelfand–Dey approach (Exercise 16.6), we choose $f(\theta)$ as the following truncated normal density:

$$f(\theta) \propto (2\pi)^{-k/2} |\hat{\Omega}|^{-1/2} \exp \left[-.5(\theta - \hat{\theta})' \hat{\Omega}^{-1} (\theta - \hat{\theta}) \right] I(\theta \in \Theta),$$

where $\theta = [\beta_0 \ \beta_1 \ \sigma^2]'$, $\hat{\theta}$ is the posterior mean of the Gibbs draws, $\hat{\Omega} = \frac{1}{R} \sum_{r=1}^R (\theta^{(r)} - \hat{\theta})(\theta^{(r)} - \hat{\theta})'$ is an approximation of the posterior covariance matrix, and $k = 3$. The support of the density, Θ , is constructed as

$$\Theta = \{\theta : (\theta - \hat{\theta})' \hat{\Omega}^{-1} (\theta - \hat{\theta}) \leq 11.34\},$$

where 11.34 is the 99th percentile of the $\chi^2(3)$ distribution. With Θ defined in this way, the normalizing constant of f is (1/.99).

Given the generated data, our analytical calculation of the log marginal likelihood was -74.0951 , whereas the methods of Gelfand–Dey and Chib produced numerical estimates of -74.0952 and -74.1118 , respectively.

Some stationary time series models

The field of time series econometrics is an enormous one and it is difficult to do it justice in only a chapter or two. There is a large range of models used with time series data. For univariate time series we have models in the autoregressive moving average (ARMA) class, state space models, models of conditional volatility in the autoregressive conditional heteroscedastic (ARCH) class, and many more. For multivariate time series data, vector autoregressive (VAR) models are the most popular, but many others exist. For both univariate and multivariate specifications there is a growing interest in nonlinear alternatives involving thresholds, regime switching, or structural breaks. In addition, with time series data, new issues arise such as those relating to the presence of nonstationary data (i.e., unit roots and cointegration). In frequentist econometrics, the presence of unit roots and cointegration causes problems since the asymptotic theory underlying, for example, estimators and test statistics, becomes quite complicated. Bayesian econometrics, which conditions upon the observed data, does not run into these problems. However, subtle issues relating to prior elicitation and identification arise when unit root and cointegration issues are relevant. In this chapter and the next, we cannot hope to cover this wide range of models and issues; but rather, we offer a selection of questions to give the reader a flavor for the Bayesian approach to time series methods. In the present chapter, we focus on stationary time series and, thus, do not worry about the complications caused by the presence of nonstationary data. In the following chapter, we consider nonstationary time series models. The reader interested in learning more about Bayesian time series methods is referred to Bauwens, Lubrano, and Richard (1999).

Before beginning the exercises, we remind the reader that several of the exercises in previous chapters involve time series data. For instance, Exercise 13.4 shows how to carry out Bayesian inference in the linear regression model with $AR(p)$ errors. This is, of course, a time series model and the methods derived in that exercise can be used to handle the

univariate AR(p) model or allow for AR dynamics in the errors of any of the regression-type models considered in this book. Exercises 11.11 and 11.12 discuss how to model structural change in time series models.

Exercise 17.1 (Using the AR(p) model to understand the properties of a series) This exercise is loosely based on Geweke (1988). Let y_t for $t = 1, \dots, T$ indicate observations on a time series variable. Assume that y_t follows an AR(p) process:

$$y_t = \beta_0 + \beta_1 y_{t-1} + \dots + \beta_p y_{t-p} + \epsilon_t, \quad (17.1)$$

where $\epsilon_t \stackrel{\text{iid}}{\sim} N(0, h^{-1})$. Many important properties of y_t depend on the roots of the polynomial $1 - \sum_{i=1}^p \beta_i z^i$, which we will denote by r_i for $i = 1, \dots, p$. Geweke (1988) lets y_t be the log of real GDP and sets $p = 3$ and, for this choice, focuses on the features of interest: $C = \{\beta : \text{two of } r_i \text{ are complex}\}$ and $D = \{\beta : \min |r_i| < 1\}$, where $\beta = (\beta_0, \beta_1, \beta_2, \beta_3)'$. Note that C and D are regions whose bounds are complicated non-linear functions of β_1, β_2 , and β_3 . If the AR coefficients lie in the region defined by C then real GDP exhibits an oscillatory response to a shock and if they lie in D then y_t exhibits an explosive response to a shock.

(a) Assuming a prior of the form

$$p(\beta_0, \dots, \beta_p, h) \propto \frac{1}{h}, \quad (17.2)$$

derive the posterior for β . To simplify things, you may ignore the (minor) complications relating to the treatment of initial conditions. Thus, assume the dependent variable is $y = (y_{p+1}, \dots, y_T)'$ and treat y_1, \dots, y_p as fixed initial conditions.

(b) Using an appropriate data set (e.g., the US real GDP data set provided on the Web site associated with this book), write a program that calculates the posterior means and standard deviations of β and $\min |r_i|$.

(c) Extend the program of part (b) to calculate the probability that y_t is oscillatory [i.e., $\Pr(\beta \in C|y)$] and the probability that y_t is explosive [i.e., $\Pr(\beta \in D|y)$] and calculate these probabilities using your data set.

Solution

(a) This model can be written as a linear regression model in matrix notation as

$$y = X\beta + \epsilon, \quad (17.3)$$

where $\epsilon = (\epsilon_{p+1}, \dots, \epsilon_T)'$ and X is the $(T - p) \times (p + 1)$ matrix containing an intercept and p lags of the dependent variable. For instance, if $p = 3$

$$X = \begin{bmatrix} 1 & y_3 & y_2 & y_1 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & y_{T-1} & y_{T-2} & y_{T-3} \end{bmatrix}.$$

Written in this way, if we condition on the initial observations, the AR(p) model is seen to be simply the normal linear regression model. The easiest way to answer this exercise

is to note that the prior in (17.2) is the standard noninformative prior for the normal linear regression model and is a limiting case of the natural conjugate prior. Thus, the answer to this exercise is exactly the same as the answer to Exercise 10.1 with particular prior hyperparameter values used. In particular, you can set $\underline{\nu} = 0$ and $\underline{Q}^{-1} = cI$. The noninformative prior in (17.2) is obtained by letting the scalar c go to zero. Alternatively, multiplying the normal linear regression model likelihood in (10.4) by the prior in (17.2) yields the relevant posterior. Manipulations exactly analogous to the solution to Exercise 10.1 can be done to confirm that the posterior is normal-gamma. To be precise, the posterior for β and h is $NG(\bar{\beta}, \bar{Q}, \bar{s}^{-2}, \bar{\nu})$, where $\bar{\nu} = T$,

$$\begin{aligned}\bar{Q} &= (X'X)^{-1}, \\ \bar{\beta} &= (X'X)^{-1} X'y,\end{aligned}$$

and $\bar{\nu}\bar{s}^2$ are the standard OLS residuals (denoted SSE in the introductory material of Chapter 10).

Using the properties of the normal-gamma distribution (see Theorem 8 of the Appendix), it follows that the marginal posterior for β is multivariate t :

$$\beta|y \sim t(\bar{\beta}, \bar{s}^2\bar{Q}, \bar{\nu}). \quad (17.4)$$

(b) From part (a), the posterior for β has a convenient form and means and standard deviations can be calculated using analytical formulas based on (17.4). That is, using the properties of the t -distribution, we have $E(\beta|y) = \bar{\beta}$ and $\text{Var}(\beta|y) = \frac{\bar{\nu}\bar{s}^2}{\bar{\nu}-2}\bar{Q}$. However, $\min|r_i|$ is a complicated nonlinear function of β and no analytical results exist for its posterior mean and variance. However, Monte Carlo integration (see Exercise 11.3) can be done by randomly drawing from (17.4), using a subroutine to calculate the solutions to $1 - \sum_{i=1}^p \beta_i z^i = 0$ and, from these, evaluating $\min|r_i|$ at each of the draws. This will produce a random sample from the posterior of $\min|r_i|$ that can be used to calculate its posterior mean and standard deviation (or any other function of its posterior). A program that does this is available on the Web site associated with this book.

Using the log of the quarterly US real GDP series provided on the Web site, which runs from 1947Q1 through 2005Q2, we obtain the results in Table 17.1. The Monte Carlo integration program was run for 10,000 replications.

Table 17.1: Posterior results for the AR(3) model.

Parameter	Mean	Std. Dev.
β_0	0.010	0.004
β_1	1.300	0.066
β_2	-0.196	0.108
β_3	-0.104	0.066
$\min r_i $	1.002	0.002

(c) $\Pr(\beta \in C|y)$ and $\Pr(\beta \in D|y)$ can also be calculated in the Monte Carlo integration program. Formally, since $\Pr(\beta \in C|y) = E[I(\beta \in C|y)]$ and $\Pr(\beta \in D|y) = E[I(\beta \in D|y)]$, where $I(\cdot)$ is the indicator function, the probabilities can be written as expected values of functions of the model parameters. In our computer code, calculating these expected values is equivalent to simply calculating the proportion of draws from the posterior that are in C (or D). Our Monte Carlo integration program, using 10,000 replications, calculates $\Pr(\beta \in C|y) = .021$ and $\Pr(\beta \in D|y) = .132$. Thus, it is very unlikely that real GDP exhibits oscillatory behavior and only slightly more likely that it is explosive.

Exercise 17.2 (The threshold autoregressive (TAR) model) Dynamics of many important macroeconomic variables can potentially vary over the business cycle. This has motivated the development of many models where different autoregressive representations apply in different regimes. Threshold autoregressive (TAR) models are a class of simple and popular regime-switching models. Potter (1995) provides an early exposition of these models in macroeconomics and Geweke and Terui (1993) is an early Bayesian treatment. This exercise asks you to derive Bayesian methods for a simple variant of a TAR model.

Consider a two-regime TAR model for a time series variable y_t for $t = p + 1, \dots, T$ (where $t = 1, \dots, p$ are used as initial conditions):

$$\begin{aligned} y_t &= \beta_{10} + \beta_{11}y_{t-1} + \dots + \beta_{1p}y_{t-p} + \epsilon_t \text{ if } y_{t-1} \leq \tau, \\ y_t &= \beta_{20} + \beta_{21}y_{t-1} + \dots + \beta_{2p}y_{t-p} + \epsilon_t \text{ if } y_{t-1} > \tau, \end{aligned} \quad (17.5)$$

where $\epsilon_t \stackrel{\text{iid}}{\sim} N(0, h^{-1})$. We will use the notation $\beta = (\beta'_1, \beta'_2)'$, where

$$\beta_j = (\beta_{j0}, \beta_{j1}, \dots, \beta_{jp})', \quad j = 1, 2.$$

For all parts of this question, you may proceed conditionally on the first p observations and, thus, ignore the (minor) complications caused by initial conditions [see Exercise 17.1(a) for more detail].

(a) Assuming that τ is known (e.g., $\tau = 0$) and normal-gamma priors are used [i.e., the joint prior for β and h is $NG(\underline{\beta}, \underline{Q}, \underline{s}^{-2}, \underline{\nu})$], derive the posterior for the model defined by (17.5).

(b) Using an appropriate data set (e.g., the real GDP data set available on the Web site associated with this book), write a program and carry out Bayesian inference in the TAR model using your results from part (a). Note that when working with GDP, TAR models are usually specified in terms of GDP growth. Hence, if you are working with a GDP series, you should define y_t as its first difference.

(c) Repeat parts (a) and (b) assuming that τ is an unknown parameter.

Solution

(a) The solution to this question follows immediately from noting that the TAR with known τ can be written as a normal linear regression model. That is, analogously to the solution to Exercise 17.1(a), we can write the TAR as

$$y = X\beta + \epsilon,$$

where $y = (y_{p+1}, \dots, y_T)'$, $\epsilon = (\epsilon_{p+1}, \dots, \epsilon_N)'$ and X is the $(T - p) \times 2(p + 1)$ matrix with t th row given by $[D_t, D_t y_{t-1}, \dots, D_t y_{t-p}, (1 - D_t), (1 - D_t) y_{t-1}, \dots, (1 - D_t) y_{t-p}]$, where D_t is a dummy variable that equals 1 if $y_{t-1} \leq \tau$ and equals 0 if $y_{t-1} > \tau$. Thus, we can immediately use the results from Exercise 10.1 to find that the joint posterior for β and h is $NG(\bar{\beta}, \bar{Q}, \bar{s}^{-2}, \bar{\nu})$, where $\bar{\nu} = T - p + \nu$,

$$\bar{Q} = (\underline{Q}^{-1} + X'X)^{-1},$$

$$\bar{\beta} = \bar{Q} (\underline{Q}^{-1} \underline{\beta} + X'X \hat{\beta}),$$

and

$$\bar{s}^2 = \frac{\nu s^2 + SSE + (\hat{\beta} - \underline{\beta})' X' X \bar{Q} \underline{Q}^{-1} (\hat{\beta} - \underline{\beta})}{\bar{\nu}}.$$

Note that, in the preceding equation, we are using the same notation defined at the beginning of Chapter 10, where $\hat{\beta}$ is the standard OLS estimator and SSE the accompanying sum of squared residuals.

(b) A Matlab program that provides the solution to this part of the exercise is available on the Web site associated with this book. The basic structure of this program is that it loads in the data, specifies prior hyperparameters, and then evaluates the formulae provided in the solution to part (a). For US real GDP growth from 1947Q2 through 2005Q2, we obtain the following results relating to the posteriors of β and $\sigma^2 = 1/h$. We set $p = 2$ and use the noninformative variant of the normal-gamma prior. Furthermore, we set τ to the mean of y_{t-1} and thus the two regimes can be interpreted as “below average growth” and “above average growth” regimes.

Table 17.2: Posterior results for the TAR model
with τ known.

Parameter	Mean	Std. Dev.
β_{10}	0.572	0.100
β_{11}	0.289	0.129
β_{12}	0.008	0.092
β_{20}	0.213	0.232
β_{21}	0.394	0.130
β_{22}	0.218	0.093
σ^2	0.879	0.083

(c) When τ is treated as an unknown parameter, then we need methods for working with the posterior $p(\beta, h, \tau|y)$. This can be done by noting that the rules of conditional probability imply

$$p(\beta, h, \tau|y) = p(\beta, h|\tau, y) p(\tau|y).$$

The first component on the right-hand side of this equation, $p(\beta, h|\tau, y)$, can be analyzed using the methods of parts (a) and (b). That is, conditioning upon τ is equivalent to

assuming it known and the answer to part (a) shows how to analyze the posterior for β and h for known τ . Hence, we can focus on the second component, $p(\tau|y)$. But, the intuition that conditioning on τ is equivalent to assuming it known can be carried further to provide us with a way of obtaining $p(\tau|y)$. Note first that, for a given value for τ , the marginal likelihood can be calculated in the standard way:

$$p(y|\tau) = \int \int L(\beta, h|\tau) p(\beta, h|\tau) d\beta dh, \quad (17.6)$$

where $L(\beta, h|\tau)$ is the likelihood function for a given value of τ . However, as we have seen from part (a), the TAR model (conditional on τ) is simply a normal linear regression model. Thus, with the normal-gamma natural conjugate prior, standard textbook expressions exist for (17.6) – conditional on τ [see, e.g., Exercise 10.4, Koop (2003), p. 41, or Poirier (1995), p. 543]. If we use the notation from the solution to part (a), then we can draw on these earlier derivations to write (17.6) as

$$p(y|\tau) = \frac{\Gamma(\frac{\bar{v}}{2}) (\nu \bar{s}^2)^{\frac{\bar{v}}{2}}}{\Gamma(\frac{\nu}{2}) \pi^{\frac{T-p}{2}}} \left(\frac{|\bar{Q}|}{|\underline{Q}|} \right)^{\frac{1}{2}} (\nu \bar{s}^2)^{-\frac{\bar{v}}{2}}, \quad (17.7)$$

although we stress that X depends on τ (and, thus, so will \bar{Q} and \bar{s}^2). Hence (17.7) will be different for every possible value of τ .

Since Bayes' theorem implies

$$p(\tau|y) \propto p(y|\tau) p(\tau), \quad (17.8)$$

we can combine (17.7) with a prior for τ to obtain the marginal posterior $p(\tau|y)$. Of course, any prior for τ can be used. A common choice is a restricted noninformative one. This treats every possible value for τ as, a priori, equally likely, which implies each regime contains a minimum number of observations (e.g., 15 percent of the observations).

Note that, even though y_{t-1} is a continuous variable, τ will be a discrete random variable since there are a finite number of ways of dividing a given data set into two regimes. Hence, in practice, posterior inference in the TAR with an unknown threshold can be performed by evaluating (17.7) and, then (17.8) for every allowable threshold value. This provides us with $p(\tau|y)$. If interest centers on β and/or h , we can obtain their posterior properties by taking a weighted average over all possible thresholds, using the normalized $p(\tau|y)$ as weights. The validity of this procedure follows from the fact that, if $\tau \in \{\tau_1, \dots, \tau_{T^*}\}$ denotes the set of possible threshold values, then

$$p(\beta, h|y) = \sum_{i=1}^{T^*} p(\beta, h|\tau = \tau_i, y) p(\tau = \tau_i|y).$$

Thus, insofar as we are willing to use the normal-gamma natural conjugate prior, Bayesian inference in the TAR model can be performed using only textbook analytical results for the normal linear regression model with no posterior simulation required.

A Matlab program that extends the program in part (b) to allow for an unknown threshold is provided on the Web site associated with this book. Here we note one aspect of our

program that is commonly used with regime-switching models. The assumption about the errors implies that we have added enough lags to make the errors independent of one another. Thus, $y_t|X_t, \beta, h$ and $y_s|X_s, \beta, h$ are independent of one another for $t \neq s$. This implies that the ordering of rows in our regression model is irrelevant. Thus, we can order all our variables (dependent and explanatory variable) by y_{t-1} without changing the posterior. That is, variables can be ordered so that the first observation of all variables is the one with the lowest value for GDP growth last quarter, the second observation has the second highest value for GDP growth last quarter, etc.. If we order our data in this way, then the TAR is equivalent to a structural break (also called a changepoint) model (see Exercises 11.11 and 11.12). Such an ordering makes the step “evaluate (17.7) for every allowable threshold” particularly easy.

Table 17.3: Posterior results for the TAR model with unknown threshold τ .

Parameter	Mean	Std. Dev.
β_{10}	0.542	0.119
β_{11}	0.269	0.139
β_{12}	0.373	0.101
β_{20}	0.249	0.242
β_{21}	0.384	0.127
β_{22}	0.196	0.100
σ^2	0.896	0.083

Posterior properties of β and σ^2 , using the US real GDP growth data, are provided in Table 17.3. Figure 17.1 plots the posterior of τ . Since we are calculating marginal likelihoods, an informative prior is required. We use a weakly informative prior, which reflects our belief that AR coefficients are likely to be well within the stationary region and the error variance is likely to be approximately 1 (since our data are measured as a percentage). However, because we are quite uncertain about these prior means, we attach moderately large prior variances to them. Prior hyperparameter values consistent with these beliefs are $\underline{\beta} = 0$, $\underline{Q} = .25I$, $\underline{s}^2 = 1$, and $\underline{\nu} = 5$. In a serious empirical study, the researcher would likely elicit the prior in a more sophisticated manner and/or carry out a prior sensitivity analysis. The posterior properties of β and h are not that different from those found in part (b), despite the fact that we are now averaging over all the thresholds (rather than selecting a single value for the threshold and treating it as known). However, the posterior for the thresholds (see Figure 17.1) is very irregular. Appreciable posterior probability is found for many possible thresholds. The posterior mode is roughly the same as the average quarterly growth rate of GDP (0.85%). This threshold value would divide the data into a “below average growth” regime and an “above average growth” regime, as was done in part (a), but many other threshold values also receive some posterior support from the data.

