Martin A. Tanner

Tools for Statistical Inference

Methods for the Exploration of Posterior Distributions and Likelihood Functions

Second Edition

With 47 Figures

Springer-Verlag
New York Berlin Heidelberg London Paris
Tokyo Hong Kong Barcelona Budapest

Martin A. Tanner
Department of Biostatistics
University of Rochester Medical Center
Rochester, NY 14642, USA

Mathematics Subject Classification: 62F15, 62Fxx, 62Jxx

Library of Congress Cataloging-in-Publication Data.
Tanner, Martin Abba, 1957–,
Tools for Statistical Inference: Methods for the Exploration of Posterior Distributions and Likelihood Functions
Martin A. Tanner. – 2nd ed.
p. cm. – (Springer series in statistics)
Includes bibliographical references and index.
ISBN 0-387-94031-6
1. Bayesian statistical decision theory. 2. Mathematical statistics. I. Title. II. Series.
QA279.5.T36 1993
519.5'42—dc20 93-3270

Printed on acid-free paper.

Production coordinated by Christin R. Ciresi and managed by SPS, Bangalore, India; manufacturing coordinated by Jacqui Ashri.
Typesetting: Macmillan India Ltd., Bangalore-25, India.

9 8 7 6 5 4 3 2 1

ISBN-13: 978-1-4684-0194-3 e-ISBN-13:978-1-4684-0192-9
DOI:10.1007/978-1-4684-0192-9

Springer Series in Statistics

Springer Series in Statistics

Andersen/Borgan/Gill/Keiding: Statistical Models Based on Counting Processes.

Anderson: Continuous-Time Markov Chains: An Applications-Oriented Approach.

Andrews/Herzberg: Data: A Collection of Problems from Many Fields for the Student and Research Worker.

Anscombe: Computing in Statistical Science through APL.

Berger: Statistical Decision Theory and Bayesian Analysis, 2nd edition.

Bolfarine/Zacks: Prediction Theory for Finite Populations.

Brémaud: Point Processes and Queues: Martingale Dynamics.

Brockwell/Davis: Time Series: Theory and Methods, 2nd edition.

Choi: ARMA Model Identification.

Daley/Vere-Jones: An Introduction to the Theory of Point Processes.

Dzhaparidze: Parameter Estimation and Hypothesis Testing in Spectral Analysis of Stationary Time Series.

Farrel: Multivariate Calculation.

Federer: Statistical Design and Analysis for Intercropping Experiments

Fienberg/Hoaglin/Kruskal/Tanur (Eds.): A Statistical Model: Frederick Mosteller's Contributions to Statistics, Science, and Public Policy.

Goodman/Kruskal: Measures of Association for Cross Classifications.

Grandell: Aspects of Risk Theory.

Hall: The Bootstrap and Edgeworth Expansion.

Härdle: Smoothing Techniques: With Implementation in S.

Hartigan: Bayes Theory.

Heyer: Theory of Statistical Experiments.

Joliffe: Principal Component Analysis.

Kotz/Johnson (Eds.): Breakthroughs in Statistics Volume I.

Kotz/Johnson (Eds.): Breakthroughs in Statistics Volume II.

Kres: Statistical Tables for Multivariate Analysis.

Leadbetter/Lindgren/Rootzén: Extremes and Related Properties of Random Sequences and Processes.

Le Cam: Asymptotic Methods in Statistical Decision Theory.

Le Cam/Yang: Asymptotics in Statistics: Some Basic Concepts.

Manoukian: Modern Concepts and Theorems of Mathematical Statistics.

Manton/Singer/Suzman : Forecasting the Elderly Populations.

Miller, Jr.: Simultaneous Statistical Inference, 2nd edition.

Mosteller/Wallace: Applied Bayesian and Classical Inference: The Case of *The Federalist Papers.*

Pollard: Convergence of Stochastic Processes.

Pratt/Gibbons: Concepts of Nonparametric Theory.

Read/Cressie: Goodness-of-Fit Statistics for Discrete Multivariate Data.

(*continued after index*)

Preface

This book provides a unified introduction to a variety of computational algorithms for likelihood and Bayesian inference. In this second edition, I have attempted to expand the treatment of many of the techniques discussed, as well as include important topics such as the Metropolis algorithm and methods for assessing the convergence of a Markov chain algorithm.

Prerequisites for this book include an understanding of mathematical statistics at the level of Bickel and Doksum (1977), some understanding of the Bayesian approach as in Box and Tiao (1973), experience with conditional inference at the level of Cox and Snell (1989) and exposure to statistical models as found in McCullagh and Nelder (1989). I have chosen not to present the proofs of convergence or rates of convergence since these proofs may require substantial background in Markov chain theory which is beyond the scope of this book. However, references to these proofs are given. There has been an explosion of papers in the area of Markov chain Monte Carlo in the last five years. I have attempted to identify key references – though due to the volatility of the field some work may have been missed. Hopefully enough information has been presented to allow the reader to impute what is missing!

I would like to thank my colleagues Hendricks Brown, Brad Carlin, Alan Gelfand, Lu Cui, Jack Hall, John Kolassa, Chris Ritter, and Wing Wong for their comments/suggestions/input into this work. I wish to thank the American Statistical Association for allowing me to reprint material which has appeared in the *Journal of the American Statistical Association*. I also wish to thank the Biometrika Trustees and the Royal Statistical Society for their permission to reprint material which has appeared in *Biometrika* and *Applied Statistics*, respectively. This work was supported by NIH grant CA35464. I take full responsibility for all errors and omissions.

בריך רחמנא דסייען

Rochester, New York
March 1993

M. Tanner

Contents

CHAPTER 5

CHAPTER 6

CHAPTER 1

Introduction

The goal of this book is to provide a unified introduction to a variety of computational algorithms that can be used as part of a Bayesian (posterior) analysis or as part of a likelihood analysis. These algorithms are tools and may be categorized using several taxonomies. The reader may find it useful to review these taxonomies as an aid to understanding how these tools complement, contrast and extend each other.

i. *Function*: *Maximize*, *Marginalize and Simulate*

The Newton–Raphson, *EM* (expectation maximization) and Monte Carlo *EM* (*MCEM*) algorithms are used to *locate the mode* or *modes* of the likelihood function or of the posterior density. Moreover, the Newton–Raphson algorithm as well as the *EM* and *MCEM* algorithms (with additional effort) provide information regarding the curvature of the likelihood at a mode. Other algorithms (e.g. Laplace's method, data augmentation and the Gibbs sampler) can be used to *obtain a marginal* of the likelihood function or of the posterior function. Alternatively, acceptance/rejection, data augmentation, the Gibbs sampler and the Metropolis algorithm can be used to *obtain a sample of parameter values* from the likelihood function or from the posterior density. This sample may then be used to obtain an estimate of a lower dimensional marginal of the likelihood or posterior, to estimate a moment of this marginal or, more generally, to estimate some functional of the likelihood or posterior (e.g. the content and boundary of the highest posterior density region or the predictive distribution).

ii. *Random* vs. *Deterministic Algorithms*

Alternatively, one can categorize these algorithms into non-Monte Carlo, noniterative Monte Carlo and iterative Monte Carlo methods. The non-Monte Carlo tools (e.g. Newton–Raphson, *EM* and Laplace) do not require the input of a stream of (pseudo) random numbers. The importance sampling and rejection/acceptance algorithms require a stream of (pseudo) random numbers as input and produce a sample from the likelihood function or from the posterior density as output. They are not iterative in nature. The Monte Carlo *EM* algorithm, data augmentation, the Gibbs sampler and the Metropolis algorithm all require a random input stream and require iteration to realize a sample from the likelihood function or from the posterior distribution of interest.

iii. *Augmentation* vs. *Nonaugmentation Algorithms*

In the *EM*, Monte Carlo *EM*, data augmentation and poor man's data augmentation algorithms, the data analyst augments the observed data with latent data to simplify the computations in the analysis. This latent data may be "missing" data (e.g. a missing data point in a two-way anova with one entry per cell), parameter values (e.g. the degrees of freedom of the t error distribution) or values of sufficient statistics. Readers of this book should understand that augmentation means more than filling in missing observations and that the augmentation vs. nonaugmentation taxonomy is not simply "missing data" vs. "observed data". Thus, in the *EM* algorithm, one augments the observed data with latent data such that one complicated maximization is replaced by an iterative series of simple maximizations. In the data augmentation algorithm, one augments the observed data with latent data to facilitate the sampling operation. The aim is to replace a complicated simulation by an iterative series of simple simulations.

Chapter 2 reviews the definitions of likelihood function and posterior density, the method of maximum likelihood, the Newton–Raphson algorithm for locating the mode of an objective function and provides a brief review of normal-based frequentist and Bayesian inference. Chapter 3 motivates and illustrates Laplace's method as a technique for calculating higher-order approximations to the likelihood or to the posterior. This methodology allows for non-normal approximations to the likelihood or to the posterior density. The noniterative Monte Carlo, importance sampling and rejection/acceptance methods are presented as tools which yield a sample from the exact likelihood or from the posterior. Thus, the algorithms in Chapters 2 and 3 do not make use of any augmentation of the observed data to define the algorithm. These methods are applied directly to the likelihood or to the posterior density.

Chapter 4 presents the first of the augmentation algorithms – the *EM* algorithm. We motivate and illustrate this algorithm, and also present various methods for calculating standard errors in the context of the *EM* algorithm. In Chapter 5 the data augmentation algorithm is motivated and illustrated and issues relating to assessment of convergence are presented. Also presented are the poor man's data augmentation algorithms which reduce the computational burden of the data augmentation algorithm, at the expense of providing a sample from an approximation to the likelihood or to the posterior. Importance sampling alternatives (e.g. *SIR* and sequential imputation) are also presented.

Since the data augmentation algorithm (Chapter 5) can be viewed as a two-component version of the Gibbs sampler, we choose to follow the data augmentation algorithm with a presentation of general Markov chain Monte Carlo methods: the Gibbs sampler and the Metropolis algorithm (Chapter 6). In fact, one may argue that the Gibbs sampler is an augmentation algorithm in that one augments the observed data with parameters (the latent data). In other words, rather than operating on the likelihood or the posterior $p(\theta_1, \theta_2, \ldots, \theta_d | Y)$, the Gibbs sampler requires that one augments the observed data Y with parameters to obtain, for example, $p(\theta_1, | \theta_2, \ldots, \theta_d, Y)$. Simply put, in augmentation one puts "stuff" on the right-hand side of the vertical bar. The Metropolis algorithm is not an augmentation algorithm, since it works directly with $p(\theta_1, \theta_2, \ldots, \theta_d | Y)$, modulo the normalizing constant. Both the Gibbs sampler and the Metropolis algorithm yield a Markov chain whose equilibrium distribution (under certain regularity conditions) is the likelihood function or the posterior desnsity of interest. Thus, both algorithms provide the data analyst with a correlated sample from the function of interest. Also discussed are techniques for assessing convergence and approaches for handling the situation where one cannot directly sample from the conditional distributions as required in the Gibbs sampler.

The frequentist, in contrast to the Bayesians or to the likelihoodists (i.e. those who do base their inference on the shape of the likelihood), may have little interest in the bulk of this book. First, the frequentist may be put off by the use of the posterior function in the presentation of the algorithms throughout this book. In this regard, it should be understood that while we show how the *EM* algorithm can be used to maximize the posterior, one can also use this algorithm to maximize the likelihood – though in any given situation the posterior calculation may be simpler due to the specification of the prior. Clearly, the mode of the likelihood and the curvature at the mode are of great value to the frequentist's *large sample* (*normal-based*) *inference approach.* However, the frequentist will have little interest in the marginalization uses of data augmentation, the Gibbs sampler and the Metropolis algorithm, since integrating out parameters of the likelihood may not be a sensible *frequentist* calculation. Thus, while to the Bayesian or to the likelihoodist tools such as data augmentation, Gibbs sampler, and Metropolis are of immense value to the analysis of a small sample problem, they are of little direct help to the frequentist since they do not determine the distribution of an estimator (e.g. the maximum likelihood estimator) or of a test statistic (e.g. likelihood ratio) *in repeated* (*small*) *samples of data.* With this concern in mind, the final section of this book illustrates how the Gibbs sampler, in conjunction with saddlepoint methods, may be used to simulate sufficient statistics within the conditional frequentist inference paradigm. In this way, we illustrate the application of Markov chain algorithms to small sample frequentist inference.

We now consider three examples which will help to motivate these various tools for statistical inference and which will be examined later in this book.

EXAMPLE. *Censored Regression Data*
The Stanford Heart Transplant Program began in October 1967. The patients in the program were admitted after review by a committee and they then waited for donor hearts to become available. Some patients died while waiting for a transplant, but most received a transplant. The data presented in Table 1.1 summarize survival time in days after transplant for 45 patients. (See Miller and Halpern (1982) for a more extensive version of this data set.) Available information for each patient include: survival time, an indication of whether the patient is dead or alive as of February 1980, the age of the patient in years at the time of transplant and a mismatch score which attempts to quantify the dissimilarity between donor and recipient tissues with respect to HLA antigens.

Table 1.1. Stanford Heart Transplant Data, February 1980.

Patient no.	Survival time, days	Status*	Age at transplant	Mismatch score**
1	15	1	54	1.11
2	3	1	40	1.66
3	46	1	42	0.61
4	623	1	51	1.32
5	126	1	48	0.36
6	64	1	54	1.89
7	1350	1	54	0.87
8	23	1	56	2.05
9	279	1	49	1.12
10	1024	1	43	1.13
11	10	1	56	2.76
12	39	1	42	1.38
13	730	1	58	0.96
14	1961	1	33	1.06
15	136	1	52	1.62
16	1	1	54	0.47
17	836	1	44	1.58
18	60	1	64	0.69
19	3695	0	40	0.38
20	1996	1	49	0.91

Table 1.1. (continued)

Patient no.	Survival time, days	Status*	Age at transplant	Mismatch score**
21	0	1	41	0.87
22	47	1	62	0.87
23	54	1	49	2.09
24	51	1	50	—
25	2878	1	49	0.75
26	3410	0	45	0.98
27	44	1	36	0.0
28	994	1	48	0.81
29	51	1	47	1.38
30	1478	1	36	1.35
31	254	1	48	1.08
32	897	1	46	—
33	148	1	47	—
34	51	1	52	1.51
35	323	1	48	1.82
36	3021	0	38	0.98
37	66	1	49	0.66
38	2984	0	32	0.19
39	2723	1	32	1.93
40	550	1	48	0.12
41	66	1	51	1.12
42	65	1	45	1.68
43	227	1	19	1.02
44	2805	0	48	1.20
45	25	1	53	1.68

Source: Miller and Halpern (1982).
* Status: dead, 1; alive, 0.
** — denotes missing score

Suppose we wish to regress log survival time on age. We notice that five of the individuals in our reduced version of the data set are still alive as of February 1980. In other words, the failure times for these subjects are *censored*. We do not have a survival time for these people – we only know that a given person survived beyond the recorded event time. Thus, we cannot directly use standard regression routines (e.g. REGRESS in MINITAB) to analyze these data. To analyze these data, we must resort to more sophisticated methods.

In this book, we will discuss two approaches to the analysis of such data. In the nonaugmentation approach, we construct the loglikelihood function for a given model for the observed data. We then maximize this function to estimate the parameters of the model. Unfortunately, the loglikelihood may be quite complicated. For example, in the case of censored regression data with normal errors, the loglikelihood will contain a sum of logarithms of normal survivor functions. In an alternative approach, a data augmentation approach, we augment the data with unobserved or *latent* data and compute the expected value of the loglikelihood for the augmented data. We then maximize this augmented loglikelihood function. This *EM* algorithm of augmenting the data and maximizing the expected value of the augmented loglikelihood is then iterated. In the case of censored regression data with normal errors, this data augmentation approach reduces to an iterated series of least-squares computations to estimate the regression coefficients. By augmenting the data, we replace the complicated analysis required in the observed data approach by a series of simple analyses.

EXAMPLE. *Latent Class Analysis*

The data in Table 1.2 represent responses of 3181 participants in the 1972, 1973 and 1974 General Social Survey. The responses are cross classified by year of study (3 levels) and a dichotomous response (yes/no) to each of three questions.

All three questions begin:

"Please tell me whether or not you think it should be possible for a pregnant woman to obtain a legal abortion if"

Question *A*:

she is married and does not want any more children

Question *B*:

the family has a very low income and cannot afford any more children

Question *C*:

she is not married and does not want to marry the man.

The traditional latent class model (Goodman, 1974) supposes that the four manifest variables (responses to questions *A*, *B*, and *C* and year) are conditionally independent given a dichotomous unobserved (latent) variable (e.g. the respondents' true attitude toward abortion – pro/anti). That is, if the value of the dichotomous latent variable is known for a given participant, knowledge of the participant's response to a given question provides no further information regarding the responses to either of the other two questions. Note that in this example, there is no "missing" data, as was the case in the previous example. It is the *model* which characterizes the unobserved (latent) data.

In Chapter 5, we will analyze these data using the data augmentation algorithm. In this data augmentation approach, one simulates latent data (conditional on simulated values of the parameters of the model) and then simulates parameter values (conditional on the simulated latent data). This procedure is then iterated to realize a *sample* of parameter values from the exact likelihood. We then use this sample of parameter values as our basis of inference.

Table 1.2. Subjects in the 1972–1974 General Social Surveys, Cross-Classified by Year of Survey and Responses to Three Questions on Abortion Attitudes.

Year (D)	Response to A	Response to B	Response to C	Observed count
1972	Yes	Yes	Yes	334
	Yes	Yes	No	34
	Yes	No	Yes	12
	Yes	No	No	15
	No	Yes	Yes	53
	No	Yes	No	63
	No	No	Yes	43
	No	No	No	501
1973	Yes	Yes	Yes	428
	Yes	Yes	No	29
	Yes	No	Yes	13
	Yes	No	No	17
	No	Yes	Yes	42
	No	Yes	No	53
	No	No	Yes	31
	No	No	No	453
1974	Yes	Yes	Yes	413
	Yes	Yes	No	29
	Yes	No	Yes	16
	Yes	No	No	18
	No	Yes	Yes	60
	No	Yes	No	57
	No	No	Yes	37
	No	No	No	430

Source: Haberman (1979, p. 559).

EXAMPLE. *Hierarchical Models*

The data in Table 1.3 represent the weights of 30 young rats measured weekly for five weeks. Note that there are no "missing" data in this data set – all rats were weighed according to schedule. However, we will still use augmentation methods to analyze these data. In Chapter 6, we will apply the two-stage hierarchical model to these data:

First stage:

$$Y_{ij} \sim N(\alpha_i + \beta_i x_{ij}, \sigma^2)$$

Second Stage:

$$\begin{pmatrix}\alpha_i\\ \beta_i\end{pmatrix} \sim N\left\{\begin{pmatrix}\alpha_C\\ \beta_C\end{pmatrix}, \Sigma\right\},$$

where $i = 1, \ldots, 30$, $j = 1, \ldots, 5$; x_{ij} is the age in days of the ith rat for

Table 1.3. Rat Population Growth Data.

Rat	Week 1	Week 2	Week 3	Week 4	Week 5	Rat	Week 1	Week 2	Week 3	Week 4	Week 5
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

Source: Gelfand et al. (1990)

measurement j; $x_{i1} = 8$, $x_{i2} = 15$, $x_{i3} = 22$, $x_{i4} = 29$, $x_{i5} = 36$, $i = 1, \ldots, 30$; and Y_{ij} is the weight of the ith rat at measurement j. This model implies that at the "first stage" the data Y_{ij} follow a normal distribution with mean $\alpha_i + \beta_i x_{ij}$ and variance σ^2. At the "second stage", the individual specific parameters α_i and β_i are assumed to be drawn from a bivariate normal distribution. Note that in contrast to the previous example, this model does not hypothesize any missing or latent data. Rather, in our analysis of the data, the parameters α_c, β_c and Σ play the role of the "missing" data. In our analysis, we augment the observed data with values for these parameters and then simulate values for α_i, β_i, and σ. Given values for α_i and β_i, we simulate the "missing" data, i.e. α_c, β_c, and Σ. This procedure is then iterated to realize a sample of parameter values from the exact posterior distribution, which we will use for such tasks as approximating various marginals of the posterior.

CHAPTER 2

Normal Approximations to Likelihoods and to Posteriors

In this chapter, we review several elementary concepts and methods from mathematical statistics. In Section 2.1, the likelihood and loglikelihood functions are defined and illustrated. The definition and an illustration of the posterior density is also presented. In Section 2.2, the method of maximum likelihood is defined and the Newton–Raphson algorithm is presented as an algorithm for computing maximum likelihood estimates. Section 2.3 presents frequentist and Bayesian justification for using the normal approximation to the likelihood or to the posterior as the basis for inference. The delta method is defined and illustrated in Section 2.4 and Section 2.5 reviews the notion of the Highest Posterior Density region, a Bayesian approach to confidence regions and significance levels.

2.1. Likelihood/Posterior Density

Let Y denote the data, which can be scalar, vector valued or matrix valued and suppose that

$$Y \sim f(Y|\theta)$$

where$f(\cdot\,|\,\cdot)$ is a density function indexed by a parameter θ (scalar or vector).

Definition 2.1.1. (Little and Rubin, 1987) Given the data Y, the likelihood function $L(\theta|Y)$ is any function of θ proportional to$f(Y|\theta)$.

The likelihood is a set of functions which differ by a factor that does not depend on θ. In other words, the likelihood function is defined up to a multiplicative constant. It is only the relative value (shape) of the likelihood which is of importance.

Definition 2.1.2. (Little and Rubin, 1987) The loglikelihood function $l(\theta|Y)$ is the natural logarithm of $L(\theta|Y)$.

EXAMPLE. *Univariate Sample from a Normal Population*
The joint density of n independent and identically distributed observations from a normal population with mean μ and variance σ^2 is given by

$$f(Y|\mu,\sigma^2) = (2\pi\sigma^2)^{-n/2} \exp\left[-\frac{1}{2}\sum_{i=1}^{n}\left(\frac{y_i-\mu}{\sigma}\right)^2\right].$$

For a given data set Y, the loglikelihood is

$$l(\mu,\sigma^2|Y) = \frac{-n}{2}\log_e \sigma^2 - \frac{1}{2}\sum_{i=1}^{n}\frac{(y_i-\mu)^2}{\sigma^2}.$$

The loglikelihood is a function of μ, σ^2 for fixed Y.

EXAMPLE. *Randomly Censored Data*
Let $T_1, T_2, \ldots, T_n$ be iid with density f, cdf F and survivor function $S = 1 - F$. Thus, $S(t)$ represents the probability of surviving beyond time t. Let $C_1, \ldots, C_n$ be iid with density g and cdf G. T_i is the survival time for the ith individual and C_i is the censoring time associated with T_i. Subjects in a clinical trial may be censored due to loss to follow up (i.e. they move elsewhere), drop out (due to adverse effects the treatment is discontinued), or termination of the study. We observe $(Y_1, \delta_1), \ldots, (Y_n, \delta_n)$, where

$$Y_i = \min(T_i, C_i)$$

and

$$\delta_i = \begin{cases} 1 & T_i \leqslant C_i \\ 0 & T_i > C_i \end{cases}.$$

The pair (y_i, δ_i) has likelihood

$$f(y_i)^{\delta_i}\, S(y_i)^{1-\delta_i}[1 - G(y_i)]^{\delta_i} g(y_i)^{1-\delta_i}$$

and the likelihood of the entire sample is

$$\prod_u f(y_i)\prod_c S(y_i)\prod_c g(y_i)\prod_u[1 - G(y_i)],$$

where "u" denotes the uncensored observations and "c" denotes the censored observations. When the censoring distribution does not involve the unknown lifetime parameters, the last two products can be treated as constants. This is a reasonable assumption in a situation where patients randomly enter a study and a decision is made to terminate the study at a later date independent of the results of the study. The assumption may not be reasonable if patients drop out of the study due to the toxicity of the therapy.

We now define the posterior density. In the Bayesian approach one conditions on the observed data and takes the parameters as random variables. Suppose that θ has a probability distribution $p(\theta)$. Then we have that

$$p(Y\,|\,\theta)p(\theta) = p(Y,\theta) = p(\theta\,|\,Y)p(Y) \ .$$

Conditional on the observed data Y, the distribution of θ is

$$p(\theta\,|\,Y) = \frac{p(Y\,|\,\theta)p(\theta)}{p(Y)} \ . \tag{2.1.1}$$

Expression (2.1.1) is known as Bayes' theorem. Given the data Y, $p(Y\,|\,\theta)$ may be regarded as a function of θ, which we recall is the likelihood function of θ given Y and write as $L(\theta\,|\,Y)$. In this way, we have:

Definition 2.1.3. Given a likelihood $L(\theta\,|\,Y)$ and a prior probability density $p(\theta)$, the posterior density for θ is given as

$$p(\theta\,|\,Y) = cp(\theta)L(\theta\,|\,Y) \ ,$$

where

$$c^{-1} = \int_\Theta p(\theta)L\theta\,|\,Y)d\theta \ .$$

The quantity $p(\theta)$, which describes what is known about θ without knowledge of the data, is called the *prior distribution of* θ. The quantity $p(\theta\,|\,Y)$, which describes what is known about θ given the data Y, is called the *posterior density of* θ.

Example. *Linear Model*
The linear model is given as

$$y = x\theta + \varepsilon$$

where y is an $n \times 1$ vector of observations, θ is $d \times 1$ vector of unknown regression coefficients, x is an $n \times d$ matrix of known constants, ε is an $n \times 1$ vector of independent errors, where $\varepsilon_i \sim N(0, \sigma^2)$ and σ^2 is unknown. In this case,

$$f(Y\,|\,\theta,\sigma^2) = \left(\frac{1}{\sqrt{2\pi}}\right)^n \sigma^{-n} \exp\left[\frac{-1}{2\sigma^2}(y - x\theta)^T(y - x\theta)\right]$$

$$= \left(\frac{1}{\sqrt{2\pi}}\right)^n \sigma^{-n} \exp\left\{\frac{-1}{2\sigma^2}[vs^2 + (\theta - \hat{\theta})^T x^T x(\theta - \hat{\theta})]\right\} ,$$

where

$$\hat{\theta} = (x^T x)^{-1} x^T y, \qquad v = n - d$$

$$\mathrm{s}^2 = (y - \hat{y})^T(y - \hat{y})/v \text{ and } \hat{y} = x\hat{\theta} \ .$$

Note that $\hat{\theta}$ is the least-squares estimate of θ, s^2 is the mean square error due to regression, and $\hat{y}$ is the predicted value of y at x.

We now present three major results relating to the posterior distribution for the linear model. These results will be used throughout the remainder of this book. The first result details the nature of the posterior distribution of θ and σ^2, – i.e. the posterior factors (neatly) into the product of an inverse chi-square distribution and a conditional normal distribution. The second result gives the marginal posterior distribution of θ having integrated out σ^2. The third result gives the posterior distribution of the quadratic form.

Result 2.1.1. Under the prior $p(\sigma^2,\theta) \propto \sigma^{-2}$, $p(\theta,\sigma^2 \mid Y) = p(\sigma^2 \mid s^2)p(\theta \mid \hat{\theta},\sigma^2)$ where the marginal distribution of σ^2 is $vs^2\chi_v^{-2}$ and the conditional distribution of θ, given σ^2, is $N(\hat{\theta},\sigma^2(x^Tx)^{-1})$.

Result 2.1.2. Under the prior $p(\sigma^2,\theta) \propto \sigma^{-2}$

$$p(\theta \mid Y) = \frac{\Gamma\left(\frac{v+d}{2}\right)|x^Tx|^{1/2}s^{-d}}{\{\Gamma(1/2)\}^d\Gamma(v/2)v^{d/2}}\left[1 + \frac{(\theta-\hat{\theta})^Tx^Tx(\theta-\hat{\theta})}{vs^2}\right]^{-(v+d)/2}$$

i.e. the multivariate t distribution.

Result 2.1.3. Under the prior $p(\sigma^2,\theta) \propto \sigma^{-2}$, the posterior distribution of

$$\frac{(\theta-\hat{\theta})^T(x^Tx)(\theta-\hat{\theta})}{ds^2}$$

follows the F distribution with d and v degrees of freedom.

2.2. Maximum Likelihood

Definition 2.2.1. (Little and Rubin, 1987) A maximum likelihood estimate (MLE) of θ is a value of θ that maximizes the likelihood $L(\theta \mid Y)$ or, equivalently, the loglikelihood $l(\theta \mid Y)$.

In situations where the maximum likelihood estimate $(\hat{\theta})$ is unique, $\hat{\theta}$ is the value of θ best supported by the data. Alternatively, $\hat{\theta}$ is the value of θ which makes the observed value of the data most likely. It is noted that if $g(\theta)$ is a 1–1 function of θ and $\hat{\theta}$ is an MLE of θ, then $g(\hat{\theta})$ is an MLE of $g(\theta)$.

Suppose the likelihood is differentiable, unimodal, bounded above, and θ is of dimension d. The MLE is obtained by

1. Differentiating the loglikelihood with respect to θ to obtain the $d \times 1$ vector $\partial l(\theta \mid Y)/\partial\theta$.
2. Setting the d simultaneous *likelihood equations* 0, i.e. $\partial l(\theta \mid Y)/\partial\theta = 0$.
3. Solving for θ.

When the loglikelihood is quadratic, the likelihood equations are linear in θ and the solution is relatively easy to obtain. In more general situations where a closed-form solution of the likelihood equations cannot be found, one may use an iterative numerical method to locate an MLE. One such algorithm, the Newton–Raphson algorithm, is described below.

2.2.1. Newton–Raphson

To motivate the Newton–Raphson algorithm, consider the Taylor series expansion of $l(\theta \mid Y)$ about $\theta^{(i)}$:

$$l(\theta \mid Y) = l(\theta^{(i)} \mid Y) + (\theta - \theta^{(i)})^T \left.\frac{\partial l(\theta \mid Y)}{\partial \theta}\right|_{\theta^{(i)}}$$

$$+ \frac{1}{2}(\theta - \theta^{(i)})^T \left.\frac{\partial^2 l(\theta \mid Y)}{\partial \theta^2}\right|_{\theta^{(i)}} (\theta - \theta^{(i)}) + \text{remainder} \ .$$

When θ is close to $\theta^{(i)}$, the remainder is negligible. The stationary point of the quadratic approximation

$$l(\theta^{(i)} \mid Y) + (\theta - \theta^{(i)})^T \left.\frac{\partial l(\theta \mid Y)}{\partial \theta}\right|_{\theta^{(i)}} + \frac{1}{2}(\theta - \theta^{(i)})^T \left.\frac{\partial^2 l(\theta \mid Y)}{\partial \theta^2}\right|_{\theta^{(i)}} (\theta - \theta^{(i)})$$

is given by

$$\left.\frac{\partial^2 l(\theta \mid Y)}{\partial \theta^2}\right|_{\theta^{(i)}} (\theta - \theta^{(i)}) = -\left.\frac{\partial l(\theta \mid Y)}{\partial \theta}\right|_{\theta^{(i)}}$$

which implies

$$\theta = \theta^{(i)} - \left[\left.\frac{\partial^2 l(\theta \mid Y)}{\partial \theta^2}\right|_{\theta^{(i)}}\right]^{-1} \left.\frac{\partial l(\theta \mid Y)}{\partial \theta}\right|_{\theta^{(i)}} .$$

This suggests the following scheme. Let $\theta^{(0)}$ denote an initial guess of $\hat{\theta}$ and let $\theta^{(i)}$ denote the guess at the ith iteration. The Newton–Raphson algorithm is given as

$$\theta^{(i+1)} = \theta^{(i)} + \left(-\left.\frac{\partial^2 l(\theta \mid Y)}{\partial \theta^2}\right|_{\theta^{(i)}} \right)^{-1} \left[\left.\frac{\partial l(\theta \mid Y)}{\partial \theta}\right|_{\theta^{(i)}} \right] .$$

If the loglikelihood is a quadratic function of θ, then convergence is obtained after one iteration. If the loglikelihood is concave and unimodal then the sequence $\theta^{(1)}, \theta^{(2)}, \ldots$ converges to $\hat{\theta}$. An alternative approach, the method of scoring is given by

$$\theta^{(i+1)} = \theta^{(i)} + \left[E\left(\left. \frac{-\partial^2 l(\theta \mid Y)}{\partial \theta^2} \right| \theta \right) \Bigg|_{\theta^{(i)}} \right]^{-1} \left\{ \left.\frac{\partial l(\theta \mid Y)}{\partial \theta}\right|_{\theta^{(i)}} \right\} .$$

For future reference, we will let $S(\theta \mid Y) = \partial l(\theta \mid Y)/\partial \theta$ and refer to $S(\theta \mid Y)$ as the score function. Cox and Hinkley (1974) present an example where the

MLE is not a solution of the likelihood equations. They present another example, where the MLE is not unique. In general, the loglikelihood may have multiple maxima and/or saddle points.

2.2.2. Examples

a. *Univariate Normal Data*

The loglikelihood is given by

$$l(\mu, \sigma^2 \mid Y) = \frac{-n}{2} \ln \sigma^2 - \frac{1}{2} \sum_{i=1}^{n} \frac{(y_i - \mu)^2}{\sigma^2} .$$

The likelihood equations are given by

$$\frac{\partial l}{\partial \mu} = \sum_{i=1}^{n} y_i - n\mu = 0$$

$$\frac{\partial l}{\partial \sigma^2} = \frac{\sum_{i=1}^{n} (y_i - \mu)^2}{2\sigma^4} - \frac{n}{2\sigma^2} .$$

Hence, $\hat{\mu} = \bar{y}$ and $\hat{\sigma}^2 = \Sigma_{i=1}^{n} (y_i - \bar{y})^2/n$. It remains to show that this stationary point is a maximum using the second derivative of the loglikelihood.

b. *Logistic Regression*

The data in Table 2.1 are taken from Mendenhall et al. (1989). The first column represents the number of days of radiotherapy received by each of 24

Table 2.1.

Days	Response	Days	Response
21	1	51	1
24	1	55	1
25	1	25	0
26	1	29	0
28	1	43	0
31	1	44	0
33	1	46	0
34	1	46	0
35	1	51	0
37	1	55	0
43	1	56	0
49	1	58	0

Mendenhall et al. (1989).

patients. The second column represents the absence (1) or presence (0) of disease at a site three years after treatment. A problem of interest is to use the covariate (days) to predict outcome at three years.

To begin, we wish to fit a logistic regression model (McCullagh and Nelder, 1989) to the data:

$$\log_e\left(\frac{P_i}{1-P_i}\right) = \alpha + \beta x_i$$

where x_i is the covariate for patient i and P_i is the corresponding probability of success (no disease). The logit model implies $P_i = e^{\alpha+\beta x_i}/(1 + e^{\alpha+\beta x_i})$. This model specifies that the log-odds of success is linearly related to the number of days the subject received radiotherapy. The intercept α represents the log-odds of success for zero days, while the slope β represents the change in the log-odds of success for every unit increase in the covariate.

The loglikelihood is given as

$$\sum_{i=1}^{24}\left[\ y_i \log_e(P_i) + (1-y_i)\log_e(1-P_i)\ \right] =$$

$$\sum_{i=1}^{24}\left[\ y_i(\alpha+\beta x_i) - \log_e(1+e^{\alpha+\beta x_i})\right]$$

Note that

$$\frac{\partial l(\theta\,|\,Y)}{\partial\alpha} = \sum_{i=1}^{24}\left(y_i - \frac{e^{\alpha+\beta x_i}}{1+e^{\alpha+\beta x_i}}\right) = \sum_{i=1}^{24}(y_i - P_i)$$

$$\frac{\partial l(\theta\,|\,Y)}{\partial\beta} = \sum_{i=1}^{24}\left(y_i x_i - \frac{x_i e^{\alpha+\beta x_i}}{1+e^{\alpha+\beta x_i}}\right) = \sum_{i=1}^{24} x_i(y_i - P_i)$$

and

$$\frac{\partial^2 l(\theta\,|\,Y)}{\partial\alpha^2} = -\sum_{i=1}^{24}\frac{e^{\alpha+\beta x_i}}{(1+e^{\alpha+\beta x_i})^2} = -\sum_{i=1}^{24} P_i(1-P_i)$$

$$\frac{\partial^2 l(\theta\,|\,Y)}{\partial\alpha\partial\beta} = -\sum_{i=1}^{24}\frac{x_i e^{\alpha+\beta x_i}}{(1+e^{\alpha+\beta x_i})^2} = -\sum_{i=1}^{24} x_i P_i(1-P_i)$$

$$\frac{\partial^2 l(\theta\,|\,Y)}{\partial\beta^2} = -\sum_{i=1}^{24}\frac{x_i^2 e^{\alpha+\beta x_i}}{(1+e^{\alpha+\beta x_i})^2} = -\sum_{i=1}^{24} x_i^2 P_i(1-P_i)\ .$$

Since $\partial^2 l(\theta\,|\,Y)/\partial\theta^2$ does not depend on Y, the Newton–Raphson algorithm and the method of scoring are equivalent for this model.

Implementing the Newton–Raphson algorithm for this data set, we find that $\hat{\alpha} = 3.819$ and $\hat{\beta} = -0.087$ after four iterations from the starting point (0.1,0.1). Over the range of the data, the effect of an additional day of treatment is to *decrease* the odds of success at 3 years by about 8%.

2.3. Normal-Based Inference

Let $\hat{\theta}$ be the MLE of θ based for the data set Y. Then in certain situations

$$(\theta - \hat{\theta}) \sim N(0, C) \tag{2.3.1}$$

where C is the $d \times d$ variance–covariance matrix of $(\theta - \hat{\theta})$.

To frequentists, θ is fixed (though unknown), while $\hat{\theta}$ is random. Given the model and the value θ, (2.3.1) shows that in repeated samples $\hat{\theta}$ will be normally distributed with mean θ and variance–covariance matrix C.

To Bayesians and to likelihoodists $\hat{\theta}$ is fixed (conditional on the data Y) and θ is random. Given the model and the data, (2.3.1) indicates that the posterior density of θ is normal with mean $\hat{\theta}$ and variance–covariance matrix C.

The frequentist justification of (2.3.1) is an expansion of $S(\hat{\theta} \mid Y)$ in a Taylor series about the fixed, but unknown constant, θ. Under some regularity conditions, asymptotically,

$$J(\theta)(\theta - \hat{\theta}) \approx S(\theta \mid Y)$$

where $J(\theta) = E(-\partial^2 l(\theta \mid Y)/\partial\theta^2 \mid \theta)$ and $J(\theta)$ is called the *expected Fisher information matrix*. Thus, $(\hat{\theta} - \theta) \approx J^{-1}(\theta)S(\theta \mid Y)$. Under certain regularity conditions, in repeated samples $S(\theta \mid Y)$ is asymptotically normal with mean 0 and variance–covariance matrix $J(\theta)$. It follows that $(\theta - \hat{\theta})$ is normally distributed with mean 0 and variance–covariance matrix $C = J^{-1}(\theta)$. We evaluate $J^{-1}(\theta)$ at $\hat{\theta}$ and ignore the resulting error.

It is important to note that $J^{-1}(\theta)$ is the variance of the asymptotic distribution, not the limit of the exact variance.

The Bayesian justification of (2.3.1) is an expansion of the loglikelihood (log-posterior) about the fixed value $\hat{\theta}$ (conditional on Y),

$$l(\theta \mid Y) = l(\hat{\theta} \mid Y) + (\theta - \hat{\theta})^T S(\hat{\theta} \mid Y) - \tfrac{1}{2}(\theta - \hat{\theta})^T I(\hat{\theta} \mid Y)(\theta - \hat{\theta}) + r(\theta \mid Y)$$

where $I(\hat{\theta} \mid Y) = -\partial^2 l(\theta \mid Y)/\partial\theta^2 \mid_{\hat{\theta}}$ and $I(\theta \mid Y)$ is called the *observed Fisher information matrix*. Since $S(\hat{\theta} \mid Y) = 0$, assuming that $r(\theta \mid Y)$ can be neglected, the likelihood/posterior density is proportional to the multivariate normal density with mean $\hat{\theta}$ and variance–covariance matrix $C = I^{-1}(\hat{\theta} \mid Y)$.

Based on the normal approximation, a $100\alpha\%$ confidence interval for θ (scalar) is given by $\hat{\theta} \pm zC^{1/2}$ where z is the value such that the area under the normal distribution from $-z$ to $+z$ is $100\alpha\%$, and C is either $I^{-1}(\hat{\theta} \mid Y)$ or $J^{-1}(\hat{\theta})$.

For a vector θ of dimension d, a $100\alpha\%$ confidence ellipsoid is given by

$$(\theta - \hat{\theta})^T C^{-1}(\theta - \hat{\theta}) \leqslant \chi^2$$

where χ^2 is the value such that the area under the chi-square distribution of d degrees of freedom from 0 to χ^2 is $100\alpha\%$.

Efron and Hinkley (1978) discuss the relative merits of I^{-1} and J^{-1}. In some sense, I^{-1} is preferred since it gives a better indication of the curvature of the *observed* loglikelihood. In some situations, I^{-1} may be easier to compute than J^{-1}.

Example. *Exponential Lifetimes With Random Censoring*
In this case, suppose $T_1, \ldots, T_n$ are iid with exponential density $f(t) = \theta e^{-\theta t}$ and survivor function $S(t) = e^{-\theta t}$, $(t > 0)$. Let n_μ be the number of uncensored observations in the sample. The loglikelihood is given as

$$l(\theta \mid Y) = n_\mu \log_e \theta - \theta \sum_{i=1}^{n} y_i$$

where y_i is the observed event time. The score function is

$$\frac{n_\mu}{\theta} - \sum_{i=1}^{n} y_i,$$

which implies that $\hat{\theta} = n_\mu / \Sigma_{i=1}^n y_i$. Note that $-\partial^2 l / \partial \theta^2 = n_\mu / \theta^2 = I(\theta \mid Y)$, while $J(\theta) = n / \theta^2 p(T < C)$. Hence, a 95% confidence interval for θ is given by

$$\hat{\theta} \pm 1.96 \sqrt{\frac{\hat{\theta}^2}{n_\mu}} .$$

Example. *Logistic Regression (Continued)*
For the logistic regression data in Section 2.2, the upper-left entry in $I^{-1}(\hat{\alpha}, \hat{\beta} \mid Y)$ is 3.345, while the lower-right entry is 0.00185. Hence, we have standard error of $\hat{\alpha} = 1.829$, and standard error of $\hat{\beta} = 0.043$. A 95% confidence interval for β obtained via $\hat{\beta} \pm 1.96 \times \mathrm{SE}(\hat{\beta})$ is $(-0.171, -0.003)$. Thus, because the interval does not overlap zero, there appears to be evidence of a day effect on the odds of success at 3 years.

2.4. The δ-Method (Propagation of Errors)

Suppose that the distribution of $\theta - \hat{\theta}$ can be approximated by a normal distribution with mean 0 and variance–covariance matrix B. Let f be a function defined on an open subset of a d-dimensional space taking values in an r-dimensional space and let f be differentiable at θ. The approximating distribution to $f(\hat{\theta}) - f(\theta)$ is the normal distribution with mean 0 and variance–covariance matrix $(\partial f / \partial \theta)^T B (\partial f / \partial \theta)$.

Example. *Asymptotic Variance of the Log Odds Ratio*
Consider the log odds ratio for a multinomially distributed 2 × 2 table with observed proportions P_{ij}, which is given by

$$\ln\left(\frac{P_{11} P_{22}}{P_{21} P_{12}}\right) = \ln(P_{11}) + \ln(P_{22}) - \ln(P_{21}) - \ln(P_{12}) .$$

Note that $B = (D - P^T P)/N$, where D is the diagonal matrix with vector $P = (P_{11}, P_{22}, P_{12}, P_{21})$ along its main diagonal and N is the total sample size. Also note that $\partial f/\partial P = CD^{-1}$, where $C = [1, 1, -1, -1]$. The variance of the log odds ratio is then given by

$$\begin{aligned} &[CD^{-1}(D - P^T P)D^{-1}C^T]/N \\ &= [CD^{-1}C^T - (CD^{-1}P^T)PD^{-1}C^T]/N \\ &= \left(\frac{1}{P_{11}} + \frac{1}{P_{22}} + \frac{1}{P_{12}} + \frac{1}{P_{21}}\right)\Big/ N \, . \end{aligned}$$

See Bishop, Fienberg and Holland (1975) for further examples of the use of the δ-method.

2.5. Highest Posterior Density Regions

From the Bayesian perspective, inferential problems about the parameter vector θ can be addressed via the posterior density $p(\theta \mid Y)$. However, when θ is of even moderate dimension, e.g. six, it may be difficult to grasp its shape. Thus, methods which summarize the structure of the posterior are of value. The Highest Posterior Density region attempts to capture a comparatively small region of the parameter space which contains most of the mass of the posterior distribution. Intuitively, the HPD region is motivated as the region where the probability density of every point inside the region is at least as large as that of any point outside the region. More formally, we have:

Definition 2.5. (Box and Tiao, 1973) A region R in parameter space Θ is called the highest posterior density region of content α if

a. $\Pr(\theta \in R \mid Y) = \alpha$

b. For $\theta_1 \in R$, $\theta_2 \notin R$, $p(\theta_1 \mid Y) \geqslant p(\theta_2 \mid Y)$.

Thus, consider the two regions R and R' (see Figure 2.1). R is an HPD region of content α, while R' is not.

Box and Tiao (1973) point out that for a given probability content α, the HPD region has smallest volume in parameter space. In addition, θ_0 is covered by the HPD region of content α iff

$$P_\theta[p(\theta \mid Y) \geqslant p(\theta_0 \mid Y) \mid Y] \leqslant \alpha$$

where $p(\theta \mid Y)$ is treated as a random variable.

Example (*Linear Model*):
Consider the linear model given as

$$y = x\theta + \varepsilon, \, .$$

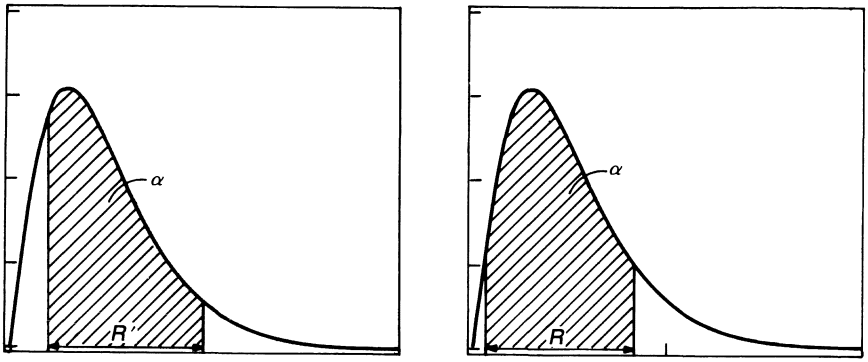


Figure 2.1.

where y is an $n \times 1$ vector of observations, x is an $n \times d$ matrix of known constants, θ is a $d \times 1$ vector of unknown regression coefficients, ε is an $n \times 1$ vector of independent errors, where $\varepsilon_i \sim N(0, \sigma^2)$ and σ is unknown.

If we define $Q(\theta) = (\theta - \hat{\theta})^T(x^T x)(\theta - \hat{\theta})$, then

a. The posterior of $Q(\theta)/ds^2$ is $F(d, v)$, where $s^2 = (y - \hat{y})^T(y - \hat{y})/v$, $v = n - d$, and $\hat{y} = x\hat{\theta}$.

b. The posterior density $p(\theta \mid Y)$ is a monotonically decreasing function of $Q(\theta)$.

In this way, the content of the smallest HPD region containing θ_0 is equal to

$$\begin{aligned} &P_\theta[p(\theta \mid Y) \geqslant p(\theta_0 \mid Y) \mid Y] \\ &= P_\theta[(\theta_0 - \hat{\theta})^T x^T x(\theta_0 - \hat{\theta}) \geqslant (\theta - \hat{\theta})^T x^T x(\theta - \hat{\theta}) \mid Y] \\ &= \Pr\left[\frac{(\theta_0 - \hat{\theta})^T x^T x(\theta_0 - \hat{\theta})}{ds^2} \geqslant F(d, v)\right] \\ &= 1 - \text{frequentist significance level for testing } \theta = \theta_0. \end{aligned}$$

In this small sample context, the Bayesian (under a noninformative prior) and the frequentist approaches yield the same results. Box and Tiao (1973) point out that these results supply a Bayesian justification for the analysis of variance. More generally, Wasserman (1989) provides a Bayesian interpretation to the likelihood region. He proves that likelihood regions are robust in the sense that their posterior probability content is relatively insensitive to contaminations of the prior.

CHAPTER 3

Nonnormal Approximations to Likelihoods and to Posteriors

In this chapter we continue our presentation of methods which are applied directly to the likelihood or to the posterior. However, in this chapter we work with higher-order approximations to these functions. In Section 3.1, we discuss conjugate priors' as well as the role of numerical integration as methods to obtain an exact or close approximation to the marginal of the posterior or of the likelihood. Section 3.2 approaches the problem of higher-order approximations to the posterior or to the likelihood from the point of view of Laplace's method. Section 3.3 presents the methods of Monte Carlo, importance sampling and rejection/acceptance to realize a sample from the function of interest. Iterative Monte Carlo methods are presented in chapters 5 and 6.

3.1. Conjugate Priors and Numerical Integration

In the previous chapter, we approximated the posterior or the likelihood by a normal distribution and then used the normal distribution as our basis of inference. Unfortunately, for a given model and a given data set, the normal approximation may be quite poor and may lead to inappropriate inferences. Alternatively, we can integrate out nuisance parameters and use the resulting marginal posterior or likelihood as the basis of inference.

To obtain the marginal density of $(\theta_1, \ldots, \theta_l)$, where $l < d$, we integrate over $(\theta_{l+1}, \ldots, \theta_d)$ to obtain

$$p(\theta_1, \ldots, \theta_l | Y) = \int p(\theta_1, \ldots, \theta_d | Y)\, d\theta_{l+1} \ldots d\theta_d \, .$$

More generally, we may be interested in evaluating the integral

$$\int q(\theta)p(\theta|Y)\,d\theta_I \tag{3.1.1}$$

where I is the complement of the subscripts of the components of θ of interest in $(1, 2, \ldots, d)$.

If the likelihood is based on a random sample of size n from an exponential family distribution with density

$$g(\theta)h(y)\exp\left\{\sum_{j=1}^{m} \phi_j(\theta)t_j(y)\right\},$$

then a conjugate prior density of the form

$$p(\theta) \propto [g(\theta)]^b \exp\left\{\sum_{j=1}^{m} \phi_j(\theta)a_j\right\} \tag{3.1.2}$$

combines with the likelihood to yield a posterior density having the same form as (3.1.2), but with $b, a_1, \ldots, a_m$ replaced by

$$b' = b + n, \qquad a_j' = a_j + \sum_{i=1}^{n} t_j(y_i), \qquad j = 1, \ldots, m\ .$$

As noted in Smith et al. (1985), there are two limitations to this approach. First, only a limited number of sampling models belong to the exponential family. Second, even when working in an exponential family framework, it may be unrealistic to attempt to represent prior information in the conjugate family form.

As an alternative approach, Smith et al. (1985) and Naylor and Smith (1982) discuss their experience with methods from the numerical analysis literature to integrate out numerically the nuisance parameters. Their approach is to use Cartesian product Gaussian quadrature rules (Davis and Rabinowitz, 1984) for multivariate parameter vectors. Such an approach assumes that the posterior density may be well approximated by the product of a multivariate normal distribution and a polynomial in $\theta_1, \ldots, \theta_k$ and that the parameters are nearly orthogonal. To justify the latter condition, Naylor and Smith (1982) transform the component parameters to a new, orthogonal set of parameters.

More formally, in the one-dimensional case, if $h(t)$ is a suitably regular function and

$$g(t) = h(t)(2\pi\sigma^2)^{-1/2}\exp\left\{-\frac{1}{2}\left[\frac{t-\mu}{\sigma}\right]^2\right\},$$

then Naylor and Smith (1982) show that

$$\int_{-\infty}^{\infty} g(t)\,dt \approx \sum_{i=1}^{n} m_i g(z_i)\ , \tag{3.1.3}$$

where

$$m_i = w_i \exp(t_i^2)\sqrt{2}\sigma,\ z_i = \mu + \sqrt{2}\sigma t_i\ . \tag{3.1.4}$$

To implement this approach, tables of t_i, w_i and $w_i \exp(t_i^2)$ are required (Naylor and Smith, 1982). The error in the approximation will be small if $h(t)$ is approximately a polynomial. Maximum likelihood values make reasonable guesses to μ and σ in (3.1.4). In the case of a k-dimensional parameter vector, a simple extension of (3.1.3) is

$$\int \ldots \int g(t_1, \ldots, t_k)\mathrm{d}t_1 \cdots dt_k \approx \sum_{i_k} m_{i_k}^{(k)} \cdots \sum_{i_2} m_{i_2}^{(2)} \sum_{i_1} m_{i_1}^{(1)} g(z_{i_1}^{(1)} \cdots z_{i_k}^{(k)}) \ ,$$

where $m_{i_j}^{(j)}, z_{i_j}^{(j)}$ are found via (3.1.4), but estimates of the marginal posterior mean and variance of θ_i substituted for μ and σ^2.

Example. *Logistic Regression* (*Continued*)
Figure 3.1 presents the β marginal (solid line) of the likelihood obtained via numerical integration for the logistic regression data of Section 2.2. In particular, if $L(\alpha, \beta | Y)$ is the likelihood, then to obtain $L(\beta^* | Y)$ (i.e. the

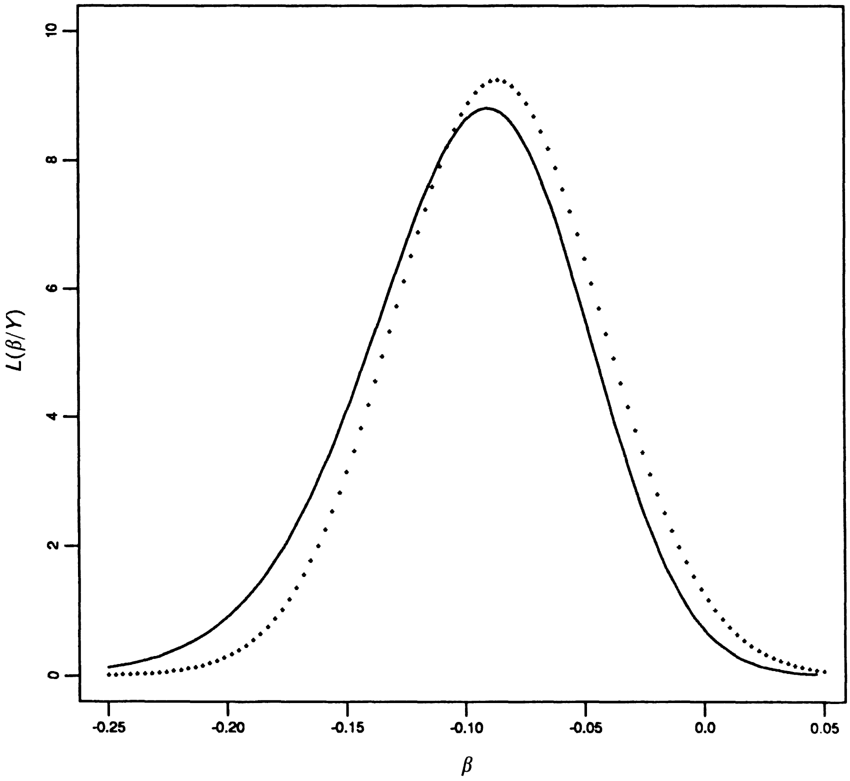


Figure 3.1. The beta marginal (solid: exact; dotted: normal approximation). The curves are normalized to integrate to unity.

β marginal evaluated at the point β^*) we used the trapezoidal rule:

$$\frac{\alpha_m - \alpha_0}{m}\left[\frac{1}{2}L(\alpha_0, \beta^* \mid Y) + L(\alpha_1, \beta^* \mid Y) + \cdots + \frac{1}{2}L(\alpha_m, \beta^* \mid Y)\right],$$

where $\alpha_0, \ldots, \alpha_m$ defines a grid of equispaced points along the α axis. This computation was repeated for β values over $[-0.25, 0.05]$ to obtain $L(\beta^{(1)} \mid Y)$, $L(\beta^{(2)} \mid Y), \ldots, L(\beta^{(n)} \mid Y)$. The trapezoidal rule was then used to obtain the normalizing constant to yield Figure 3.1. The dotted line represents the normal approximation to the marginal. As can be seen, the true marginal of β is slightly skewed, which is not detected in the normal approximation. The difference is minor in terms of the mode (new mode = -0.088). However, the exact HPD region of content 95% is $(-0.185, -0.015)$, rather than $(-0.171, -0.002)$ based on the normal approximation to the likelihood.