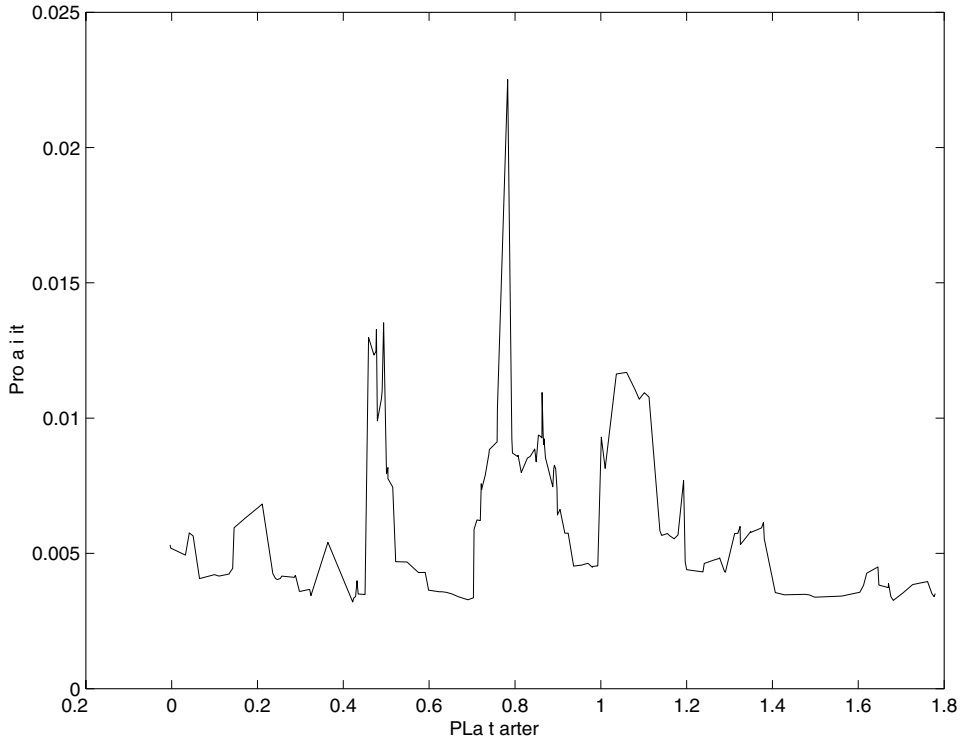


Figure 17.1 — Posterior distribution of the threshold parameter τ .

Exercise 17.3 (Extensions of the basic TAR model, 1: other threshold triggers) There are many extensions of the TAR that have been found to be useful in empirical work. In Exercise 17.2, we assumed that the first lag of the dependent variable (last quarter's GDP growth) triggered the regime switch. However, in general, it might be another variable, z , that is the *threshold trigger* and it may take longer than one period to induce the regime switch. Thus, we use the same assumptions and definitions as in Exercise 17.2, except we replace (17.5) by

$$\begin{aligned} y_t &= \beta_{10} + \beta_{11}y_{t-1} + \cdots + \beta_{1p}y_{t-p} + \epsilon_t \text{ if } z_{t-d} \leq \tau, \\ y_t &= \beta_{20} + \beta_{21}y_{t-1} + \cdots + \beta_{2p}y_{t-p} + \epsilon_t \text{ if } z_{t-d} > \tau, \end{aligned} \quad (17.9)$$

where d is the *delay parameter* and z_{t-d} is either an exogenous variable or a function of the lags of the dependent variable.

(a) Assuming that d is an unknown parameter with a noninformative prior over $1, \dots, p$ [i.e., $\Pr(d = i) = 1/p$ for $i = 1, \dots, p$] and normal-gamma priors are used [i.e., the joint prior for β and h is $NG(\underline{\beta}, \underline{Q}, \underline{s}^{-2}, \underline{\nu})$], derive the posterior for the model defined by (17.9).

(b) Using an appropriate data set (e.g., the real GDP growth data set available on the Web site associated with this book), write a program and carry out Bayesian inference for this model using your results from part (a). Set $p = 4$ and

$$z_{t-d} = \frac{\sum_{d=1}^p y_{t-d}}{d},$$

so that (if you are using quarterly real GDP growth data) the threshold trigger is average GDP growth over the last d quarters.

Solution

(a) The solution to this exercise is a straightforward extension of the solution to Exercise 17.2 parts (a) and (c). That is, it follows immediately from noting that, if we condition on τ and d , then this TAR model can be written as a normal linear regression model. Note that, since z_{t-d} is either exogenous or a function of lags of the dependent variable, we can also condition on it in the same manner as we condition on X . As in Exercise 17.2(c), we can write the TAR as

$$y = X\beta + \epsilon,$$

where X is the $(T - p) \times 2(p + 1)$ matrix with t th row given by

$$[D_t, D_t y_{t-1}, \dots, D_t y_{t-p}, (1 - D_t), (1 - D_t) y_{t-1}, \dots, (1 - D_t) y_{t-p}],$$

and where D_t is now a dummy variable that equals 1 if $z_{t-d} \leq \tau$ and equals 0 if $z_{t-d} > \tau$. Thus, as in the solution to Exercise 17.2(a), the posterior for β and h (now conditional on τ and d) is $NG(\bar{\beta}, \bar{Q}, \bar{s}^{-2}, \bar{\nu})$, where $\bar{\beta}$, \bar{Q} , \bar{s}^{-2} , and $\bar{\nu}$ are defined as in that solution except that X , and thus the posterior hyperparameters, now depend on τ and d .

Analogous to the solution to Exercise 17.2(c), we can write

$$p(\beta, h, \tau, d|y) = p(\beta, h|\tau, d, y) p(\tau, d|y).$$

The preceding paragraph describes methods of posterior inference for $p(\beta, h|\tau, d, y)$ and therefore we can focus on $p(\tau, d|y)$. As in Exercise 17.2(c), we can also use the facts that

$$p(y|\tau, d) = \int \int L(\beta, h|\tau, d) p(\beta, h|\tau, d) d\beta dh$$

and $p(\tau, d|y) \propto p(y|\tau, d) p(\tau, d)$ to note that, by calculating the marginal likelihoods for the normal linear regression model (conditional on τ and d) for every possible value of τ and d , we can build up the posterior $p(\tau, d|y)$. The relevant formula for the marginal likelihood is given in (17.7).

(b) A Matlab program that extends the program of Exercise 17.2(c) to allow for an unknown delay (and general threshold trigger) is provided on the Web site associated with this book. Since we are calculating marginal likelihoods, we use the same weakly informative prior described in the solution to Exercise 17.4(c). As we stressed in that solution, in a serious empirical study, the researcher would likely elicit the prior in a more sophisticated manner and/or carry out a prior sensitivity analysis. Posterior properties of β and σ^2 , using the US real GDP growth data, are provided in Table 17.4. Figure 17.2 plots the posterior of d . Note

Table 17.4: Posterior results for the TAR model with unknown threshold and delay.

Parameter	Mean	Std. Dev.
β_{10}	0.255	0.226
β_{11}	0.080	0.174
β_{12}	-0.330	0.196
β_{13}	-0.340	0.167
β_{14}	-0.264	0.166
β_{20}	0.557	0.159
β_{21}	0.253	0.087
β_{22}	0.270	0.080
β_{23}	-0.110	0.076
β_{24}	-0.008	0.072
σ^2	0.806	0.076

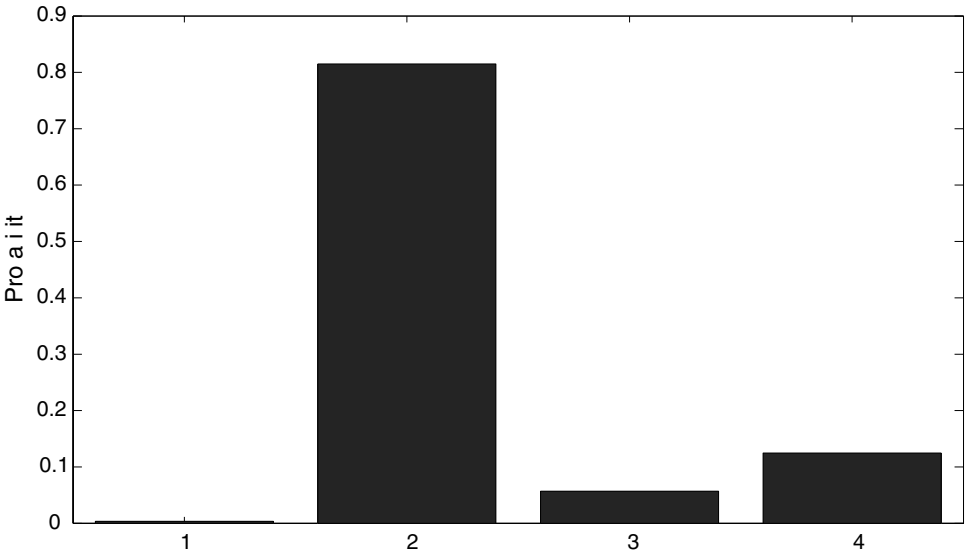


Figure 17.2 — Posterior distribution of d .

that, since different values of d imply different threshold triggers (with different interpretations), the interpretation of the threshold differs across d . Hence, it would make little sense to plot the posterior of τ and so we do not do so [although one could plot $p(\tau|y, d)$ for a choice of d]. Interestingly, the posterior for d allocates most of the probability to $d = 2$, indicating that it is not last quarter’s GDP growth that triggers a regime shift (as the model of Exercise 17.2 implies), but rather GDP growth averaged over the past two quarters. The finding that GDP growth two quarters ago plays an important role in regime shifting in US GDP has been found by others [e.g., Potter (1995)].

Note also that the solution to this question can be interpreted as a standard Bayesian posterior analysis, where d is an unknown parameter whose posterior we have derived. Alternatively (and equivalently) it can be an example of Bayesian model averaging where we are presenting results (e.g., in Table 17.4) that are averaged over different models that are defined by different threshold triggers. Koop and Potter (1999) present an empirical paper that focuses on the second interpretation.

Exercise 17.4 (Extensions of the basic TAR model, 2: switches in the error variance)

Recently, there has been much interest in the volatility of macroeconomic variables and, in particular, whether the error variance exhibits regime-switching behavior [see, e.g., Zha and Sims (2006)]. Accordingly, we can extend the models of Exercise 17.2 and 17.3 to

$$\begin{aligned} y_t &= \beta_{10} + \beta_{11}y_{t-1} + \cdots + \beta_{1p}y_{t-p} + \sqrt{h_1^{-1}}\epsilon_t \text{ if } z_{t-d} \leq \tau, \\ y_t &= \beta_{20} + \beta_{21}y_{t-1} + \cdots + \beta_{2p}y_{t-p} + \sqrt{h_2^{-1}}\epsilon_t \text{ if } z_{t-d} > \tau, \end{aligned} \quad (17.10)$$

where all definitions and assumptions are the same as in Exercises 17.2 and 17.3 except that we now assume $\epsilon_t \stackrel{\text{iid}}{\sim} N(0, 1)$.

(a) Assume that d is an unknown parameter with a noninformative prior over $1, \dots, p$ [i.e., $\Pr(d = i) = 1/p$ for $i = 1, \dots, p$] and normal-gamma priors are used in each regime [i.e., the joint prior for β_j and h_j is $NG(\beta_j, Q_j, s_j^{-2}, \nu_j)$ for $j = 1, \dots, 2$]. Derive the posterior for the model defined by (17.10).

(b) Using an appropriate data set (e.g., the real GDP growth data set available on the Web site associated with this book), write a program and carry out Bayesian inference for this model using your results from part (a). Set $p = 4$ and

$$z_{t-d} = \frac{\sum_{d=1}^p y_{t-d}}{d}.$$

Solution

(a) The posterior for the parameters of this model can be obtained by noting that, conditional on τ and d , the model breaks down into two linear regression models with natural conjugate priors. Thus, standard results (see Chapter 10, Exercise 10.1) for the regression model can be combined to produce a conditional posterior. The unconditional posterior can be obtained by using the same methods as in the solution to Exercise 17.3. Precise details follow.

We are interested in the posterior $p(\beta_1, h_1, \beta_2, h_2, \tau, d|y)$, which (given the assumed prior and fact that the errors are assumed to be independent) can be written as

$$p(\beta_1, h_1, \beta_2, h_2, \tau, d|y) = p(\beta_1, h_1|y^1, \tau, d) p(\beta_2, h_2|y^2, \tau, d) p(\tau, d|y),$$

where y^j denotes the data in the j th regime for $j = 1, 2$ (which depends on τ and d). Here, $p(\beta_j, h_j|y^j, \tau, d)$ is simply the posterior for the normal linear regression model with

natural conjugate prior using the data in the j th regime (for $j = 1, 2$). That is, we can write each regime as

$$y^j = X^j \beta_j + \epsilon^j,$$

where y^j and ϵ^j contain all the observations on the dependent variable and errors, respectively, for $j = 1, 2$. The matrix X^j contains an intercept and p lags of the dependent variable for observations in the j th regime. We can immediately use the results from Exercise 10.1 to find that the joint posterior for β_j and h_j (conditional on τ and d) is $NG(\bar{\beta}_j, \bar{Q}_j, \bar{s}_j^{-2}, \bar{\nu}_j)$, where $\bar{\nu}_j = T_j + \underline{\nu}_j$, T_j is the number of observations in the j th regime,

$$\bar{Q}_j = \left(\underline{Q}_j^{-1} + X^{j'} X^j \right)^{-1},$$

$$\bar{\beta}_j = \bar{Q}_j \left(\underline{Q}_j^{-1} \underline{\beta}_j + X^{j'} X^j \hat{\beta}_j \right),$$

and

$$\bar{s}_j^2 = \frac{\nu_j s_j^2 + SSE_j + \left(\hat{\beta}_j - \underline{\beta}_j \right)' X^{j'} X^j \bar{Q}_j \underline{Q}_j^{-1} \left(\hat{\beta}_j - \underline{\beta}_j \right)}{\bar{\nu}}.$$

Note that, in the preceding equation, we are adapting our previous notation so that $\hat{\beta}_j$ and SSE_j are now the OLS estimate and OLS sum of squared residuals, respectively, using data from regime j . This completes our description of $p(\beta_1, h_1 | y^1, \tau, d)$ and $p(\beta_2, h_2 | y^2, \tau, d)$.

In our answer to Exercise 17.3(a) we noted that $p(\tau, d | y) \propto p(y | \tau, d) p(\tau, d)$. Since $p(y | \tau, d)$ was simply the marginal likelihood for the normal linear regression model (conditional on τ and d), we could obtain the requisite posterior by combining the standard analytical formula for this marginal likelihood with $p(\tau, d)$. Here we can adopt the same strategy with one slight extension, by noting that the assumptions of our model imply

$$p(y | \tau, d) = p(y^1 | \tau, d) p(y^2 | \tau, d).$$

Thus, we can calculate marginal likelihoods for the normal linear regression model in each of our two regimes (defined by τ and d) and simply multiply them together to yield the required $p(y | \tau, d)$. As in Exercise 17.3, we must evaluate this marginal likelihood for every possible threshold and delay parameter but, because these are both discrete random variables, this can be done in a straightforward fashion. The relevant formula for the marginal likelihood (see Exercise 17.2) in each regime, adapted to our present notation, is

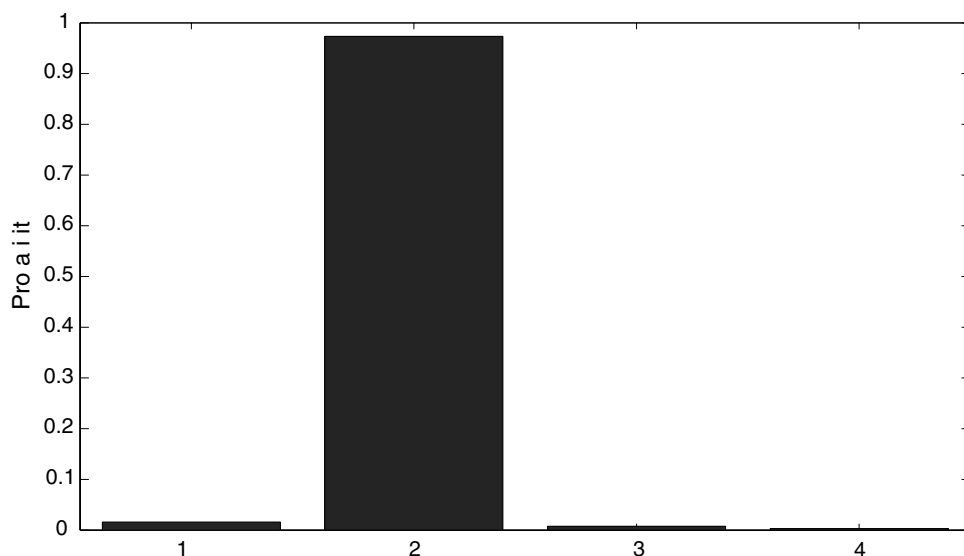
$$p(y^j | \tau, d) = \frac{\Gamma\left(\frac{\bar{\nu}_j}{2}\right) \left(\frac{\nu_j s_j^2}{2}\right)^{\frac{\bar{\nu}_j}{2}}}{\Gamma\left(\frac{\nu_j}{2}\right) \pi^{\frac{T_j}{2}}} \left(\frac{|\bar{Q}_j|}{|\underline{Q}_j|}\right)^{\frac{1}{2}} (\bar{\nu}_j s_j^2)^{-\frac{\bar{\nu}_j}{2}}.$$

(b) The Matlab program that carries out posterior inference in the TAR with regime switching in the error variance is provided on the Web site associated with this book. We ran this program, with our US real GDP growth data set, using the same weakly informative prior for the coefficients in each regime described in the solution to Exercise 17.2(b).

Table 17.5: Posterior results for the TAR model with switches in error variance.

Parameter	Mean	Std. Dev.
β_{10}	0.128	0.227
β_{11}	0.075	0.197
β_{12}	-0.400	0.198
β_{13}	-0.337	0.186
β_{14}	-0.248	0.169
β_{20}	0.568	0.135
β_{21}	0.246	0.077
β_{22}	0.271	0.077
β_{23}	-0.113	0.071
β_{24}	-0.084	0.068
σ_1^2	1.022	0.246
σ_2^2	0.772	0.078

Table 17.5 presents posterior results for the regression coefficients and error variances in each regime. In this data set, allowing for regime switching in the errors makes little difference for the AR coefficients (compare Table 17.4 with Table 17.5). Note also that Figure 17.3 indicates the same pattern as in Figure 17.2: It is average GDP growth over the last two quarters that triggers the change in regime. However, there is some evidence

Figure 17.3 — Posterior distribution of d .

that the error variances in the two regimes are different from one another. In particular, the error variance in the expansionary regime has a point estimate that is substantially smaller than the point estimate of the error variance in the recessionary regime. However, posterior standard deviations associated with these point estimates are fairly large.

As a final thought, it is worth noting that the results for Exercises 17.2–17.4 can be extended in a straightforward manner to the cases of many regimes, regimes with different AR lag orders, TAR models where the threshold triggers depend on unknown parameters, etc.

Exercise 17.5 (Autoregressive conditional heteroscedasticity) Particularly in finance, interest often centers of the volatility of a time series. There are many popular models for volatility. This exercise discusses one of the more popular ones: autoregressive conditional heteroscedasticity (ARCH). If y_t is an observed time series variable and x'_t a $k \times 1$ vector of explanatory variables (for $t = 1, \dots, T$), then the regression model with ARCH(p) errors is given by

$$y_t = x_t \beta + \epsilon_t, \quad (17.11)$$

where $\epsilon_t \stackrel{\text{iid}}{\sim} N(0, \sigma_t^2)$ and

$$\sigma_t^2 = \alpha_0 + \alpha_1 \epsilon_{t-1}^2 + \dots + \alpha_p \epsilon_{t-p}^2. \quad (17.12)$$

Since variances must be positive, we impose the restrictions $\alpha_j > 0$ for $j = 0, \dots, p$. Denote this positivity restriction by $P = \{\alpha : \alpha_j > 0 \text{ for } j = 0, \dots, p\}$, where $\alpha = (\alpha_0, \alpha_1, \dots, \alpha_p)'$. Furthermore, let the roots of $1 - \sum_{i=1}^p \alpha_i z^i$ be denoted by r_i for $i = 1, \dots, p$. We will restrict (17.12) to be stationary by imposing $\min |r_i| > 1$. Denote the stationary region as $S = \{\alpha : \min |r_i| > 1\}$.

(a) Derive a posterior simulator that carries out Bayesian inference in the normal linear regression model with ARCH errors. Use the following diffuse prior (subject to the positivity and stationarity restrictions):

$$p(\beta, \alpha) \propto I(P) I(S). \quad (17.13)$$

As an aside, the interested reader is referred to Geweke (1988) for a discussion of prior elicitation in ARCH models, including a discussion of several other diffuse priors. You may proceed conditionally on the first p observations and thus ignore the (minor) complications caused by initial conditions [see Exercise 17.1(a) for more detail].

(b) Using an appropriate data set, write a program and carry out Bayesian inference for this model using your results from part (a). The Web site associated with this book provides a monthly New York Stock Exchange (NYSE) stock return data set and the solution to this exercise uses these data. These data contain monthly stock returns (exclusive of dividends) from the NYSE from 1952 through 1995. If you are using the NYSE data set, let x_t contain only an intercept and set $p = 4$.

Solution

(a) The posterior is obtained by multiplying prior by likelihood. By using the properties of the normal distribution (and conditioning on y_1, \dots, y_p), the log likelihood is (dropping irrelevant constants)

$$\begin{aligned} \ln L(\beta, \alpha) &= -\frac{1}{2} \sum_{t=p+1}^T \ln(\sigma_t^2) - \frac{1}{2} \sum_{t=p+1}^T \frac{\epsilon_t^2}{\sigma_t^2} \\ &= -\frac{1}{2} \sum_{t=p+1}^T \ln \left(\alpha_0 + \alpha_1 (y_{t-1} - x'_{t-1}\beta)^2 + \dots + \alpha_p (y_{t-p} - x'_{t-p}\beta)^2 \right) \\ &\quad - \frac{1}{2} \sum_{t=p+1}^T \frac{(y_t - x'_t\beta)^2}{\alpha_0 + \alpha_1 (y_{t-1} - x'_{t-1}\beta)^2 + \dots + \alpha_p (y_{t-p} - x'_{t-p}\beta)^2}. \end{aligned} \quad (17.14)$$

The difficulty of posterior inference in the ARCH model can be seen from (17.14). The likelihood does not belong in any convenient class of distributions. Hence, even with a convenient prior, it does not lend itself to analytical results, nor to direct Monte Carlo integration or Gibbs sampling. Note that, unlike many other extensions of the basic regression model, even the conditional posterior for β does not have a simple form. This arises since (17.12) depends on the errors and these depend on β . Thus, β enters not only the conditional mean, but also the conditional variance of the model. Several approaches to posterior simulation are possible [see, e.g., Geweke (1988a) and Bauwens et al. (1999a), Chapter 7]. Here we describe a random walk chain Metropolis–Hastings [M–H] algorithm (see the introductory material in Chapter 11 and Exercise 11.18). Let $\theta = (\beta', \alpha')'$ and $\theta^{(r)}$ for $r = 1, \dots, R$ denote the posterior simulator output. The random walk M–H algorithm proceeds by generating candidate draws, θ^* , according to

$$\theta^* = \theta^{(r-1)} + z,$$

where z is the increment random variable, which we take to be $N(0, \Sigma)$. These draws are accepted with probability

$$\min \left[\frac{p(\theta = \theta^*|y)}{p(\theta = \theta^{(r-1)}|y)}, 1 \right].$$

If θ^* is accepted then $\theta^{(r)} = \theta^*$; otherwise the posterior simulator sets $\theta^{(r)} = \theta^{(r-1)}$. Note that, since the posterior appears in both the numerator and denominator of the acceptance probability, the integrating constant will cancel out and, hence, we do not need to fully know the posterior, but only its kernel. In our case, the kernel of the posterior is (17.13) times (17.14).

(b) A Matlab program that implements the random walk chain M–H algorithm described in the solution to part (a) is available on the Web site associated with this book. This algorithm requires the selection of Σ . There are many ways of doing this and, typically, some

experimentation is required to find a value that yields a reasonable acceptance probability. Remember that, with the M–H algorithm you do not want an acceptance probability that is either too low or too high. If it is too low, then virtually all candidate draws are rejected and results will, in essence, be based on very few draws. If it is too high, this is evidence that the random walk chain algorithm is taking candidate draws too close to the current draw. In this case, the algorithm might not be exploring the entire posterior distribution in a reasonable number of draws. Accepting somewhere between a quarter and a half of the draws is often taken as a reasonable benchmark to aim for. We have achieved this by initially setting $\Sigma = cI$ and experimenting with the scalar c until we obtain an acceptance probability that is not very near zero or one. We then use this choice for Σ to produce an estimate of the posterior covariance matrix: $\widehat{\text{Var}}(\theta|y)$. We then set $\Sigma = c\widehat{\text{Var}}(\theta|y)$ and experimented again with different values for c . We finally used $c = 0.5$ in a long run (1,000 burn-in replications and 100,000 included replications) to produce our final results. This choice for Σ resulted in 37.1 percent of the candidate draws being accepted.

Table 17.6 presents posterior means and standard deviations for the parameters of the ARCH(4) model using the NYSE stock return data.

Table 17.6: Posterior results for a regression model with ARCH errors.

Parameter	Mean	Std. Dev.
β	0.666	0.166
α_0	9.343	1.526
α_1	0.067	0.041
α_2	0.080	0.055
α_3	0.240	0.089
α_4	0.140	0.081

Exercise 17.6 (Vector autoregressive models) The previous exercises in this chapter involved only a single variable. However, researchers in macroeconomics and finance are often interested in the relationships among many variables. Of the many multivariate time series models that have been developed, variants of the vector autoregressive (VAR) model are the most popular. The VAR(p) model can be written as

$$y_t = a_0 + \sum_{j=1}^p A_j y_{t-j} + \epsilon_t, \quad (17.15)$$

where y_t for $t = 1, \dots, T$ is an $M \times 1$ vector containing observations on M time series variables, ϵ_t is an $M \times 1$ vector of errors, a_0 is an $M \times 1$ vector of intercepts and A_j is an $M \times M$ matrix of coefficients. We assume $\epsilon_t \stackrel{\text{iid}}{\sim} N(0, \Sigma)$ and let $A = (a \ A_1 \ \cdots \ A_p)'$.

(a) Using a common noninformative prior,

$$p(A, \Sigma) \propto |\Sigma|^{-\frac{M+1}{2}},$$

derive the posterior of the VAR model and discuss Bayesian estimation.

(b) In some cases, the researcher may be interested in restricted versions of the VAR of part (a). For instance, instead of every variable depending on the same number of lags of all variables (as the unrestricted VAR does), the researcher might want to have a different number of lags in each equation and/or each variable depending on lags of only some of the other variables. Discuss how Bayesian inference can be carried out in such a model.

Solution

(a) The VAR model can be written as

$$y_t = A'x_t + \epsilon_t,$$

where $x_t = (1 \ y'_{t-1} \ \cdots \ y'_{t-p})'$. In matrix form we can then write

$$Y = XA + E,$$

where Y and E are $T \times M$ matrices with t th rows given by y'_t and ϵ'_t , respectively. X is a $T \times k$ matrix with t th rows given by x'_t , where $k = 1 + Mp$. E has a matrix-variate normal distribution (see Appendix Definition 12):

$$E \sim MN(0, \Sigma \otimes I_T).$$

Using the properties of the matrix-variate normal distribution (see Appendix Definition 12), we can obtain the likelihood function as

$$L(A, \Sigma) \propto |\Sigma|^{-\frac{1}{2}T} \exp \left\{ -\frac{1}{2} \text{tr} [\Sigma^{-1} (Y - XA)' (Y - XA)] \right\}.$$

Multiplying this by the prior yields the posterior:

$$p(A, \Sigma | y) \propto |\Sigma|^{-\frac{1}{2}(T+M+1)} \exp \left\{ -\frac{1}{2} \text{tr} [\Sigma^{-1} (Y - XA)' (Y - XA)] \right\}.$$

Analogous to derivations for the Normal linear regression model (see introduction to Chapter 10), we can write the term in the exponential in terms of OLS quantities. In particular, if we define

$$\hat{A} = (X'X)^{-1} X'Y$$

and

$$S = Y'Y - Y'X (X'X)^{-1} X'Y,$$

we can use the fact that

$$(Y - XA)' (Y - XA) = S + (A - \hat{A})' X'X (A - \hat{A})$$

to write the posterior as

$$\begin{aligned} p(A, \Sigma | y) &\propto |\Sigma|^{-\frac{1}{2}(T+M+1)} \exp \left\{ -\frac{1}{2} \text{tr} \left[\Sigma^{-1} \left[S + (A - \hat{A})' X' X (A - \hat{A}) \right] \right] \right\} \\ &= |\Sigma|^{-\frac{1}{2}(T+M+1)} \left[\exp \left\{ -\frac{1}{2} \text{tr} [\Sigma^{-1} S] \right\} \right] \\ &\quad \left[\exp \left\{ -\frac{1}{2} \text{tr} \left[\Sigma^{-1} (A - \hat{A})' X' X (A - \hat{A}) \right] \right\} \right]. \end{aligned}$$

Examining this as a function of Σ^{-1} , we can see that the marginal posterior of Σ^{-1} is Wishart. By examining this as a function of A , a matric-variate normal kernel can be found. Thus, we have

$$A | y, \Sigma \sim MN \left(\text{vec} [\hat{A}], \Sigma \otimes [X' X]^{-1} \right)$$

and

$$\Sigma^{-1} | y \sim W(S^{-1}, T).$$

Thus, Bayesian estimation of the normal VAR with a noninformative prior can be carried out using analytical results based on OLS quantities. This is the multivariate generalization of the normal linear regression model with a (noninformative version of a) natural conjugate prior (see Chapter 10, Exercise 10.1). Standard results for the natural conjugate prior for the VAR can be used to derive many other results. For instance, the marginal posterior for A is matric-variate t [see, e.g., Bauwens et al. (1999), Chapter 9]. Kadiyala and Karlsson (1997) provide a discussion of a variety of other priors for VAR models. They show, however, that analytical results are not available for most of the priors the researcher may want to use. Gibbs sampling is commonly used with such extensions.

(b) The VAR with different explanatory variables in each equation is equivalent to a SUR (see Chapter 11, Exercise 11.10). In particular, we can write

$$y_{mt} = x'_{mt} \beta_m + \epsilon_{mt},$$

with $t = 1, \dots, T$ observations for $m = 1, \dots, M$ variables, where y_{mt} is the t th observation on the m th variable, x_{mt} is a k_m vector containing the t th observation of the vector of explanatory variables relevant for the m th variable, and β_m is the accompanying k_m vector of regression coefficients.

We can stack all equations into vectors or matrices as $y_t = (y_{1t}, \dots, y_{Mt})'$, $\epsilon_t = (\epsilon_{1t}, \dots, \epsilon_{Mt})'$,

$$\beta = \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_M \end{pmatrix},$$

and

$$X_t = \begin{pmatrix} x'_{1t} & 0 & . & . & 0 \\ 0 & x'_{2t} & 0 & . & . \\ . & . & . & . & . \\ . & . & . & . & 0 \\ 0 & . & . & 0 & x'_{Mt} \end{pmatrix}$$

and define $k = \sum_{m=1}^M k_m$. Using this notation, we can write the restricted VAR as

$$y_t = X_t \beta + \epsilon_t.$$

We now stack all the observations together as

$$y = \begin{pmatrix} y_1 \\ \vdots \\ y_T \end{pmatrix},$$

$$\epsilon = \begin{pmatrix} \epsilon_1 \\ \vdots \\ \epsilon_T \end{pmatrix},$$

and

$$X = \begin{pmatrix} X_1 \\ \vdots \\ X_T \end{pmatrix}$$

and write

$$y = X\beta + \epsilon.$$

Thus, ϵ is $N(0, \Omega)$, where Ω is an $NM \times NM$ block-diagonal matrix given by

$$\Omega = \begin{pmatrix} \Sigma & 0 & . & . & 0 \\ 0 & \Sigma & . & . & . \\ . & . & . & . & . \\ . & . & . & . & 0 \\ 0 & . & . & 0 & \Sigma \end{pmatrix}.$$

It can be seen that the restricted VAR can be written as a Normal linear regression model with general error covariance matrix (see Chapter 13, Exercise 13.1) and the results derived in that chapter can be directly used. Analytical results do not exist for this model and, thus, we need to think in terms of posterior simulation.

Perhaps the easiest way of developing methods of posterior simulation is to note that, since Ω is a positive definite matrix, a standard theorem in matrix algebra implies that

a matrix P exists with the property that $P\Omega P' = I_N$. If we multiply both sides of the regression model by P , we obtain a transformed model:

$$y^* = X^*\beta + \epsilon^*, \quad (17.16)$$

where $y^* = Py$, $X^* = PX$, and $\epsilon^* = P\epsilon$. It can be verified that ϵ^* is $N(0, I_{TM})$. Hence, conditional upon Ω , we have a normal linear regression model with known error covariance matrix. Thus, for suitably chosen priors (e.g., normal or noninformative), $p(\beta|y, \Omega)$ will be normal. If we can find a convenient form for $p(\Omega|y, \beta)$, a Gibbs sampler can be set up for carrying out posterior inference in this restricted VAR.

A commonly used prior for this model is an extension of the independent normal-gamma prior used with the normal linear regression model (see Chapter 11, Exercise 11.8), the independent normal-Wishart prior:

$$p(\beta, \Sigma^{-1}) = p(\beta) p(\Sigma^{-1}),$$

where

$$p(\beta) = \phi(\beta|\underline{\beta}, \underline{V})$$

and

$$p(\Sigma^{-1}) = f_W(\Sigma^{-1}|\underline{H}, \underline{\nu}).$$

For prior elicitation, the most important properties of the Wishart are that $E(\Sigma^{-1}) = \underline{\nu}\underline{H}$ and noninformativeness is achieved by setting $\underline{\nu} = 0$ and $\underline{H}^{-1} = 0_{M \times M}$.

Using this prior, we can derive the conditional posteriors $p(\beta|y, \Sigma^{-1})$ and $p(\Sigma^{-1}|y, \beta)$ in a straightforward fashion using methods presented in solutions to other problems. That is, the fact that $p(\beta|y, \Sigma^{-1})$ is normal involves the same proof as was provided in Chapter 10, Exercise 10.1, using the transformed model in (17.16). In fact, the proof is a simplified version of our earlier proof since in the transformed model the error covariance matrix is I_{TM} (and, hence, we can employ use here the Exercise 10.1 proof with h set equal to 1). Repeating the Exercise 10.1 proof and then transforming back to the original model results in

$$\beta|y, \Sigma^{-1} \sim N(\bar{\beta}, \bar{V}),$$

where

$$\bar{V} = \left(\underline{V}^{-1} + \sum_{t=1}^T X_t' \Sigma^{-1} X_t \right)^{-1}$$

and

$$\bar{\beta} = \bar{V} \left(\underline{V}^{-1} \underline{\beta} + \sum_{t=1}^T X_t' \Sigma^{-1} y_t \right).$$

The posterior for Σ^{-1} conditional on β can also be derived using solutions to our previous exercises, such as Chapter 11, Exercise 11.10, and Chapter 12, Exercise 12.3.

Alternatively, since we are conditioning on β , this conditional posterior is exactly the same as the one derived in Chapter 2, Exercise 2.14(b) (with appropriate relabeling of prior hyperparameters and noting that the known mean for each observation will now be $X_t\beta$). Using any of these earlier derivations, we can confirm that

$$\Sigma^{-1}|y, \beta \sim W(\bar{H}, \bar{\nu}),$$

where

$$\bar{\nu} = T + \underline{\nu}$$

and

$$\bar{H} = \left[\underline{H}^{-1} + \sum_{t=1}^T (y_t - X_t\beta)(y_t - X_t\beta)' \right]^{-1}.$$

A Gibbs sampler that sequentially draws from the normal $p(\beta|y, \Omega)$ and the Wishart $p(\Sigma^{-1}|y, \beta)$ can be programmed in a straightforward fashion. As with any Gibbs sampler, the resulting posterior simulator output can be used to calculate posterior properties of any function of the parameters, calculate marginal likelihoods (for model comparison), and/or perform prediction.

Some nonstationary time series models

In this chapter we present various exercises that involve nonstationary time series. In particular, we focus on the topics of unit roots and cointegration. In one sense, there is nothing new in the Bayesian treatment of nonstationary variables. For instance, unit root issues are often addressed in the context of AR (or ARMA) models. Cointegration issues are often addressed using restricted versions of VARs. In Chapter 17, we discussed AR and VAR models and the methods derived there were not dependent on the variables being stationary. However, with nonstationary data, some important contrasts exist between Bayesian and frequentist approaches and some important issues of prior elicitation arise. Hence, we devote a separate chapter to nonstationary time series models.

Exercises 18.1 and 18.2 illustrate some of the differences between Bayesian and frequentist results with unit root variables. Exercise 18.1 derives the finite-sample posterior distribution for the AR(1) model with AR coefficient θ . For the reader with no frequentist training in time series, we note that it contrasts sharply with the sampling distribution of the maximum likelihood estimator. The latter is quite complicated and its asymptotic distribution differs markedly depending on whether $\theta < 1$, $\theta = 1$, or $\theta > 1$. Exercise 18.2 reproduces the enlightening Monte Carlo experiment of Sims and Uhlig (1991), which demonstrates the differences between posterior and sampling distributions in finite samples. These differences also persist asymptotically because the convergence of the asymptotic sampling distribution is not uniform [see Kwan (1998)]. Note also that adding a constant or trend complicates the asymptotic sampling distribution in a nontrivial manner. Most of the remaining questions (see Exercises 18.3–18.5) discuss various approaches to prior elicitation in the presence of unit roots. The chapter ends with an exercise involving cointegration.

An issue that arises in time series models is the treatment of initial conditions. For instance, in an AR(1) model with data from $t = 1, \dots, T$, $y_0 = 0$ will appear in the likelihood function. In the previous chapter, we ignored this issue by simply using data from $t = 2, \dots, T$ and using the (observed) y_1 as the initial condition. With nonstationary time

series, treatment of initial conditions can be more important. It is still common to use the approximate (or conditional) likelihood function built from the density $p(y_1, \dots, y_T | \theta, y_0)$. However, for stationary models, it is also possible to obtain the exact likelihood function obtained from the joint density of the sample and the initial condition. Assuming stationarity, we have $y_0 \sim N[0, \sigma^2 / (1 - \theta^2)]$, where σ^2 is the error variance, but if a unit root is present ($\theta = 1$) the variance of this initial condition is undefined. The conditional (on y_0) approach lends itself to analytical results (when using a conjugate prior) and it will be adopted in the following exercises. The exact likelihood function introduces a nonlinearity in the autoregressive coefficient and relies on the assumed stationarity of the process. If the initial conditions bring a lot of information to the likelihood function, they may also produce pathologies when the AR coefficient is near the unit root. The same holds for the constant term. See Bauwens et al. (1999), pp. 169–174 for further details.

The reader interested in exploring any of these issues is referred to Bauwens et al. (1999). Chapter 6 provides an excellent summary of issues relating to unit roots. Cointegration is discussed in Chapter 9. Koop, Strachan, van Dijk, and Villani (2006) survey the Bayesian cointegration literature. A few papers that illustrate some of the issues arising with unit roots include those by Phillips (1991), Schotman (1994), Sims (1988), and Sims and Uhlig (1991).