Figure 3.2 presents the $-\alpha/\beta$ (i.e. TD_{50}) marginal. The solid line represents the normal approximation obtained via the δ-method. The dotted curve was obtained by reparameterizing the likelihood in terms of $\rho(= -\alpha/\beta)$ and

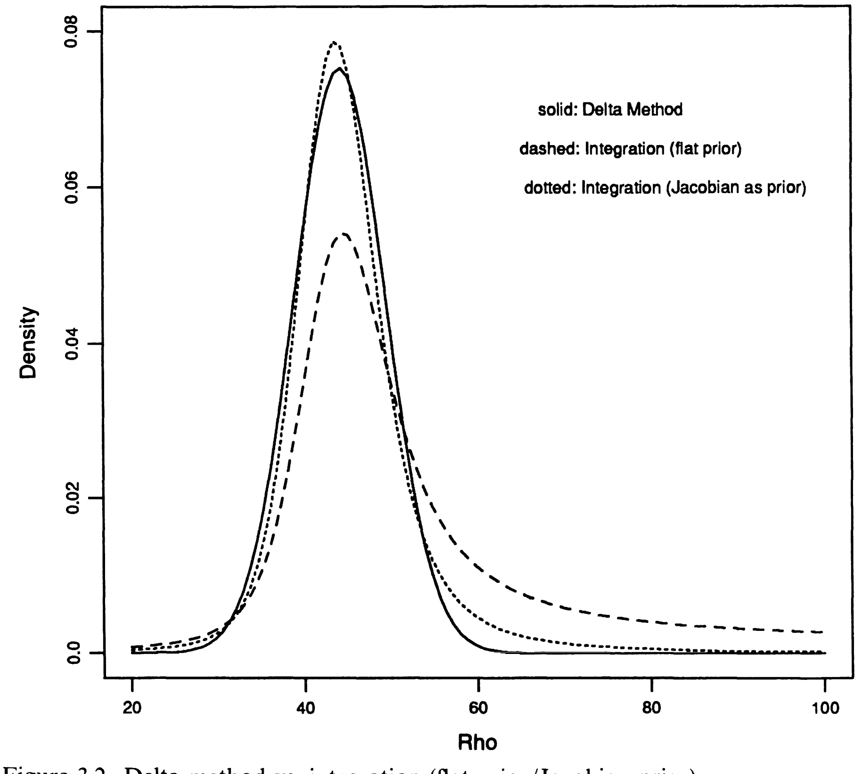


Figure 3.2. Delta method vs. integration (flat prior/Jacobian prior).

α and then numerically integrating out α. In this case, the prior on ρ was the Jacobian of the transformation. In other words, the dotted curve was obtained via:

$$\int L(\rho, \alpha \mid Y) |\alpha| / \rho^2 \, d\alpha \ .$$

The dashed line represents the marginal having adopted a flat prior on ρ, i.e. via:

$$\int L(\rho, \alpha \mid Y) \, d\alpha \ .$$

As can be seen, the mode of each curve is located at about 44 days. The integrated curve (Jacobian prior) is slightly more skewed than the normal approximation. There is evidence that the integrated curve (with a flat prior) is not proper, i.e. has infinite mass. In particular, evaluating the marginal at $-\alpha/\beta = 200, 300, \ldots, 2000$ indicates that the marginal is flat. In this regard, it is noted that the dashed curve in Fig. 3.2 was truncated at $-\alpha/\beta = 100$ and scaled to integrate to 1.0. Truncating at a larger value would scale the curve downward. For related work, see Buonaccorsi and Gatsonis (1988).

3.2. Posterior Moments and Marginalization Based on Laplace's Method

3.2.1. Moments

Suppose $-h(\theta)$ is a smooth, bounded unimodal function, with a maximum at $\hat{\theta}$, and θ is a scalar. By Laplace's method (Tierney and Kadane, 1986), the integral

$$I = \int f(\theta) \exp[-nh(\theta)] \, d\theta$$

can be approximated by

$$\hat{I} = f(\hat{\theta}) \sqrt{\frac{2\pi}{n}} \, \sigma \exp[-nh(\hat{\theta})] \ ,$$

where

$$\sigma = \left[\left. \frac{\partial^2 h}{\partial \theta^2} \right|_{\hat{\theta}} \right]^{-1/2} .$$

As presented in Mosteller and Wallace (1964), Laplace's method is to expand about $\hat{\theta}$ to obtain:

$$I \approx \int f(\hat{\theta}) \exp\left\{ -n \left[h(\hat{\theta}) + (\theta - \hat{\theta}) h'(\hat{\theta}) + \frac{(\theta - \hat{\theta})^2}{2} h''(\hat{\theta}) \right] \right\} d\theta \ .$$

Recalling that $h'(\hat{\theta}) = 0$, we have

$$\begin{aligned} I &\approx \int f(\hat{\theta}) \exp\left\{ - n\left[h(\hat{\theta}) + \frac{(\theta - \hat{\theta})^2}{2} h''(\hat{\theta}) \right] \right\} d\theta \\ &= f(\hat{\theta}) \exp[- nh(\hat{\theta})] \int \exp\left[\frac{- n(\theta - \hat{\theta})^2}{2\sigma^2} \right] d\theta \\ &= f(\hat{\theta}) \sqrt{\frac{2\pi}{n}}\, \sigma \exp[- nh(\hat{\theta})] \ . \end{aligned}$$

Intuitively, if $\exp[- nh(\theta)]$ is very peaked about $\hat{\theta}$, then the integral can be well approximated by the behavior of the integrand near $\hat{\theta}$. More formally, it can be shown that

$$I = \hat{I} \left\{ 1 + O\left(\frac{1}{n} \right) \right\} \ .$$

To calculate moments of posterior distributions, we need to evaluate expressions such as:

$$E\{ g(\theta) \} = \frac{\int g(\theta) \exp\{ - nh(\theta)\} \, d\theta}{\int \exp\{ - nh(\theta)\} \, d\theta} \ . \qquad (3.2.1)$$

Tierney and Kadane (1986) provide two approximations to $E\{g(\theta)\}$. The term n refers to the sample size.

Result 3.2.1. $E\{g(\theta)\} = g(\hat{\theta})[1 + O(1/n)]$

To see this, apply Laplace's method to the numerator of (3.2.1) with $f = g$, to obtain the approximation

$$g(\hat{\theta}) \sqrt{\frac{2\pi}{n}}\, \sigma \exp[- nh(\hat{\theta})] \ .$$

Next, apply Laplace's method to the denominator of (3.2.1) with $f = 1$, to obtain the approximation

$$\sqrt{\frac{2\pi}{n}}\, \sigma \exp[- nh(\hat{\theta})] \ .$$

Tierney and Kadane (1986) show that the resulting ratio $g(\hat{\theta})$ has error $O(1/n)$. Mosteller and Wallace (1964) present related results.

Result 3.2.2. $E\{g(\theta)\} = \dfrac{\sigma^*}{\sigma} \{\exp[- nh^*(\theta^*)]\}/\{\exp[- nh(\hat{\theta})]\}$
$\times [1 + O(1/n^2)]$

To see this, first apply Laplace's method to the numerator of (3.2.1) with $f = 1$, g positive, and $- nh^*(\theta) = - nh(\theta) + \log(g(\theta))$, where θ^* is the mode of $- h^*(\theta)$. Next, apply Laplace's method to the denominator with $f = 1$.

Tierney and Kadane (1986) show that the resulting ratio has error $O(1/n^2)$. Again, Mosteller and Wallace (1964) present related results.

For multivariate θ, $\Sigma^* = \left[\left.\frac{\partial^2 h^*}{\partial\theta^2}\right|_{\theta^*}\right]^{-1}$, $\Sigma = \left[\left.\frac{\partial^2 h}{\partial\theta^2}\right|_{\hat{\theta}}\right]^{-1}$ and

$$E[g(\theta)] = \left(\frac{\det\Sigma^*}{\det\Sigma}\right)^{1/2} \frac{\exp[-nh^*(\theta^*)]}{\exp[-nh(\hat{\theta})]}\left\{1 + O\left(\frac{1}{n^2}\right)\right\}.$$

We will return to these results in Section 5.6 when we discuss poor man's data augmentation.

3.2.2. Marginalization

Partition the $d \times 1$ vector θ as (θ_1, θ_2), where θ_1 is a scalar and θ_2 is $(d-1)\times 1$. Let $\hat{\theta}$ maximize the posterior $p(\theta|Y)$. For a given value of θ_1, let $\hat{\theta}_2^* = \hat{\theta}_2^*(\theta_1)$ maximize $p(\theta_1, \theta_2|Y)$ with θ_1 fixed, and let

$$\hat{\Sigma}^* = \left[\left.\frac{-\partial^2 \log p(\theta_1, \theta_2|Y)}{\partial\theta_2^2}\right|_{\hat{\theta}_2^*}\right]^{-1}$$

which is a $(d-1)\times(d-1)$ matrix. Tierney and Kadane (1986) approximate $p(\theta_1|Y)$ with

$$\hat{p}(\theta_1|Y) \propto (\det\hat{\Sigma}^*)^{1/2} p(\theta_1, \hat{\theta}_2^*|Y) = (\det\hat{\Sigma}^*)^{1/2}\ (\text{profile posterior of } \theta_1)\,.$$

To obtain the marginal of an arbitrary function $g(\theta)$, Tierney, Kass and Kadane (1989) approximate $p(\gamma|Y)$ with

$$\hat{p}(\gamma|Y) \propto \left[\frac{\det[\Sigma(\gamma)]}{\det[(D_g)^T\Sigma(\gamma)(D_g)]}\right]^{1/2} p[\hat{\theta}(\gamma)|Y]\ ,$$

where $\hat{\theta}(\gamma)$ maximizes the posterior $p(\theta|Y)$ subject to the constraint that $g(\theta) = \gamma$, $\Sigma(\gamma)$ is the inverse Hessian of the log posterior evaluated at $\hat{\theta}(\gamma)$ and D_g is the gradient or Jacobian of g evaluated at $\hat{\theta}(\gamma)$. This approximation may be poor in small samples. See Leonard, Hsu and Tsui (1989), as well as Wong and Li (1992) for alternative approximations.

Example. *Stanford Heart Transplant Data*

Turnbull, Brown and Hu (1974) proposed a Pareto model for data from the Stanford Heart Transplant Program. Each patient has an exponential lifetime (i.e. constant hazard) upon entering the program, with the hazard rate varying from patient to patient according to a gamma distribution with parameters p and λ. In this way, the marginal lifetime density is Pareto. A transplant, if it occurs, multiplies the patient's hazard rate by a factor τ, resulting in the likelihood

$$\prod_{i=1}^{n}\frac{p\lambda^p}{(\lambda + x_i)^{p+1}}\prod_{i=n+1}^{N}\left(\frac{\lambda}{\lambda + x_i}\right)^p\prod_{j=1}^{m}\frac{\tau p\lambda^p}{(\lambda + y_j + \tau z_j)^{p+1}}\prod_{j=m+1}^{M}\left(\frac{\lambda}{\lambda + y_j + \tau z_j}\right)^p$$

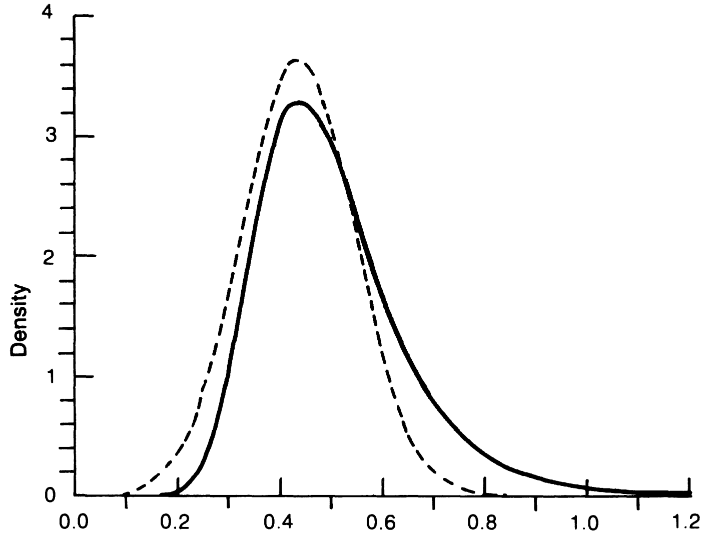


Figure 3.3. Marginal posterior densities for p. (—) Laplace and 20-point adaptive Gauss–Hermite approximations; (· · ·) asymptotic normal approximation (Tierney and Kadane 1986, Figure 1).

where x_i is the survival time in days for $N = 30$ nontransplant patients, of which $n = 26$ died and 4 were censored, and y_j, z_j are the time to transplant and survival time (beyond transplant), respectively, for the transplant patients ($M = 52$) of whom ($m = $)34 died and 18 were censored.

Figure 3.3 presents the marginal density of p. The solid line is based on the Laplace approximation, while the dashed line is the asymptotic normal approximation.

Superimposed on the solid line is the marginal density obtained by numerical integration (Gauss–Hermite quadrature, see Naylor and Smith, 1982). There is a noticeable discrepancy between the normal approximation and other approaches. Note that the marginal is somewhat skewed, even though the total sample size is moderate.

Example. *Logistic Regression* (*Continued*)

Figure 3.4 presents the β marginal (starred line) of the likelihood obtained via the Tierney and Kadane (1986) marginalization procedure. The solid line represents the marginal of the likelihood obtained via numerical integration in Section 3.1. The dotted line is the normal approximation. There is a slight discrepancy between the normal approximation and the other approaches. However, the Laplace approach and the numerically integrated marginal are quite congruent. The Tierney and Kadane (1986) marginal for β was obtained

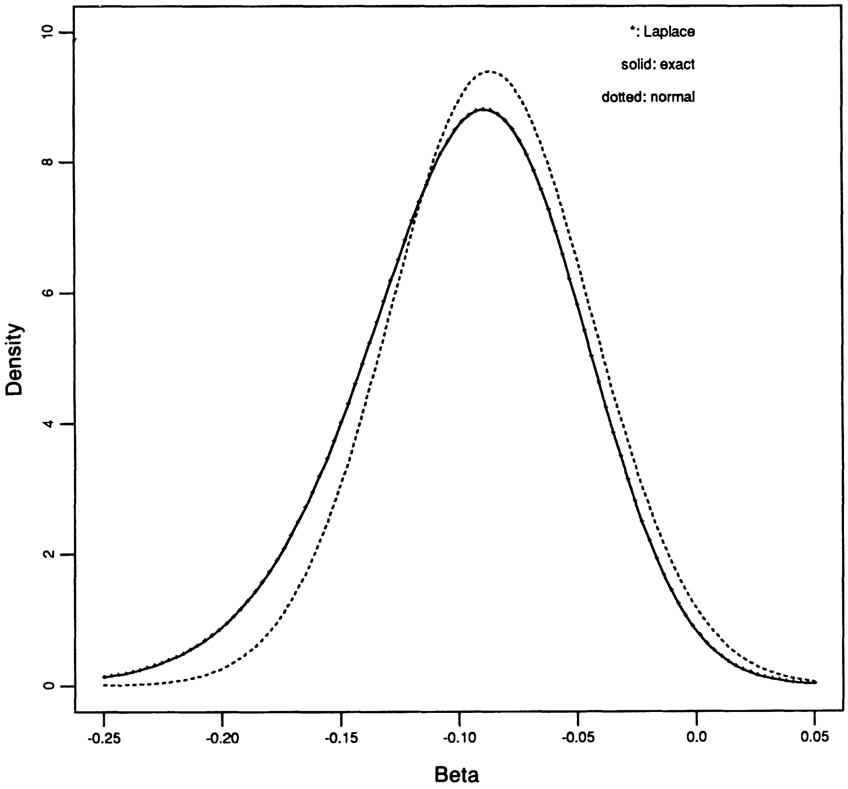


Figure 3.4. Comparison: exact marginal, normal approximation and Laplace approximation.

by calculating the profile likelihood for β (weighted by $\hat{\Sigma}^*$) on a grid of points between -0.25 and 0.05. The curve was then normalized to integrate to unity by calculating the area under this curve and then dividing through by this constant. The trapezoidal rule was used to calculate the normalizing constant.

Figure 3.5 presents the $-\alpha/\beta$ marginal. The starred line was obtained by applying the Tierney and Kadane (1986) marginalization procedure to the likelihood parameterized in terms of α and ρ, where $\rho = -\alpha/\beta$, with a flat prior on ρ. The solid line represents the marginal obtained via numerical integration with a flat prior on ρ. The normal approximation (δ-method) is given by the dotted line. The Laplace approach and numerical integration yield virtually identical curves in this example.

It is noted that in this case we have applied the Tierney and Kadane (1986) marginalization procedure to the reparameterized likelihood $L(\alpha, \rho)$ supposing a flat prior on ρ, thus leading to an improper posterior (the improper

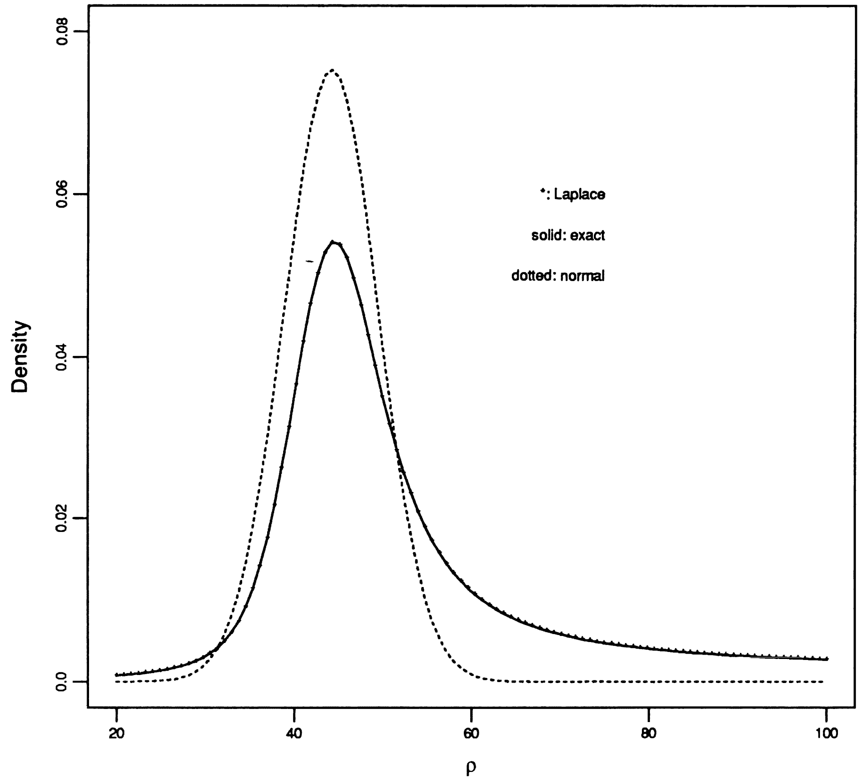


Figure 3.5. Comparison: exact marginal, normal approximation, Laplace approximation.

posterior in the above figure was truncated at $\rho = 100$). The Tierney, Kass and Kadane (1989) approach would assign the prior $|\alpha|/\rho^2$, the Jacobian of the transformation, to ρ and would result in a more symmetric curve, which is closer to the delta method (normal) solution.

3.3. Monte Carlo Methods

3.3.1. Monte Carlo

Consider the problem of approximating the integral

$$J(y) = \int f(y|x)g(x)\,dx = E_g[f(y|x)] < +\infty,$$

where y and x may be vectors.

If $g(x)$ is a density, then the Monte Carlo method (after the famous casino in Monaco) approximates $J(y)$ as

$$\hat{J}(y) = \frac{1}{n} \sum_{i=1}^{n} f(y|x_i) \ ,$$

where $x_1, \ldots, x_n \overset{\text{iid}}{\sim} g(x)$. By the Law of Large Numbers,

$$\hat{J}(y) \overset{\text{a.s}}{\rightarrow} J(y)$$

(Geweke, 1989). The estimated Monte Carlo standard error of $\hat{J}(y)$ is

$$\frac{1}{\sqrt{n}} \sqrt{\frac{\sum_{i=1}^{n} [f(y|x_i) - \hat{J}(y)]^2}{n-1}} \ .$$

Extensive use will be made of the method of Monte Carlo when we consider the data augmentation algorithm (Chapter 5).

Example. *Area of the Quarter Circle*
Let $f(y|x) = \sqrt{1 - x^2}$ and let $g(x)$ be the density of the uniform distribution on $[0, 1]$, i.e. we wish to compute one-quarter of the area of the unit circle ($\pi/4 = 0.7854\ldots$)

Five thousand pseudo-random uniform deviates were generated using the RANDOM command in MINITAB. $\hat{J}$ was then taken to be

$$\hat{J} = \frac{1}{5000} \sum_{i=1}^{5000} \sqrt{1 - x_i^2} = 0.7865 \ .$$

The standard error of $\hat{J}$ is

$$\sqrt{\frac{2/3 - \left(\frac{\pi}{4}\right)^2}{5000}} = 0.0032 \ .$$

There are a number of techniques which can improve the precision of the Monte Carlo estimate. A general discussion of variance-reduction techniques, such as antithetic variates and control variates, is given in Rubenstein (1981).

3.3.2. Composition

Suppose $f(y|x)$ is a density where x and y may be vectors. To obtain a sample $y_1, \ldots, y_m \overset{\text{iid}}{\sim} J(y) = \int f(y|x)g(x)\,dx$, one may use the method of composition:

1. Draw $x^* \sim g(x)$.
2. Draw $y^* \sim f(y|x^*)$.

Steps 1 and 2 are repeated m times. The pairs $(x_1, y_1), \ldots, (x_m, y_m)$ are an iid sample from the joint density $h(x, y) = f(y|x)g(x)$, while the quantities $y_1, \ldots, y_m$ are an iid sample from the marginal $J(y)$.

When x is a discrete random variable, taking on values 0, 1, 2, ..., select an integer (i) with probability $g(i)$ and draw y^* from $f_i(y)$. The resulting deviate is an observation from the distribution with density

$$\sum_{j=1}^{\infty} f_j(y)g(j) \ .$$

When x takes on a finite number of values, n, select an integer (i) with probability $1/n$, draw y^* from $f_i(y)$ and then assign mass $g(i)$ to y^*. The resulting deviate (with assigned mass $g(i)$) is drawn from the distribution with density

$$\sum_{j=1}^{n} f_j(y)g(j) \ .$$

The method of composition is a key technique underlying the data augmentation algorithm of Chapter 5.

Example. *Predictive Distribution*
Let $Y = (y_1, \ldots, y_n)$ be the observed data and let y_f denote a (univariate) future observation. Then the predictive distribution is given by

$$p(y_f | Y) = \int p(y_f | Y, \theta) p(\theta | Y) d\theta \ .$$

To compute an approximation to $p(y_f | Y)$ via the Monte Carlo method:

1. Draw a sample $\theta_1, \ldots, \theta_m \overset{\text{iid}}{\sim} p(\theta | Y)$.
2. Let

$$\hat{p}(y_f | Y) = \frac{1}{m} \sum_{i=1}^{m} p(y_f | Y, \theta_i) \ .$$

To sample from $p(y_f | Y)$, apply the method of composition:

1. Draw θ^* from $p(\theta | Y)$.
2. Draw y^* from $p(y_f | Y, \theta^*)$.

Steps 1 and 2 are repeated m times to yield the desired sample. This sample may then be used to compute the posterior distribution of any functional of the future observation.

As an example of a predictive distribution, consider the linear model

$$y = x\theta + \varepsilon$$

where y is an $n \times 1$ vector of associations, x is an $n \times d$ matrix of known constants, θ is a $d \times 1$ vector of unknown regression coefficients, and ε is an

$n \times 1$ vector of errors, where $\varepsilon_i \sim N(0, \sigma^2)$ and σ is unknown. As noted in Sections 2.1 and 2.5, under the prior $p(\theta, \sigma^2) \propto \sigma^{-2}$,

$$p(\theta, \sigma^2 | Y) = p(\sigma^2 | s^2) p(\theta | \hat{\theta}, \sigma^2) \ ,$$

where the marginal distribution of σ^2 is given by $\nu s^2 \chi_\nu^{-2}$ ($\nu = n - d$ and $s^2 = (y - \hat{y})(y - \hat{y})/\nu$). The conditional marginal distribution of θ given σ^2 is

$$N[\hat{\theta}, \sigma^2 (x^T x)^{-1}]$$

where $\hat{y} = x\hat{\theta}$ and $\hat{\theta} = (x^T x)^{-1} x^T y$. Sampling from $p(\sigma^2 | s^2)$ requires the generation of a chi-square deviate, while sampling from $p(\theta | \hat{\theta}, \sigma^2)$ requires the generation of a multivariate normal deviate conditional on the generated value of σ^2.

Note that in this case, $p(y_f | Y, \sigma^2, \theta)$ is the normal distribution with mean $x_f \theta$ and variance σ^2. We can approximate

$$p(y_f | Y) = \int p(y_f | Y, \sigma^2, \theta) p(\theta, \sigma^2 | Y) d\theta \, d\sigma^2$$

via the Monte Carlo method. First draw an observation from $p(\theta, \sigma^2 | Y)$:

1a. Draw σ_*^2 from $p(\sigma^2 | s^2)$,
1b. Draw θ^* from $p(\theta | \hat{\theta}, \sigma_*^2)$.

Steps (1a) and (1b) are repeated to obtain a sample $(\theta_1, \sigma_1^2), \ldots, (\theta_m, \sigma_m^2)$ from $p(\theta, \sigma^2 | Y)$. The approximation $\hat{p}(y_f | Y)$ is then given by

$$\frac{1}{m} \sum_{i=1}^{m} \phi(y_f | x_f \theta_i, \sigma_i^2) \ ,$$

where $\phi(x | a, b)$ is the (univariate) normal density function with mean a and variance b, i.e. $\exp[-\frac{1}{2}(x - a)^2 / b] / \sqrt{2\pi b}$.

To draw σ_*^2 from $p(\sigma^2 | s^2)$, we draw a chi-square deviate on $n - d$ degrees of freedom, form the reciprocal of this number and then multiply by $(n - d)s^2$. To draw θ^* from $p(\theta | \hat{\theta}, \sigma_*^2)$, we draw a deviate from the multivariate normal distribution $N[\hat{\theta}, \sigma_*^2 (x^T x)^{-1}]$. Thus, the tuple (σ_*^2, θ^*) is an observation drawn from $p(\theta, \sigma^2 / Y)$.

To draw a sample from $p(y_f | Y)$, apply the method of composition:

1a. Draw σ_*^2 from $p(\sigma^2 | s^2)$.
1b. Draw θ^* from $p(\theta | \hat{\theta}, \sigma_*^2)$.
2. Draw y_f from $N(x_f \theta^*, \sigma_*^2 I)$.

Steps (1a), (1b) and (2) are repeated to obtain the required sample.

Thompson and Miller (1986) use the method of composition to sample future paths from a time series.

3.3.3. Importance Sampling and Rejection/Acceptance

Consider the problem of calculating the integral

$$J(y) = \int f(y | x) g(x) dx \ .$$

Previously, it was assumed that one can sample directly from $g(x)$. Importance sampling is of use when one cannot directly sample from $g(x)$. Let $I(x)$ be a density which is easy to sample from and which approximates $g(x)$. The method of importance sampling approximates $J(y)$ as

1. Draw $x_1, \ldots, x_m \overset{\text{iid}}{\sim} I(x)$.
2. $\tilde{J}(y) = \sum_{i=1}^{m} w_i f(y|x_i) \Big/ \sum_{i=1}^{m} w_i$,

where $w_i = g(x_i)/I(x_i)$.

Importance sampling gives more weight to regions where $I(x) < g(x)$ and downweights regions where $I(x) > g(x)$ to correctly calculate $J(y)$, given a sample from $I(x)$.

Geweke (1989) has shown that if the support of $I(x)$ includes the support of $g(x)$, the x_i's are an iid sample from $I(x)$, and $J(y)$ exists and is finite, then

$$\tilde{J}(y) \overset{\text{a.s}}{\to} J(y).$$

The first condition is sensible, for if the support of $I(x)$ is strictly contained in the support of $g(x)$, then there would be no hope of generating deviates in the complement of the support of $I(x)$. The rate of convergence depends on how closely $I(x)$ mimics $g(x)$. As noted by Geweke (1989), it is important that the tails of $I(x)$ do not decay faster than the tails of $g(x)$.

Furthermore, it can be shown (Geweke, 1989) that the Monte Carlo standard error of $\tilde{J}(y)$ is estimated by

$$\sqrt{\sum_{i=1}^{m} \{f(y|x_i) - \tilde{J}(y)\}^2 w_i^2 / \sum_{i=1}^{m} w_i}.$$

In this way, the standard error of $\tilde{J}(y)$ is inflated (the effective Monte Carlo sample size is decreased) if $I(x)$ poorly approximates $g(x)$.

One can apply the method of composition to the integral

$$J(y) = \int f(y|x)\left[\frac{g(x)}{I(x)}\right] I(x)\, dx$$

to obtain a sample from the marginal of y. Note that

1. draw x^* from $I(x)$
2. draw y^* from $f(y|x^*)$

yields an observation from the joint density $f(y|x)I(x)$. Assigning mass $w_i/\sum_{j=1}^{m} w_j$ to each of the $y_1, \ldots, y_m$, where $w_i = g(x_i)/I(x_i)$, yields a weighted sample from an approximation to $J(y)$. One can carry along these weights in all subsequent analyses of $y_1, \ldots, y_m$. Alternatively one can draw $y_1^*, \ldots, y_n^*$ from the discrete distribution over $(y_1, \ldots, y_m)$ placing mass w_i on y_i to obtain an iid sample. Crucial to this method is the necessity that the support of $g(x)$ is contained in the support of $I(x)$. Smith and Gelfand (1992) call this sample a "weighted bootstrap" sample from an approximation to

$J(y)$. They prove that the approximation improves as m increases. Rubin (1987a) uses this idea as part of the *SIR* algorithm (see Section 5.7). In this regard, note that this sampling approach only requires that $g(x)$ be known up to a proportionality constant, since the constant cancels out in $w_i/\sum_{i=1}^{m} w_i$. This observation is key, since we have avoided the need to perform the integration to standardize $g(x)$. Note however, that $\sum_{i=1}^{m} w_i/m$ is a consistent estimator of $\int g(x)\,dx$.

In the situation where there is a finite known constant $M > 0$ such that $g(x)/I(x) \leq M$, for all x, one can realize an iid sample from $J(y)$ exactly. The algorithm is given as:

1. Generate x^* from $I(x)$.
2. Generate u from the uniform (0, 1) distribution.
3. If $u \leq g(x^*)/[M \cdot I(x^*)]$, then accept x^*; otherwise, repeat steps 1–3.
4. Draw y^* from $f(y|x^*)$.

Steps (1)–(3) above form a *rejection/acceptance algorithm.* The function $M \cdot I(x)$ is called the majorizing function. Ripley (1987) proves that x^* is an observation from $g(x)$. Rejection/acceptance allows one to indirectly sample from $g(x)$, if the functional form of $g(x)$ is known and a sample from an approximating distribution $I(x)$ is given. Note that the probability of accepting x^* is equal to $\int g(x)\mathrm{d}x/M$. Hence, if the approximation of $I(x)$ to $g(x)$ is poor in some regions (forcing $\int g(x)\,\mathrm{d}x/M$ to be small), then the algorithm may have a low acceptance rate.

Example. *Statistical Uses of Importance Sampling*

Suppose that $\theta_1, \ldots, \theta_m$ are an iid sample from the posterior $p(\theta|Y)$ calculated under the prior $p(\theta)$. Assigning mass proportional to the weight $w = q(\theta)/p(\theta)$ to the sampled values realizes an iid sample from the posterior $p'(\theta|Y)$ calculated under the prior $q(\theta)$. Thus, one may wish to work with a convenient prior to obtain the sample $\theta_1, \ldots, \theta_m$ and then weight these observations to realize a sample from the posterior of interest. However, since the Monte Carlo standard error of some function of the θ_i's will depend on the weights, a high variance of the weights will have the effect of decreasing the effective (Monte Carlo) sample size. Thus, for the method to be effective, one must find a trial density which is easy to sample from, yet leads to reasonably distributed weights.

To judge the sensitivity of the posterior to an observation y_i, one may wish to delete the corresponding term from the likelihood and examine the resulting posterior. Having drawn $\theta_1, \ldots, \theta_m$ from $p(\theta|Y)$, a sample from $p(\theta|Y_{-i})$, i.e. the posterior based on the reduced sample, is obtained by weighting θ_l proportional to $1/L(\theta_l|y_i)$, where $L(\theta_l|y_i)$ is the contribution to the likelihood corresponding to y_i evaluated at θ_l. As noted above, the effective (Monte Carlo) sample size will be reduced if the terms $1/L(\phi|y_i)$, as a function of ϕ, are highly variable. Tanner (1991) shows how to use sampling based algorithms for prior sensitivity and for case influence analyses. See also Kass, Tierney and Kadane (1989), Smith and Gelfand (1992) and Kong, Liu and Wong (1992). The key point is that the computation does not have to be redone from scratch for each change in the likelihood or prior.

Example. *Logit Model* (*Continued*)
To illustrate importance sampling approximate the posterior (under a flat prior) of the logit model by a normal distribution with mean $\hat{\theta}$ (MLE) and variance–covariance matrix $c\sum$, when $\sum$ is the inverse information at $\hat{\theta}$ and c is a constant. Let $I_c(\alpha, \beta) = I_c(\theta)$ denote the density of the bivariate normal distribution, let $L(\alpha, \beta \mid Y) = L(\theta \mid Y)$ denote the bivariate likelihood and let the weight $w_i = L(\theta_i \mid Y)/I_c(\theta_i)$, where θ_i is a value of θ.

Trial runs ($m = 100$) were conducted by sampling from the normal distribution with variance–covariance matrices $\sum$, $1.5\sum$ and $2.0\sum$. It was found that the coefficient of variation of the weights w_i, ($i = 1, \ldots, 100$) was minimized from the covariance matrix $1.5\sum$.

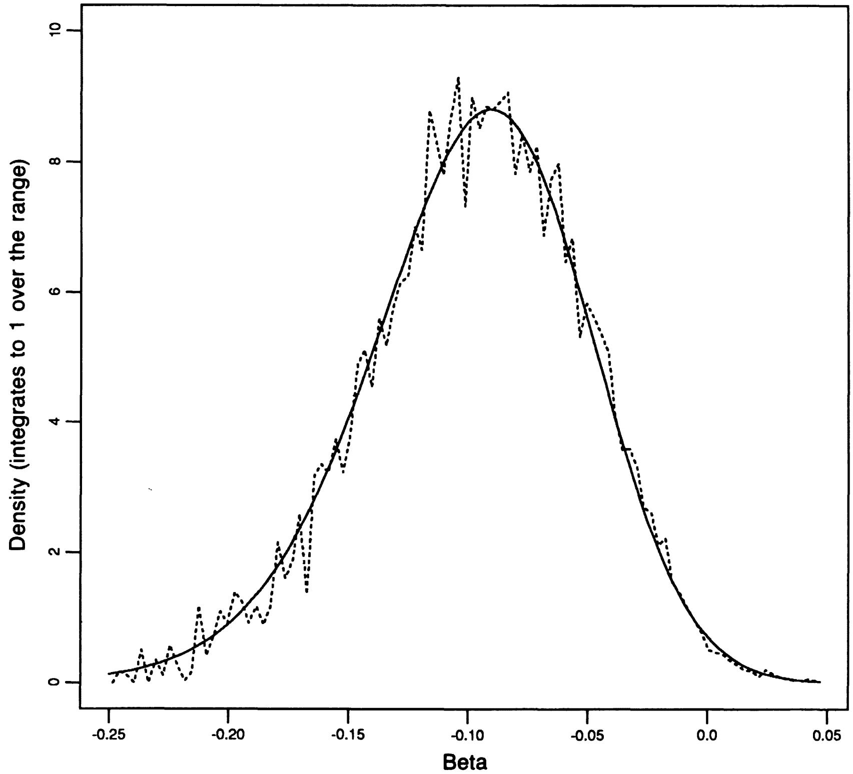


Figure 3.6. Comparison: exact vs. importance sampling. solid: exact; dotted: importance sampling.

A sample of size 10,000 was then drawn from this bivariate normal distribution. The mass $w_i/\sum_{j=1}^{10000} w_j$ was assigned to the bivariate normal realization θ_i, where $w_i = L(\theta_i | Y)/I(\theta_i)$.

The resulting β marginal (dashed line) is presented in Figure 3.6, along with the marginal obtained via numerical integration (solid line). As can be seen, the importance sampling approach follows the marginal. Using a smaller number of bins or adopting a kernel estimator would further smoothen the estimate. Alternatively, if the normalized version of $L(\beta | \alpha, Y)$ is available, one can use the Rao–Blackwell estimator (see Section 6.2.1)

$$\frac{\sum_{j=1}^{10000} w_j L(\beta | \alpha_j, Y)}{\sum_{j=1}^{10000} w_j},$$

where the normalized conditionals are used in this mixture.

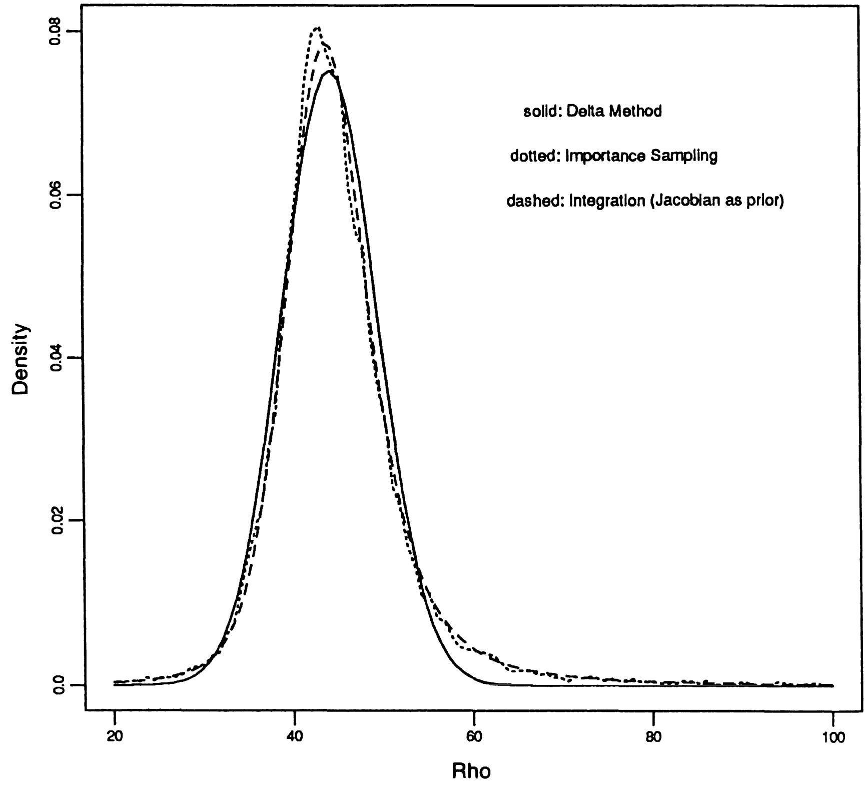


Figure 3.7. Delta method vs. integration and importance sampling (all curves integrate to 1 over the displayed range).

The posterior of $-\alpha/\beta$ is easily obtained by computing $-\alpha/\beta$ for all the simulated bivariate normal deviates and assigning mass $w_i/\sum_{j=1}^{10,000} w_j$ to $-\alpha_i/\beta_i$. This estimate of the marginal is presented in Figure 3.7 (dotted line) along with the δ-method approximation (solid line). The dashed line represents the marginal obtained via numerical integration (Jacobian prior). The importance sampling is quite congruent with the result obtained via numerical integration under the Jacobian prior.

As noted in Section 3.1, by reparametrizing the likelihood in terms of α and $\rho = -\alpha/\beta$, i.e. $L(\alpha, -\alpha/\rho|Y)$, and assuming a flat prior on $-\alpha/\beta$, one obtains the nonproper marginal. However, if the flat prior is put on (α, β), leading to the prior $\beta^2/|\alpha|$ on ρ, i.e. the Jacobian of the transformation, the more symmetric marginal is obtained.

Sampling from an approximation $I(\alpha, \beta)$ to $L(\alpha, \beta|Y)$ and computing the weights

$$\frac{L(\alpha_i, \beta_i|Y)}{I(\alpha_i, \beta_i)} = \frac{L(\alpha_i, -\alpha_i/\rho_i|Y)|\alpha_i|/\rho_i^2}{I(\alpha_i, -\alpha_i/\rho_i)|\alpha_i|/\rho_i^2},$$

where $\rho_i = -\alpha_i/\beta_i$, yields the proper marginal due to the cancellation of the Jacobian. The nonproper marginal can be obtained by multiplying w_i by $\rho_i^2/|\alpha_i|$. However, in this case it seems more appropriate to put the flat prior on (α, β) rather than on ρ. Posterior distributions under other priors may be obtained by multiplying w_i by the appropriate term.

CHAPTER 4

The *EM* Algorithm

In the previous chapters, we examined various methods which are applied directly to the likelihood or to the posterior density. In this and the following chapters, we examine the data augmentation algorithms, including the *EM* algorithm, the data augmentation algorithm and the Gibbs sampler. All of these data augmentation algorithms share a common approach to problems: rather than performing a complicated maximization or simulation, one augments the observed data with "stuff" (latent data) which simplifies the calculation and subsequently performs a series of simple maximizations or simulations. This "stuff" can be the "missing" data or parameter values. The principle of data augmentation can then be stated as follows: Augment the observed data Y with latent data Z so that the augmented posterior distribution $p(\theta|Y, Z)$ is "simple". Make use of this simplicity in maximizing/marginalizing/calculating/sampling the observed posterior $p(\theta|Y)$.

4.1. Introduction

To motivate the *EM* algorithm, consider the situation depicted in Figure 4.1. The x's in the plot represent failure event times, while the o's represent right-censored event times. Given a current guess to the slope and intercept, we can impute a value for each of the censored event times. In particular, assuming normally distributed errors, this imputation would be based on the conditional normal distribution, conditional on the fact that the (unobserved) failure event time is larger than the (observed) censored time. Having "filled-in" the "holes", update the estimates of the slope and intercept and iterate to

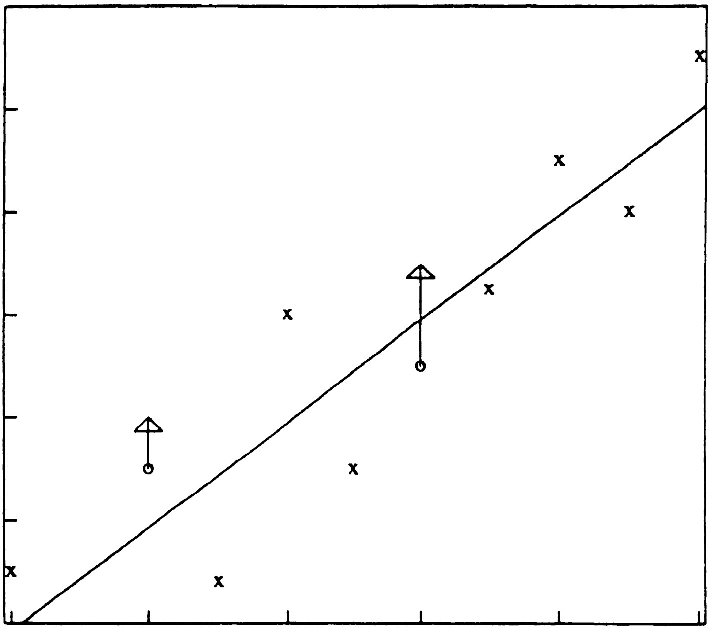


Figure 4.1. Right censored event time data.

obtain the parameter estimates. Thus, a complicated analysis is replaced by a series of simpler analyses. This situation is treated in detail later in this section.

This idea of filling-in missing values with estimated values and then updating parameter estimates is an *EM* algorithm in the situation where the log-augmented posterior is linear in the latent data Z. Estimates can be severely biased when this simple approach is applied in general (Little and Rubin, 1983). In more general situations, expected sufficient statistics, rather than individual observations, are computed at each iteration. In the most general situation, the expected log-augmented posterior is computed at each iteration.

More specifically, the *EM* algorithm is an iterative method for locating the posterior mode. Each iteration consists of two steps: The E-step (expectation step) and the M-step (maximization step). Formally, let θ^i denote the current guess to the mode of the observed posterior $p(\theta|Y)$; let $p(\theta|Y, Z)$ denote the augmented posterior, i.e. the posterior of the augmented data; and let $p(Z|\theta^i, Y)$ denote the conditional predictive distribution of the latent data Z, conditional on the current guess to the posterior mode. In the most general setting, the E-step consists of computing

$$Q(\theta, \theta^i) = \int_Z \log[p(\theta|Z, Y)]p(Z|\theta^i, Y)\mathrm{d}Z \ ,$$

i.e. the expectation of $\log p(\theta|Z, Y)$ with respect to $p(Z|\theta^i, Y)$. In the M-step the Q function is maximized with respect to θ to obtain θ^{i+1}. The algorithm is iterated until $\|\theta^{i+1} - \theta^i\|$ or $|Q(\theta^{i+1}, \theta^i) - Q(\theta^i, \theta^i)|$ is sufficiently small.

Example. *Genetic Linkage Model* (*Rao, 1973*)
Suppose 197 animals (Y)are distributed into four categories as follows:

$$Y = (y_1, y_2, y_3, y_4) = (125, 18, 20, 34)$$

with cell probabilities

$$\left(\frac{1}{2} + \frac{\theta}{4}, \frac{1}{4}(1 - \theta), \frac{1}{4}(1 - \theta), \frac{\theta}{4}\right) .$$

Augment the observed data by splitting the first cell into two cells with probabilities $\frac{1}{2}$ and $\theta/4$. The augmented data are given by

$$(x_1, x_2, x_3, x_4, x_5)$$

such that

$$x_1 + x_2 = 125$$
$$x_3 = y_2$$
$$x_4 = y_3$$
$$x_5 = y_4 \ .$$

Notice that the observed posterior (under a flat prior) is proportional to

$$(2 + \theta)^{y_1}(1 - \theta)^{y_2 + y_3}\theta^{y_3}$$

while the augmented posterior (under a flat prior) is proportional to

$$\theta^{x_2 + x_5}(1 - \theta)^{x_3 + x_4} \ .$$

By working with the augmented posterior, we realize a simplification in functional form.

For this genetic linkage model note that

$$Q(\theta, \theta^i) = E[(x_2 + x_5)\log(\theta) + (x_3 + x_4)\log(1 - \theta)|\theta^i, Y] \ ,$$

where $p(Z|\theta^i, Y)$ is the binomial distribution with $n = 125$ and $p = \theta^i/(\theta^i + 2)$. In this case, $Q(\theta, \theta^i)$ simplifies to

$$[E(x_2|\theta^i, Y) + x_5]\log(\theta) + (x_3 + x_4)\log(1 - \theta)$$

and is linear in the latent data.

For the M-step, note that

$$\left.\frac{\partial Q(\theta, \theta^i)}{\partial \theta}\right|_{\hat{\theta}} = 0 \Rightarrow \frac{E(x_2|\theta^i, Y) + x_5}{\hat{\theta}} - \frac{x_3 + x_4}{1 - \hat{\theta}} = 0$$

$$\Rightarrow \theta_{i+1} = \frac{E(x_2|\theta^i, Y) + x_5}{E(x_2|\theta^i, Y) + x_3 + x_4 + x_5} \ .$$

Starting at $\theta^0 = 0.5$, the *EM* algorithm converges to $\theta^* = 0.6268$ (the observed posterior mode) after four iterations. While this example is quite simple, it is sufficiently rich to illustrate a number of algorithms to be presented in Chapters 5 and 6.

EXAMPLE. *Simple Linear Regression with Right-Censored Data.*
We consider the motorette data found in Schmee and Hahn (1979). Ten motorettes were tested at each of the four temperatures: 150°, 170°, 190° and 220° in degrees °C. The time to failure in hours is given in Table 4.1.

A star indicates that a motorette was taken off the study without failing at the event time indicated. For these data we will fit the model:

$$t_i = \beta_0 + \beta_1 v_i + \sigma\varepsilon_i$$

where $\varepsilon_i \sim N(0, 1)$; $v_i = 1000/(\text{temperature} + 273.2)$ and $t_i = \log_{10}$ (*i*th failure time).

Reorder the data so that the first m observations are uncensored (i.e. a failure is observed at t_i) and the remaining $n - m$ are censored (c_i denotes a censored event time). The log-augmented posterior (under a flat prior) is given (up to a constant) by

$$-n\log\sigma - \sum_{i=1}^{m}(t_i - \beta_0 - \beta_1 v_i)^2/2\sigma^2 - \sum_{i=m+1}^{n}(Z_i - \beta_0 - \beta_1 v_i)^2/2\sigma^2 \ ,$$

where Z_j is the (unobserved) failure time for case j. The conditional predictive distribution $p(Z_i|\beta_0, \beta_1\ \sigma, c_i)$ is the conditional normal distribution – conditional on the fact that the unobserved failure time Z_j is greater than c_j. Hence, the E-step requires one to compute:

$$-n\log\sigma - \frac{1}{2\sigma^2}\sum_{i=1}^{m}(t_i - \beta_0 - \beta_1 v_i)^2 - \frac{1}{2\sigma^2}\sum_{i=m+1}^{n}[E(Z_i^2|\beta_0, \beta_1, \sigma, Z_i > c_i)$$
$$- 2(\beta_0 + \beta_1 v_i)\,E(Z_i|\beta_0, \beta_1, \sigma, Z_i > c_i) + (\beta_0 + \beta_1 v_i)^2] \ .$$

Table 4.1.

150°	170°	190°	220°
8064*	1764	408	408
8064*	2772	408	408
8064*	3444	1344	504
8064*	3542	1344	504
8064*	3780	1440	504
8064*	4860	1680*	528*
8064*	5196	1680*	528*
8064*	5448*	1680*	528*
8064*	5448*	1680*	528*
8064*	5448*	1680*	528*

Source: Schmee and Hahn (1979).

Note that

$$E(Z_i^2|\beta_0,\beta_1,\sigma,Z_i>c_i)=\mu_i^2+\sigma^2+\sigma(c_i+\mu_i)H\left(\frac{c_i-\mu_i}{\sigma}\right)$$

and

$$E(Z_i|\beta_0,\beta_1,\sigma,Z_i>c_i)=\mu_i+\sigma H\left(\frac{c_i-\mu_i}{\sigma}\right),$$

where $\mu_i=\beta_0+\beta_1 v_i$, $H(x)=\phi(x)/\{1-\Phi(x)\}$, and $\phi(x)$ and $\Phi(x)$ are the density and cdf of the standard normal distribution.

To see this last result, note that

$$\begin{aligned}
&E(Z_i|\beta_0,\beta_1,\sigma,Z_i>c_i)\\
&=\beta_o+\beta_1 v_i+\sigma E\left(\varepsilon_i|\varepsilon_i>\frac{c_i-\beta_o-\beta_1 v_i}{\sigma}\right)\\
&=\beta_0+\beta_1 v_i+\sigma\left[\int_{\frac{c_i-\beta_0-\beta_1 v_i}{\sigma}}^{\infty}w\phi(w)\,dw\right]\Bigg/\left[1-\Phi\left(\frac{c_i-\beta_0-\beta_1 v_i}{\sigma}\right)\right]\\
&=\beta_0+\beta_1 v_i+\sigma\phi\left(\frac{c_i-\beta_0-\beta_1 v_i}{\sigma}\right)\Bigg/\left[1-\Phi\left(\frac{c_i-\beta_0-\beta_1 v_i}{\sigma}\right)\right]\\
&=\beta_0+\beta_1 v_i+\sigma H\left(\frac{c_i-\beta_0-\beta_1 v_i}{\sigma}\right).
\end{aligned}$$

The *M*-step is then

$$\frac{\partial Q}{\partial\beta_0}=0\quad\Rightarrow\quad\sum_{i=1}^{m}(t_i-\beta_0-\beta_1 v_i)+\sum_{i=m+1}^{n}\{E(Z_i)-\beta_0-\beta_1 v_i\}=0$$

$$\frac{\partial Q}{\partial\beta_1}=0\quad\Rightarrow\quad\sum_{i=1}^{m}v_i(t_i-\beta_0-\beta_1 v_i)+\sum_{i=m+1}^{n}v_i\{E(Z_i)-\beta_0-\beta_1 v_i\}=0$$

$$\frac{\partial Q}{\partial\sigma^2}=0\quad\Rightarrow\quad\frac{\sum_{i=1}^{m}(t_i-\beta_0-\beta_1 v_i)^2}{\sigma^4}$$

$$+\frac{\sum_{i=m+1}^{n}E(Z_i^2)-2(\beta_0+\beta_1 v_i)E(Z_i)+(\beta_0+\beta_1 v_i)^2}{\sigma^4}-\frac{n}{\sigma^2}=0\,.$$

To obtain β^{i+1}, replace c_j by $E(Z_j|\beta_0^i,\beta_1^i,\sigma^i,Z_j>c_j)$ and apply least squares. To obtain σ_{i+1}^2, compute

$$\sigma_{i+1}^2=\frac{\sum_{j=1}^{m}(t_j-\mu_j^i)^2}{n}+\frac{\sigma_i^2\sum_{j=m+1}^{n}\left[1+\left(\frac{c_j-\mu_j^i}{\sigma_i}\right)H\left(\frac{c_j-\mu_j^i}{\sigma_i}\right)\right]}{n},$$

where $\mu_j^i=\beta_0^i+\beta_1^i v_j$.

Applying the *EM* algorithm to the data of Schmee and Hahn, we arrive at

$$\hat{\beta}_0 = -6.019, \qquad \hat{\beta}_1 = 4.311 \quad \text{and} \quad \hat{\sigma} = 0.2592$$

after 16 iterations.

4.2. Theory

To begin, notice that

$$\log[p(\theta|Y)] = \log[p(\theta|Z, Y)] - \log[p(Z|\theta, Y)] + \log[p(Z|Y)] \ .$$

Now integrate both sides of this equation with respect to the conditional predictive distribution $p(Z|Y, \phi)$ to obtain:

$$\log[p(\theta|Y)] = \int_Z \log[p(\theta|Y, Z)]p(Z|\phi, Y)\,\mathrm{d}Z - \int_Z \log[p(Z|\theta, Y)]p(Z|\phi, Y)\,\mathrm{d}Z$$
$$+ \int_Z \log[p(Z|Y)]\,p(Z|\phi, Y)\,\mathrm{d}Z \ .$$

Dempster, Laird and Rubin (1977) define the Q function by

$$Q(\theta, \theta^*) = \int_Z \log[p(\theta|Z, Y)]\,p(Z|\theta^*, Y)\,\mathrm{d}Z$$

and the H function by

$$H(\theta, \theta^*) = \int_Z \log[p(Z|\theta, Y)]p(Z|\theta^*, Y)\,\mathrm{d}Z \ .$$

Let $K(\theta, \theta^*) = \int_Z \log[p(Z|Y)]p(Z|\theta^*, Y)\mathrm{d}Z$. Note that

$$\log[p(\theta^{i+1}|Y)] - \log[p(\theta^i|Y)] = Q(\theta^{i+1}, \theta^i) - Q(\theta^i, \theta^i)$$
$$- [H(\theta^{i+1}, \theta^i) - H(\theta^i, \theta^i)] + \{K(\theta^{i+1}, \theta^i) - K(\theta^i, \theta^i)\} \ .$$

Dempster, Laird and Rubin (1977) define an *EM* algorithm to select θ^{i+1} so that $Q(\theta, \theta^i)$ is maximized as a function of θ. These authors define a *GEM* (generalized *EM*) algorithm to select θ^{i+1} so that $Q(\theta^{i+1}, \theta^i) > Q(\theta^i, \theta^i)$. From formula (le.6.6) of Rao (1973) it follows that

$$H(\theta^{i+1}, \theta^i) - H(\theta^i, \theta^i) \leq 0 \ .$$

Moreover, $K(\theta^{i+1}, \theta^i) - K(\theta^i, \theta^i) = 0$. Hence, we have shown:

Theorem 4.2.1. Every *EM* or *GEM* algorithm increases the posterior $p(\theta|Y)$ at each iteration, i.e. $p(\theta^{i+1}|Y) \geq p(\theta^i|Y)$, equality holding iff $Q(\theta^{i+1}, \theta^i) = Q(\theta^i, \theta^i)$.

In addition, the following results are available. See Wu (1983) and Little and Rubin (1987) for further details.

Theorem 4.2.2. Suppose a sequence of *EM* iterates θ^i satisfy:

1. $\left.\dfrac{\partial Q(\theta, \theta^i)}{\partial \theta}\right|_{\theta = \theta^{i+1}} = 0$.

2. θ^i converge to some value θ^* and let $p(Z|Y, \theta)$ be "sufficiently" smooth.

Then it follows that

$$\left.\frac{\partial \log p(\theta | Y)}{\partial \theta}\right|_{\theta = \theta^*} = 0 \ .$$

In other words, if the iterates θ^i converge, they converge to a stationary point of $p(\theta|Y)$. This implies that when there are multiple stationary points (local minima, maxima or saddle points), the algorithm may not converge to the global maximum.

Dempster, Laird and Rubin (1977) show that the *EM* algorithm converges at a linear rate, with the rate depending on the proportion of information about θ in $p(\theta|Y)$ which is observed. This implies that the convergence can be quite slow if a large portion of the data are missing. A more rigorous treatment of this statement is given in Section 4.4.5.