Exercise 18.1 (The AR(1) model with a flat prior) Let y_t for $t = 1, \dots, T$ represent T observations on a time series variable and set $y_0 = 0$. Assume $\epsilon_t \stackrel{\text{iid}}{\sim} N(0, 1)$. (The assumption that the variance of ϵ_t is known and equal to unity is merely made for convenience.)

Consider the AR(1) model

$$y_t = \theta y_{t-1} + \epsilon_t, \quad (18.1)$$

where θ is an unknown parameter.

- (a) Given a flat prior, find the posterior distribution of θ .
- (b) Find the predictive distribution of y_{T+1} using the prior in part (a).

Solution

- (a) Note that

$$y_t | y_{t-1}, \dots, y_1, \theta \sim N(\theta y_{t-1}, 1).$$

The likelihood function is

$$\begin{aligned} L(\theta) &= \prod_{t=1}^T \phi(y_t; \theta y_{t-1}, 1) \\ &= (2\pi)^{-\frac{T}{2}} \exp \left[-\frac{1}{2} \sum_{t=1}^T (y_t - \theta y_{t-1})^2 \right]. \end{aligned}$$

The maximum likelihood estimate of θ is

$$\hat{\theta} = \frac{\sum_{t=1}^T y_t y_{t-1}}{\sum_{t=1}^T y_{t-1}^2}. \quad (18.2)$$

Noting that

$$\sum_{t=1}^T (y_t - \theta y_{t-1})^2 = \sum_{t=1}^T (y_t - \hat{\theta} y_{t-1})^2 + (\hat{\theta} - \theta)^2 w,$$

where

$$w = \sum_{t=1}^T y_{t-1}^2,$$

we see that

$$\begin{aligned} L(\theta) &= (2\pi)^{-\frac{T}{2}} \exp \left[-\frac{1}{2} \left\{ \sum_{t=1}^T (y_t - \hat{\theta} y_{t-1})^2 + (\hat{\theta} - \theta)^2 w \right\} \right] \\ &\propto \phi(\theta | \hat{\theta}, w^{-1}). \end{aligned} \quad (18.3)$$

Thus the posterior distribution of θ is $N(\hat{\theta}, w^{-1})$.

Note that this exercise is a special case of Exercise 17.1(a). We provide the derivations for this special case since we will use it in the next exercise.

(b) $\theta|y$ is $N(\hat{\theta}, w^{-1})$ implies that $\theta = \hat{\theta} + u$, where $u|y$ is $N(0, w^{-1})$ and $y = (y_1, \dots, y_T)'$. Similarly, $y_{T+1}|y, \theta \sim N(\theta y_T, 1)$ implies $y_{T+1} = \theta y_T + \epsilon_{T+1}$, where $\epsilon_{T+1}|y, \theta$ is $N(0, 1)$. Hence, $y_{T+1} = \hat{\theta} y_T + \epsilon_{T+1} + y_T u$ and the properties of the normal distribution imply $(\epsilon_{T+1} + y_T u) \sim N(0, 1 + y_T^2 w^{-1})$. Therefore,

$$y_{T+1}|y \sim N(\hat{\theta} y_T, 1 + y_T^2 w^{-1}).$$

Exercise 18.2 (Sims and Uhlig [1991]) Using the setup of Exercise 18.1, implement the following experimental design of Sims and Uhlig (1991). Consider 31 values of θ ranging from 0.80 to 1.10 in increments of 0.01. For each value of θ , generate 10,000 random vectors $\epsilon = (\epsilon_1, \dots, \epsilon_T)'$, each comprising $T = 100$ standard normal random variables. Using (18.1), generate a $T \times 1$ vector y in each case. Thus, 310,000 data series for y are available (i.e., 10,000 series for each of the 31 values for θ).

Now here is the analysis to perform. For each of these time series, compute $\hat{\theta}$ using (18.2). Then, for each of the 31 values for θ , construct a histogram of the $\hat{\theta}$'s corresponding to a particular value for θ using bins: $(-\infty, 0.795)$, $[0.795, 0.805)$, \dots , $[1.095, 1.105)$, $[1.105, \infty)$. Note that this will involve the construction of 31 different histograms. Line up these histograms to produce a surface that is the joint p.d.f. of θ and $\hat{\theta}$ under a uniform prior defined over the region $[0.80, 1.10]$. Slicing this surface along a given value for θ

produces the sampling distribution of $\hat{\theta}$, the maximum likelihood estimator. Slicing this surface along a given value for $\hat{\theta}$ produces the posterior distribution for θ corresponding to the realized value of $\hat{\theta}$.

(a) Create the three-dimensional plot described in the exercise and carry out a “helicopter tour,” examining it from different angles.

(b) Create a two-dimensional plot that slices along $\theta = 1$ to obtain an approximate sampling distribution for the maximum likelihood estimator in the unit root case.

(c) Create a two-dimensional plot that slices along $\hat{\theta} = 1$ to get the exact posterior distribution for θ at the point $\hat{\theta} = 1$.

Solution

A Matlab program that creates the three figures requested is available on the Web site associated with this book. The three figures are produced here as Figures 18.1–18.3. With regards to (a), note that Figure 18.1 provides a view from just one angle. If you have software with the appropriate graphing capabilities (e.g., Matlab), you can produce Figure 18.1 on the computer screen and then rotate it to obtain a view of any aspect of the distribution you want. This is what Sims and Uhlig (1991) refer to as a “helicopter tour.” Note that the small upturn in the histogram for small values of $\hat{\theta}$ is caused by the graph effectively truncating the posterior at 0.795. That is, the upturn is associated with the bin $(-\infty, 0.795)$, which can receive appreciable posterior support for data sets generated with θ near 0.80. With regards to parts (b) and (c), note that the posterior distribution is symmetric and centered over $\hat{\theta} = 1$ (as we would expect from Exercise 18.1), but the sampling distribution of the MLE is skewed and downward biased.

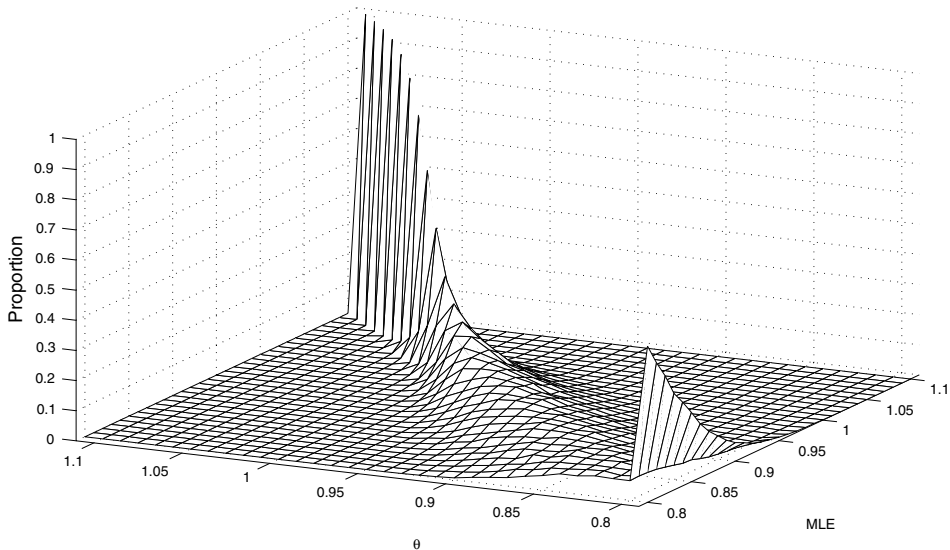
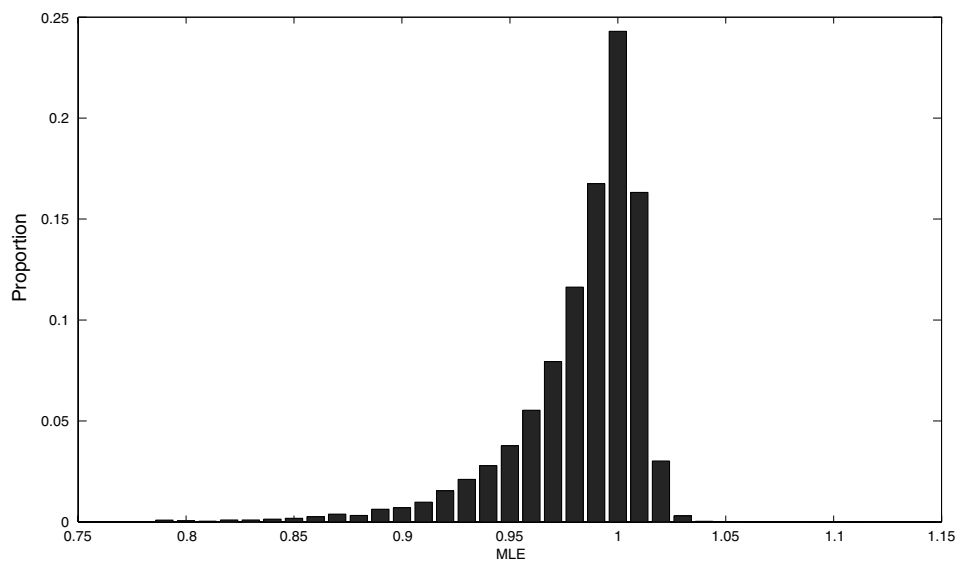
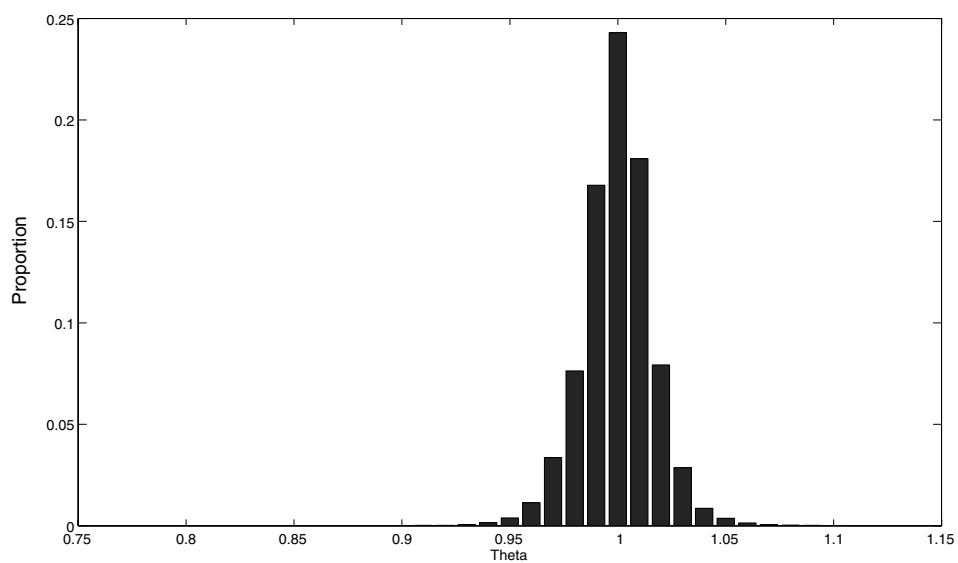


Figure 18.1 — Joint histogram involving θ and MLE.

Figure 18.2 — Sampling distribution when $\theta = 1$.Figure 18.3 — Posterior distribution when $\text{MLE} = 1$.

Exercise 18.3 (Unit roots and Jeffreys' prior) Let $y = (y_1, \dots, y_T)'$, $\epsilon_t \stackrel{\text{iid}}{\sim} N(0, \sigma^2)$ for $t = 1, \dots, T$ and assume y_0 is given and σ^2 is unknown. Consider the following AR(1) model:

$$y_t = \theta y_{t-1} + \epsilon_t. \quad (18.4)$$

Find Jeffreys' prior for θ .

Solution

Conditional on y_0 , the likelihood function is

$$\begin{aligned} L(\theta, \sigma^2) &= \prod_{t=1}^T \phi(y_t | \theta y_{t-1}, \sigma^2) \\ &= (2\pi\sigma^2)^{-\frac{T}{2}} \exp \left[-\frac{1}{2\sigma^2} \sum_{t=1}^T (y_t - \theta y_{t-1})^2 \right]. \end{aligned} \quad (18.5)$$

Taking second derivatives of the log of (18.5) yields the elements of the Hessian matrix:

$$\frac{\partial^2 \ln L(\theta, \sigma^2)}{\partial \theta^2} = -\frac{1}{\sigma^2} \sum_{t=1}^T y_{t-1}^2, \quad (18.6)$$

$$\frac{\partial^2 \ln L(\theta, \sigma^2)}{\partial (\sigma^2)^2} = \frac{T}{2\sigma^4} - \frac{1}{\sigma^6} \sum_{t=1}^T (y_t - \theta y_{t-1})^2, \quad (18.7)$$

and

$$\frac{\partial^2 \ln L(\theta, \sigma^2)}{\partial \theta \partial (\sigma^2)} = -\frac{1}{\sigma^4} \sum_{t=1}^T (y_t - \theta y_{t-1}) y_{t-1}. \quad (18.8)$$

To compute Jeffreys' prior, we must calculate the negative of the sampling expectation (conditional on y_0) of these quantities.

First consider (18.6). Noting that (18.4) can be written as

$$y_t = \theta^t y_0 + \sum_{i=0}^{t-1} \theta^i \epsilon_{t-i},$$

it follows that

$$\begin{aligned} E(y_t^2 | y_0) &= E \left(\theta^t y_0 + \sum_{i=0}^{t-1} \theta^i \epsilon_{t-i} \right)^2 \\ &= \theta^{2t} y_0^2 + \sum_{i=0}^{t-1} \theta^{2i} E(\epsilon_{t-i}^2) \\ &= \theta^{2t} y_0^2 + \sigma^2 \frac{1 - \theta^{2t}}{1 - \theta^2}, \end{aligned} \quad (18.9)$$

where the last line uses a formula for the sum of a finite geometric series. In particular, we use the result that $1 + \theta + \theta^2 + \cdots + \theta^n = (1 - \theta^{n+1})/(1 - \theta)$, which holds provided $\theta \neq 1$. In the unit root case where $\theta = 1$, the analogue of (18.9) is

$$\begin{aligned} E(y_t^2 | y_0) &= E \left(y_0 + \sum_{i=0}^{t-1} \epsilon_{t-i} \right)^2 \\ &= y_0^2 + \sum_{i=0}^{t-1} E(\epsilon_{t-i}^2) \\ &= y_0^2 + t\sigma^2. \end{aligned} \quad (18.10)$$

For the case where $\theta \neq 1$, using (18.9), we obtain the negative of the expectation (conditional on y_0) of (18.6):

$$\begin{aligned} -E \left[\frac{\partial^2 \ln L(\theta, \sigma^2)}{\partial \theta^2} | y_0 \right] &= \frac{1}{\sigma^2} \sum_{t=1}^T E(y_{t-1}^2 | y_0) \\ &= \sum_{t=1}^T \left[\theta^{2(t-1)} \left(\frac{y_0}{\sigma} \right)^2 + \frac{1 - \theta^{2(t-1)}}{1 - \theta^2} \right] \\ &= \left[\left(\frac{y_0}{\sigma} \right)^2 \left(\frac{1 - \theta^{2T}}{1 - \theta^2} \right) + \left(\frac{1}{1 - \theta^2} \right) \left(T - \frac{1 - \theta^{2T}}{1 - \theta^2} \right) \right]. \end{aligned} \quad (18.11)$$

For the case where $\theta = 1$, we can use (18.10) to find

$$\begin{aligned} -E \left[\frac{\partial^2 \ln L(\theta, \sigma^2)}{\partial \theta^2} | y_0 \right] &= \frac{1}{\sigma^2} \sum_{t=1}^T E(y_{t-1}^2 | y_0) \\ &= \frac{1}{\sigma^2} \sum_{t=1}^T [y_0^2 + (t-1)\sigma^2] \\ &= T \left[\left(\frac{y_0}{\sigma} \right)^2 + \frac{1}{2}(T-1) \right]. \end{aligned} \quad (18.12)$$

Next consider (18.7) and (18.8). Noting that $\epsilon_t = y_t - \theta y_{t-1}$, we obtain the negative of the expectation (conditional on y_0) of (18.7) and (18.8), respectively:

$$\begin{aligned} -E \left[\frac{\partial^2 \ln L(\theta, \sigma^2)}{\partial (\sigma^2)^2} | y_0 \right] &= -\frac{T}{2\sigma^4} + \frac{1}{\sigma^6} \sum_{t=1}^T E(\epsilon_t)^2 \\ &= -\frac{T}{2\sigma^4} + \frac{T}{\sigma^4} = \frac{T}{2\sigma^4} \end{aligned} \quad (18.13)$$

and

$$-E \left[\frac{\partial^2 \ln L(\theta, \sigma^2)}{\partial \theta \partial (\sigma^2)} | y_0 \right] = -\frac{1}{\sigma^4} \sum_{t=1}^T E(\epsilon_t y_{t-1}) = 0. \quad (18.14)$$

Jeffreys' prior is proportional to the square root of the determinant of the sample information matrix. In our case, for $\theta \neq 1$, Jeffreys' prior, which we denote by $p_P(\cdot)$, is proportional to the square root of the product of (18.11) and (18.13):

$$p_P(\theta, \sigma^2) \propto \left[\left(\frac{y_0}{\sigma} \right)^2 \left(\frac{1 - \theta^{2T}}{1 - \theta^2} \right) + \left(\frac{1}{1 - \theta^2} \right) \left(T - \frac{1 - \theta^{2T}}{1 - \theta^2} \right) \right]^{\frac{1}{2}} \left[\frac{T}{2\sigma^4} \right]^{\frac{1}{2}}, \quad (18.15)$$

for $\theta \neq 1$.

For $\theta = 1$, the analogous calculation involves (18.12) and (18.13) and yields

$$\begin{aligned} p_P(\theta, \sigma^2) &= \left[T \left[\left(\frac{y_0}{\sigma} \right)^2 + \frac{1}{2} (T - 1) \right] \right]^{\frac{1}{2}} \left[\frac{T}{2\sigma^4} \right]^{\frac{1}{2}} \\ &\propto \frac{T}{\sigma^2} \left[\left(\frac{y_0}{\sigma} \right)^2 + \frac{1}{2} (T - 1) \right]^{\frac{1}{2}}. \end{aligned} \quad (18.16)$$

The prior given in (18.15) and (18.16) was advocated by Phillips (1991) and the discussion that accompanied this paper reflects its controversial nature. Hence, we use the subscript P for Phillips (rather than Jeffreys') because it is not clear whether Jeffreys would have advocated this prior. One controversial aspect about this prior is that it is improper and shoots off to ∞ as $\theta \rightarrow \infty$. Note also that, with dependent data (unlike a random sample), there is a crucial difference between the information matrix for a single observation and the information matrix for the entire sample. Here, since the latter is used, sample size T and the initial value y_0 appear in Phillips' prior.

Exercise 18.4 (Berger and Yang [1994]) Berger and Yang (1994) tried deriving a reference prior [in the sense of Bernardo (1979)] for the model in Exercise 18.3 and found they could only do so when $|\theta| < 1$ but not when $|\theta| \geq 1$. In the end, they advocated use of the following prior density:

$$p_{BY}(\theta) = \begin{cases} \frac{1}{2\pi\sqrt{1-\theta^2}} & \text{if } |\theta| < 1, \\ \frac{1}{2\pi|\theta|\sqrt{\theta^2-1}} & \text{if } |\theta| \geq 1. \end{cases} \quad (18.17)$$

The prior over $|\theta| < 1$ agrees with Jeffreys' prior in the stationary case [see Thornber (1967)]. Show that $\Pr(|\theta| < 1) = \Pr(|\theta| \geq 1) = 1/2$ and hence that this prior is proper.

Solution

Using properties of the arcsin function, we obtain

$$\int_0^1 \frac{1}{\sqrt{1-\theta^2}} d\theta = \frac{\pi}{2}$$

and hence

$$\begin{aligned}
 \Pr(|\theta| < 1) &= \int_{-1}^1 \frac{1}{2\pi\sqrt{1-\theta^2}} d\theta \\
 &= 2 \int_0^1 \frac{1}{2\pi\sqrt{1-\theta^2}} d\theta \\
 &= \frac{1}{\pi} \int_0^1 \frac{1}{\sqrt{1-\theta^2}} d\theta \\
 &= \frac{1}{\pi} \frac{\pi}{2} = \frac{1}{2}.
 \end{aligned}$$

To obtain $\Pr(|\theta| \geq 1)$, define $\gamma = 1/\theta$ and consider $d\theta = -\gamma^{-2}d\gamma$. Using the change-of-variable technique, we have

$$\begin{aligned}
 \Pr(|\theta| \geq 1) &= \int_{|\theta| \geq 1} \frac{1}{2\pi|\theta|\sqrt{\theta^2-1}} d\theta \\
 &= \int_{-1}^1 \frac{1}{\gamma^2 2\pi|\gamma^{-1}|\sqrt{\gamma^{-2}-1}} d\gamma \\
 &= \int_{-1}^1 \frac{1}{2\pi\sqrt{1-\gamma^2}} d\gamma \\
 &= \frac{1}{2}.
 \end{aligned}$$

Therefore, (18.17) integrates to unity. These properties make (18.17) attractive for testing $|\theta| < 1$ versus $|\theta| \geq 1$.

Exercise 18.5 (Kadane, Chan, and Wolfson [1996]) Consider extending the conjugate normal analysis of Exercise 18.1 to allow a piecewise normal prior distribution. Specifically, we could allow one piece to be a normal distribution over the range $(-\infty, 1)$, a second normal distribution might pertain to $(1, \infty)$, and conceivably there might be a lump of prior mass at $\theta = 1$. The prior density for θ in this situation can be expressed as

$$\begin{aligned}
 p_{\text{KCW}}(\theta) &= \underline{\omega}_1 \frac{\phi(\theta|\underline{\mu}_1, \underline{\sigma}_1^2) I(\theta < 1)}{\Phi\left(\left[1 - \underline{\mu}_1\right] \underline{\sigma}_1^{-1}\right)} \\
 &\quad + \underline{\omega}_2 \frac{\phi(\theta|\underline{\mu}_2, \underline{\sigma}_2^2) I(\theta > 1)}{1 - \Phi\left(\left[1 - \underline{\mu}_2\right] \underline{\sigma}_2^{-1}\right)} + \underline{\omega}_3 I(\theta = 1),
 \end{aligned} \tag{18.18}$$

where $\Phi(\cdot)$ denotes the c.d.f. of an $N(0, 1)$ random variable and $I(\cdot)$ is the indicator function. The hyperparameters are $\underline{\omega}_i$ for $i = 1, 2, 3$ (which are constrained to be nonnegative and sum to one), $\underline{\mu}_i$, and $\underline{\sigma}_i^2$, $i = 1, 2$. Show that the marginal posterior distribution is also piecewise normal in the same form and derive expressions for the updated hyperparameters.

Note: if the error precision in our time series process is taken to be stochastic with a Gamma prior, rather than assumed to be equal to unity, nothing essential changes. The prior is piecewise normal-gamma, as is the posterior.

Solution

One way to approach this problem is to represent the prior in (18.18) hierarchically by adding a component indicator variable, much like the solution to Exercises 15.5 and 15.6. To this end, let $c \in \{1, 2, 3\}$ be a discrete random variable with $\Pr(c = j) = \underline{w}_j$, $j = 1, 2, 3$. We can thus write a conditional version of the prior in (18.18) as

$$p_{\text{KCW}}(\theta|c) = \begin{cases} \frac{\phi(\theta|\underline{\mu}_1, \underline{\sigma}_1^2)I(\theta < 1)}{\Phi([1 - \underline{\mu}_1]\underline{\sigma}_1^{-1})} & \text{if } c = 1, \\ \frac{\phi(\theta|\underline{\mu}_2, \underline{\sigma}_2^2)I(\theta > 1)}{1 - \Phi([1 - \underline{\mu}_2]\underline{\sigma}_2^{-1})} & \text{if } c = 2, \\ I(\theta = 1) & \text{if } c = 3. \end{cases}$$

Thus, marginalized over the prior $p(c)$, it follows that the unconditional prior for θ , $p_{\text{KCW}}(\theta)$, is exactly the same as in (18.18).

We are interested in obtaining $p(\theta|y)$. We know

$$p(\theta|y) = \sum_{j=1}^3 p(\theta|c = j, y)\Pr(c = j|y),$$

which shows that the posterior for θ is a mixture of the component densities $p(\theta|c = j, y)$, $j = 1, 2, 3$. For $j = 1, 2$, the posterior densities $p(\theta|c = j, y)$ are truncated normal, following similar logic to that of Exercise 15.7. Specifically,

$$p(\theta|c = 1, y) = \frac{\phi(\theta; \bar{\mu}_1, \bar{\sigma}_1^2)}{\Phi[(1 - \bar{\mu}_1)\bar{\sigma}_1^{-1}]}I(\theta < 1),$$

where

$$\bar{\sigma}_1^2 = (w + \underline{\sigma}_1^{-2})^{-1}, \quad \bar{\mu}_1 = \bar{\sigma}_1^2(w\hat{\theta} + \underline{\sigma}_1^{-2}\underline{\mu}_1), \quad \text{and } w \equiv \sum_{t=1}^T y_{t-1}^2.$$

Similarly,

$$p(\theta|c = 2, y) = \frac{\phi(\theta; \bar{\mu}_2, \bar{\sigma}_2^2)}{1 - \Phi[(1 - \bar{\mu}_2)\bar{\sigma}_2^{-1}]}I(\theta > 1),$$

where

$$\bar{\sigma}_2^2 = (w + \underline{\sigma}_2^{-2})^{-1}, \quad \bar{\mu}_2 = \bar{\sigma}_2^2(w\hat{\theta} + \underline{\sigma}_2^{-2}\underline{\mu}_2).$$

Finally, given the point prior employed within the third component,

$$p(\theta|c = 3, y) = I(\theta = 1).$$

Hence, the posterior will be in exactly the same form as the prior in (18.18), where the posterior means and standard deviations within each component are updated using the formulas given here. To finish characterizing $p(\theta|y)$, it remains to calculate the updated posterior weights $\Pr(c = j|y)$, $c = 1, 2, 3$. To this end, we note that

$$\begin{aligned} p(c|y) &\propto p(y|c)p(c) \\ &= \left[\int_{-\infty}^{\infty} p(y, \theta|c) d\theta \right] p(c) \\ &= \left[\int_{-\infty}^{\infty} p(y|\theta)p(\theta|c) d\theta \right] p(c), \end{aligned}$$

where the last line indicates that the data density does not depend on c , given θ . As shown in Exercise (18.1), the likelihood function $L(\theta)$ can be decomposed into a term involving the least squares residuals and a quadratic form in θ . Specifically, these results implied

$$p(y|\theta) \propto \exp \left(-(1/2) \sum_{t=1}^T (y_t - \hat{\theta} y_{t-1})^2 \right) \exp \left[-(1/2)(\theta - \hat{\theta})^2 w \right].$$

When working out the calculations for $p(c|y)$, the term $\exp \left(-(1/2) \sum_{t=1}^T (y_t - \hat{\theta} y_{t-1})^2 \right)$ moves outside of the integral and is absorbed into the normalizing constant. As such, we can ignore this term when performing the needed calculations. We thus have

$$p(c|y) \propto p(c) \int_{-\infty}^{\infty} \exp \left[-(1/2)(\theta - \hat{\theta})^2 w \right] p(\theta|c) d\theta.$$

This expression can be evaluated for each of the three values of c . For the case where $c = 1$, for example, we have (letting $\Psi_1 \equiv \Phi \left(\left[1 - \underline{\mu}_1 \right] \underline{\sigma}_1^{-1} \right)$ for notational simplicity)

$$\begin{aligned} \Pr(c = 1|y) &\propto \frac{\omega_1}{\Psi_1} \int_{-\infty}^{\infty} \left(\exp \left[-(1/2)(\theta - \hat{\theta})^2 w \right] \right. \\ &\quad \times \left. [1/\underline{\sigma}_1] \exp \left[-[1/2\underline{\sigma}_1^2](\theta - \underline{\mu}_1)^2 \right] I(\theta < 1) \right) d\theta \\ &\propto \frac{\omega_1}{\Psi_1 \underline{\sigma}_1} \exp \left[-(1/2)(\hat{\theta} - \underline{\mu}_1)^2 / (\underline{\sigma}_1^2 + w^{-1}) \right] \\ &\quad \times \int_{-\infty}^1 \exp \left[-[1/2\bar{\sigma}_1^2](\theta - \bar{\mu}_1)^2 \right] d\theta \\ &\propto \frac{\omega_1 \bar{\sigma}_1}{\Psi_1 \underline{\sigma}_1} \exp \left[-(1/2)(\hat{\theta} - \underline{\mu}_1)^2 / (\underline{\sigma}_1^2 + w^{-1}) \right] \Phi \left[\frac{(1 - \bar{\mu}_1)}{\bar{\sigma}_1} \right]. \end{aligned}$$

The second line in this derivation follows from completing the square on θ . By similar calculations, we obtain

$$\Pr(c = 2|y) \propto \frac{\omega_2 \bar{\sigma}_2}{\Psi_2 \underline{\sigma}_2} \exp \left[-(1/2)(\hat{\theta} - \underline{\mu}_2)^2 / (\underline{\sigma}_2^2 + w^{-1}) \right] \left(1 - \Phi \left[\frac{(1 - \bar{\mu}_2)}{\bar{\sigma}_2} \right] \right),$$

where $\Psi_2 \equiv 1 - \Phi \left(\left[1 - \underline{\mu}_2 \right] \underline{\sigma}_2^{-1} \right)$. Finally,

$$\Pr(c = 3|y) \propto \underline{\omega}_3 \exp \left(-(1/2)(\hat{\theta} - 1)^2 w \right).$$

These final expressions provide updated formulas for the (unnormalized) component weights. Normalization follows trivially upon dividing each of these quantities by the sum of the unnormalized weights. These results, combined with the derived conditionals $p(\theta|c, y)$ and the formula $p(\theta|y) = \sum_{j=1}^3 p(\theta|c = j, y) \Pr(c = j|y)$ provide the marginal posterior of interest.

Exercise 18.6 (Cointegration) Cointegration is an important topic in nonstationary time series econometrics. If cointegration is present among several unit root variables it can be shown that a vector error correction model (VECM) exists that is a restricted version of a VAR(p). That is, if y_t for $t = 1, \dots, T$ is an $M \times 1$ vector containing observations on M time series variables with unit roots and ϵ_t is an $M \times 1$ vector of errors, then the VECM is

$$\Delta y_t = \Pi y_{t-1} + \sum_{i=1}^{p-1} \Psi_i \Delta y_{t-i} + \Phi d_t + \epsilon_t, \quad (18.19)$$

where the matrix of long-run multipliers, Π , can be written as $\Pi = \alpha\beta'$, where α and β are both full-rank $M \times r$ matrices and where $0 \leq r \leq M$ is the number of cointegrating relationships. d_t is a k_d vector containing deterministic terms (e.g., an intercept, deterministic trends, dummy variables, etc.) and Φ is $M \times k_d$. Assume $\epsilon_t \stackrel{\text{iid}}{\sim} N(0, \Sigma)$.

Given the interest in cointegration, many authors have derived Bayesian methods for working with the VECM. This exercise is loosely based on Geweke (1996). However, before getting into the exercise, a warning is called for. Important statistical issues arise because Π is potentially of reduced rank. A global identification issue can be seen by noting that $\Pi = \alpha\beta'$ and $\Pi = \alpha S S^{-1} \beta'$ are identical for any nonsingular S . This indeterminacy is surmounted in Geweke (1996a) by imposing *linear normalization*, where $\beta = [I_r \quad B']'$. Even if global identification is imposed, a local identification issue occurs at the point $\alpha = 0$ (i.e., at this point β does not enter the model). As discussed in Kleibergen and van Dijk (1994, 1998), this local identification problem can cause serious problems, when certain noninformative priors are used, for posterior computation (e.g., the Gibbs sampler can have an absorbing state) and for Bayesian inference (i.e., the posterior might not integrate to one and hence is not a valid p.d.f.). The approach of Geweke (1996a) does yield a valid posterior and Gibbs sampler since it imposes the linear normalization and works with informative normal priors. However, there are some drawbacks to this linear normalization [see Strachan and Inder (2004)] and normal priors may not be attractive in some empirical contexts. Koop et al. (2006) survey the Bayesian cointegration literature and the interested reader is referred to this paper for discussion of some of the complications relating to identification and prior elicitation.

In this exercise, impose the linear normalization and, thus, work in the parameterization involving α, B, Σ^{-1} , and C , where $C = (\Psi_1 \cdots \Psi_{d-1} \Phi)$ is an $M \times k$ matrix with

$k = k_d + pM$. Also let $A = (\alpha \ C')$ and use the following priors [where $f_{MN}(\cdot)$ denotes the matrix-variate normal p.d.f.; see Appendix Definition 12]:

$$p(A|\Sigma) = f_{MN}(A|\underline{A}, \Sigma \otimes \underline{V}_A)$$

and

$$p(\Sigma^{-1}, B) = f_W(\Sigma^{-1}|\underline{H}, \underline{\nu}).$$

Note that we are using a natural conjugate prior for A and Σ^{-1} , conditional on B , and a noninformative prior for B . In the previous formulae, \underline{V}_A is a $(k+r) \times (k+r)$ positive definite matrix and \underline{V}_{B1} and \underline{V}_{B2} are $r \times r$ and $(M-r) \times (M-r)$ positive definite matrices, respectively.

Derive the conditional posteriors $p(B|y, A, \Sigma)$, $p(A|y, B, \Sigma)$, and $p(\Sigma^{-1}|y, B)$ and, thus, show how a Gibbs sampler can be set up to carry out posterior inference in the VECM.

Solution

The basic idea in developing a Gibbs sampler for the VECM with the prior given in the question is that, conditional on B , the model becomes equivalent to the unrestricted VAR of Chapter 17, Exercise 17.6(a). In Exercise 17.6(a), we used a different prior; hence, the derivations required in the present exercise are not exactly the same. Nevertheless, they are quite similar. It is only the derivation of $p(B|y, A, \Sigma)$ that is new. The latter posterior conditional can be derived using a suitable transformation of the VECM and methods for the restricted VAR of Exercise 17.6(b). Exact details are provided in the following material.

Let $z_t = \beta' y_{t-1}$, and $Z_t = (z'_t, \Delta y'_{t-1}, \dots, \Delta y'_{t-p-1}, d'_t)'$, the VECM can be written as

$$\Delta y_t = A' Z_t + \epsilon_t.$$

In matrix form we can then write

$$\Delta Y = ZA + E,$$

where ΔY and E are $T \times M$ matrices with t th rows given by $\Delta y'_t$ and ϵ'_t , respectively. Z is a $T \times (k+r)$ matrix with t th rows given by Z'_t . E has a matrix-variate normal distribution:

$$E \sim MN(0, \Sigma \otimes I_T).$$

Thus, conditional upon β (or, equivalently, B), the VECM has the same form as the unrestricted VAR (or a multivariate regression model). Hence, the proof is virtually identical to that provided in Exercise 17.6(a) and we will not provide all the details (see also Chapter 2, Exercise 2.14). Briefly, combining the matrix-variate normal likelihood with the normal-Wishart prior (and treating β as fixed), we obtain the posterior. Writing the posterior in terms of OLS quantities and rearranging yields the normal-Wishart posterior (conditional on β):

$$A|y, \beta, \Sigma \sim MN(\text{vec}[\bar{A}], \Sigma \otimes \bar{V}_A)$$

and

$$\Sigma^{-1}|y, \beta \sim W\left(\bar{S}^{-1}, \underline{\nu} + T\right).$$

In the previous formulae we use the notation

$$\begin{aligned}\bar{V}_A &= (\underline{V}_A^{-1} + Z'Z)^{-1}, \\ \bar{A} &= \bar{V}_A \left(\underline{V}_A^{-1} \underline{A} + Z'Z\hat{A} \right), \\ \hat{A} &= (Z'Z)^{-1} Z' \Delta Y,\end{aligned}$$

and

$$\bar{S} = \underline{H}^{-1} + (\Delta Y - Z\bar{A})' \left[I_T - Z \left(\bar{V}_A^{-1} + Z'Z \right)^{-1} Z' \right] (\Delta Y - Z\bar{A}).$$

Thus, $p(A|y, B, \Sigma)$ follows a matrix-variate normal distribution with mean matrix \bar{A} and covariance matrix $\Sigma \otimes \bar{V}_A$ and $p(\Sigma^{-1}|y, B)$ is Wishart.

It remains to derive $p(B|y, A, \Sigma)$. To achieve this end, it proves useful to write the VECM in another way. Begin by rearranging (18.19) as

$$\Delta y_t - \sum_{i=1}^{p-1} \Psi_i \Delta y_{t-i} - \Phi d_t = \alpha \beta' y_{t-1} + \epsilon_t,$$

and denoting the left-hand side of this equation as x_t^* we obtain

$$x_t^* = \alpha \beta' y_{t-1} + \epsilon_t.$$

Stacking (in the same fashion as earlier in the solution to this exercise), we can write this model as

$$X^* = Y_{-1} \beta \alpha' + E, \quad (18.20)$$

where X^* and Y_{-1} are $T \times M$ matrices with t th rows give by x_t^* and y'_{t-1} , respectively. E is as defined previously in this solution.

Intuitively, if α were a square nonsingular matrix we could postmultiply both sides of this equation by α'^{-1} and the resulting transformed model would simply be a multivariate regression and standard methods could be used (since we are conditioning on A , x_t^* can be treated as known). This, of course, will not work since α is not a square matrix. However, following Geweke (1996a), we can do something similar. First construct the $M \times M$ matrix (which a standard theorem in matrix algebra tells us is nonsingular):

$$C = [\alpha^+, \alpha^0],$$

where α^+ (an $M \times r$ matrix) is the Moore–Penrose generalized inverse of α' (i.e., $\alpha' \alpha^+ = I_r$) and α^0 [an $M \times (M - r)$ matrix] is orthogonal to α^+ . If we postmultiply (18.20) by C we obtain

$$[X^* \alpha^+, X^* \alpha^0] = [Y_{-1} \beta, 0] + [E \alpha^+, E \alpha^0]. \quad (18.21)$$

There is one final issue that needs to be addressed: We are interested in posterior inference on the identified parameter B , rather than β (recall that $\beta = [I_r \ B']'$). But this can be dealt with by imposing the identification restriction and pulling over these elements to the right-hand side of (18.21). We do this by defining the partition

$$Y_{-1} = [Y_{-1}^1, Y_{-1}^2],$$

where Y_{-1}^1 is $T \times r$ and Y_{-1}^2 is $T \times (M - r)$. Write (18.21) as

$$\begin{bmatrix} \tilde{X}_1 \\ \tilde{X}_2 \end{bmatrix} = [Y_{-1}^2 B, 0] + \begin{bmatrix} \tilde{E}_1 \\ \tilde{E}_2 \end{bmatrix}, \quad (18.22)$$

where $\tilde{X}_1 = X^* \alpha^+ - Y_{-1}^1$, $\tilde{X}_2 = X^* \alpha^0$, $\tilde{E}_1 = E \alpha^+$, and $\tilde{E}_2 = E \alpha^0$.

Now note that (18.22) is a multivariate normal regression model, except that restrictions are placed on some of the coefficients (i.e., the final r columns of the matrix of explanatory variables are zeros). But a multivariate normal regression model with restrictions on the coefficients is a restricted VAR (i.e., a SUR) and Exercise 17.6(b) shows us how to work with these. Additional details are provided in the following.