4.3. *EM* in the Exponential Family

Suppose the augmented data $X = (Y, Z)$ are distributed as

$$f(X|\theta) = b(X) \exp[\theta^{\mathrm{T}} s(X)]/a(\theta)$$

i.e. the regular exponential family, where θ is a $d \times 1$ vector, and $s(X)$ is a $1 \times d$ vector of sufficient statistics. In this case, the Q function is given as

$$Q(\theta, \theta^i) = \int_{\mathrm{Z}} \log p(\theta | Y, Z) p(Z | Y, \theta^i) \, \mathrm{d}Z$$

$$= \int_Z \{\log[b(X)] + \theta^T s(X) - \log[a(\theta)]\} \, p(Z|\theta^i, Y) \, dZ$$

$$= \int_Z \log[b(X)] p(Z|Y, \theta^i) \, dZ + \theta^T \int_Z s(X) p(Z|Y, \theta^i) \, dZ - \log[a(\theta)] \ .$$

Now notice that

$$\int_Z \log[b(X)] \, p(Z|\theta^i, Y) \, dZ$$

does not depend on θ, so there is no need to compute it in the E-step. Thus, the E-step consists of computing

$$E[s(X)|Y,\theta^i] = \int_Z s(X)\, p(Z|Y,\theta^i)\, dZ = s^i \ .$$

The M-step consists of maximizing $Q(\theta,\theta^i)$ or

$$-\log[a(\theta)] + \theta^T s^i \ .$$

In this way,

$$\frac{\partial Q(\theta,\theta^i)}{\partial\theta} = 0 \quad \Rightarrow \quad \frac{-\partial \log[a(\theta)] - \theta^T s^i}{\partial\theta} = 0 \Rightarrow \frac{\partial \log a(\theta)}{\partial\theta} = s^i \ .$$

Now recall that

$$a(\theta) = \int_X b(X)\exp\{\theta^T s(X)\}\, dX \ .$$

So

$$\begin{aligned}\frac{\partial \log a(\theta)}{\partial\theta} &= \left\{\int_X b(X)\frac{\partial \exp[\theta^T s(X)]}{\partial\theta} dX\right\}\Big/ a(\theta)\\ &= \left\{\int_X s(X) b(X)\exp[\theta^T s(X)]\, dX\right\}\Big/ a(\theta)\\ &= E[s(X)|\theta] \ .\end{aligned}$$

Thus, in the regular exponential family, maximizing $Q(\theta,\theta^i)$ is equivalent to solving

$$E[s(X)|\theta] = s^i$$

for θ.

4.4. Standard Errors in the Context of *EM*

The output of the *EM* algorithm is the mode of the posterior distribution $p(\theta|Y)$. To define the normal approximation to $p(\theta|Y)$, one must compute the Hessian (matrix) of $\log p(\theta|Y)$. Several approaches are available, which are now discussed.

4.4.1. Direct Computation/Numerical Differentiation

Having arrived at the observed posterior mode, θ^*, one can potentially evaluate

$$-\left.\frac{\partial^2 \log p(\theta|Y)}{\partial\theta^2}\right|_{\theta^*} .$$

In practice, however, this may be tedious to code or difficult to evaluate for a given data set.

An alternative approach is to numerically differentiate $\partial \log p(\theta|Y)/\partial\theta$ to obtain the Hessian. Meilijson (1989) suggests that one can perturb $\theta^* = (\theta_1^*, \ldots, \theta_d^*)$ by adding (a small amount) $\varepsilon > 0$ to one coordinate and evaluate $\partial \log p(\theta|Y)/\partial\theta$ at the perturbed parameter $\tilde{\theta}$. The ith row of the Hessian is approximately

$$\frac{1}{\varepsilon}\left[\left.\frac{\partial \log p(\theta|Y)}{\partial\theta}\right|_{\tilde{\theta}} - \left.\frac{\partial \log p(\theta|Y)}{\partial\theta}\right|_{\theta^*}\right] .$$

The entire Hessian requires d such operations. Care must be taken in the selection of ε to ensure that the derivatives are not distorted by the round-off errors.

4.4.2. Missing Information Principle

Note that

$$\log p(\theta|Y) = \log p(\theta|Y,Z) - \log p(Z|Y,\theta) + C \ .$$

This implies that

$$\frac{-\partial^2 \log p(\theta|Y)}{\partial\theta^2} = -\frac{\partial^2 \log p(\theta|Y,Z)}{\partial\theta^2} + \frac{\partial^2 \log p(Z|Y,\theta)}{\partial\theta^2} .$$

Integrating both sides of this equation with respect to $p(Z|Y,\theta)$ we obtain

$$\begin{aligned} -\frac{\partial^2 \log p(\theta|Y)}{\partial\theta^2} &= -\int_Z \frac{\partial^2 \log p(\theta|Y,Z)}{\partial\theta^2} p(Z|\theta, Y)\, dZ \\ &\quad + \int_Z \frac{\partial^2 \log p(Z|Y,\theta)}{\partial\theta^2} p(Z|Y,\theta)\, dZ \\ &= \left.\frac{-\partial^2 Q(\theta,\phi)}{\partial\theta^2}\right|_{\phi=\theta} - \left.\frac{-\partial^2 H(\theta,\phi)}{\partial\theta^2}\right|_{\phi=\theta} , \end{aligned} \qquad (4.4.1)$$

where Q and H are defined in Section 4.2. Referring to $-\partial^2 Q/\partial\theta^2$ as the complete information and to $-\partial^2 H/\partial\theta^2$ as the missing information, we have the Missing Information Principle (Louis, 1982):

Observed Information = Complete Information − Missing Information .

Orchard and Woodbury (1972) present an analogous expression for expected information by taking expectations over Y.

4.4.3. Louis' Method

A basic result due to Louis (1982) is that $-\partial^2 H/\partial\theta^2 = \operatorname{var}[\partial \log p(\theta|Y,Z)/\partial\theta]$. Thus, it follows that:

$$-\frac{\partial^2 \log p(\theta|Y)}{\partial\theta^2} = -\int_Z \frac{\partial^2 \log p(\theta|Y,Z)}{\partial\theta^2} p(Z|Y,\theta) dZ - \operatorname{var}\left\{\frac{\partial \log p(\theta|Y,Z)}{\partial\theta}\right\}. \tag{4.4.2}$$

where the variance is with respect to $p(Z|Y,\theta)$. We now prove this result. Those readers not interested in the proof may skip to the example.

To prove Louis' result, we first need to show:

$$\frac{\partial \log p(\theta|Y)}{\partial\theta} = \int_Z \frac{\partial \log p(\theta|Y,Z)}{\partial\theta} p(Z|\theta, Y)\, dZ \tag{4.4.3}$$

i.e. the observed score is equal to the expected score of the log-augmented posterior. To see this, note that

$$\frac{\partial \log p(\theta|Y)}{\partial\theta} = \frac{\partial \log p(\theta|Y,Z)}{\partial\theta} - \frac{\partial \log p(Z|Y,\theta)}{\partial\theta}$$

which implies that

$$\frac{\partial \log p(\theta|Y)}{\partial\theta} = \int_Z \frac{\partial \log p(\theta|Y,Z)}{\partial\theta} p(Z|Y,\theta)\, dZ$$
$$- \int_Z \frac{\partial \log p(Z|Y,\theta)}{\partial\theta} p(Z|Y,\theta)\, dZ .$$

Regarding the last term,

$$\int_Z \frac{\partial \log p(Z|Y,\theta)}{\partial\theta} p(Z|Y,\theta)\, dZ = \int_Z \frac{\partial p(Z|Y,\theta)}{\partial\theta} dZ .$$

Interchanging derivative and integral,

$$\int_Z \frac{\partial\, p(Z|Y,\theta)}{\partial\theta} dZ = \frac{\partial \int_Z p(Z|Y,\theta)\, dZ}{\partial\theta} = 0 ,$$

since $p(Z|Y,\theta)$ integrates to unity over the space Z. The result (4.4.3) follows.

To prove Louis' result, note that

$$\frac{\partial^2 \log p(\theta|Y)}{\partial\theta^2} = \frac{\partial \int_Z (\partial \log p(\theta|Y,Z)/\partial\theta)\, p(Z|Y,\theta)\, dZ}{\partial\theta}$$
$$= \int_Z \frac{\partial^2 \log p(\theta|Y,Z)}{\partial\theta^2} p(Z|Y,\theta)\, dZ$$
$$+ \int_Z \frac{\partial \log p(\theta|Y,Z)}{\partial\theta} \frac{\partial\, p(Z|Y,\theta)}{\partial\theta} dZ$$
$$= (*) + (\Delta)$$

by the chain rule. Now

$$\Delta = \int_Z \frac{\partial \log p(\theta|Y,Z)}{\partial\theta} \frac{[\partial p(Z|Y,\theta)/\partial\theta]}{p(Z|Y,\theta)} p(Z|Y,\theta)\,\mathrm{d}Z$$

$$= \int_Z \frac{\partial \log p(\theta|Y,Z)}{\partial\theta} \frac{\partial \log p(Z|Y,\theta)}{\partial\theta} p(Z|Y,\theta)\,dZ\ .$$

Since

$$\frac{\partial \log p(Z|Y,\theta)}{\partial\theta} = \frac{\partial \log \mathrm{p}(\theta|Y,Z)}{\partial\theta} - \frac{\partial \log p(\theta|Y)}{\partial\theta}\ ,$$

it follows that

$$\Delta = \int_Z \frac{\partial \log p(\theta|Y,Z)}{\partial\theta}\left[\frac{\partial \log p(\theta|Y,Z)}{\partial\theta} - \frac{\partial \log p(\theta|Y)}{\partial\theta}\right] p(Z|Y,\theta)\,dZ$$

$$= \int_Z \left(\frac{\partial \log p(\theta|Y,Z)}{\partial\theta}\right)^2 p(Z|Y,\theta)\,dZ$$

$$- \left[\int_Z \frac{\partial \log p(\theta|Y,Z)}{\partial\theta} p(Z|Y,\theta)\,dZ\right]^2$$

since

$$\int_Z \frac{\partial \log p(\theta|Y,Z)}{\partial\theta} p(Z|Y,\theta)\,dZ = \frac{\partial \log p(\theta|Y)}{\partial\theta}\ .$$

Note that (4.4.1) and (4.4.2) imply that $\Delta = -\,\partial^2 H(\theta,\phi)/\partial\theta^2|_{\phi=\theta}$.

Finally, note that at the posterior mode $\hat\theta$, $\partial \log p(\theta|Y)/\partial\theta|_{\hat\theta} = 0$. Hence, at $\hat\theta$

$$\Delta = \int_Z \left(\frac{\partial \log p(\theta|Y,Z)}{\partial\theta}\right)^2 p(Z|Y,\hat\theta)\,dZ\ .$$

EXAMPLE. *Genetic Linkage (Continued)*
For this model,

$$\frac{\partial \log p(\theta|Y,Z)}{\partial\theta} = \frac{x_2+x_5}{\theta} - \frac{x_3+x_4}{1-\theta}$$

while

$$\left.\frac{-\,\partial^2 Q(\theta,\hat\theta)}{\partial\theta^2}\right|_{\hat\theta} = \frac{E(x_2|\hat\theta,Y)+x_5}{\hat\theta^2} + \frac{x_3+x_4}{(1-\hat\theta)^2}\ .$$

For the data set listed in the example of Section 4.1

$$-\left.\frac{\partial^2 Q(\theta,\hat\theta)}{\partial\theta^2}\right|_{\hat\theta} = \frac{29.83}{0.6268^2} + \frac{38}{(1-0.6268)^2} = 435.3\ .$$

Now

$$\operatorname{var}\left(\frac{\partial \log p(\theta|Y,Z)}{\partial\theta}\bigg|_{\hat\theta}\right)=\frac{\operatorname{var}(x_2|\hat\theta)}{\hat\theta^2}=125\left(\frac{\hat\theta}{2+\hat\theta}\right)\left(\frac{2}{2+\hat\theta}\right)\bigg/\hat\theta^2$$

$$=22.71/0.6268^2=57.8\ .$$

Hence, it follows that

$$\frac{-\partial^2\log p(\theta|Y)}{\partial\theta^2}\bigg|_{\hat\theta}=435.3-57.8=377.5$$

and the standard error of $\hat\theta$ is equal to $\sqrt{(1/377.5)}=0.05$.

4.4.4. Simulation

In some situations it may be difficult to analytically compute

$$\int_Z\frac{\partial^2\log p(\theta|Y,Z)}{\partial\theta^2}p(Z|\hat\theta,Y)\,dZ\ .$$

If one can sample from $p(Z|\theta, Y)$, this integral may be approximated by the sum

$$\frac{1}{m}\sum_{j=1}^{m}\frac{\partial^2\log p(\theta|Y,z_j)}{\partial\theta^2}\ ,$$

given $z_1,z_2,\ldots,z_m\overset{\text{iid}}{\sim}p(Z|\hat\theta,Y)$, Rubin (1987b) refers to the z_i's as multiple imputations.

Similarly, one can approximate

$$\operatorname{var}\left(\frac{\partial\log p(\theta|Y,Z)}{\partial\theta}\bigg|_{\hat\theta}\right)$$

via

$$\frac{1}{m}\sum_{j=1}^{m}\left(\frac{\partial\log p(\theta|Y,z_j)}{\partial\theta}\right)^2-\left[\frac{1}{m}\sum_{j=1}^{m}\frac{\partial\log p(\theta|Y,z_j)}{\partial\theta}\right]^2\ ,$$

where

$$z_1,\ldots,z_m\sim p(Z|\theta^i,Y)\ .$$

Example. *Genetic Linkage* (*Continued*)

In the genetic linkage example, $p(Z|\theta^i, Y)$ is the binomial distribution with parameters $n = 125$ and $p = \theta^i/(2+\theta^i)$. Having converged to the posterior mode $\hat\theta = 0.6268$, 10,000 imputations were drawn from the binomial distribution with parameters $n = 125$ and $p = \hat\theta/(\hat\theta + 2)$. Computations were performed using the RANDOM command of MINITAB. The estimated variance of the augmented score

$$\frac{1}{10{,}000}\sum_{j=1}^{10{,}000}\left(\frac{\partial \log p(\theta|Y,z_j)}{\partial\theta}\right)^2$$

was found to be equal to 57.4. The exact value is 57.8.

4.4.5. Using *EM* Iterates

In this approach, we show how to adjust the Hessian of the Q function using the *EM* iterates θ^i to obtain the variance–covariance matrix. A one-dimensional version of this approach is given in Smith (1977). The multivariate version is given in Meng and Rubin (1991).

The *EM* algorithm defines a mapping $M:\theta\rightarrow\theta$ from the parameter space θ onto itself such that

$$\theta^{i+1} = M(\theta^i)\ .$$

If the *EM* iterates converge to θ^*, then

$$\theta^* = M(\theta^*)\ .$$

Dempster, Laird and Rubin (1977) show that

$$\left\{\frac{\partial M(\theta)}{\partial\theta}\bigg|_{\theta^*}\right\}\left\{\frac{\partial^2 Q(\theta,\theta^*)}{\partial\theta^2}\bigg|_{\theta^*}\right\} = \frac{\partial^2 H(\theta,\theta^*)}{\partial\theta^2}\bigg|_{\theta^*}\ . \tag{4.4.4}$$

These authors use this result to argue that in a neighborhood of θ^*, the rate of convergence of the *EM* algorithm is given by

$$\left(\frac{\partial^2 H}{\partial\theta^2}\right)\left(\frac{\partial Q^2}{\partial\theta^2}\right)^{-1}$$

i.e. the ratio of the missing information to the complete information.

From (4.4.4) and the decomposition

$$\frac{-\partial^2\log p(\theta|Y)}{\partial\theta^2}\bigg|_{\theta^*} = \frac{-\partial^2 Q(\theta,\theta^*)}{\partial\theta^2}\bigg|_{\theta^*} - \frac{-\partial^2 H(\theta,\theta^*)}{\partial\theta^2}\bigg|_{\theta^*}$$

given in Section 4.4.2, it follows that

$$\frac{-\partial^2\log p(\theta|Y)}{\partial\theta^2}\bigg|_{\theta^*} = \left[I - \frac{\partial M(\theta)}{\partial\theta}\bigg|_{\theta^*}\right]\left[\frac{-\partial^2 Q(\theta,\theta^*)}{\partial\theta^2}\bigg|_{\theta^*}\right]$$

where I is the $d\times d$ identity matrix. It is easy to check that

$$\left[\frac{-\partial^2\log p(\theta|Y)}{\partial\theta^2}\bigg|_{\theta^*}\right]^{-1} = \left[\frac{-\partial^2 Q(\theta,\theta^*)}{\partial\theta^2}\bigg|_{\theta^*}\right]^{-1} + \left[\frac{\partial^2 Q(\theta,\theta^*)}{\partial\theta^2}\bigg|_{\theta^*}\right]^{-1}$$
$$\times\left[I - \frac{\partial M(\theta)}{\partial\theta}\bigg|_{\theta^*}\right]^{-1}\left[\frac{\partial M(\theta)}{\partial\theta}\bigg|_{\theta^*}\right]\ .$$

Meng and Rubin (1991) note that

$$r_{ij} = \left.\frac{\partial M_j(\theta)}{\partial \theta_i}\right|_{\theta^*} = \lim_{\theta_i \to \theta_i^*} \frac{M_j(\theta_1^*, \ldots, \theta_i, \ldots, \theta_d^*) - M_j(\theta^*)}{\theta_i - \theta_i^*}$$

$$= \lim_{t \to \infty} \frac{M_j(\theta_1^*, \ldots, \theta_i^t, \ldots, \theta_d^*) - (\theta_j^*)}{\theta_i^t - \theta_i^*} = \lim_{t \to \infty} r_{ij}^t \, .$$

In other words, these authors use the *EM* iterates to differentiate numerically $M(\theta)$. In this way, to compute r_{ij}^t, having obtained θ^* and θ^t, Meng and Rubin (1991) suggest:

1. Fix $i = 1$ and form $\theta^t(i) = (\theta_1^*, \ldots, \theta_i^t, \ldots, \theta_d^{*\prime})$. Evaluate $\tilde{\theta}^{t+1}(i) = M[\theta^t(i)]$.
2. Form

$$r_{ij}^t = \frac{\tilde{\theta}_j^{t+1}(i) - \theta_j^*}{\theta_i^t - \theta_i^*}$$

for $j = 1, \ldots, d$.

3. Repeat steps 1 and 2 for $i = 2, \ldots, d$.

To implement these three steps, d evaluations of the mapping M are required. Meng and Rubin (1991) obtain r_{ij} when the sequence $r_{ij}^{t^*}, r_{ij}^{t^*+1}, \ldots$ is stable for some t^*. As these authors illustrate, different values of t^* may be used for different r_{ij} elements. Care, however, must be taken in coding to avoid problems with round-off error in computing the finite difference in step 3. In addition, Meng and Rubin (1991) point out that the present estimate of $\partial^2 \log p(\theta|Y)/\partial\theta^2 |_{\hat{\theta}}$ may not be symmetric. A possible solution to this problem would be to replace the estimated variance–covariance matrix $\hat{V}$ with $\frac{1}{2}(\hat{V} + \hat{V}^T)$, though this idea has not been investigated.

Meng and Rubin argue that the present approach will be of use if calculation of the variance of the augmented score (Approach 3) is tedious or intractable. In such a case, it would be of interest to compare the present approach with the simulation technique presented in Section 4.4.4. Clearly, the simulation approach moves out of the context of the classic *EM* as it requires access to a random number generator.

To prove (4.4.4), first note that

$$\left.\frac{\partial Q(\theta, \theta_1)}{\partial \theta}\right|_{\theta_2} = \left.\frac{\partial Q(\theta, \theta^*)}{\partial \theta}\right|_{\theta^*} + (\theta_2 - \theta^*)\left.\frac{\partial^2 Q(\theta, \theta^*)}{\partial \theta^2}\right|_{\theta^*}$$

$$+ (\theta_1 - \theta^*)\left.\frac{\partial Q(\theta, \phi)}{\partial \theta \partial \phi}\right|_{\substack{\theta = \theta^* \\ \phi = \theta^*}} + \cdots .$$

Substituting $\theta_1 = \theta^t$ and $\theta_2 = \theta^{t+1}$, we have

$$0 = (\theta^{t+1} - \theta^*)\left.\frac{\partial^2 Q(\theta, \theta^*)}{\partial \theta^2}\right|_{\theta^*} + (\theta^t - \theta^*)\left.\frac{\partial Q(\theta, \phi)}{\partial \theta \partial \phi}\right|_{\substack{\theta = \theta^* \\ \phi = \theta^*}} + \cdots .$$

Since $\theta^{t+1} = M(\theta^t)$ and $\theta^* = M(\theta^*)$, we have in the limit

$$0 = \left[\frac{\partial M(\theta)}{\partial \theta}\bigg|_{\theta^*}\right]\left[\frac{\partial^2 Q(\theta,\theta^*)}{\partial \theta^2}\bigg|_{\theta^*}\right] + \frac{\partial Q(\theta,\phi)}{\partial\theta\partial\phi}\bigg|_{\substack{\theta=\theta^*\\ \phi=\theta^*}} .$$

However, as noted in Section 4.4.3,

$$\frac{\partial Q(\theta,\phi)}{\partial\theta\partial\phi}\bigg|_{\substack{\theta=\theta^*\\ \phi=\theta^*}} = \frac{-\partial^2 H(\theta,\theta^*)}{\partial\theta^2}\bigg|_{\theta^*} .$$

Hence,

$$\left[\frac{\partial M(\theta)}{\partial \theta}\bigg|_{\theta^*}\right]\left[\frac{\partial^2 Q(\theta,\theta^*)}{\partial \theta^2}\bigg|_{\theta^*}\right] = \frac{\partial^2 H(\theta,\theta^*)}{\partial\theta^2}\bigg|_{\theta^*} .$$

4.5. Monte Carlo Implementation of the *E*-Step

Given the current guess to the posterior mode, θ^i, the *E*-step requires the computation

$$Q(\theta,\theta^i) = \int_Z \log p(\theta|Y,Z)p(Z|\theta^i, Y)dZ \ .$$

To facilitate the *E*-step, one may apply the method of Monte Carlo to calculate the Q function. In particular, the Monte Carlo *E*-step is given as

a. Draw $z_1,\ldots,z_m \overset{\text{iid}}{\sim} p(Z|Y,\theta^i)$.

b. Let $\hat{Q}_{i+1}(\theta,\theta^i) - \frac{1}{m}\sum_{j=1}^{m} \log p(\theta|z_j,Y)$.

In the M-step, $\hat{Q}$ is maximized to obtain θ^{i+1}.

Two important considerations in implementing Monte Carlo *EM* are the specification of m and the monitoring of convergence. Regarding the specification of m, it is inefficient to start with a large value of m when θ^i is far from the mode. Rather, one may increase m as the current approximation moves closer to the true value (θ^*).

One may monitor the convergence of the algorithm by plotting θ^i vs. iteration number i. After a certain number of iterations the plot will reveal random fluctuation about the line $\theta = \theta^*$. At such a point, one may terminate the algorithm or continue with a large value of m to decrease the system variability.

Note that within the context of Monte Carlo *EM*, the observed Fisher Information matrix of the observed data at the posterior mode θ^* can be estimated via

$$\frac{1}{m}\sum_{j=1}^{m} \frac{\partial^2 \log p(\theta \mid Y, z_j)}{\partial \theta^2}\bigg|_{\theta^*} + \frac{1}{m}\sum_{j=1}^{m} \left(\frac{\partial \log p(\theta \mid Y, z_j)}{\partial \theta}\bigg|_{\theta^*} \right)^2 .$$

As will be seen in Section 5.5, a slight modification of Monte Carlo *EM* is available – the poor man's data augmentation algorithm – which results in an approximation to the entire posterior distribution, rather than just the maximizer. See also Wei and Tanner (1990a).

EXAMPLE. *Genetic Linkage* (*Continued*)
Given the current guess to the posterior mode, θ^i, the conditional predictive distribution $p(Z|\theta^i, Y)$ is the binomial distribution with $n = 125$ and $p = \theta^i/(2 + \theta^i)$. As noted in the example of Section 4.1, the log-augmented posterior $\log p(\theta| Y, Z)$, under a flat prior, is

$$(Z + x_5)\log(\theta) + (x_3 + x_4)\log(1 - \theta) \ .$$

Having sampled $z_1, \ldots, z_m \overset{\text{iid}}{\sim} B_i\,(125;\, \theta^i/2 + \theta^i)$, the Q function is calculated as

$$(ave + x_5)\log(\theta) + (x_3 + x_4)\log(1 - \theta) \ ,$$

where $ave = \Sigma_{j=1}^{m} z_j/m$. It follows that

$$\theta^{i+1} = \frac{ave + x_5}{ave + x_5 + x_3 + x_4} \ .$$

Table 4.2. History of MCEM

Iteration	Theta
1	0.5833
2	0.6222
3	0.6192
4	0.6321
5	0.6153
6	0.6259
7	0.6238
8	0.6245
9	0.6270
10	0.6265
11	0.6264
12	0.6270

Table 4.2 presents the history of an implementation of the Monte Carlo *EM* algorithm for these data. The algorithm was initiated at $\theta^0 = 0.4$ and m was equal to 10 (1000) for iterations 1–8 (9–12).

As can be seen from Table 4.2, the process with $m = 10$ stabilized by the eighth iteration. From iterations 9–12, the mode is found to be equal to 0.627. The mode to four places is 0.6268.

Example. *Censored Regression (Continued)*
As noted in the example of Section 4.1, the log-augmented posterior under a flat prior is (up to a constant)

$$-n\log\sigma - \sum_{j=1}^{k}(t_j - \beta_0 - \beta_1 v_j)^2/2\sigma^2 - \sum_{j=k+1}^{n}(Z_j - \beta_o - \beta_1 v_j)^2/2\sigma^2 .$$

The conditional predictive distribution $p(Z_j|\beta_0^i, \beta_1^i, \sigma_i, c_j)$ is the conditional normal distribution, conditional on the fact that the unobserved failure time Z_j is greater than the observed right-censored time c_j. That is,

$$Z_j \sim \frac{\phi(x)}{1 - \Phi\left(\dfrac{c_j - \beta_0^i - \beta_1^i v_j}{\sigma_i}\right)} ,$$

where $\phi(x)$ and $\Phi(x)$ are the density and cdf of the standard normal distribution, respectively, and β_o^i, β_1^i and σ_i are the current estimates of the intercept, slope and scale parameter, respectively.

A failure time is inputed by sampling from the appropriate conditional normal distribution. Thus, the integration $\int_c^\infty x\phi(x)dx$ in *EM* is replaced by random draws from the conditional normal in Monte Carlo *EM*.

Having imputed failure times for $n - k$ censored observations to obtain an augmented data set $\underset{\sim}{\mu}_1$, the process is repeated to obtain m augmented data sets. It follows that in the *M*-step

$$\underset{\sim}{\beta}^{i+1} = (V^T V)^{-1} V^T\left(\sum_{j=1}^{m} \underset{\sim}{\mu}_j/m\right)$$

and

$$\sigma_{i+1}^2 = \sum_{j=1}^{m}(\underset{\sim}{\mu}_j - V\underset{\sim}{\beta}^i)^T(\underset{\sim}{\mu}_j - V\underset{\sim}{\beta}^i)/mn ,$$

where V is the $n \times 2$ design matrix with 1's in the first column and the v_j's in the second column.

For the Schmee and Hahn (1979) data, the algorithm was initiated with $\beta_0^0 = -4.931$, $\beta_1^0 = 3.747$ and $\sigma^0 = \sqrt{0.0247}$. The value of m was equal to 50 (5000) for iterations 1–14 (15–18). The values $\beta_0 = -6.02$, $\beta_1 = 4.31$ and $\sigma^2 = 0.067$ from the final three iterations agree with the values quoted in the example of Section 4.1.

Table 4.3. History of MCEM

Iteration	β_0	β_1	σ^2 ($\times 100$)
1	− 5.27	3.93	3.36
2	− 5.61	4.10	4.07
3	− 5.64	4.12	4.65
4	− 5.77	4.18	5.08
5	− 5.81	4.20	5.31
6	− 5.84	4.22	5.41
7	− 5.85	4.23	5.62
8	− 5.86	4.23	5.70
9	− 5.94	4.27	5.72
10	− 5.88	4.24	5.82
11	− 5.97	4.29	5.93
12	− 5.88	4.24	5.80
13	− 5.78	4.19	5.77
14	− 5.94	4.27	5.67
15	− 5.97	4.28	5.84
16	− 5.96	4.28	5.89
17	− 5.96	4.28	5.89
18	− 5.96	4.28	5.89

Table 4.3 gives history of the Monte Carlo *EM* algorithm implemented under the noninformative prior $p(\beta, \sigma^2) \propto 1/\sigma^2$. The same starting point and values of m were used as in the case given above.

Sinha, Tanner and Hall (1992) apply MCEM to address the analysis of grouped data via a marginal likelihood approach by imputing ranks.

4.6. Acceleration of *EM* (Louis' Turbo *EM*)

As noted in Dempster, Laird, and Rubin (1977), the convergence rate of the *EM* algorithm is linear and this rate is governed by the fraction of missing information. Thus, when the proportion of missing data is high, convergence can be quite slow. In this regard, Louis (1982) has proposed a device for accelerating the convergence of the *EM* algorithm.

Recall the Newton–Raphson iteration:

$$\theta^{m+1} = \theta^m + \left[\left. \frac{-\partial^2 \log p(\theta|Y)}{\partial\theta^2} \right|_{\theta^m} \right]^{-1} \left. \frac{\partial \log p(\theta|Y)}{\partial\theta} \right|_{\theta^m}$$

Newton–Raphson has a quadratic convergence rate in a neighborhood of the mode. The goal is to be able to use *EM* "type" quantities to perform a Newton–Raphson step, i.e. achieve quadratic convergence near the mode.