To see the relationship with the restricted VAR most clearly, it is helpful to stack all the columns in (18.22) to put the model in the same form as the solution to Exercise 17.6(b). Letting $\tilde{x}_j = \text{vec}(\tilde{X}_j)$ and $\tilde{\epsilon}_j = \text{vec}(\tilde{E}_j)$ for $j = 1, 2$ and $b = \text{vec}(B)$, we can write (18.22) as

$$\begin{bmatrix} \tilde{x}_1 \\ \tilde{x}_2 \end{bmatrix} = \begin{bmatrix} I_r \otimes Y_{-1}^2 \\ 0 \end{bmatrix} b + \begin{bmatrix} \tilde{\epsilon}_1 \\ \tilde{\epsilon}_2 \end{bmatrix}$$

or

$$\tilde{x} = \tilde{Z}b + \tilde{\epsilon}, \quad (18.23)$$

where $\tilde{\epsilon}$ is $N(0, \tilde{\Sigma})$ with $\tilde{\Sigma} = C' \Sigma C \otimes I_T$. In this form it can be explicitly seen that this model has exactly the same format as the restricted VAR of Exercise 17.6(b) (see also Chapter 2, Exercise 2.14). Accordingly, we will not repeat the steps necessary to derive the posterior for b (conditional on A and Σ). Using the noninformative limiting case of either of those previous derivations, we find

$$b|y, A, \Sigma \sim N(\bar{b}, \bar{V}_b),$$

where

$$\bar{V}_b = [\tilde{Z}' (\tilde{\Sigma} \otimes I_T) \tilde{Z}]^{-1},$$

which, given the particular structure of \tilde{Z} , reduces to

$$\bar{V}_b = (\tilde{\Sigma}^{11})^{-1} \otimes (Y_{-1}^{2'} Y_{-1}^2)^{-1},$$

where $\tilde{\Sigma}^{11}$ is the $r \times r$ upper left-hand block of $\tilde{\Sigma}^{-1}$. Furthermore,

$$\bar{b} = \bar{V}_b \tilde{Z}' (\tilde{\Sigma} \otimes I_T) \tilde{x}.$$

Thus, Bayesian inference in the VECM with this prior using the linear normalization can be performed using a Gibbs sampler involving only standard distributions (the normal and the Wishart). It is a minor extension to add an informative prior for B and, if this is done, marginal likelihoods using Gibbs sampler output can be obtained in the standard ways (see Chapter 16, Exercises 16.6–16.8). If marginal likelihoods are calculated for every possible value of r then the researcher can select a preferable cointegrating rank (i.e., test for cointegration) or average over different cointegrating ranks in a Bayesian model averaging exercise. We remind the reader that Koop et al. (2006) survey the Bayesian cointegration literature and discuss some other (more satisfactory) approaches to prior elicitation and identification.

Appendix

In this section, the p.d.f.'s used in the book are defined and some of their properties are described. Except for the gamma p.d.f. (which we parameterize in a different way), we follow the notational conventions described in Abadir and Magnus (2002). There are, of course, many other p.d.f.'s that arise in other models. Poirier (1995, chapter 3) discusses many of these. Probably the most complete source of information on probability distribution is the *Distributions in Statistics* series; see Johnson, Kotz, and Balakrishnan (1994, 1995, 2000) and Johnston, Kotz, and Kemp (1993).

Following standard practice, we introduce two sorts of notation. The first is for the p.d.f. or probability function itself. The second is the symbol “ \sim ,” meaning “is distributed as.”

Definition 1: The uniform distribution

A continuous random variable Y has a *uniform distribution* over the interval $[a, b]$, denoted $Y \sim U(a, b)$, if its p.d.f. is given by

$$f_U(y|a, b) = \begin{cases} \frac{1}{b-a} & \text{if } a \leq y \leq b, \\ 0 & \text{otherwise,} \end{cases}$$

where $-\infty < a < b < \infty$.

Theorem 1: Mean and variance of the uniform distribution

If $Y \sim U(a, b)$ then $E(Y) = (a + b)/2$ and $\text{var}(Y) = (b - a)^2/12$.

Definition 2: The gamma and related distributions

A continuous random variable Y has a *gamma* distribution with mean $\mu > 0$ and degrees of freedom $\nu > 0$, denoted by $Y \sim \gamma(\mu, \nu)$, if its p.d.f. is

$$f_\gamma(y|\mu, \nu) \equiv \begin{cases} c_\gamma^{-1} y^{\frac{\nu-2}{2}} \exp\left(-\frac{y\nu}{2\mu}\right) & \text{if } 0 < y < \infty, \\ 0 & \text{otherwise,} \end{cases}$$

where the integrating constant is given by $c_\gamma = (2\mu/\nu)^{\frac{\nu}{2}} \Gamma(\frac{\nu}{2})$ and $\Gamma(a)$ is the gamma function (see Poirier, 1995, p. 98). It is also common to parameterize the gamma in terms of $\alpha = \nu/2$ and $\beta = 2\mu/\nu$, in which case we denote the distribution as $Y \sim G(\alpha, \beta)$. The associated density function under this parameterization is denoted by $f_G(y|\alpha, \beta)$, where

$$f_G(y|\alpha, \beta) \equiv \begin{cases} c_G^{-1} y^{\alpha-1} \exp(-y/\beta) & \text{if } 0 < y < \infty, \\ 0 & \text{otherwise,} \end{cases}$$

and $c_G = \beta^\alpha \Gamma(\alpha)$.

Theorem 2: Mean and variance of the gamma distribution

If $Y \sim G(\alpha, \beta)$ then $E(Y) = \alpha\beta$ and $\text{Var}(Y) = \alpha\beta^2$. If $Y \sim \gamma(\mu, \nu)$, then $E(Y) = \mu$ and $\text{Var}(Y) = 2\mu^2/\nu$.

Notes: Distributions related to the gamma include the *chi-square distribution*, which is a gamma distribution with $\nu = \mu$. It is denoted by $Y \sim \chi^2(\nu)$. The *exponential distribution* is a gamma distribution with $\nu = 2$.

The *inverted gamma distribution* has the property that, if Y has an inverted gamma distribution, then $1/Y$ has a gamma distribution. In this book we denote the inverted gamma density as $Y \sim IG(\alpha, \beta)$. Though different parameterizations exist (particularly for how β enters the density), we utilize the following form in this text:

$$Y \sim IG(\alpha, \beta) \Rightarrow p(y) = [\Gamma(\alpha)\beta^\alpha]^{-1} y^{-(\alpha+1)} \exp(-1/[y\beta]).$$

The mean of this inverted gamma is $E(Y) = [\beta(\alpha - 1)]^{-1}$, for $\alpha > 1$, and the variance is $\text{Var}(Y) = [\beta^2(\alpha - 1)^2(\alpha - 2)]^{-1}$ for $\alpha > 2$. Further properties of the inverted gamma are given, for example, on page 111 of Poirier (1995).

Definition 3: The multivariate normal distribution

A continuous k -dimensional random vector $Y = (Y_1, \dots, Y_k)'$ has a *normal distribution* with mean μ (a k vector) and variance Σ (a $k \times k$ positive definite matrix), denoted $Y \sim N(\mu, \Sigma)$, if its p.d.f. is given by

$$\phi(y|\mu, \Sigma) = \phi(y; \mu, \Sigma) = \frac{1}{2\pi^{\frac{k}{2}}} |\Sigma|^{-\frac{1}{2}} \exp \left[-\frac{1}{2} (y - \mu)' \Sigma^{-1} (y - \mu) \right].$$

In some cases throughout the book we add a subscript to this notation, such as $\phi_k(y|\mu, \Sigma)$, to remind the reader that y is a k -dimensional vector. In other instances, we make use of the notation $\phi(y; \mu, \Sigma)$ (typically suppressing the dimensionality subscript) to explicitly denote that μ and Σ are the mean and covariance matrix of y rather than simply arbitrary parameters of the density.

The cumulative distribution function of the multivariate normal, evaluated at the point y^* , is denoted by $\Phi(y^*|\mu, \Sigma)$ or, if $\mu = 0, \Sigma = I$, by $\Phi(y^*)$.

Note: The special case where $k = 1, \mu = 0$, and $\Sigma = 1$ is referred to as the *standard normal distribution*.

Theorem 3: Marginals and conditionals of the multivariate normal distribution

Suppose the k vector $Y \sim N(\mu, \Sigma)$ is partitioned as

$$Y = \begin{pmatrix} Y_{(1)} \\ Y_{(2)} \end{pmatrix},$$

where $Y_{(i)}$ is a k_i vector for $i = 1, 2$ with $k_1 + k_2 = k$ and μ and Σ have been partitioned conformably as

$$\mu = \begin{pmatrix} \mu_{(1)} \\ \mu_{(2)} \end{pmatrix}$$

and

$$\Sigma = \begin{pmatrix} \Sigma_{(11)} & \Sigma_{(12)} \\ \Sigma'_{(12)} & \Sigma_{(22)} \end{pmatrix}.$$

Then the following results hold:

- The marginal distribution of $Y_{(i)}$ is $N(\mu_{(i)}, \Sigma_{(ii)})$ for $i = 1, 2$.
- The conditional distribution of $Y_{(1)}$ given $Y_{(2)} = y_{(2)}$ is $N(\mu_{(1|2)}, \Sigma_{(1|2)})$, where

$$\mu_{(1|2)} = \mu_{(1)} + \Sigma_{(12)}\Sigma_{(22)}^{-1}(y_{(2)} - \mu_{(2)})$$

and

$$\Sigma_{(1|2)} = \Sigma_{(11)} - \Sigma_{(12)}\Sigma_{(22)}^{-1}\Sigma'_{(12)}.$$

The conditional distribution of $Y_{(2)}$ given $Y_{(1)} = y_{(1)}$ can be obtained by reversing subscripts 1 and 2 in the previous formulas.

Theorem 4: Linear combinations of normals are normal

Let $Y \sim N(\mu, \Sigma)$ be a k -dimensional random vector and A be a fixed (nonrandom) $m \times k$ matrix with $\text{rank}(A) = m$; then $AY \sim N(A\mu, A\Sigma A')$.

Definition 4: The multivariate t -distribution

A continuous k -dimensional random vector $Y = (Y_1, \dots, Y_k)'$ has a t -distribution with mean μ (a k -vector), scale matrix Σ (a $k \times k$ positive definite matrix), and ν (a positive scalar referred to as a *degrees-of-freedom* parameter), denoted $Y \sim t(\mu, \Sigma, \nu)$, if its p.d.f. is given by

$$f_t(y|\mu, \Sigma, \nu) = \frac{1}{c_t} |\Sigma|^{-\frac{1}{2}} [\nu + (y - \mu)' \Sigma^{-1} (y - \mu)]^{-\frac{\nu+k}{2}},$$

where

$$c_t = \frac{\pi^{\frac{k}{2}} \Gamma\left(\frac{\nu}{2}\right)}{\nu^{\frac{\nu}{2}} \Gamma\left(\frac{\nu+k}{2}\right)}.$$

Notes: The special case where $k = 1$, $\mu = 0$, and $\Sigma = 1$ is referred to as the *Student t -distribution* with ν degrees of freedom. Tables providing percentiles of the Student t are

available in most econometrics and statistics textbooks. The case where $\nu = 1$ is referred to as the *Cauchy distribution*.

Theorem 5: Mean and variance of the t -distribution

If $Y \sim t(\mu, \Sigma, \nu)$ then $E(Y) = \mu$ if $\nu > 1$ and $\text{Var}(Y) = \frac{\nu}{\nu-2}\Sigma$ if $\nu > 2$.

Note that the mean and variance only exist if $\nu > 1$ and $\nu > 2$, respectively. This implies, for instance, that the mean of the Cauchy does not exist even though it is a valid p.d.f. and hence its median and other quantiles exist. The matrix Σ is not exactly the same as the variance matrix and hence is given another name: the *scale matrix*.

Theorem 6: Marginals and conditionals of the multivariate t -distribution

Suppose the k vector $Y \sim t(\mu, \Sigma, \nu)$ is partitioned as in Theorem B.8 as are μ and Σ . Then the following results hold:

- The marginal distribution of $Y_{(i)}$ is $t(\mu_{(i)}, \Sigma_{(ii)}, \nu)$ for $i = 1, 2$.
- The conditional distribution of $Y_{(1)}$ given $Y_{(2)} = y_{(2)}$ is $t(\mu_{(1|2)}, \Sigma_{(1|2)}, \nu + k_1)$, where

$$\mu_{(1|2)} = \mu_{(1)} + \Sigma_{(12)}\Sigma_{(22)}^{-1}(y_{(2)} - \mu_{(2)}),$$

$$\Sigma_{(1|2)} = h_{(1|2)} \left[\Sigma_{(11)} - \Sigma_{(12)}\Sigma_{(22)}^{-1}\Sigma'_{(12)} \right],$$

and

$$h_{(1|2)} = \frac{1}{\nu + k_2} \left[\nu + (y_{(2)} - \mu_{(2)})' \Sigma_{(22)}^{-1} (y_{(2)} - \mu_{(2)}) \right].$$

The conditional distribution of $Y_{(2)}$ given $Y_{(1)} = y_{(1)}$ can be obtained by reversing subscripts 1 and 2 in the previous formulas.

Theorem 7: Linear combinations of t 's are t

Let $Y \sim t(\mu, \Sigma, \nu)$ be a k -dimensional random vector and A be a fixed (nonrandom) $m \times k$ matrix with rank equal to m ; then $AY \sim t(A\mu, A\Sigma A', \nu)$.

Definition 5: The normal-gamma distribution

Let Y be a k -dimensional random vector and h a scalar random variable. If the conditional distribution of Y given h is normal and the marginal distribution for h is gamma then (Y, h) is said to have a *normal-gamma distribution*. Formally, if $Y|h, \mu, \Sigma \sim N(\mu, h^{-1}\Sigma)$, and $h|m, \nu \sim \gamma(m, \nu)$ then $\theta = (Y', h)'$ has a normal-gamma distribution denoted $\theta \sim NG(\mu, \Sigma, m, \nu)$. The corresponding p.d.f. is denoted $f_{NG}(\theta|\mu, \Sigma, m, \nu)$.

Theorem 8: Marginal distributions involving the normal-gamma

If $\theta = (Y', h)' \sim NG(\mu, \Sigma, m, \nu)$, then the marginal for Y is given by $Y \sim t(\mu, m^{-1}\Sigma, \nu)$. By definition the marginal for h is given by $h \sim \gamma(m, \nu)$.

Definition 6: The Wishart distribution

Let H be an $N \times N$ positive definite (symmetric) random matrix, A be a fixed (nonrandom) $N \times N$ positive definite matrix, and $\nu > 0$ be a scalar degrees-of-freedom parameter. Then H has a Wishart distribution, denoted $H \sim W(A, \nu)$, if its p.d.f. is given by

$$f_W(H|A, \nu) = \frac{1}{c_W} |H|^{\frac{\nu-N-1}{2}} |A|^{-\frac{\nu}{2}} \exp \left[-\frac{1}{2} \text{tr}(A^{-1}H) \right],$$

where

$$c_W = 2^{\frac{\nu N}{2}} \pi^{\frac{N(N-1)}{4}} \prod_{i=1}^N \Gamma \left(\frac{\nu + 1 - i}{2} \right).$$

If $N = 1$, then the Wishart reduces to a gamma distribution [i.e., $f_W(H|A, \nu) = f_G(H|\nu A, \nu)$ if $N = 1$].

Theorem 9: Means, variances, and covariances relating to the Wishart distribution

If $H \sim W(A, \nu)$ then $E(H_{ij}) = \nu A_{ij}$, $\text{Var}(H_{ij}) = \nu (A_{ij}^2 + A_{ii}A_{jj})$ for $i, j = 1, \dots, N$ and $\text{Cov}(H_{ij}, H_{km}) = \nu (A_{ik}A_{jm} + A_{im}A_{jk})$ for $i, j, k, m = 1, \dots, N$, where subscripts i, j, k, m refer to elements of matrices.

Definition 7: The binomial distribution

A discrete random variable Y has a *binomial distribution* with parameters T and p , denoted $Y \sim B(T, p)$, if its probability function is given by

$$f_B(y|T, p) = \begin{cases} \frac{T!}{(T-y)!y!} p^y (1-p)^{T-y} & \text{if } y = 0, 1, \dots, T, \\ 0 & \text{otherwise,} \end{cases}$$

where $0 \leq p \leq 1$ and T is a positive integer. The *Bernoulli* distribution is a special case of the binomial when $T = 1$.

Theorem 10: Mean and variance of the binomial distribution

If $Y \sim B(T, p)$ then $E(Y) = Tp$ and $\text{Var}(Y) = Tp(1-p)$.

Note: This distribution is used in cases where an experiment, the outcome of which is either "success" or "failure," is repeated independently T times. The probability of success in an experiment is p . The distribution of the random variable Y , which counts the number of successes, is $B(T, p)$.

Definition 8: The Poisson distribution

A discrete random variable Y has a *Poisson distribution* with parameter λ , denoted $Y \sim Po(\lambda)$, if its probability function is given by

$$f_{Po}(y|\lambda) = \begin{cases} \frac{\lambda^y \exp(-\lambda)}{y!} & \text{if } y = 0, 1, 2, \dots, \\ 0 & \text{otherwise,} \end{cases}$$

where λ a positive real number.

Theorem 11: Mean and variance of the Poisson distribution

If $Y \sim Po(\lambda)$ then $E(Y) = \lambda$ and $Var(Y) = \lambda$.

Definition 9: The multinomial distribution

A discrete N -dimensional random vector $Y = (Y_1, \dots, Y_N)'$ has a *multinomial distribution* with parameters T and p , denoted $Y \sim M(T, p)$, if its probability function is given by

$$f_M(y|T, p) = \begin{cases} \frac{T!}{y_1! \dots y_N!} p_1^{y_1} \dots p_N^{y_N} & \text{if } y_i = 0, 1, \dots, T \text{ and } \sum_{i=1}^N y_i = T \\ 0 & \text{otherwise,} \end{cases}$$

where $p = (p_1, \dots, p_N)'$, $0 \leq p_i \leq 1$ for $i = 1, \dots, N$, $\sum_{i=1}^N p_i = 1$, and T is a positive integer.

Definition 10: The Dirichlet and beta distribution

Let $Y = (Y_1, \dots, Y_N)'$ be a vector of continuous non-negative random variables with the property that $Y_1 + \dots + Y_N = 1$. Then Y has a *Dirichlet distribution*, denoted $Y \sim D(\alpha)$, if its p.d.f. is given by

$$f_D(Y|\alpha) = \left[\frac{\Gamma(a)}{\prod_{i=1}^N \Gamma(\alpha_i)} \right] \prod_{i=1}^N y_i^{\alpha_i - 1},$$

where $\alpha = (\alpha_1, \dots, \alpha_N)'$, $\alpha_i > 0$ for $i = 1, \dots, N$ and $a = \sum_{i=1}^N \alpha_i$. The *beta distribution*, denoted by $Y \sim \beta(\alpha_1, \alpha_2)$, is the Dirichlet distribution for the case $N = 2$. Its p.d.f. is denoted by $f_B(Y|\alpha_1, \alpha_2)$.

Note that in the case $N = 2$, the restriction $Y_1 + Y_2 = 1$ can be used to remove one of the random variables. Thus, the beta distribution is a univariate distribution.

Theorem 12: Means and variances of the Dirichlet distribution

Suppose $Y \sim D(\alpha)$, where α and a are as given in Definition 10; then for $i, j = 1, \dots, N$,

- $E(Y_i) = \frac{\alpha_i}{a}$,
- $Var(Y_i) = \frac{\alpha_i(a - \alpha_i)}{a^2(a + 1)}$ and
- $Cov(Y_i, Y_j) = -\frac{\alpha_i \alpha_j}{a^2(a + 1)}$.