Let θ_{EM}^m denote the update to the posterior mode, when the *EM* algorithm is started at θ^m. Expanding $\partial Q(\theta,\theta^m)/\partial\theta$ about θ^m we have

$$0=\left.\frac{\partial Q(\theta,\theta^m)}{\partial\theta}\right|_{\theta_{EM}^m}\approx\left.\frac{\partial Q(\theta,\theta^m)}{\partial\theta}\right|_{\theta^m}+\left[\left.\frac{\partial^2 Q(\theta,\theta^m)}{\partial\theta^2}\right|_{\theta^m}\right](\theta_{EM}^m-\theta^m)\ .$$

Hence, in the limit we have

Result 4.6.1.

$$\left.\frac{\partial Q(\theta,\theta^m)}{\partial\theta}\right|_{\theta^m}=\left.\frac{-\,\partial^2 Q(\theta,\theta^m)}{\partial\theta^2}\right|_{\theta^m}(\theta_{EM}^m-\theta^m)$$

As noted in Section 4.4.3, we have:

Result 4.6.2.

$$\frac{\partial\log p(\theta|Y)}{\partial\theta}=\int\frac{\partial\log p(\theta|Y,Z)}{\partial\theta}p(Z|Y,\theta)\,dZ=\left.\frac{\partial Q(\theta,\theta')}{\partial\theta}\right|_{\theta'=\theta}\ .$$

Combining Results 4.6.1 and 4.6.2 with the Newton–Raphson step, we have:

$$\theta^{m+1}=\theta^m+\left[\left.\frac{-\,\partial^2\log p(\theta|Y)}{\partial\theta^2}\right|_{\theta^m}\right]^{-1}$$
$$\times\left[-\int\frac{\partial^2\log p(\theta|Y,Z)}{\partial\theta^2}p(Z|Y,\theta^m)\,dZ|_{\theta^m}\right](\theta_{EM}^m-\theta^m)\ .$$

Louis (1982) points out that this formula is an example of Aitken's acceleration method. Louis also notes that this approximation is the most useful in a neighborhood of the mode. In this regard, further experience is required to develop rules about when to perform such an update.

Note that $[-\,\partial^2\log p(\theta|Y)/\partial\theta^2]\,|_{\theta^m}$ may be replaced by

$$-\left.\int\frac{\partial^2\log p(\theta|Y,Z)}{\partial\theta^2}p(Z|\theta^m,Y)\,dZ\right|_{\theta^m}-\mathrm{var}\left[\left.\frac{\partial\log p(\theta|Y,Z)}{\partial\theta}\right|_{\theta^m}\right]\ .$$

Finally, it is noted that the Monte Carlo *EM* algorithm can be accelerated in an analogous manner. Jamshidian and Jennrich (1993) discuss an approach to accelerating the *EM* algorithm based on conjugate gradients.

EXAMPLE. *Genetic Linkage* (*Continued*)

To illustrate the Louis' Turbo *EM* algorithm, initiate the *EM* algorithm at $\theta^0=0.6$ to obtain:

$$\theta^1=0.623188$$

$$\theta^2=0.626338$$

$$\theta^3 = 0.626757$$

$$\theta^3_{EM} = 0.626812$$

$$\theta^* = 0.6268215\ldots$$

In this case,

$$[-\partial^2 \log p(\theta|Y)/\partial\theta^2|_{\theta^3}]^{-1}\left[-\int \frac{\partial^2 \log p(\theta|Y,Z)}{\partial\theta^2} p(Z|Y,\theta^3)dZ|_{\theta^3}\right]$$

$$= \frac{434.79}{376.95} = 1.153442 \ .$$

It then follows that

$$\theta^4 = 0.626757 + 1.153442(0.626812 - 0.626757) = 0.6268205 \ .$$

CHAPTER 5

The Data Augmentation Algorithm

5.1. Introduction and Motivation

Analogous to the *EM* algorithm, the data augmentation algorithm exploits the simplicity of the likelihood function or posterior distribution of the parameter given the augmented data. In contrast to the *EM* algorithm, the present goal is to obtain the entire (normalized) likelihood or posterior distribution, not just the maximizer and the curvature at the maximizer. In large samples, it is comforting that the posterior or likelihood is consistent with the normal approximation, though in practice it is not often clear when one is in a large sample setting. In a small sample situation, the data augmentation algorithm will provide a way of improving the inference, based on the entire posterior distribution or the entire likelihood function.

The basic idea behind the data augmentation algorithm is to augment the *observed data* Y by a quantity Z which will be referred to as *latent data*. It is assumed that given both Y and Z, one can calculate or sample from the augmented posterior $p(\theta | Y, Z)$. To obtain the posterior $p(\theta | Y)$, one generates multiple values (imputations) of Z from the predictive distribution $p(Z | Y)$ and then computes the average of $p(\theta | Y, Z)$ over the imputations. Because $p(Z | Y)$ depends on $p(\theta | Y)$, one obtains an iterative algorithm to calculate $p(\theta | Y)$.

The data augmentation algorithm is motivated by two simple identities. The *posterior identity* is given by:

$$p(\theta | Y) = \int_Z p(\theta | Y, Z) p(Z | Y) dZ \ ,$$

where $p(\theta| Y)$ denotes the posterior density of the parameter θ given the observation Y, $p(Z|Y)$ denotes the predictive density of the latent data Z given Y, and $p(\theta|Y, Z)$ denotes the conditional density of θ given the augmented data $X = (Y, Z)$, i.e. the augmented posterior. The *predictive identity* is given by:

$$p(Z|Y) = \int_{\Theta} p(Z|\phi, Y)p(\phi|Y)d\phi \ ,$$

where $p(Z|\phi, Y)$ is the conditional predictive distribution. Substituting the predictive identity into the posterior identity and interchanging the order of integration, $p(\theta|Y)$ satisfies the following integral equation

$$g(\theta) = \int K(\theta, \phi)g(\phi)d\phi \ ,$$

where

$$K(\theta, \phi) = \int p(\theta|Z, Y)p(Z|\phi, Y)dZ \qquad (5.1.1)$$

and $g(\theta) = p(\theta| Y)$. The method of successive substitution can now be used to solve (5.1.1). In other words, starting with an initial approximation $g_0(\theta)$, calculate successively

$$g_{i+1}(\theta) = (Tg_i)(\theta) \qquad (5.1.2)$$

where

$$Tf(\theta) = \int K(\theta, \phi)f(\phi)d\phi \ .$$

Tanner and Wong (1987) adopt the method of Monte Carlo to perform the integration in (5.1.1). In particular, applying the method of Monte Carlo to the posterior identity yields the following iterative scheme:

a. Generate a sample $z_1, \ldots, z_m$ from the current approximation to the predictive distribution $p(Z|Y)$.
b. Update the current approximation to $p(\theta|Y)$ to be the mixture of augmented posteriors of θ, given the augmented data from (a), i.e.

$$g_{i+1}(\theta) = \frac{1}{m} \sum_{j=1}^{m} p(\theta|z_j, Y) \ .$$

Given the current guess to the posterior, $(g_i(\theta))$, the method of composition is applied to the predictive identity to generate a sample of latent data.

a1. Generate θ^* from $g_i(\theta)$.
a2. Generate z from $p(Z|\theta^*, Y)$, where θ^* is the value generated in (a1).

These two steps are then repeated m times to obtain: $z_1, \ldots, z_m$. Rubin (1987b) refers to the quantities $z_1, \ldots, z_m$ as *multiple imputations*. In this regard, we call step (a) the *imputation step*, while step (b) is referred to as the *posterior step*. The data augmentation algorithm consists of iterating between the imputation and the posterior steps.

Two important practical considerations in the implementation of the data augmentation algorithm are the monitoring of convergence and the selection of the number of imputations (m) to be performed per iteration. When m is large, the two steps (a) and (b) will provide a close approximation to the iteration of (5.1.2). Further comments are presented in Section 5.4. Regularity conditions under which the algorithm converges are given in Section 5.5.

EXAMPLE. *Genetic Linkage (Continued)*
As noted in Chapter 4, the augmented posterior for the genetic linkage model is of the beta form

$$p(\theta \mid Y, Z) \propto \theta^{x_2+x_5}(1-\theta)^{x_3+x_4} .$$

The conditional predictive distribution $p(Z \mid \theta, Y)$ is the binomial distribution with $n = 125$ and $p = \theta/(\theta + 2)$. Thus, the algorithm is given as

a. Imputation Step

a1. Draw θ^* from the current estimate of the posterior.
a2. Generate x_2 by drawing from the binomial distribution with parameters $[125, \theta^*/(\theta^* + 2)]$.

Repeat steps (a1) and (a2) m times to yield $x_2^{(1)} \ldots x_2^{(m)}$.

b. *Posterior Step*

Set the posterior density of θ equal to the mixture of beta distributions, mixed over the m imputed values of x_2; i.e.

$$p(\theta \mid Y) = \frac{1}{m} \sum_{i=1}^{m} Be(v_1^{(i)}, v_2^{(i)})(\theta)$$

where

$$v_1^{(i)} = x_2^{(i)} + x_5 + 1, \qquad v_2^{(i)} = x_3 + x_4 + 1$$

and

$$Be(v_1, v_2)(\theta) = \frac{\Gamma(v_1 + v_2)}{\Gamma(v_1)\Gamma(v_2)} \theta^{v_1-1}(1-\theta)^{v_2-1} .$$

In this example, the prior for θ is assumed to be uniform in (0,1). Steps (a1), (a2) and (b) are to be iterated until convergence of the algorithm is achieved.

In step (a1) of the first iteration, one would draw a value for θ from a reasonable starting distribution (e.g. uniform). This value for θ would then be used to generate a binomial deviate in step (a2). Steps (a1) and (a2) would then be repeated m times to generate m imputations. In subsequent iterations, the current approximation to the posterior is a mixture of beta distributions.

To sample a value for θ from the mixture of betas:

$$\frac{1}{m}\sum_{j=1}^{m} Be(v_1^{(i)}, v_2^{(i)})(\theta) \ ,$$

one randomly selects a distribution (e.g. $Be(v_1^{(j)}, v_2^{(j)})(\theta)$) from this mixture and then draws a θ value from this beta distribution.

Figure 5.1 presents the posterior density estimates of θ for this example. In particular, the normal approximation with $\hat{\mu} = 0.63$ and $\hat{\sigma} = 0.05$ (solid line) is plotted along with the true posterior distribution (dotted line)

$$p(\theta | Y) \propto (2 + \theta)^{y_1}(1 - \theta)^{y_2 + y_3}\theta^{y_4}$$

and the estimated posterior (dashed line) obtained by plotting the mixture of the beta distribution at the final iteration in which $m = 1600$. In the density scale, all three estimates are congruent. If possible, it is preferable to use the mixture as the approximation to the posterior rather than using the histogram of the generated θ values. However, as in the next example, this may not always be possible, though the approach due to Chen (1992) – see Section 5.9.3 – may help.

Alternatively, a second version of these data is examined in which the sample size is reduced by a factor of 10, though the cell proportions are approximately unchanged; i.e. $Y = (13, 2, 2, 3)$. The resulting posterior density estimates are plotted in Figure 5.2. In this case, although the true posterior density and the estimated posterior density are congruent, the validity of the

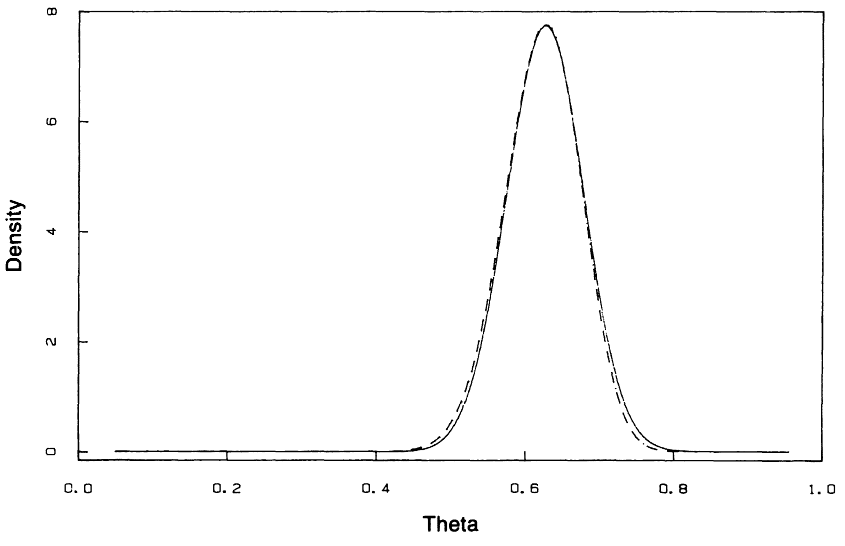


Figure 5.1. Posterior distribution of θ.

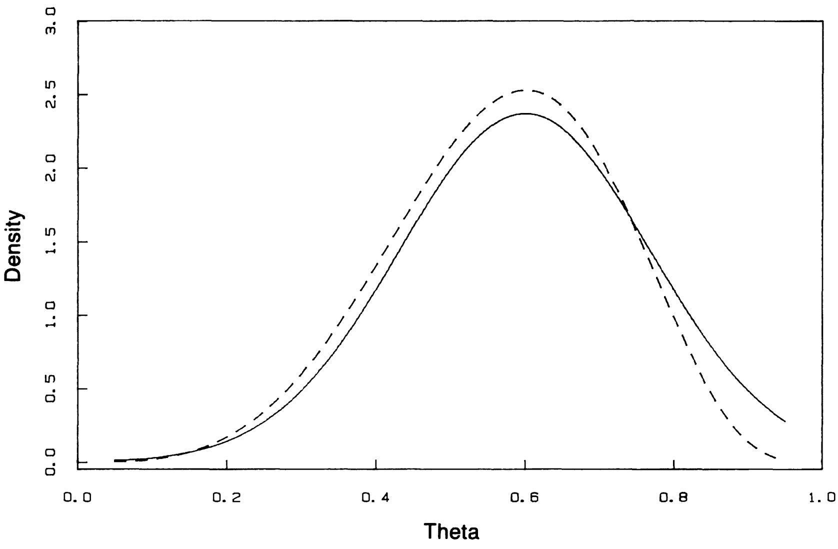


Figure 5.2. Posterior distribution of θ.

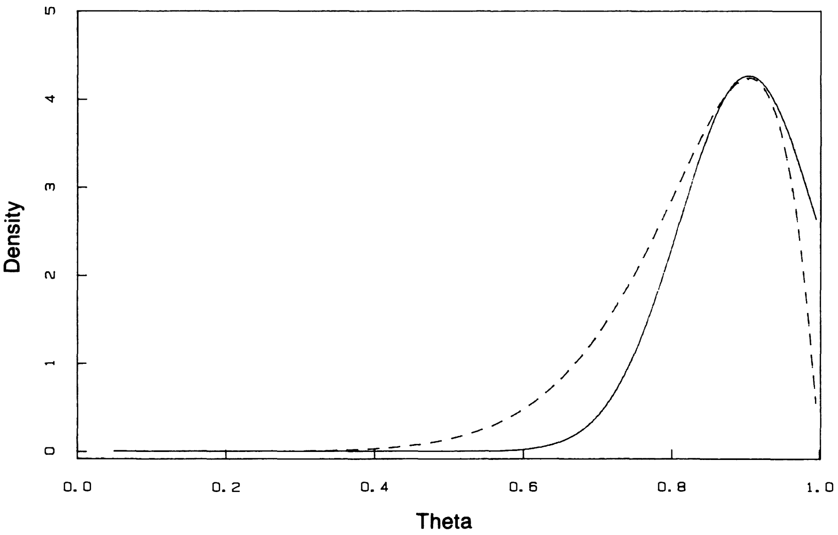


Figure 5.3. Posterior distribution of θ.

normal approximation may be in doubt, even when viewed on the density scale.

An even more dramatic illustration is given in Figure 5.3 where $Y = (14, 0, 1, 5)$. In cases with such a dramatic departure from normality, one or two iterations of the data augmentation algorithm indicates the inadequacy of the normal approximation.

Table 5.1. Twelve Observations from a Bivariate Normal Distribution.

1	1	−1	−1	2	2	−2	−2	*	*	*	*
1	−1	1	−1	*	*	*	*	2	2	−2	−2

*Value not observed (missing at random).

Example. *Bivariate Normal Covariance Matrix*

Suppose the data in Table 5.1 represent 12 observations from the bivariate normal distribution with $\mu_1 = \mu_2 = 0$, correlation coefficient ρ and variances σ_1^2 and σ_2^2. Note that in the four pairs of observations, two pairs have correlation 1 and the remaining two pairs have correlation -1. Thus, one can expect a nonunimodal posterior distribution for ρ in this data set. In such a case, the maximum likelihood estimate and the associated standard error will clearly be misleading. Furthermore, note that the information regarding σ_1^2 and σ_2^2 in the eight incomplete observations cannot be ignored because information regarding σ_1^2 and σ_2^2 is of use in making inference regarding ρ.

In this example, the conditional predictive distribution ($p(Z|\theta, Y)$) is the conditional normal. Under the prior on Σ given as

$$p(\sum) \propto |\sum|^{-(d+1)/2} ,$$

where d is the dimension of the observation vector, the augmented posterior of Σ is an inverted Wishart distribution (Box and Tiao, 1973). Thus, the data augmentation algorithm is as given below.

a. Imputation Step

a1. Draw Σ from the current estimate of the observed posterior.
a2. Generate latent data as follows.

If x_1 is known, then generate the unobserved observation from

$$N\left[\rho\frac{\sigma_2}{\sigma_1}x_1, \sigma_2^2(1-\rho^2)\right] .$$

If x_2 is known, then generate the unobserved observation from

$$N\left[\rho\frac{\sigma_1}{\sigma_2}x_2, \sigma_1^2(1-\rho^2)\right] .$$

Repeat steps (a1) and (a2) m times.

b. Posterior Step

Set the posterior density of Σ equal to the mixture of inverted Wishart distributions.

Regarding the implementation of the algorithm, it is noted that the algorithm of Odell and Feiveson (1966) can be used to generate observations

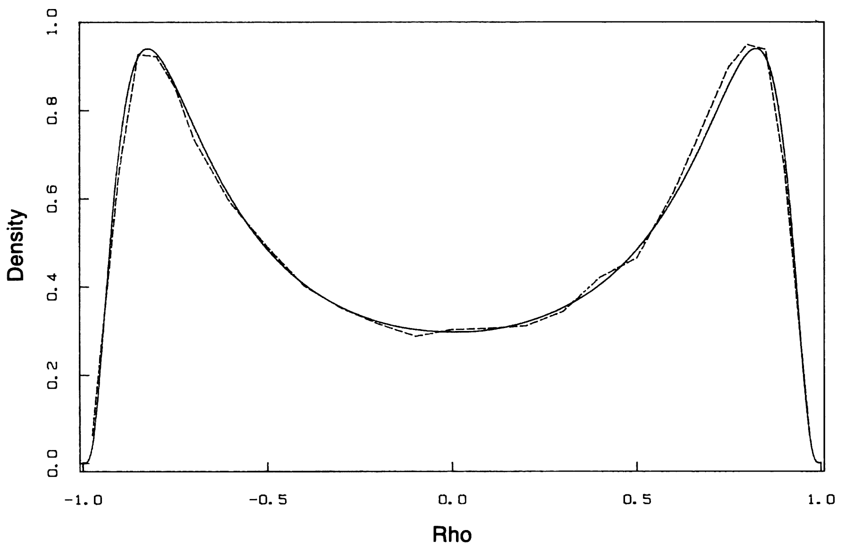


Figure 5.4. Posterior distribution of ρ.

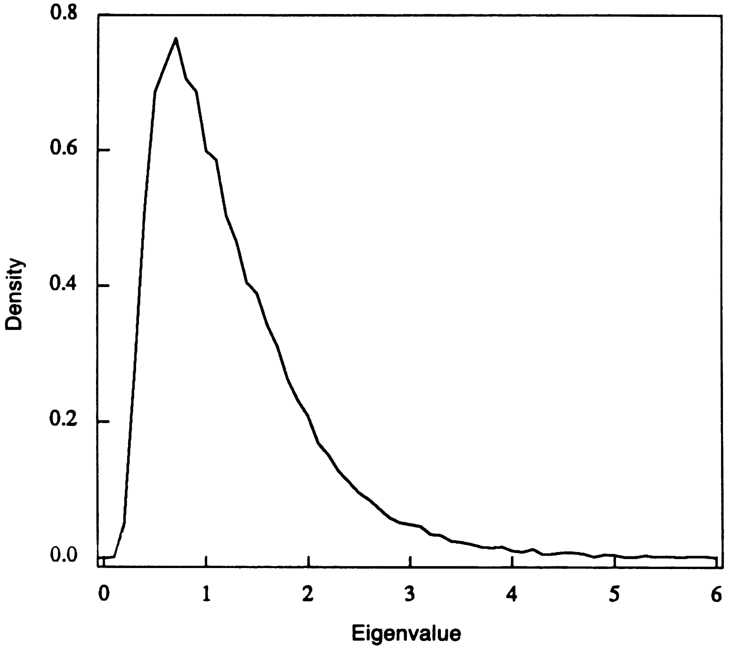


Figure 5.5. Posterior distribution of smallest eigenvalue.

from the inverted Wishart distribution. The amount of computation in this algorithm is not extensive, since the computation is of order $d/(d+1)/2$, which does not depend on the sample size.

In Figure 5.4, the histogram of the imputed correlation coefficients based on pooling iterations 10 to 15 ($m = 6400$) is presented. In addition, the true posterior of the correlation coefficient, which is proportional to $(1-\rho^2)^{4.5}/(1.25-\rho^2)^8$, is also plotted. As is evident from the plot, the estimated posterior distribution recovers the bimodal nature of the true distribution.

Finally, it is noted that the data augmentation algorithm can be used to examine the posterior distribution of any functional of the covariance matrix. For example, the posterior distribution of the smallest eigenvalue of the covariance matrix (Tiao and Fienberg, 1969) may be examined by simply computing the smallest eigenvalue of each of the observations sampled from the mixture of inverted Wishart distributions computed in step b of the algorithm (see Figure 5.5).

Example. *Censored Regression* (*Continued*)
To apply the data augmentation algorithm , we need to know how to sample from an augmented posterior and how to sample from the conditional predictive distribution. As in Section 4.1, we assume that the errors follow the standard normal distribution. See Wei and Tanner (1990c).

In the case of no censoring, the posterior density function of $(\sigma^2, \beta_0, \beta_1)$ can be factored exactly into the product of the marginal posterior density of σ^2 and the conditional marginal density of (β_0, β_1) given σ^2. In fact, for the noninformative prior, it is well known (Box and Tiao, 1973) that the marginal posterior distribution of σ^2 is that of the distribution of the random variable $(n-d)s^2/\chi^2_{n-d}$, where χ^2_v is a chi-square random variable with degrees of freedom v. The conditional marginal posterior distribution of (β_0, β_1) is a conditional bivariate normal distribution centered at the least-squares estimate of (β_0, β_1) for the given data set, conditional on σ^2.

The factorization in the previous paragraph is a convenient tool for the sampling of $(\sigma^2, \beta_0, \beta_1)$ from an augmented posterior. In particular, for an augmented data set, one first generates σ^2_* from $p(\sigma^2 \mid s^2)$, then generates a value of (β_0, β_1) from the conditional bivariate normal distribution.

To impute a failure time for the right-censored observation given the values σ_* and (β_0, β_1), one simply draws from the conditional distribution of the error, conditional on the fact that the failure time must be larger than the observed event time (see Figure 5.6.) Had the data been left censored, to impute a failure time one would draw from the conditional error distribution, conditional on the fact that the failure time must be *smaller* than the observed event time. Had the data been interval censored, to impute a failure time, one would draw from the conditional error distribution, conditional on the fact that the failure must have occurred in a particular interval.

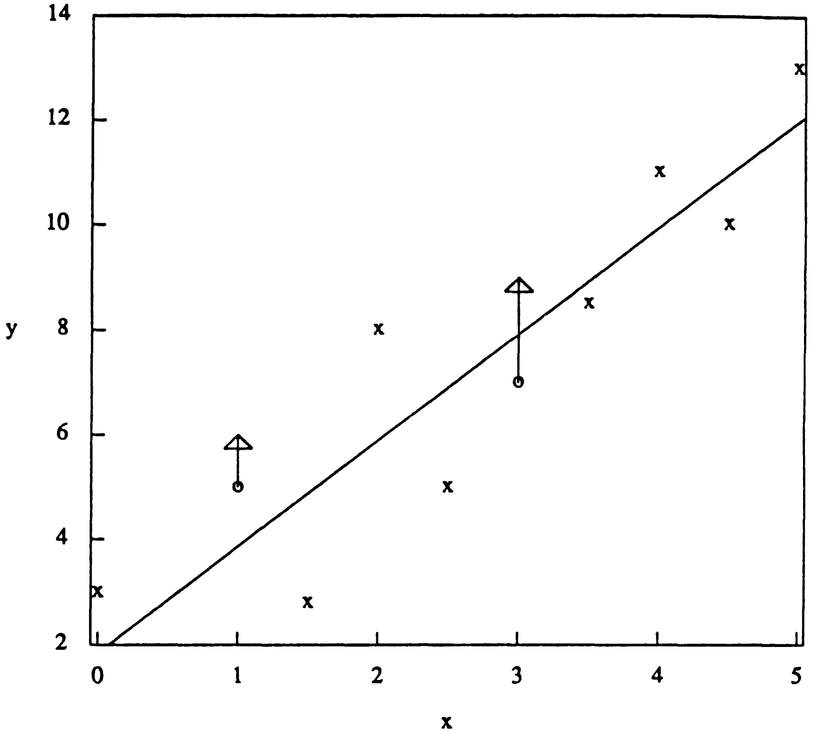


Figure 5.6. Right censored event data.

In the present situation of right-censored normal data, we draw from the distribution

$$\frac{\phi(s)}{1 - \Phi(z_i)},$$

where $\phi(s)$ and $\Phi(s)$ are the density and cdf of the standard normal distribution, respectively, and z_i is the value

$$z_i = \frac{c_i - \beta_0^* - \beta_1^* x_i}{\sigma_*}.$$

Having presented an approach for generating the tuple $(\sigma, \beta_0, \beta_1)$ from an augmented posterior, as well as for generating the latent data, the data augmentation algorithm for the regression analysis of censored data with normal errors is given below.

a1. Generate (β_0, β_1) and σ^2 from the current guess to the posterior.
a2. Generate latent data from the conditional normal distributions.

Repeat steps (a1) and (a2) m times.

b. Update the posterior as a mixture of augmented data posteriors, mixed over the latent data, generated in part (a) of the algorithm.

In general, one may stop the data augmentation algorithm at any point and realize a sample from $p(\theta | Y)$, rather than from $g_i(\theta)$. In particular, by attaching weights (w_j) proportional to

$$\frac{p(\theta_j | Y)}{g_i(\theta_j)}$$

to the sample values $\theta_1, \theta_2, \ldots, \theta_n$ from $g_i(\theta)$, one obtains a sample from the desired distribution. However, as pointed out in the rejoinder in Tanner and Wong (1987), because the variance of any function of these θ and weight values will depend on the values

$$\frac{p(\theta_j | Y)^2}{g_i(\theta_j)},$$

it is important to iterate the data augmentation algorithm to help eliminate outliers which would inflate the (Monte Carlo) variance of the estimate. Section 5.4 discusses how the w_j can be used to assess convergence of the algorithm.

Figure 5.7 presents the posterior density of β_1 calculated via the data augmentation algorithm (solid and long-dash lines), along with the normal aproximation based on the maximum likelihood solution (dotted line) and the true posterior density (short-dash line). The true posterior was calculated by numerical integration. A histogram of the 300 values of β_1 from the 12–15

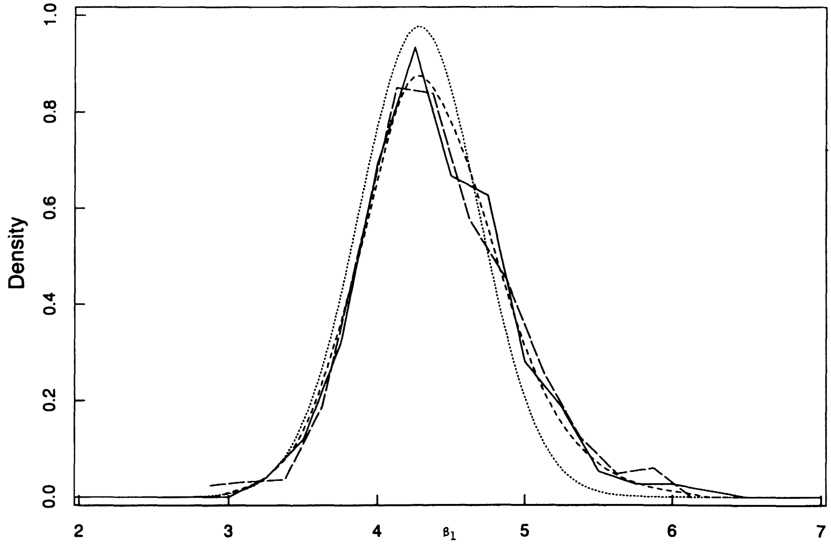


Figure 5.7. Posterior distribution of β.

iterations ($m = 100$) is given by the solid curve. The corresponding curve for the weighted values (weighted according to p/g) is given by the long-dash curve. As can be seen, the normal approximation did not capture the slight skew of the true posterior. Both the weighted and unweighted data augmentation curves detected this skew. If a more precise estimate is required, then the value of m could be increased and the algorithm iterated further. Chib (1992) uses data augmentation to analyze a censored regression model arising in econometrics.

5.2. Computing and Sampling from the Predictive Distribution

Having obtained a sample from $p(\theta | Y)$, calculating the posterior distribution of any functional of θ is straightforward. In this section, we focus on the predictive distribution for a future observation. In the following section, we consider the content and boundary of the highest posterior density region.

Suppose y_f is a (univariate) future observation from the model

$$y_f = v_f^T \beta + \sigma \varepsilon_f \ .$$

As noted in Section 3.3.2, the predictive distribution for y_f is given by

$$p(y_f | Y) = \int \phi(y_f | v_f^T \beta, \sigma^2) p(\sigma^2, \beta | Y) d\sigma^2 d\beta \ .$$

When the errors follow the normal distribution, $\phi(x | \mu, \sigma^2)$ is the density of the normal distribution.

Following the Monte Carlo method, given a sample of size m from $p(\sigma^2, \beta | Y)$, one may evaluate the expression

$$\frac{1}{m} \sum_{i=1}^{m} \phi(y_f | v_f^T \beta_i, \sigma_i^2)$$

over the m β and σ^2 values to obtain an estimate of the predictive distribution.

One may sample from $p(y_f | Y)$ by first drawing (σ_0^2, β_0) from $p(\sigma^2, \beta | Y) d\sigma^2 d\beta$, and then generate a random number from $\phi(x | \sigma_0^2, \beta_0)$. Repeating this two-stage algorithm yields a random sample from the exact distribution. One may then use these samples to estimate the predictive distribution using a density estimator such as the kernel estimator.

Gelfand and Smith (1990) show that the first approach to the estimation of the density has a smaller mean squared error and is, therefore, the preferred method. However, the second approach does yield a sample from the predictive distribution which can be used to compute the posterior of a variety of quantities as the Kullback–Leibler divergence measure (Cook and Weisberg, 1982).

EXAMPLE. *Censored Regression* (*Continued*)
Having applied the data augmentation algorithm to this data set, we have in the final iteration a sample from $p(\beta_0, \beta_1, \sigma^2 | Y)$. Thus, it is straightforward to calculate the predictive distribution for a future observation.

Figure 5.8 presents two curves representing the predictive distributions at temperatures 180° and 130° respectively, for the motorette data and matching normal curves. (These matching normal curves are not the normal approximation given by the maximum likelihood analysis.) The densities were constructed by drawing histograms of the values generated from the appropriate normal distribution. As can be seen from Figure 5.8, both curves are slightly more skewed than a normal curve, with the curve corresponding to the lower temperature being more skewed. The means of the two distributions are slightly larger than the corresponding quantities computed via maximum likelihood (as reported in Schmee and Hahn, 1979). On the log scale, the maximum likelihood means are 3.49 and 4.67, respectively, while the data augmentation values are 3.52 and 4.72, respectively. Moreover, the maximum likelihood approach underestimates the variability of the predicted values. For example, the interval obtained by computing the 5th and 95th percentiles of the appropriate predictive distribution at 130° is (on the log scale) given by [4.17, 5.31]. The corresponding normal approximation based interval reported in Schmee and Hahn (1979) is [4.44, 4.90]. This comment reflects the remarks in Aitkin (1980) and Naylor and Smith (1982) that the downward bias that arises in the maximum likelihood estimate of σ produces misleading statements of uncertainty.

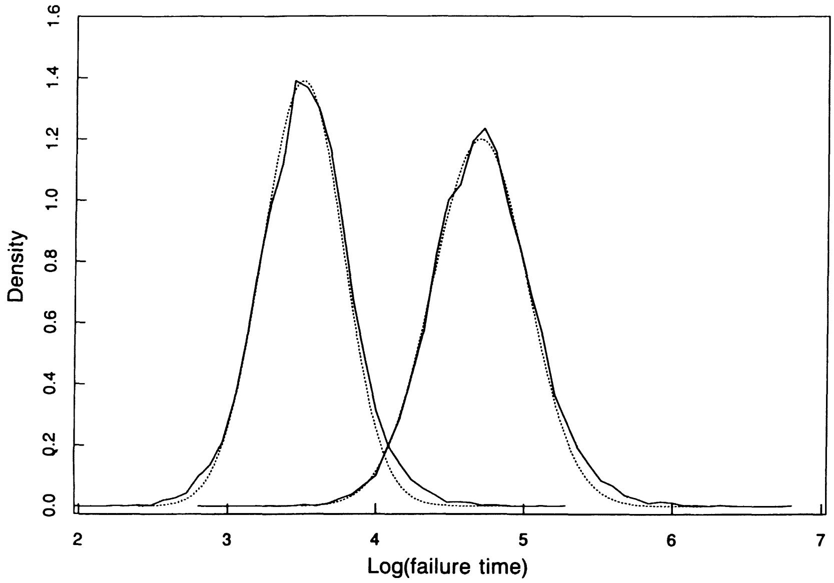


Figure 5.8. Predictive distribution at 180° and 130°.

5.3. Calculating the Content and Boundary of the HPD Region

As noted in Section 2.5, it is sometimes of interest to delineate a region of the parameter space which contains most of the mass under the posterior. Box and Tiao (1973, p. 122) point out that for a given probability content $1 - \alpha$, the region of highest posterior density (HPD) has smallest volume in parameter space. In addition, these authors point out that a point θ_0 is covered by the highest posterior density region of content $1 - \alpha$ if and only if $P_\theta[p(\theta| Y) \geq p(\theta_0| Y)| Y] \leq 1 - \alpha$, where the density $p(\theta| Y)$ is treated as a random variable. In practice, the content may be computed through the posterior distribution, by some monotonic function of the posterior, or by relying on a χ^2 approximation to $-2\ln[p(\theta| Y)/p(\hat{\theta}| Y)]$, where $\hat{\theta}$ is the posterior mode.

5.3.1. Calculating the Content

The computation of $P_\theta[p(\theta| Y) \geq p(\theta_0| Y)| Y]$ may not be easy, especially for a high-dimensional parameter space. However, in the context of data augmentation, this probability is easily estimated. Given a sample of size m from $p(\theta| Y)$, estimate the probability as

$$\frac{\text{Number of samples such that } p(\theta| Y) \geq \mathrm{p}(\theta_0| Y)}{m}$$

where θ_0 is a fixed value, and $p(\theta| Y)$ is evaluated over the m values of θ obtained from the final iteration of data augmentation.

5.3.2. Calculating the Boundary

To compute the boundary of the HPD Region of content $1 - \alpha$ (see Figure 5.9).

1. Locate the 100α percentile of the $p(\theta| Y)$ values evaluated at the θ values obtained from the final iteration of the algorithm.
2. Label this value as $p^*(\theta| Y)$.
3. Locate those θ values such that the value of the posterior density evaluated at these points is contained in the interval

$$\{p^*(\theta| Y) - \varepsilon, p^* (\theta| Y) + \varepsilon\} .$$

4. Label this set as S_ε.

If $p(\theta| Y)$ unimodal and θ is one dimensional, then for a sufficiently large value of ε there are two distinct groups of value in S_ε:

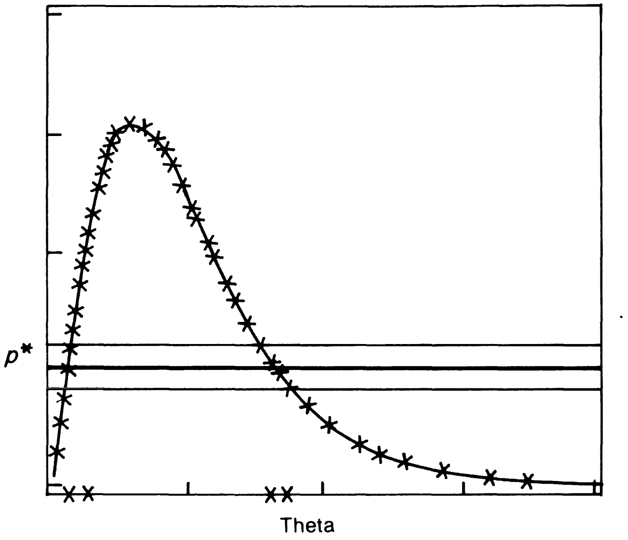


Figure 5.9. Calculating the boundary of an HPD region.

The left-hand endpoint of the HPD in such a case is given by the θ value in the left-hand group of values in S_ε whose posterior is closest to $p^*(\theta | Y)$. The right-hand endpoint is the corresponding θ value in the right-hand group.

In higher dimensions, the boundary of the HPD region of content $1 - \alpha$ is approximated by drawing a scatter plot of points in S_ε. One may wish to start with a value of ε large enough to include 3–4% of the θ values and then construct scatter plots for decreasing values of ε. A plot based on a small value of ε will have a smaller bias, but will be sparser than a plot based on a larger value of ε. Such a plot may be enhanced by including a smooth curve or surface through the points.

Suppose that θ is partitioned as (θ_1, θ_2), where θ_1 is $k \times 1$ and θ_2 is $(d - k) \times 1$ and interest focuses on θ_1. To compute the content of the HPD region, one must be able to evaluate $p(\theta_1 | Y)$. In some situations, one may be able to compute $p(\theta_1 | Y)$ analytically, by analytic integration of the posterior. Alternatively, it is possible that

$$\frac{1}{m} \sum_{i=1}^{m} p(\theta_1, \theta_2 | Y, z_i)$$

may be integrated analytically, while the integration of $p(\theta | Y)$ may not be straightforward. This is illustrated in the following example. See also Wei and Tanner (1990b).

Example. *Censored Regression (Continued)*
The data augmentation algorithm was implemented with $m = 10{,}000$ and allowed to run for 15 iterations. Given the output from the final iteration, the

computation of the content of the smallest HPD region which includes the value $\sigma^2 = 0.16$, as well as the computation of the boundary of 95% HPD region for σ^2 is straightforward. Note that $p(\sigma^2 | Y)$ is readily available since $p(\sigma^2 | Y) = \int p(\sigma^2, \beta | Y) d\beta$ may be approximated by

$$\frac{1}{m} \sum_{i=1}^{m} \int p(\sigma^2, \beta | \text{augmented data set } i) \, d\beta \propto \frac{1}{m} \sum_{i=1}^{m} \frac{s_i^{38}}{(\sigma^2)^{20}} \exp\left(- \frac{19 s_i^2}{\sigma^2} \right),$$

where s_i^2 is the least-squares estimate of σ^2 given the ith augmented data set, due to the inverse chi-square/conditional normal factorization.

Given the output from the final iteration of the data augmentation algorithm, one may evaluate the proportion of samples such that $p(\sigma^2 | Y) \geq p(\sigma^2 = 0.16 | Y)$. For this example, the content of the smallest HPD region which includes the value $\sigma^2 = 0.16$ is equal to 0.94, with a Monte Carlo standard error of 0.0024. In addition, the HPD region of content 95% for σ^2 is given by the interval (0.032, 0.17). This interval was obtained by first evaluating $p(\sigma^2 | Y)$ over the σ^2 values obtained in the final iteration of the data augmentation algorithm, next locating the 5th percentile of these $p(\sigma^2 | Y)$ values and, finally, locating the "right" and "left" σ^2 values whose posterior are closest to the 5th percentile. It is noted that the output from the data augmentation algorithm indicates at what height to slice the posterior, and it also provides points which lie on or near the boundary.

The dashed line in Figure 5.10 presents the distribution of the posterior ratio values, $-2\ln[p(\sigma^2 | Y)/p(\hat{\sigma}^2 | Y)]$, evaluated over the sample σ^2 values

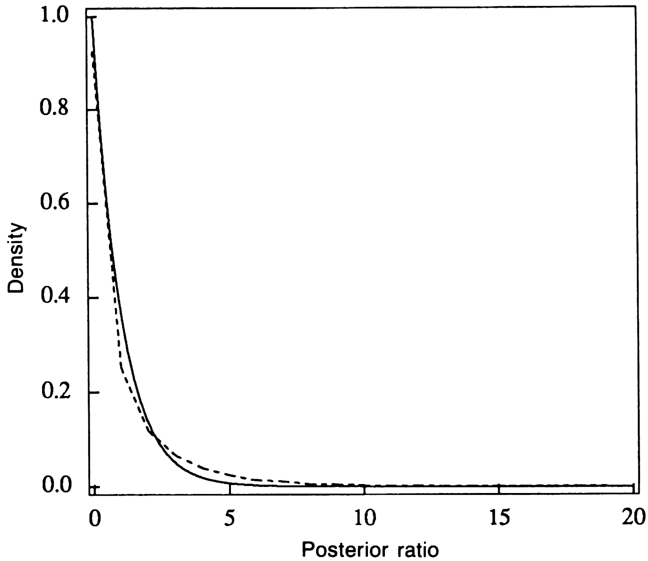


Figure 5.10. Distribution of posterior ratio values.

obtained from the final iteration of the data augmentation algorithm. The quantity $\hat{\sigma}^2$ is the marginal posterior maximizer, which in this case equals 0.0659. The solid line in Figure 5.10 is the corresponding χ_1^2 approximation. While the solid and dashed curves appear to be in good agreement, it is noted

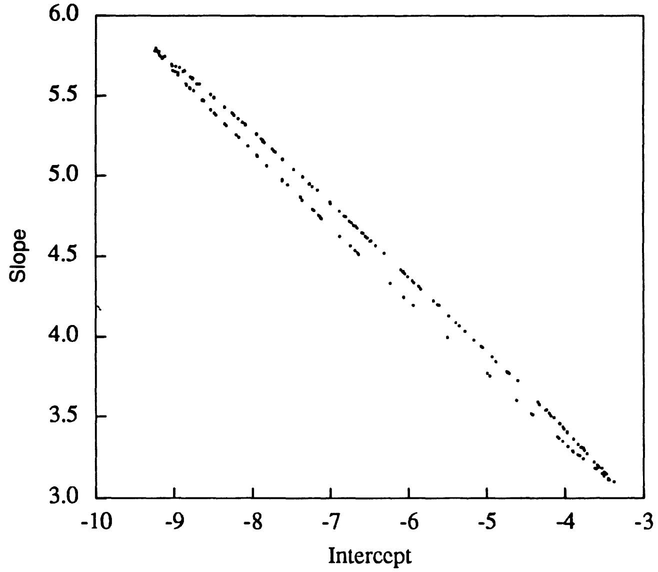


Figure 5.11(a). 95% HPD Region for the slope and intercept, 200 points.

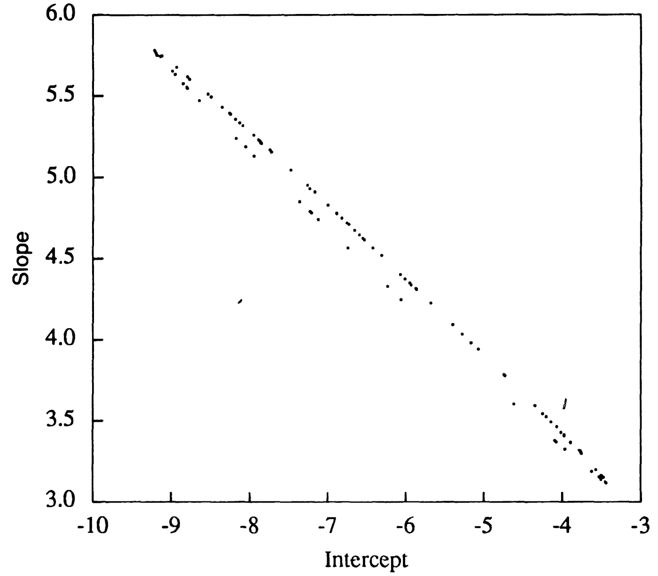


Figure 5.11(b). 95% HPD Region for the slope and intercept, 100 points.

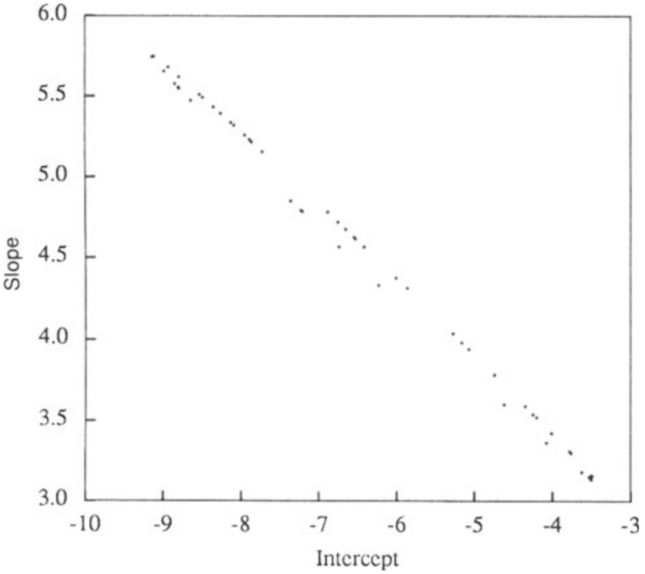


Figure 5.11(c). 95% HPD Region for the slope and intercept, 50 points.

that the area to the right of $-2\ln[p(\sigma^2 = 0.16|Y)/p(\hat{\sigma}^2|Y)]$ under the data augmentation curve is 0.06, while the corresponding value based on the χ^2 approximation is 0.03.

Similar computations regarding the intercept β_0 and the slope β_1 are facilitated by the observation that $p(\beta_0, \beta_1|Y)$ is estimated as

$$\frac{1}{m}\sum_{i=1}^{m}\frac{1}{s_i^2}\left[1 + \frac{(\beta - \hat{\beta}_i)^T X^T X(\beta - \hat{\beta}_i)}{38 s_i^2}\right]^{-20},$$

where $\hat{\beta}_i$ is the least-squares estimate of β given the ith augmented data set and s_i^2 is the corresponding estimate of σ^2. This follows from the inverse chi-square/conditional normal factorization. Figures 5.11(a)–(c) present the 95% HPD region for β_0 and β_1 based on a set S_ε containing 200, 100 and 50 points, respectively. All three plots highlight the high degree of correlation between the two parameters. One may compensate for the sparseness of the last figure by passing a smooth curve through the points.

5.4. Remarks on the General Implementation of the Data Augmentation Algorithm

Two important practical considerations in the implementation of the data augmentation algorithm are the monitoring of convergence and the selection of the number of imputations (m) to be performed per iteration. Tanner and Wong (1987) point out that it is helpful to monitor graphically the progress of the data augmentation algorithm, for example, using selected percentiles of the estimated posterior distribution. If one is interested in first and second moments, then these moments, rather than extreme tail behavior, may be monitored. The data augmentation algorithm (for a fixed value of m) may be iterated until the fluctuations in such a plot indicate that the process has become stationary. At such a point, the algorithm may be terminated or the value of m increased to improve the precision (with respect to the Monte Carlo variation) of the estimate of the functional of the posterior of interest. In this way, one may start with a smaller value of m and then increase the value of m at various junctures in the iteration process to realize computational savings.

Example. *Genetic Linkage (Continued)*

The observed data was taken to be (13, 1, 2, 3). The lines in Figure 5.12 represent the upper, middle and lower quartiles at each iteration. At the initial iteration, m was taken to be 20. The algorithm was then run through 40 iterations, at which point it appeared (see Figure 5.12) that the process had become stationary. The sample size was then increased to 400 and the algorithm proceeded through 20 further iterations. From Figure 5.12 we see that the effect of increasing m has been to reduce substantially the system variability. The final 10 iterations were run with $m = 1600$.

In an alternative approach to monitoring convergence of the algorithm (see Wei and Tanner, 1990c), one monitors the distribution of the weights (w_j) proportional to:

$$\frac{p(\theta_j | Y)}{g_i(\theta_j)}$$

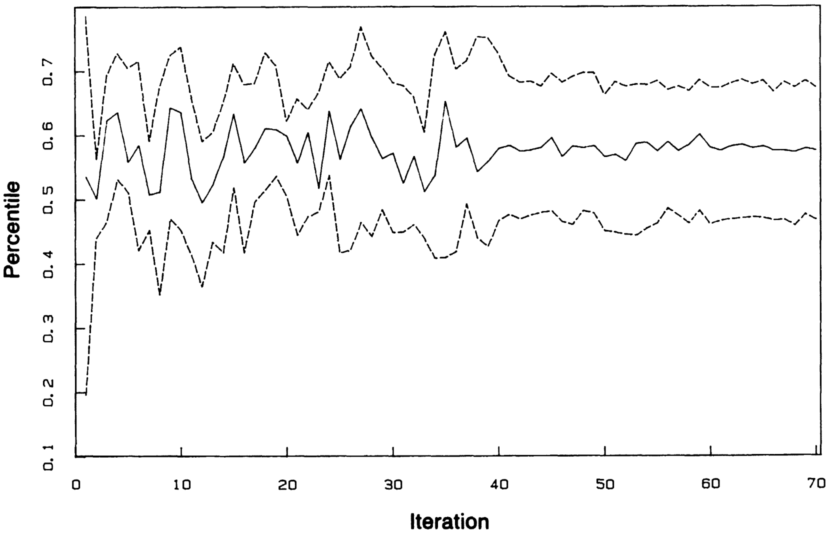


Figure 5.12. Upper, middle and lower quartiles versus iteration.

given the current approximation to the posterior g_i. Clearly, if the current estimate of the posterior is "close" to the posterior, then the distribution of the weights will be degenerate about a constant. In fact, if the weights are normalized to sum to unity, then this constant will be $1/m$. Thus, an alternative graphical approach to monitoring the convergence of the algorithm would be to construct a series of plots, each of which presents the distribution of the weights at a given iteration. A numerical approach would consist of computing some functional of the distribution of the weights, such as the standard deviation, and then to monitor this value as the iterations increase. As the data augmentation algorithm is iterated, for a fixed value of m, the standard deviation will decrease in magnitude. Eventually, the standard deviation will begin to fluctuate about a value. At such a point one may wish to increase m or to terminate the algorithm depending on the required degree of precision of the estimate of a functional of the posterior. Note that the information on the convergence of *all* the marginals is included in the series of distribution plots of the weights.

EXAMPLE. *Censored Regression* (*Continued*)
Six histograms are presented in Figure 5.13. The top-left histogram corresponds to the weights (w_j) proportional to:

$$\frac{p(\theta_j \mid Y)}{g_2(\theta_j)}$$

obtained after the second iteration, with $m = 100$. The standard deviation of the (normalized) weights ($\Sigma w_j = 1$) used to construct this plot is 0.017. The remaining histograms correspond to the weights obtained after iterations 5, 7, 10, 12 and 15, all with $m = 100$, with standard deviations of 0.006, 0.006, 0.004, 0.003, and 0.003, respectively. As one can see from the plots, by the 12th iteration, the distribution of the weights (normalized to sum to unity) appears to have degenerated about the value 1/100. In successive iterations, the distributions of the weights tend to fluctuate about this constant. Having arrived at a stationary process, the parameter values.from iterations 13–15 were then pooled. The unweighted average of these 300 β_1 values is 4.40, with a standard error of 0.03. It is noted that if one requires a more precise estimate, then the iterations could be continued with a larger value of m.

5.5. Overview of the Convergence Theory of Data Augmentation

For the examples considered in this chapter, the data augmentation algorithm was seen to converge. In fact, one can state several general results regarding the convergence of the algorithm. See Tanner and Wong (1987) for more details.

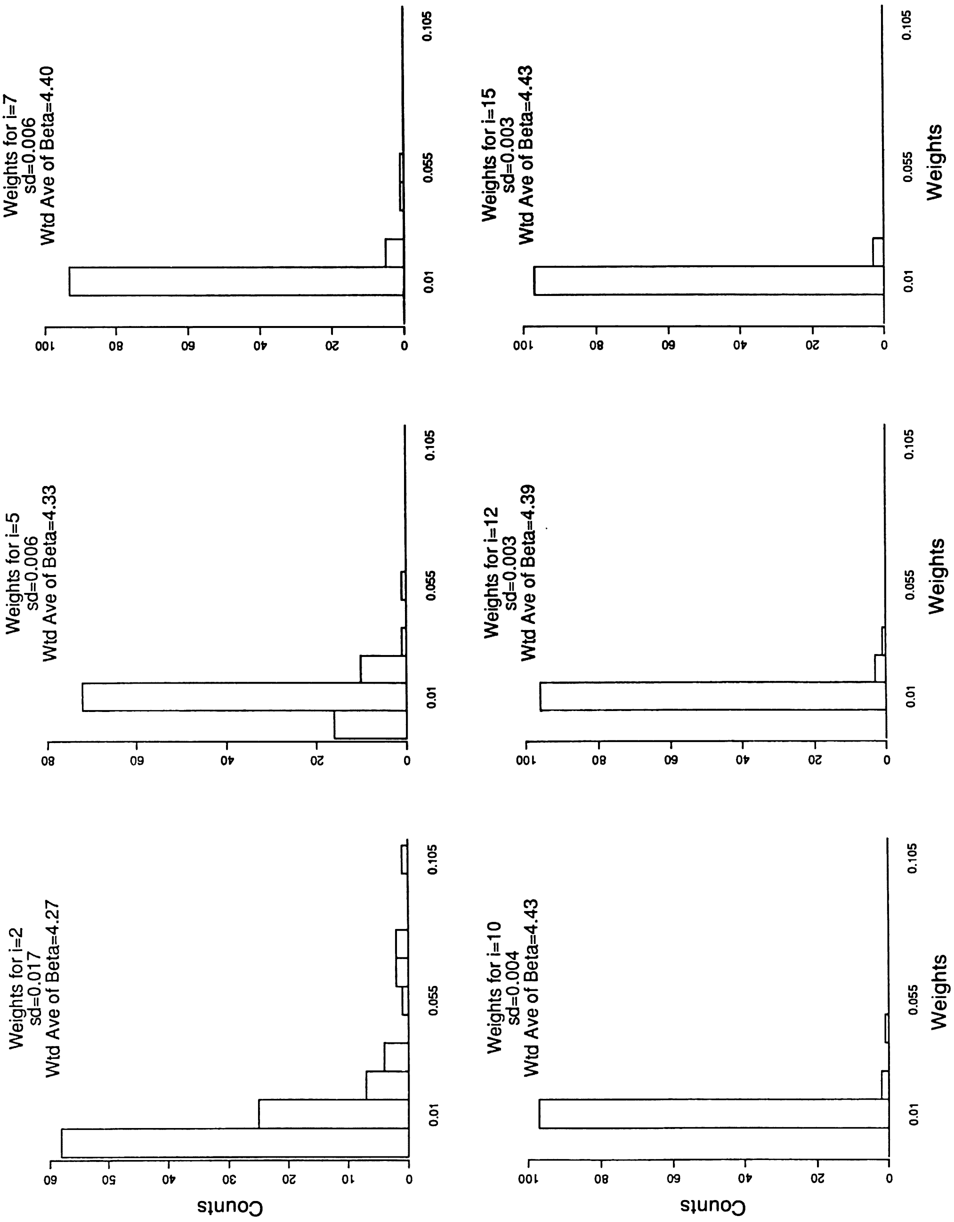


Figure 5.13. Distribution of weights across iterations.

Assume that Θ is a connected subspace of $\Re^d$. Let L_1 be the space of Lebesgue integrable functions of θ and $\|f\| = \int |f(\theta)| d\theta$ for $f \in L_1$. The posterior density satisfies

$$g(\theta) = \int K(\theta, \phi) g(\phi) d\phi \tag{5.5.1}$$

where $g(\theta) \geq 0$, $\int g(\theta) d\theta = 1$ and $K(\theta, \phi)$ is defined as in Section 5.1. Let $g^*(\theta) = p(\theta | Y)$.

The first result shows that the L_1 distances from g^* are nonincreasing in the iterations.

Result 5.5.1 $\|g_{i+1} - g^*\| \leq \|g_i - g^*\|$.

The next result guarantees the uniqueness of the solution to (5.5.1).

Result 5.5.2 Under the regularity condition, g^* is the only density that satisfies (5.5.1).

The final result relates to convergence of $g_{i+1}(\theta) = Tg_i(\theta)$, where $Tf(\theta) = \int K(\theta, \phi) f(\phi) d\phi$.

Result 5.5.3 Under the regularity condition and assuming that the starting value g_0 satisfies $\sup_\theta |g_0(\theta)/g^*(\theta)| < \infty$, $\|g_i - g^*\| \to 0$.

In Section 5.6, we discuss approaches for obtaining good starting points for data augmentation.