Definition 11: The Pareto distribution

A continuous random variable Y has a *Pareto distribution* if its p.d.f. is given by

$$f_{Pa}(y|\gamma, \lambda) = \begin{cases} \frac{\lambda \gamma^\lambda}{y^{\lambda+1}} & \text{if } y \geq \gamma, \\ 0 & \text{otherwise.} \end{cases}$$

Definition 12: The matric-variate normal distribution

A $p \times q$ matrix Y of random variables is said to have a matric-variate normal distribution, denoted $Y \sim MN(\text{vec}[M], Q \otimes P)$ with parameters M (a $p \times q$ matrix), P (a positive definite $p \times p$ matrix), and Q (a positive definite $q \times q$ matrix), if and only if $\text{vec}(Y) \sim N(\text{vec}[M], Q \otimes P)$. Thus, its density is given by

$$f_{MN}(Y|M, Q \otimes P) = [2\pi^{pq} |P|^q |Q|^p]^{-\frac{1}{2}} \exp \left\{ -\frac{1}{2} \text{tr} [Q^{-1} (Y - M)' P^{-1} (Y - M)] \right\}.$$

Note that $\text{vec}(\cdot)$ takes the columns of a matrix and stacks them in order into a vector.

Bibliography

- Abadir, K. M., R. D. H. Heijmans, and J. R. Magnus (2006). *Statistics, Econometric Exercises Series, Volume 2*, Cambridge University Press, New York.
- Abadir, K. M. and J. R. Magnus (2002). Notation in econometrics: A proposal for a standard, *Econometrics Journal*, 5, 76–90.
- Achcar, J. A. and A. F. M. Smith (1990). Aspects of reparameterization in approximate Bayesian inference, in *Bayesian and Likelihood Methods in Statistics and Econometrics* (eds. S. Geisser, J. S. Hodges, S. J. Press, and A. Zellner), North-Holland, Amsterdam, 439–452.
- Albert, J. and S. Chib (1993). Bayesian analysis of binary and polychotomous response data, *Journal of the American Statistical Association*, 88, 669–679.
- Amemiya, T. (1985). *Advanced Econometrics*, Harvard University Press, Cambridge.
- Andrews, D. F. and C. L. Mallows (1974). Scale mixtures of normal distributions, *Journal of the Royal Statistical Society, Series B*, 36, 99–102.
- Azzalini, A. and A. Dalla Valle (1996). The multivariate skew-normal distribution, *Biometrika* 83, 715–726.
- Bauwens, L., M. Lubrano, and J.-F. Richard (1999). *Bayesian Inference in Dynamic Econometric Models*, Oxford University Press, Oxford.
- Berger, J. (1985). *Statistical Decision Theory and Bayesian Analysis*, 2nd edition, Springer-Verlag, New York.
- Berger, J. O. and J. M. Bernardo (1989). Estimating a product of means: Bayesian analysis with reference priors, *Journal of the American Statistical Association*, 84, 200–207.
- Berger, J. O. and J. M. Bernardo (1992). On the development of reference priors (with discussion), in *Bayesian Statistics 4* (eds. J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith), Oxford University Press, Oxford, 35–60.
- Berger, J. O. and L. R. Pericchi (1996). The intrinsic Bayes factor for model selection and prediction, *Journal of the American Statistical Association*, 91, 109–122.

- Berger, J. and R. Wolpert (1988). *The Likelihood Principle*, 2nd edition, Institute of Mathematical Statistics, Hayward, CA.
- Berger, J. and R.-Y. Yang (1994). Noninformative priors and Bayesian testing for the AR(1) model, *Econometric Theory*, 10, 461–482.
- Bernardo, J. M. (1979). Reference prior distributions for Bayesian inference (with discussion), *Journal of the Royal Statistical Society, Series B*, 41, 113–147.
- Bernardo, J. M. (2004). Reference analysis, Universitat de València, unpublished manuscript.
- Bernardo, J. and A. F. M. Smith (1994). *Bayesian Theory*, Wiley, Chichester.
- Branco, M. D. and D. K. Dey (2002). Regression model under skew elliptical error distribution, *Journal of Mathematical Sciences* 1, 151–169.
- Carlin, B. P., A. E. Gelfand, and A. F. M. Smith (1992). Hierarchical Bayesian analysis of changepoint problems, *Applied Statistics*, 41(2), 389–405.
- Carlin, B. P. and T. A. Louis (2000). *Bayes and Empirical Bayes Methods for Data Analysis*, Chapman & Hall / CRC Press, Boca Raton.
- Carlin, B. P. and N. G. Polson (1991). Inference for nonconjugate Bayesian models using the Gibbs sampler, *The Canadian Journal of Statistics*, 19, 399–405.
- Casella, G. and R. L. Berger (2002). *Statistical Inference*, 2nd edition, Duxbury, Pacific Grove, CA.
- Casella, G. and E. George (1992). Explaining the Gibbs sampler, *The American Statistician*, 46, 167–174.
- Casella, G., J. T. G. Hwang, and C. Robert (1993). A paradox in decision-theoretic interval estimation, *Statistica Sinica*, 3, 141–155.
- Celeux, G., M. Hurn, and C. P. Robert (2000). Computational and inferential difficulties with mixture posterior distributions, *Journal of the American Statistical Association*, 95, 957–970.
- Chib, S. (1992). Bayes regression for the tobit censored regression model, *Journal of Econometrics*, 51, 79–99.
- Chib, S. (1995). Marginal likelihood from the Gibbs sampler, *Journal of the American Statistical Association*, 90, 1313–1321.
- Chib, S. (1998). Estimation and comparison of multiple change-point models, *Journal of Econometrics*, 86, 221–241.
- Chib, S. (2001). Markov chain Monte Carlo methods: Computation and inference, in *Handbook of Econometrics Volume 5* (eds. J. J. Heckman and E. Leamer), Elsevier, Amsterdam, 3569–3649.
- Chib, S. and B. Carlin (1999). On MCMC sampling in hierarchical longitudinal models, *Statistics and Computing*, 9, 17–26.
- Chib, S. and E. Greenberg (1995). Understanding the Metropolis–Hastings algorithm, *The American Statistician*, 49, 327–335.
- Chib, S. and I. Jeliazkov (2006). Inference in semiparametric dynamic models for binary longitudinal data, *Journal of the American Statistical Association*, 101, 685–700.

- Cowles, M. (1996). Accelerating Monte Carlo Markov chain convergence for cumulative-link generalized linear models, *Statistics and Computing*, 6, 101–111.
- Cowles, M. and B. Carlin (1996). Markov chain Monte Carlo convergence diagnostics: A comparative review, *Journal of the American Statistical Association*, 91, 883–904.
- Cox, D. R. and N. Reid (1987). Parameter orthogonality and approximate conditional inference, *Journal of the Royal Statistical Society, Series B*, 49, 1–39.
- Cragg, J. G. (1971). Some statistical models for limited dependent variables with application to the demand for durable goods, *Econometrica*, 39, 829–844.
- de Finetti, B. (1937). Foresight: Its logical laws, its subjective sources, translated from French in *Studies in Subjective Probability* (eds. H. E. Kyburg Jr. and H. E. Smokler), Wiley, New York, 1964.
- Deb, P., M. K. Munkin and P. K. Trivedi (2006). Bayesian analysis of the two-part model with endogeneity: application to health care expenditure, *Journal of Applied Econometrics*, forthcoming.
- Denison, D. G. T., C. C. Holmes, B. K. Mallick, and A. F. M. Smith (2002). *Bayesian Methods for Nonlinear Classification and Regression*, Wiley, New York.
- DiNardo, J. and J. L. Tobias (2001). Nonparametric density and regression estimation, *Journal of Economic Perspectives*, 15, 11–28.
- Dunn, P. K. (1999). A simple data set for demonstrating common distributions, *Journal of Statistics Education*, 7, 3. See http://www.amstat.org/publications/jse/jse/jse_archive.html.
- Efron, B. (1973). Discussion, *Journal of the Royal Statistical Society, Series B*, 35, 211.
- Fair, R. (1978). A theory of extramarital affairs, *Journal of Political Economy*, 86, 45–61.
- Fernandez, C., E. Ley, and M. Steel (2001a). Model uncertainty in cross-country growth regressions, *Journal of Applied Econometrics*, 16, 563–576.
- Fernandez, C., E. Ley, and M. Steel (2001b). Benchmark priors for Bayesian model averaging, *Journal of Econometrics*, 100, 381–427.
- Gamerman, D. (1997). *Markov chain Monte Carlo stochastic simulation for Bayesian inference*, Chapman & Hall, London.
- Garthwaite, P. H., J. B. Kadane, and A. O'Hagan (2005). Statistical methods for eliciting probability distributions, *Journal of the American Statistical Association*, 100, 680–700.
- Geisser, S. (1984). On prior distributions for binary trials (with discussion), *American Statistician* 38, 244–251.
- Gelfand, A. E. and D. Dey (1994). Bayesian model choice: Asymptotics and exact calculations, *Journal of the Royal Statistical Society, Series B*, 56, 501–514.
- Gelfand, A. E., S. Hills, A. Racine-Poon, and A. F. M. Smith (1990). Illustration of Bayesian inference in normal data models using Gibbs sampling, *Journal of the American Statistical Association*, 85, 972–985.
- Gelfand, A. E., S. K. Sahu, and B. P. Carlin (1995). Efficient parameterizations for normal linear mixed models, *Biometrika*, 82, 479–488.

- Gelman, A., J. B. Carlin, H. Stern, and D. Rubin (1995). *Bayesian Data Analysis*, Chapman & Hall, London, 2nd ed. (2004).
- George, E. and R. McCulloch (1993). Variable selection via Gibbs sampling, *Journal of the American Statistical Association*, 85, 398–409.
- George, E. and R. McCulloch (1997). Approaches for Bayesian variable selection, *Statistica Sinica*, 7, 339–373.
- Geweke, J. (1988a). Exact inference in models with autoregressive conditional heteroskedasticity, in *Dynamic Econometric Modeling* (eds. E. Berndt, H. White, and W. Barnett), Cambridge University Press, Cambridge.
- Geweke, J. (1988b). The secular and cyclical behavior of real GDP in 19 OECD countries, 1957–1983, *Journal of Business and Economic Statistics*, 6, 479–486.
- Geweke, J. (1991). Efficient simulation from the multivariate normal and Student-t distributions subject to linear constraints, in *Computer Science and Statistics: Proceedings of the Twenty-Third Symposium on the Interface* (ed. E. Keramidas), Interface Foundation of North America, Inc., Fairfax, 571–578.
- Geweke, J. (1993). Bayesian treatment of the independent Student-t linear model, *Journal of Applied Econometrics*, 8, S19–S40.
- Geweke, J. (1996a). Bayesian reduced rank regression in econometrics, *Journal of Econometrics*, 75, 121–146.
- Geweke, J. (1996b). Bayesian inference for linear models subject to linear inequality constraints, in *Modeling and Prediction: Honoring Seymour Geisser* (eds. W. O. Johnson, J. C. Lee, and A. Zellner), Springer-Verlag, New York, 248–263.
- Geweke, J. (1999). Using simulation methods for Bayesian econometric models: Inference, development, and communication (with discussion and rejoinder), *Econometric Reviews*, 18, 1–126.
- Geweke, J. (2004). Getting it right: Joint distributions tests of posterior simulators, *Journal of the American Statistical Association* 99, 799–804.
- Geweke, J. (2005). *Contemporary Bayesian Econometrics and Statistics*, Wiley, Hoboken, NJ.
- Geweke, J. (2006). Interpretation and inference in mixture models: MCMC is alive, well and simple, University of Iowa, Departments of Economics and Statistics, working paper.
- Geweke, J. and M. Keane (2001). Computationally intensive methods for integration in econometrics, in *Handbook of Econometrics Volume 5* (eds. J. J. Heckman and E. Leamer), Elsevier, Amsterdam, 3463–3568.
- Geweke, J., M. P. Keane, and D. E. Runkle (1997). Statistical inference in the multinomial multiperiod probit model, *Journal of Econometrics*, 80, 125–165.
- Geweke, J. and N. Terui (1993). Bayesian threshold autoregressive models for nonlinear time series, *Journal of Times Series Analysis*, 14, 441–454.
- Gill, J. (2002). *Bayesian Methods: A Social and Behavioral Sciences Approach*, Chapman & Hall / CRC, Boca Raton.

- Greene, W. (1999). Marginal effects in the censored regression model, *Economics Letters*, 1, 43–50.
- Gujarati, D. (2003). *Basic Econometrics*, McGraw-Hill, Boston.
- Hay, J. and R. Olsen (1984). Let them eat cake: A note on comparing alternative models of the demand for medical care, *Journal of Business and Economic Statistics*, 2, 279–282.
- Heath, D. and W. Sudderth (1976). De Finetti's theorem on exchangeable variables, *American Statistician*, 30, 188–189.
- Heckman, J. (1976). The common structure of statistical models of truncation, sample selection and limited dependent variables and a simple estimator for such models, *Annals of Economic and Social Measurement*, 5, 475–492.
- Hoeting, J., D. Madigan, A. Raftery, and C. Volinsky (1999). Bayesian model averaging: A tutorial, *Statistical Science*, 14, 382–417.
- Jeffreys, H. (1961). *Theory of Probability*, 3rd edition, Clarendon Press, Oxford.
- Johnson, N., S. Kotz, and N. Balakrishnan (1994). *Continuous Univariate Distributions*, Volume 1, 2nd edition, Wiley, New York.
- Johnson, N., S. Kotz and N. Balakrishnan (1995). *Continuous Univariate Distributions*, Volume 2, 2nd edition, Wiley, New York.
- Johnson, N., S. Kotz, and N. Balakrishnan (2000). *Continuous Multivariate Distributions*, Volume 1, 2nd edition, Wiley, New York.
- Johnson, N., S. Kotz, and A. Kemp (1993). *Univariate Discrete Distributions*, 2nd edition. Wiley, New York.
- Kadane, J., N. Chan, and L. Wolfson (1996). Priors for unit root models, *Journal of Econometrics*, 88, 99–121.
- Kadane, J. and J. Dickey (1980). Bayesian decision theory and the simplification of models, in *Evaluation of Econometric Models* (eds. J. Kmenta and J. Ramsey), Academic Press, New York.
- Kadiyala, K. and S. Karlsson (1997). Numerical methods for estimation and inference in Bayesian VAR-models, *Journal of Applied Econometrics*, 12, 99–132.
- Kass, R. and S. Vaidyanathan (1992). Approximate Bayes factors and orthogonal parameters, with application to testing equality of two Binomial proportions, *Journal of the Royal Statistical Society, Series B*, 54, 129–144.
- Kass, R. and L. Wasserman (1995). A reference Bayes test for nested hypotheses and its relationship to the Schwartz criterion, *Journal of the American Statistical Association*, 90, 928–934.
- Kass, R. and L. Wasserman (1996). The selection of prior distributions by formal rules, *Journal of the American Statistical Association*, 91, 1343–1370.
- Keane, M. (1992). A note on identification in the multinomial probit model, *Journal of Business and Economic Statistics*, 10, 193–200.
- Kleibergen, F. and H. K. van Dijk (1994). On the shape of the likelihood/posterior in cointegration models, *Econometric Theory*, 10, 514–551.

- Kleibergen, F. and H. K. van Dijk (1998). Bayesian simultaneous equations analysis using reduced rank structures, *Econometric Theory*, 14, 701–743.
- Koop, G. (2003). *Bayesian Econometrics*, Wiley, Chichester.
- Koop, G. and D. J. Poirier (1997). Learning about the cross-regime correlation in switching regression models, *Journal of Econometrics*, 78, 217–227.
- Koop, G. and D. J. Poirier (2004). Bayesian variants of some classical semiparametric regression techniques, *Journal of Econometrics*, 123, 259–282.
- Koop, G. and S. Potter (1999). Dynamic asymmetries in US unemployment, *Journal of Business and Economic Statistics*, 17, 298–312.
- Koop, G. and S. Potter (2005). Forecasting and estimating multiple change-point models with an unknown number of change points, Department of Economics, University of Leicester, working paper.
- Koop, G., R. Strachan, H. van Dijk, and M. Villani (2006). Bayesian approaches to cointegration, in *The Palgrave Handbook of Econometrics, Volume 1: Theoretical Econometrics* (eds. T. C. Mills and K. Patterson), Palgrave Macmillan, Basingstoke, Chapter 25.
- Koop, G. and J. L. Tobias (2006). Semiparametric Bayesian inference in smooth coefficient models, *Journal of Econometrics*, 134, 283–315.
- Kwan, Y. K. (1998). Asymptotic Bayesian analysis based on a limited information estimator, *Journal of Econometrics*, 88, 99–121.
- Lancaster, T. (2004). *An Introduction to Modern Bayesian Econometrics*, Blackwell, Oxford.
- Leamer, E. (1978). *Specification Searches*. Wiley, New York.
- Leamer, E. (1982). Sets of posterior means with bounded variance priors, *Econometrica*, 50, 725–736.
- LeSage, J. (1999). *Applied Econometrics Using MATLAB*, available at <http://www.spatial-econometrics.com/>.
- Leung, S. F. and S. Yu (1996). On the choice between sample selection and two-part models, *Journal of Econometrics*, 72, 197–229.
- Li, K. (1998). Bayesian inference in a simultaneous equation model with limited dependent variables, *Journal of Econometrics*, 85, 387–400.
- Li, M., D. J. Poirier, and J. L. Tobias (2003). Do dropouts suffer from dropping out? Estimation and prediction of outcome gains in generalized selection models, *Journal of Applied Econometrics*, 19(2), 203–225.
- Li, M. and J. L. Tobias (2005). Bayesian analysis of structural effects in an ordered equation system, in *Studies in Nonlinear Dynamics and Econometrics*, forthcoming.
- Lindley, D. V. (2004). That wretched prior, *Significance* 1, 85–87.
- Lindley, D. and A. F. M. Smith (1972). Bayes estimates for the linear model, *Journal of the Royal Statistical Society, Series B*, 34, 1–41.
- Madigan, D. and J. York (1995). Bayesian graphical models for discrete data, *International Statistical Review*, 63, 215–232.

- Manning, W., N. Duan, and W. Rogers (1987). Monte-Carlo evidence on the choice between sample selection and two-part models, *Journal of Econometrics*, 35, 59–82.
- Manning, W., C. Morris, and J. Newhouse (1981). A two-part model of the demand for medical care: Preliminary results from the health insurance experiment, in *Health, Economics and Health Economics* (eds. J. van der Gaag and M. Perlman), North-Holland, Amsterdam.
- McCulloch, R. and P. E. Rossi (1994). An exact likelihood analysis of the multinomial probit model, *Journal of Econometrics*, 64, 207–240.
- McCulloch, R., N. Polson and P. E. Rossi (2000). A Bayesian analysis of the multinomial probit model with fully identified parameters, *Journal of Econometrics*, 99, 173–193.
- McDonald, J. and R. Moffitt (1980). The uses of tobit analysis, *Review of Economics and Statistics*, 62, 318–321.
- McLachlan, G. and D. Peel (2000). *Finite Mixture Models*, Wiley, New York.
- Michael, J. R., W. R. Schucany, and R. W. Haas (1976). Generating random variates using transformations with multiple roots, *The American Statistician*, 30, 88–90.
- Nandram, B. and M.-H. Chen (1996). Reparameterizing the generalized linear model to accelerate Gibbs sampler convergence, *Journal of Statistical Computation and Simulation*, 54, 129–144.
- Neyman, J. and E. S. Scott (1948). Consistent estimates based on partially consistent observations, *Econometrica*, 16, 1–32.
- Nobile, A. (2000). Comment: Bayesian multinomial probit models with a normalization constraint, *Journal of Econometrics*, 99, 335–345.
- Phillips, P. C. B. (1991). To criticize the critics: An objective Bayesian analysis of stochastic trends (with discussion), *Journal of Applied Econometrics*, 6, 333–474.
- Poirier, D. J. (1973). Piecewise regression using cubic splines, *Journal of the American Statistical Association*, 68, 515–524.
- Poirier, D. J. (1995). *Intermediate Statistics and Econometrics: A Comparative Approach*, MIT Press, Cambridge.
- Poirier, D. J. (1996). Local Bayes factor approximations, in *1996 Proceedings of the American Statistical Association Section on Bayesian Statistical Science*, 334–339.
- Poirier, D. J. (1998). Revising beliefs in non-identified models, *Econometric Theory*, 14, 483–509.
- Poirier, D. J. and A. Melino (1978). A note on the interpretation of regression coefficients within a class of truncated distributions, *Econometrica* 46, 1207–1209.
- Poirier, D. J. and J. L. Tobias (2003). On the predictive distributions of outcome gains in the presence of an unidentified parameter, *Journal of Business and Economic Statistics*, 21, 258–268.
- Potter, S. M. (1995). A nonlinear approach to US GNP, *Journal of Applied Econometrics*, 10, 109–126.
- Raftery, A. E., D. Madigan, and J. A. Hoeting (1997). Bayesian model averaging for linear regression models, *Journal of the American Statistical Association*, 92(437), 179–191.

- Ramsey, F. (1926). Truth and probability, in *The Foundations of Mathematics and Other Logical Essays* (ed. R. B. Braithwaite), Routledge & Keegan Paul, London, 156–198.
- Richardson, S. and P. Green (1997). On Bayesian analysis of mixtures with an unknown number of components (with discussion), *Journal of the Royal Statistical Society, Series B*, 59, 731–792.
- Rubin, D. (1987). A noniterative sampling/importance resampling alternative to the data augmentation algorithm for creating a few imputations when fractions of missing information are modest: The SIR algorithm, *Journal of the American Statistical Association*, 82, 543–546.
- Rubin, D. (1988). Using the SIR algorithm to simulate posterior distributions (with discussion), in *Bayesian Statistics 3* (eds. J. M. Bernardo, M. H. DeGroot, D. V. Lindley, and A. F. M. Smith), Oxford University Press, Oxford, 395–402.
- Schotman, P. (1994). Priors for the AR(1) model: Parameterisation issues and time series considerations, *Econometric Theory*, 10, 579–595.
- Schwarz, G. (1978). Estimating the dimension of a model, *Annals of Statistics*, 6, 461–464.
- Shuster, J. (1968). On the inverse Gaussian distribution function, *Journal of the American Statistical Association*, 63, 1514–1516.
- Silverman, B. (1986). *Density Estimation for Statistics and Data Analysis*, Chapman & Hall, London.
- Sims, C. (1988). Bayesian skepticism on unit root econometrics, *Journal of Economic Dynamics and Control*, 12, 463–474.
- Sims, C and H. Uhlig (1991). Understanding unit rooters: A helicopter tour, *Econometrica*, 59, 1591–1600.
- Smith, A. F. M. and A. E. Gelfand (1992). Bayesian statistics without tears: A sampling-resampling perspective, *The American Statistician* 46, 84–88.
- Smith, M. and R. Kohn (1996). Nonparametric regression using Bayesian variable selection, *Journal of Econometrics*, 75, 317–343.
- Strachan, R. and B. Inder (2004). Bayesian analysis of the error correction model, *Journal of Econometrics*, 123, 307–325.
- Tanner, M. and W. Wong (1987). The calculation of posterior distributions by data augmentation, *Journal of the American Statistical Association*, 82, 528–549.
- Thornber, H. (1967). Finite sample Monte Carlo studies: An autoregressive illustration, *Journal of the American Statistical Association*, 62, 810–818.
- Tierney, L. (1994). Markov chains for exploring posterior distributions (with discussion), *Annals of Statistics*, 22, 1701–1762.
- Tierney, L. and J. Kadane (1986). Accurate approximations for posterior moments and marginal posterior densities, *Journal of the American Statistical Association*, 81, 82–86.
- Tierney, L., R. E. Kass, and J. B. Kadane (1989). Fully exponential Laplace approximations to expectations and variances of nonpositive functions, *Journal of the American Statistical Association*, 84, 710–716.

- Tobin, J. (1958). Estimation of relationships for limited dependent variables, *Econometrica*, 26, 24–36.
- van Hasselt, M. (2005). Bayesian sampling algorithms for the sample selection and two-part models, working paper, Department of Economics, Brown University.
- Vijverberg, W. (1993). Measuring the unidentified parameter of the extended Roy model of selectivity, *Journal of Econometrics*, 57, 69–89.
- Wald, A. (1950). *Statistical Decision Functions*, Wiley, New York.
- Yang, R. and J. O. Berger (1996). A catalogue of noninformative priors, Technical Report #97-42, Duke University, Institute of Statistics and Decision Sciences.
- Zellner, A. (1962). An efficient method of estimating seemingly unrelated regressions and tests for aggregation bias, *Journal of the American Statistical Association*, 57, 348–368.
- Zellner, A. (1986a). On assessing prior distributions and Bayesian regression analysis with g-prior distributions, in *Bayesian Inference and Decision Techniques: Essays in Honour of Bruno de Finetti* (eds. P. K. Goel and A. Zellner), North-Holland, Amsterdam.
- Zellner, A. (1986b). Bayesian estimation and prediction using asymmetric loss functions, *Journal of the American Statistical Association*, 81, 446–451.
- Zellner, A. (1994). Bayesian and non-Bayesian estimation using balanced loss functions, in *Statistical Decision Theory and Related Topics V* (eds. S. Gupta and J. Berger), Springer-Verlag, New York.
- Zellner, A. and P. Rossi (1984). Bayesian analysis of dichotomous quantal response models, *Journal of Econometrics*, 25, 365–393.
- Zha, T. and C. Sims (2006). Were there regime switches in US monetary policy?, *American Economic Review*, 96, 54–81.

Index

- Abadir, 2, 3, 148, 335
Acceptance/rejection sampling, 160, 162
Achcar, 98, 100
Admissibility, 38, 47, 81
Albert, 203, 211, 257
Amemiya, 225
Ancillarity, 57
Andrews, 259
AR model, 298, 320
ARCH model, 310
Asymptotics, 91
Augmented joint posterior, 204, 205
Autocorrelated errors, 198
Autocorrelation, 130, 182, 211, 214
Azzalini, 276
- Balakrishnan, 335
Bartlett's paradox, 122
Bauwens, xix, 297, 311, 314, 320
Bayes factor, 61, 69, 87, 101, 112, 284
Bayes risk, 38, 47
Bayes' theorem
 for densities, 12, 60
 for events, 3, 13, 60
Bayesian credible region, 51, 57
Bayesian critical value, 62, 68
Bayesian econometrics, 1
Bayesian inference, 11
Bayesian model averaging, 281, 284
Bayesian model selection, 281, 287
Bayesian statistics, 1
Berger, 12, 38, 47, 52, 57, 81, 82, 92, 326
Bernardo, 1, 6, 81, 91, 326
- Bernoulli, 1
Bernoulli distribution, 4, 5, 7, 12, 73, 82, 84, 263, 339
Beta distribution, 13, 19, 73, 120, 263, 340
Beta function, 13, 73
Beta-binomial distribution, 73
Bias, 43
Binomial distribution, 263, 339
Blocking steps, 181, 211
Blundell, 230
Branco, 276
Burn-in, 118, 143
- Carlin, B. P., 117, 141, 177, 181, 184, 259
Carlin, J. B., 117, 152, 266
Casella, 52, 129
Cauchy distribution, 338
Celeux, 253
Chan, 327
Changepoint, 138, 140
Chen, 214, 235
Chi-square distribution, 54, 120, 262, 336
Chib, 117, 118, 181, 203, 211, 220, 257, 290, 293, 296
Choice of prior, 79, 82
Cobb–Douglas production function, 146
Coherence principle, 2, 81
Cointegration, 330
Component indicator variables, 262
Component label vector, 265
Conditional frequentist, 57
Conditional independence, 8, 71, 169, 174
Confidence, 51, 54
Confidence interval, 57

- Conjugate analysis
 - Bernoulli, 12
 - Exponential, 22
 - Exponential family, 23
 - Multivariate normal, 25
 - Normal with known mean and unknown variance, 20
 - Normal with unknown mean and variance, 17
 - Normal with unknown mean, known variance, 14
 - Poisson, 22
 - Uniform, 22
- Conjugate families, 12
- Conjugate priors, 19, 79
- Cost function, 29, 59
 - all-or-nothing, 31, 32, 33
 - alternative squared error, 34
 - asymmetric linear, 31, 32
 - balanced, 43
 - linex, 36, 46
 - predictive, 72
 - quadratic, 31, 38, 40, 43, 75, 76
 - symmetric linear, 31
 - weighted squared error, 31, 34
- Count data, 140
- Cowles, 177
- Cragg, 240
- Dalla Valle, 276
- Data augmentation, 203, 204, 210, 212, 217, 219, 241, 276
- de Finetti, 2, 3
 - Representation theorem, 4, 6
- Deb, 240
- Decision space, 59
- Denison, 290
- Dey, 276, 293, 296
- Diagnostics, 142
 - Gibbs sampler convergence, 177, 178
- Dickey, 62
- DiNardo, 230
- Dirichlet distribution, 235, 265, 340
- Double exponential distribution, 259
- Duan, 240, 243
- Duncan, 230
- Dutch book, 2, 7
- Efron, 82
- Ellipsoid bound theorem, 111
- Empirical Bayes, 45, 198
- Exchangeability, 3, 4, 7
- Exponential density, 159
- Exponential distribution, 22, 40, 53, 78, 85, 97, 98, 100, 120, 336
- Exponential family, 12, 23
- Fair, 208
- Fernandez, 282, 284
- Frequentists, 2
- Fully exponential method, 93, 94, 98, 100, 101, 103
- g-prior, 282, 283, 290
- Gamerman, 167
- Gamma distribution, 21, 22, 40, 48, 73, 78, 86, 95, 96, 97, 103, 110, 123, 335
- Gamma function, 13, 73, 336
- Gamma-gamma distribution, 78
- Garthwaite, 79
- Geisser, 80
- Gelfand, 141, 167, 173, 184, 293, 296
- Gelman, 117, 152, 266
- Generalized inverse Gaussian distribution, 259
- Generalized tobit model, 225, 229
- George, 129, 282, 287, 290
- Geweke, 117, 118, 146, 147, 150, 196, 218, 219, 293, 298, 300, 311, 330
- Gibbs sampling, 118, 128, 130, 134, 135, 142, 144, 293
- Gill, 141
- Green, 253
- Greenberg, 118
- Greene, 222
- Gujarati, 208
- Haas, 259
- Hay, 243
- Heath, 5
- Heckman, 225
- Heijmans, 2
- Heteroscedasticity, 135, 193, 195
- Hierarchical model, 169, 210
- Hills, 173
- Hoeting, 282
- Holmes, 290
- HPD interval, 52, 54, 55, 58, 72
- HPD region, 51
- Hurdle model, 240
- Hurn, 253
- Hwang, 52
- Hypothesis testing, 59, 87
- Identification, 27, 206
- Importance sampling, 117, 124, 126
- Improper prior, 38, 81, 82, 87, 88, 112
- Incidental truncation, 225

- Independence chain, 118, 154
- Independent, 8
- Information matrix, 80, 83, 85, 86
- Intrinsic prior, 82
- Inverse Gaussian distribution, 259
- Inverse transform method, 157, 159
- Inverted Gamma distribution, 42, 86, 291, 336
- Jeffreys, 80
- Jeffreys' prior, 81–85, 324
- Jeffreys' rule, 81
- Jeliazkov, 290
- Johnson, 335
- Kadane, 62, 79, 93, 94, 98, 100, 103, 327
- Kadiyala, 314
- Karlsson, 314
- Kass, 79, 81, 82, 84, 89, 94, 102
- Keane, 117, 218
- Kemp, 335
- Kleibergen, 330
- Kohn, 289, 290
- Koop, xix, 117, 118, 143, 156, 192, 230, 284, 290, 302, 307, 320, 330, 334
- Kotz, 335
- Kwan, 91, 319
- Lancaster, xix, 117, 143
- Laplace, 1, 5
- Laplace approximation, 92, 95–97, 101, 103
- Laplace distribution, 154, 259
- Latent variable models, 203
- Leamer, 112
- Learning, 28, 114, 229, 230
- LeSage, 123
- Let's make a deal, 6
- Leung, 243
- Ley, 282, 284
- Li, K., 225
- Li, M., 230, 232, 235
- Likelihood principle, 12, 48, 57, 82
- Lindley, 79, 175
- Linear regression model, 107, 119, 121, 134, 191, 256, 265
- Lognormal distribution, 262
- Longitudinal data, *see* Panel data
- Loss function, *see* Cost function
- Lottery, 2
- Louis, 117
- Lubrano, xix, 297, 311, 314, 320
- Madigan, 282, 283
- Magnus, 2, 148, 335
- Mallick, 290
- Mallows, 259
- Manning, 240, 243
- Marginal effects, 209, 211, 219, 221
- Marginal likelihood, 11, 60, 61, 101, 281, 293, 295, 308
- Marginal-conditional simulator, 150
- Marginalization paradox, 81
- Markov chain Monte Carlo model composition, 283, 285
- McCulloch, 218, 282, 287, 290
- McDonald, 220
- McLachlan, 253
- Mean squared error, 38, 40
- Melino, 220
- Method of composition, 123, 150
- Metropolis–Hastings algorithm, 118, 152, 154, 194, 257, 284
- Michael, 259
- Missing data, 246
- Mixed model, 181
- Mixtures, 23, 72, 253
 - of binomials, 270
 - of normals, 254, 262, 265, 266, 287, 328
 - of Poissons, 260
- Model probability, 281, 283
- Moffitt, 220
- Moment generating function, 36, 46
- Monte Carlo integration, 117, 119, 121, 123, 124, 130, 290
- Morris, 240
- Multicollinearity, 114
- Multinomial distribution, 265, 340
- Multinomial probit, 216
- Munkin, 240
- Nandram, 214, 235
- Negative binomial distribution, 83, 84, 260
- Newhouse, 240
- Neyman, 81
- Nobile, 218, 229
- Noninformative prior, 80
- Nonparametric regression, 289
- Normal distribution, 120
 - half, 276
 - matric variate, 313, 331
 - multivariate, 25, 75, 130, 336
 - piecewise, 327
 - skew, 276
 - split, 63
 - truncated, 147, 200, 218, 220, 234, 268
 - univariate, 14, 17, 19, 20, 38, 43, 63, 67, 74, 86

- Normal-gamma distribution, 18, 21, 56, 76, 103, 108, 111, 192, 283, 287, 338
- Nuisance parameters, 12, 72
- Numerical standard error, 119, 121, 144, 182
- O'Hagan, 79
- Objective priors, 79, 81
- Olsen, 243
- One-sided hypotheses, 63
- Ordered probit model, 211, 214, 232
- Ordinary least squares, 108, 127, 177, 180
- Panel data, 169, 171, 210
- Pareto distribution, 22, 340
- Peel, 253
- Pericchi, 82
- Phillips, 320, 326
- Point estimation
 - all-or-nothing loss, 32, 33
 - alternative squared error loss, 34
 - asymmetric linear loss, 32, 33
 - Bayesian, 29, 30, 43, 46
 - conjugate normal analysis, 34
 - frequentist, 29
 - linex loss, 36
 - nonconjugate Normal analysis, 35
 - prediction, 71
 - quadratic loss, 31, 33
 - weighted squared error loss, 34
- Poirier, xix, 11, 42, 102, 107, 112, 122, 154, 220, 230, 259, 290, 291, 302, 335, 336
- Poisson distribution, 22, 48, 73, 84, 339
- Poisson-gamma distribution, 74
- Polson, 218, 259
- Population model, 184
- Posterior density, 11
- Posterior expected cost, 59, 60
- Posterior odds, 61
- Posterior predictive p -values, 143
- Posterior predictive density, 71, 320
- Posterior simulator
 - AR model, 298
 - ARCH model, 310
 - generalized tobit, 225, 229
 - hierarchical model, 173, 181
 - linear regression model, 134
 - mixture of binomials, 270
 - multinomial probit, 216
 - nonparametric regression model, 187, 289
 - Normal mixture model, 262, 265, 266
 - ordered outcomes model, 232
 - ordered probit, 211, 214
 - panel probit, 210
 - probit model, 204, 206
 - regression model with an unknown changepoint, 138, 140
 - regression model with autocorrelation, 198
 - regression model with heteroscedasticity, 135, 193, 195
 - regression model with linear constraints, 146
 - sample selection model, 243
 - stochastic frontier model, 236
 - SUR model, 136
 - TAR model, 300, 304, 307
 - tobit model, 219, 221
 - treatment effects model, 223
 - two-part model, 240
 - VAR model, 312
 - VECM, 330
- Potter, 290, 300, 306, 307
- Prediction, 71
- Prior density, 11
- Prior elicitation, 79, 303
- Prior odds, 61
- Probability, 2
 - long-run relative frequency, 2, 5
 - subjective, 2, 3, 55
- Probit model, 204, 206, 208, 210, 256, 257
- Proper prior, 38, 87
- QQ plot, 143
- Racine-Poon, 173
- Raftery, 282
- Ramsey, 2
- Random walk chain, 118, 154, 194, 284
- Reference prior, 81, 326
- Reparameterization, 58, 81, 85, 93, 94, 100, 214
- Richard, xix, 297, 311, 314, 320
- Richardson, 253
- Risk function, 29, 44, 46, 47
- Robert, 52, 253
- Rogers, 240, 243
- Rossi, 204, 218
- Rubin, 117, 152, 167, 266
- Runkle, 218
- Sahu, 184
- Sample selection model, 243
- Sampling distribution, 30, 71
- Sampling properties, 38, 40, 43
- Sampling-importance resampling, 167
- Savage–Dickey density ratio, 69, 198
- Schotman, 320
- Schucany, 259
- Schwarz, 102

- Scott, 81
- Shrinkage estimator, 171, 177
- Shuster, 259
- Silverman, 230
- Simple hypotheses, 62, 63, 66, 67
- Sims, 307, 319–322
- Smith, A.F.M., 1, 6, 91, 98, 100, 141, 167, 173, 175, 290
- Smith, M., 289, 290
- Steel, 282, 284
- Stern, 117, 152, 266
- Stirling, 1
- stochastic frontier model, 236
- Stochastic search variable selection, 287
- Strachan, 320, 330, 334
- Strong inconsistency, 88, 89
- Student t -distribution, 76, 120, 123, 124, 196, 254, 256, 283, 337
- Successive-conditional simulator, 150
- Sudderth, 5
- SUR model, 136, 314, 333
- Tanner, 203
- TAR model, 300, 304, 307
- Terui, 300
- Threshold triggers, 304
- Tierney, 93, 94, 98, 100, 103, 129
- Time series
 - Nonstationary, 319
 - Stationary, 297
- Tobias, 230, 232, 235, 290
- Tobin, 219
- Tobit model, 219, 221, 240
- Transition kernel, 150
- Treatment effects, 232
- Trivedi, 240
- Two-part model, 243
- Two-sided hypothesis, 66, 67
- Type I and II errors, 65, 68
- Type II maximum likelihood estimator, 45
- Uhlig, 319–322
- Uniform distribution, 22, 120, 335
- Unit root, 322, 324
- Unobserved confounding, 223
- van Dijk, 320, 330, 334
- van Hasselt, 240, 243
- VAR model, 312, 331
- Vector error correction model, 330
- Vijverberg, 230
- Villani, 320, 330, 334
- Volinsky, 282
- Wald, 38
- Wasserman, 79, 81, 82, 84, 89, 102
- Weighted bootstrap, 167
- Wishart distribution, 174, 339
- Wolfson, 327
- Wolpert, 12, 57
- Wong, 203
- Yang, 82, 326
- York, 283
- Yu, 243
- Zellner, 36, 43, 204, 283
- Zha, 307