The regularity condition required in Results 5.5.2 and 5.5.3 is:

$K(\theta, \phi)$ is uniformly bounded and is equicontinuous in θ. For any $\theta_0 \in \theta$, there is an open neighborhood U of θ_0, so $K(\theta, \phi) > 0$ for all $\theta, \phi \in U$.

The second part of this condition says that if θ and ϕ are close, then it is possible to generate some latent data pattern z from $p(Z | \phi, Y)$ such that $p(\theta | Z, Y)$ is nonzero. Rosenthal (1991) discusses rates of convergence of data augmentation on finite sample spaces.

5.6. Poor Man's Data Augmentation Algorithms

In this section, we present the poor man's data augmentation algorithms. These algorithms can be used as good starting points for the full data augmentation analysis, as well as approximations to the full analysis.

5.6.1. *PMDA* 1

The data augmentation algorithm states:

a. Generate $z_1, \ldots, z_m \sim p(Z | Y)$.
b. Update the approximation to $p(\theta | Y)$ as

$$\frac{1}{m} \sum_{i=1}^{m} p(\theta | z_i, Y) \ .$$

Note that

$$p(Z | Y) = \int_\theta p(Z | Y, \theta) p(\theta | Y) d\theta = E\{p(Z | Y, \theta)\} \ .$$

Recalling Result 3.2.1 of Section 3.2.1, we have

$$p(Z | Y) = p(Z | Y, \hat{\theta}) \left\{1 + O\left(\frac{1}{n}\right)\right\} \ ,$$

where $\hat{\theta}$ is the mode of $p(\theta | Y)$. Thus, having obtained $\hat{\theta}$, we have *PMDA* 1.

a. Generate $z_1, \ldots, z_m \sim p(Z | Y, \hat{\theta})$.
b. Approximate the posterior as

$$\frac{1}{m} \sum_{i=1}^{m} p(\theta | z_i, Y) \ .$$

PMDA 1 is a noniterative algorithm for obtaining an approximation to $p(\theta | Y)$. This approximation allows for non-normal shapes for the posterior. It is a "poor man's" algorithm, in the sense that it can be used by those who cannot afford the full data augmentation analysis. In addition, it is noted that *PMDA* 1 may provide a good starting point for data augmentation.

There is a connection between *PMDA* and Monte Carlo *EM* (*MCEM*). Recall in *MCEM* that at the $(i + 1)$th iteration, the Q function is calculated as

$$\hat{Q}_{i+1}(\theta, \theta^i) = \frac{1}{m} \sum_{j=1}^{m} \log p(\theta | z_j, Y) \ , \qquad (5.6.1)$$

where $z_j \sim p(Z | \theta^i, Y)$. When θ^i has converged to the posterior mode $\hat{\theta}$, in *MCEM*, one can drop the logs in (5.6.1), to obtain *PMDA* 1, i.e. an approximation to the entire posterior, not just the maximizer.

EXAMPLE. *Genetic Linkage* (*Continued*)
To illustrate *PMDA* 1, consider the following small sample data set for the genetic linkage model: (14, 0, 1, 5). The *MCEM* algorithm was run with $m = 5000$ for 15 iterations, yielding $\hat{\theta} = 0.9034$. Using this value of $\hat{\theta}$, 5000 samples were drawn from the conditional predictive distribution $p(Z | Y, \hat{\theta})$.

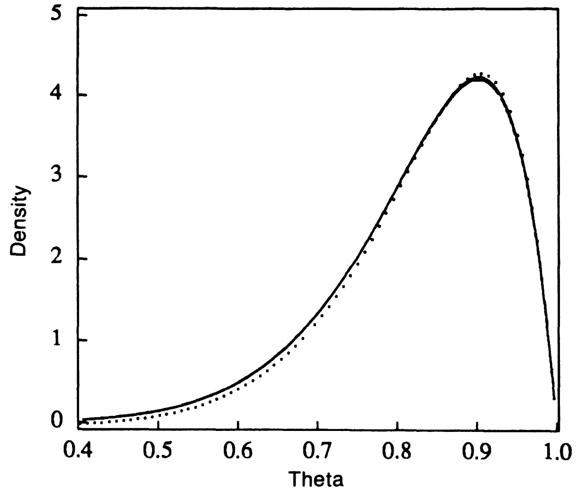


Figure 5.14. Marginal of θ. Solid – exact observed posterior, dotted – first-order approximation.

Figure 5.14 presents the mixture of augmented posteriors, mixed over the *PMDA* 1 imputations (dotted line), along with the exact observed posterior (solid line). As can be seen from Figure 5.14, *PMDA* 1 successfully recovers the highly skewed shape of the observed posterior.

5.6.2. *PMDA*-Exact

PMDA 1 is an approximation because $z_i \sim p(Z|\hat{\theta}, Y)$ rather than from $p(Z|Y)$. If it is straightforward to calculate $p(Z|Y)$, then one can use importance sampling to sample from the exact predictive distribution $p(Z|Y)$.

PMDA-Exact is given as:

a1. Generate $z_i, \ldots, z_m \sim p(Z|\hat{\theta}, Y)$.

a2. Calculate

$$w_j = \frac{p(z_j|Y)}{p(z_j|Y, \hat{\theta})}; \quad j = 1, \ldots, m \ .$$

b) Calculate the posterior as

$$\sum_{j=1}^{m} w_j p(\theta|z_j, Y) \Big/ \sum_{j=1}^{m} w_j \ .$$

5.6.3. *PMDA* 2

When $p(Z \mid Y)$ is difficult to compute, a second-order approximation is available. Again note that

$$p(Z \mid Y) = E[p(Z \mid Y, \theta)] = \int_{\Theta} p(Z \mid Y, \theta) p(\theta \mid Y) d\theta$$
$$= \int_{\Theta} p(Z \mid Y, \theta) \left[\frac{p(\theta \mid Y, Z) p(Z \mid Y)}{p(Z \mid Y, \theta)} \right] d\theta \; .$$

Recall Result 3.2.2 of Section 3.2.1, namely, via Laplace's method we have:

$$E[g(\theta)] = \left(\frac{\det \Sigma^*}{\det \Sigma} \right)^{1/2} \frac{\exp\{-nh^*(\theta^*)\}}{\exp\{-nh(\hat{\theta})\}} \left[1 + O\left(\frac{1}{n^2} \right) \right] ,$$

where

$$-nh(\theta) = \log p(\theta \mid Y), \qquad -nh^*(\theta) = -nh(\theta) + \log [g(\theta)],$$
$$\Sigma^* = \left[\frac{\partial^2 h^*}{\partial \theta^2} \bigg|_{\theta^*} \right]^{-1}, \qquad \Sigma = \left[\frac{\partial^2 h}{\partial \theta^2} \bigg|_{\hat{\theta}} \right]^{-1} ,$$

$\hat{\theta}$ is the maximizer of $-nh(\theta)$ and θ^* is the maximizer of $-nh^*(\theta)$.

In the present case,

$$-nh(\theta) = \log p(\theta \mid Y, Z) - \log p(Z \mid \theta, Y) + \log p(Z \mid Y)$$

and

$$-nh^*(\theta) = \log p(\theta \mid Y, Z) + \log p(Z \mid Y) \; .$$

Note that θ^* is the maximizer of $\log p(\theta \mid Y, Z)$. Hence, we have shown that to a second-order approximation,

$$p(Z \mid Y) \propto (\det \Sigma^*)^{1/2} \frac{p(\theta^* \mid Y, Z) p(Z \mid \hat{\theta}, Y)}{p(\hat{\theta} \mid Y, Z)} \; .$$

One can use importance sampling to sample from this second-order approximation. In particular, *PMDA* 2 is given as:

a1. Generate $z_1, \ldots, z_m \sim p(Z \mid \hat{\theta}, Y)$.
a2. Calculate

$$w_j = \frac{(\det \Sigma^*)^{1/2} p(\theta_j^* \mid Y, z_j) p(z_j \mid \hat{\theta}, Y)}{p(\hat{\theta} \mid Y, z_j) p(z_j \mid \hat{\theta}, Y)}$$
$$= \frac{(\det \Sigma^*)^{1/2} p(\theta_j^* \mid Y, z_j)}{p(\hat{\theta} \mid Y, z_j)} \; . \tag{5.6.2}$$

b. Approximate the posterior as

$$\sum_{j=1}^{m} w_j p(\theta \mid Y, z_j) \Big/ \sum_{j=1}^{m} w_j .$$

Example. *Censored Regression* (*Continued*)
In this example, $\hat{\theta}$ is the posterior mode, i.e. $(\hat{\beta}_0, \hat{\beta}_1, \hat{\sigma})$; $p(Z \mid \hat{\theta}, Y)$ is a conditional normal distribution and $p(\theta \mid Y, Z)$ is given by the inverse chi-square by conditional bivariate normal factorization.

To illustrate the *PMDA*s, we will focus on $p(\sigma^2 \mid Y)$. Following Wei and Tanner (1990a) we approximate

$$p(\sigma^2 \mid Y) = \int p(\sigma^2, \beta_0, \beta_1 \mid Y) \, d\beta_0 \, d\beta_1$$

by

$$\frac{1}{m} \sum_{i=1}^{m} \int p(\sigma^2, \beta_0, \beta_1 \mid Y, z_i) \, d\beta_0 \, d\beta_1 \propto \frac{1}{m} \sum_{i=1}^{m} \frac{s_i^{38}}{\sigma^{40}} \exp\left(\frac{-19 s_i^2}{\sigma^2}\right),$$

where s_i^2 is the least-squares estimate of σ^2 for the ith augmented data set. Thus, *PMDA* 1 is given as

a. Generate $z_1, \ldots, z_m \sim p(Z \mid Y, \hat{\theta})$.
b. Approximate $p(\sigma^2 \mid Y)$ as

$$\frac{1}{m} \sum_{i=1}^{m} \frac{s_i^{38}}{\sigma^{40}} \exp\left(\frac{-19 s_i^2}{\sigma^2}\right).$$

PMDA 2 is given as

a1. Generate $z_1, \ldots, z_m \sim p(Z \mid Y, \hat{\theta})$.
a2. Calculate w_j using (5.6.2) above.
b. Approximate $p(\sigma^2 \mid Y)$ as

$$\sum_{i=1}^{m} w_i \frac{s_i^{38}}{\sigma^{40}} \exp\left(\frac{-19 s_i^2}{\sigma^2}\right) \Big/ \sum_{j=1}^{m} w_j .$$

Figure 5.15 presents *PMDA* 1 (short dash line), *PMDA* 2 (long dash line) and the data augmentation estimate (solid line) of the marginal based on $m = 5000$. Note that all three curves are in general accord regarding the location of the mode. However, regarding the shape of the density, the normal approximation to the marginal does not seem appropriate in this case. (In fact, the normal approximation is not appropriate even on the $\log(\sigma)$ scale.) *PMDA* 1 gives a hint of the skew in the marginal posterior. *PMDA* 2 does, however, represent an improvement.

In practice, the magnitude of the error of *PMDA* 1 will not be known. Having obtained *PMDA* 1, one may wish to "check" it against *PMDA* 2. If they are in accord, there may be little reason to doubt the validity of either approximation. If, as in this case, they are in discord, it is not clear whether

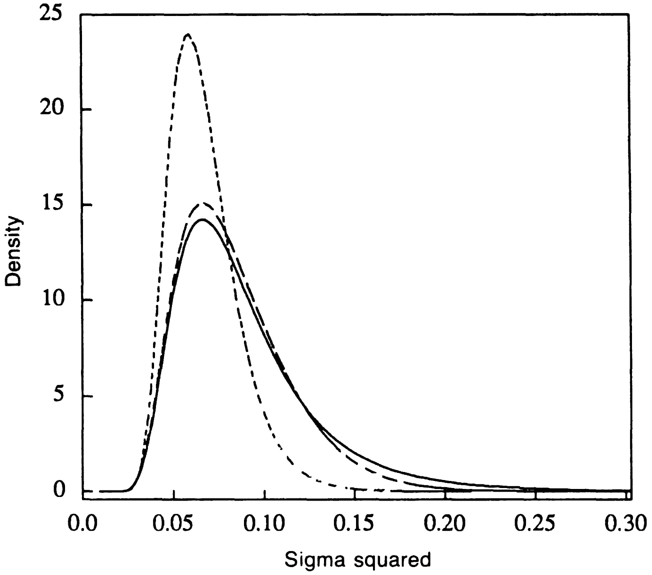


Figure 5.15. Marginal of σ^2. Short dash – first-order approximation; long dash – second-order approximation; solid – data augmentation.

the error in *PMDA* 2 can be neglected. One may then wish to proceed to the implementation of the full data augmentation algorithm using *PMDA* 2 as a starting point (i.e. $g_0(\theta)$) for the data augmentation algorithm.

5.7. Sampling/Importance Resampling (SIR)

Rubin (1987a) developed the *SIR* (noniterative) algorithm for drawing missing data patterns from $p(Z \mid Y)$. His intent was to develop methodology to allow for the supplementation of an incomplete public-use database. Standard complete-data methods can then be applied to each of the m augmented data sets to facilitate the analysis of the database. In this context, the fraction of missing information is typically modest. Moreover, one desires a technique which is adequate when m is small. Other imputation methods are presented in Section 5.8.

To begin, one requires a good approximation to $p(\theta, Z \mid Y)$,

$$h(\theta \mid Y)h(Z \mid \theta, Y) = h(\theta, Z \mid Y)$$

where $h(\theta \mid Y)$ approximates $p(\theta \mid Y)$ and $h(Z \mid \theta, Y)$ approximates $p(Z \mid \theta, Y)$.

Step 1. Draw M values of (θ, Z) from $h(\theta, Z \mid Y)$.

Step 2. Calculate

$$w_j = w(\theta_j, z_j \mid Y) \propto \frac{p(Y, z_j \mid \theta_j)p(\theta_j)}{h(\theta_j, z_j \mid Y)} \quad \text{for } j = 1, \ldots, M \ .$$

This follows since $p(\theta, Z \mid Y) \propto p(\theta)p(Y, Z \mid \theta)$.

Step 3. Draw $m < M$ values of Z from $z_1, \ldots, z_M$ with probabilities proportional to $w_1, \ldots, w_M$.

Cochran (1977, Chapter 9) presents methods for drawing such samples. The adequacy of the approximation can be assessed by examining the distribution of the weights. In particular, a highly skewed distribution indicates a poor approximation.

If one can sample from $p(Z \mid \theta, Y)$; i.e. $h(Z \mid \theta, Y) = p(Z \mid \theta, Y)$, then the weights

$$w_j \propto \frac{p(Y \mid \theta_j)p(z_j \mid \theta_j, Y)p(\theta_j)}{h(\theta_j \mid Y)p(z_j \mid \theta_j, Y)} = \frac{p(Y \mid \theta_j)p(\theta_j)}{h(\theta_j \mid Y)} \ .$$

Note in this case, the weights do not depend on z_j. Hence, in step 1, only the θ_j need be drawn. In addition, to draw the multiple imputations, draw m values of θ_j with probability proportional to w_j and for each θ_j^*, draw z_j^* from $p(Z \mid \theta_j^*, Y)$.

Note that as $M/m \to \infty$, the m pairs (z_j^*, θ_j^*) are selected with probability

$$\frac{h(\theta, Z \mid Y)w(\theta, Z \mid Y)}{\int\int h(\theta, Z \mid Y)w(\theta, Z \mid Y)d\theta\, dZ}$$

$$= \frac{p(Y, Z \mid \theta)p(\theta)}{\int\int p(Y, Z \mid \theta)p(\theta)d\theta\, dZ} = p(Z, \theta \mid Y) \ .$$

By "peeling-off" the z^* values, one obtains an iid sample of size m from $p(Z \mid Y)$.

The choice of M/m depends on the fraction of missing information. Rubin (1987-a) suggests that $M/m = 20$ should be adequate, though he does mention an adaptive scheme for selecting M/m which may warrant further study.

Gelfand and Smith (1990) modify *SIR* to obtain estimates of the functions $p(\theta \mid Y)$ or $p(Z \mid Y)$, rather than to obtain an iid sample from $p(Z \mid Y)$. In particular, they replace step 3 with weighted averages.

Step 3. a. Calculate $p(\theta \mid Y)$ via

$$\sum_{j=1}^{M} w_j p(\theta \mid Y, z_j) \Big/ \sum_{j=1}^{M} w_j \ .$$

b. Calculate $p(Z \mid Y)$ via

$$\sum_{j=1}^{M} w_j p(Z \mid Y, \theta_j) \Big/ \sum_{j=1}^{M} w_j \ .$$

By carrying along the weights, $(\theta_1, w_1), \ldots, (\theta_M, w_M)$ is a sample from $p(\theta \mid Y)$, while $(z_1, w_1), \ldots, (z_M, w_M)$ is a sample from $p(Z \mid Y)$.

Gelfand and Smith (1990) examine the performance of the *SIR* algorithm in several examples. They conclude that the performance of the algorithm depends on the specification of $h(\theta, Z \mid Y)$. This conclusion is not unexpected since this dependency is typical of importance sampling techniques.

A similar point based on theoretical considerations is made in the rejoinder of Tanner and Wong (1987). Tanner and Wong (1987) consider the estimation of the posterior moment:

$$\rho = \int a(\theta) p(\theta \mid Y) d\theta \ .$$

To estimate ρ via data augmentation, one can use

$$\hat{\rho}_{DA} = \frac{1}{M} \sum_{j=1}^{M} a(\theta_j) \ ,$$

where $\theta_1, \ldots, \theta_M \sim g_i(\theta)$. To estimate ρ via SIR, can can use

$$\hat{\rho}_{SIR} = \sum_{j=1}^{M} w_j a(\theta_j) \Big/ \sum_{j=1}^{M} w_j,$$

where $(\theta_1, z_1), \ldots, (\theta_M, z_M) \sim h(\theta, Z \mid Y)$ and $w_1, \ldots, w_M$ are the weights computed in Step 2.

It then follows that

$$E(\hat{\rho}_{DA}) = \int a(\theta) g_i(\theta) d\theta$$

and

$$E(\hat{\rho}_{SIR}) = \int a(\theta) p(\theta \mid Y) d\theta.$$

Hence, $\hat{\rho}_{SIR}$ is unbiased while $\hat{\rho}_{DA}$ is not unbiased, though the bias of $\hat{\rho}_{DA}$ decreases with increasing i since g_i converges to $p(\theta \mid Y)$ in L_1.

Moreover

$$\text{var}(\hat{\rho}_{DA}) = \frac{1}{M} \{ \int a^2(\theta) g_i(\theta) \, d\theta - (E\hat{\rho}_{DA})^2 \}$$

and

$$\text{var}(\hat{\rho}_{SIR}) = \frac{1}{M} \{ \int \int a^2(\theta) \frac{p^2(\theta, Z \mid Y)}{h(\theta, Z \mid Y)} \, dZ \, d\theta - \rho^2 \}.$$

Thus, both variances decrease with increasing M. However, for any fixed M, if $h(\theta, Z \mid Y)$ is a poor approximation to $p(\theta, Z \mid Y)$, $\text{var}(\hat{\rho}_{SIR})$ can be quite large (due to the p^2/h ratio) leading to a low Monte Carlo efficiency of SIR.

What should be done when the weights are highly skewed? In this regard, Tanner and Wong (1987) suggest that the data augmentation algorithm be used to refine the importance function of SIR so that the weights are satisfactorily distributed (hence insuring an adequate performance, i.e. low Monte Carlo standard error of the SIR algorithm). Alternatively, when $h(Z \mid \theta, Y) = p(Z \mid \theta, Y)$ one can potentially use $PMDA$ 1 or 2 as importance functions for the SIR algorithm. These ideas have not been investigated. In summary, we see that iteration corrects for inaccuracies in the starting distribution.

5.8. General Imputation Methods

5.8.1. Introduction

Imputation is a popular approach for handling nonresponse in a survey. In imputation, each missing item is "filled-in" and the survey data are then analyzed using standard techniques for complete data. Rubin (1987b) points out that imputation is advantageous in situations where the data collector and data analyst are different people and the collector has access to more information (due to confidentiality constraints) than the analyst. The major drawback to the *single* imputation approach is that the imputed values are treated as if known, thus ignoring the variability due to imputation. Rubin (1987b) suggests *multiple* imputation as a device for handling nonresponse. An important contribution of this device is that it allows for the assessment of within and between imputation variation.

In this section we assume that the scalar responses follow the model $Y_i = X_i^T\theta + \varepsilon_i$, where the ε_i's are iid random variables with mean 0 and variance σ^2, θ is the coefficient vector of dimension d and X_i is a covariate vector of known constants.

We consider four techniques for imputing a single set of n_0 missing responses $Y_{(1)}, \ldots, Y_{(n_0)}$ given the observed data

$$\begin{pmatrix} Y_1 \\ X_1 \end{pmatrix}, \ldots, \begin{pmatrix} Y_{n_1} \\ X_{n_1} \end{pmatrix} \quad \text{and} \quad X_{(1)}, \ldots, X_{(n_0)} \,.$$

Multiple imputations are obtained by repeated (independent) applications of the particular technique. Like *SIR*, these techniques are noniterative. In contrast with *SIR*, these techniques do not require the specification of the distribution of the ε_i's. However, these techniques are restricted to less complicated situations than *SIR*.

5.8.2. Hot Deck Imputation

Assume that the covariates X take on b distinct values and for each covariate there are several responses. Following Schenker and Welsh (1988), reexpress the data in b blocks as

$$\begin{matrix} \begin{pmatrix} Y_{11} \\ X_1 \end{pmatrix}, \begin{pmatrix} Y_{12} \\ X_1 \end{pmatrix}, \ldots, \begin{pmatrix} Y_{1n_{11}} \\ X_1 \end{pmatrix}, \underbrace{\begin{pmatrix} * \\ X_1 \end{pmatrix}, \ldots, \begin{pmatrix} * \\ X_1 \end{pmatrix}}_{n_{01}} \\ \vdots \\ \begin{pmatrix} Y_{b1} \\ X_b \end{pmatrix}, \begin{pmatrix} Y_{b2} \\ X_b \end{pmatrix}, \ldots, \begin{pmatrix} Y_{bn_{1b}} \\ X_b \end{pmatrix}, \underbrace{\begin{pmatrix} * \\ X_b \end{pmatrix}, \ldots, \begin{pmatrix} * \\ X_b \end{pmatrix}}_{n_{0b}} , \end{matrix}$$

where $n_1 = \sum_{j=1}^{b} n_{1_j}$ and $n_0 = \sum_{j=1}^{b} n_{0_j}$. In hot deck imputation, within each block, one is to sample independently with replacement the observed responses to form the multiple imputations for the missing responses. For each of the m augmented data sets, one computes the least-squares estimate of the parameter vector $(\hat{\theta}_j)$. The hot deck estimate $(\hat{\theta}_{HD})$ is then given as $(1/m)\sum_{j=1}^{m} \hat{\theta}_j$.

Schenker and Welsh (1988) provide regularity conditions under which

$$(\hat{\theta}_{HD} - \theta) \sim N(0, V) \ .$$

These authors suggest that V may be approximated by

$$\hat{W}\hat{W}_1^{-1}\hat{W} + \left(\frac{m+1}{m}\right)\hat{B} \ ,$$

where

$$\hat{W} = \frac{1}{m}\sum_{j=1}^{m} s_j^2(\psi_1 + \psi_0)^{-1}$$

$$\hat{W}_1 = \frac{1}{m}\sum_{j=1}^{m} s_j^2\psi_1^{-1}, \qquad \psi_1 = \sum_{j=1}^{n_1} X_j X_j^T, \qquad \psi_0 = \sum_{j=1}^{n_0} X_{(j)} X_{(j)}^T$$

and

$$\hat{B} = \frac{1}{m-1}\sum_{j=1}^{m} (\hat{\theta}_j - \hat{\theta}_{HD})(\hat{\theta}_j - \hat{\theta}_{HD})^T \ .$$

5.8.3. Simple Residual Imputation

Let $r_i = Y_i - X_i^T\hat{\theta}_1$, $1 \le i \le n_1$ where $\hat{\theta}_1$ is the least-squares estimate of θ based on the n_1 complete X, Y tuples. In this technique, one is to draw a sample of n_0 residuals, $r_{(1)}, \ldots, r_{(n_0)}$, by sampling with replacement from the set $\{r_1, \ldots, r_{n_1}\}$ and then let $Y_{(i)} = X_{(i)}^T\hat{\theta}_1 + r_{(i)}$, $1 \le i \le n_0$. As with the hot deck method, for each of the m augmented data sets, one computes the least-squares estimate $(\hat{\theta}_j)$. The simple residual imputation estimate $(\hat{\theta}_{SRI})$ is given by $(1/m)\sum_{j=1}^{m} \hat{\theta}_j$.

Schenker and Welsh (1988) give sufficient conditions under which $(\hat{\theta}_{SRI} - \theta) \sim N(0, V)$. The authors suggest that V can be estimated via $\hat{W}_1 + [(m+1)/m]\hat{B}$, where $\hat{W}_1$ and $\hat{B}$ are defined above.

In effect, this method is a semiparametric extension of *PMDA* 1. In particular, given the least-squares approximation to the posterior mode, $\hat{\theta}_1$, the *SRI* method draws latent data patterns from $\hat{p}(Z \mid \hat{\theta}_1, Y)$, where the conditional predictive distribution is nonparametrically estimated using the residuals. The estimates $\hat{\theta}_{SRI}$ is the mean of the mixture

$$\frac{1}{m}\sum_{j=1}^{m} N(\hat{\theta}_j, \hat{\Sigma}_j) \ ,$$

where $\hat{\theta}_j$ ($\hat{\Sigma}_j$) is the least-squares estimate of θ (covariance matrix) for the jth augmented data set. Thus, in effect, *SRI* replaces the augmented posterior of *PMDA* 1 with a large-sample normal approximation. Also note that the variance of the mixture (conditional on the $\hat{\Sigma}_j$'s and $\hat{\theta}_j$'s) is equal to

$$\frac{1}{m}\sum_{j=1}^{m}(\hat{\Sigma}_j + \hat{\theta}_j\hat{\theta}_j^T) - \left(\frac{1}{m}\sum_{j=1}^{m}\hat{\theta}_j\right)\left(\frac{1}{m}\sum_{j=1}^{m}\hat{\theta}_j\right)^T =$$

$$\text{Average of } [\hat{\Sigma}_1, \hat{\Sigma}_2, \ldots, \hat{\Sigma}_m] + \text{Variance of } [\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_m]. \quad (5.8.1)$$

The first term in (5.8.1) reflects the within-imputation variation. The second term reflects the between-imputation variation, which would vanish if there were no missing observations. The variance estimate given in Schenker and Welsh (1988) includes an inflation factor when estimating the between-imputation variance, i.e.

$$\text{Average of } [\hat{\Sigma}_1, \hat{\Sigma}_2, \ldots \hat{\Sigma}_m] + \left(1 + \frac{1}{m}\right)\text{Variance of } [\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_m] .$$

The inflation factor suggests that the reduction in variability obtained by increasing m is affected mainly by the $1/m$-term. Therefore, using five imputations has relative efficiency larger than 5/6, as compared with using an infinite number of imputations.

5.8.4. Normal and Adjusted Normal Imputation

In contrast to the *SRI* method, the techniques of this section require sampling of parameter values as well as latent data patterns.

In particular, recall that under a normal error model, the augmented posterior factors into product of an inverse chi-square and a conditional multivariate normal (Box and Tiao, 1973). In this way, draw σ_*^2 from $(n_1 - d)s_1^2/\chi^2_{n_1-d}$ and θ^* from $N(\hat{\theta}_1, \sigma_*^2\psi_1^{-1})$, where s_1^2 is the mean-sqare error based on the n_1 complete X, Y tuples, and $\hat{\theta}_1$ and ψ_1 are defined above. Next draw n_0 independent observations $e_{(1)}, \ldots, e_{(n_0)}$ from the standard normal distribution and let $Y_{(i)} = X_{(i)}^T\theta^* + \sigma_* e_{(i)}$, $1 \leq i \leq n_0$. For each of the m augmented data sets, compute the least-squares estimate of the parameter vector ($\hat{\theta}_j$). The normal imputation estimate ($\hat{\theta}_N$) is given as $(1/m)\sum_{j=1}^{m}\hat{\theta}_j$. The adjusted normal imputation estimate ($\hat{\theta}_{AN}$) is also obtained by averaging the augmented least-squares estimates. The latter approach differs from the former in that $Y_{(i)} = X_{(i)}^T\theta^* + \sigma_*\bar{r}_{(i)}$, $1 \leq i \leq n_0$, where $\bar{r}_{(i)}$ is the standardized version (mean zero, variance 1) of $r_{(i)}$.

Schenker and Welsh (1988) show that under certain conditions both estimators follow

$$(\hat{\theta} - \theta) \sim N(0, V)$$

where the variance is estimated via $\hat{W}_1 + [(m+1)/m]\hat{B}$.

The normal and adjusted normal estimators are noniterative adaptations of data augmentation. In particular, parameter values are generated from the current approximation to the posterior (i.e. the inverse chi-square by multivariate normal factorization). In the normal imputation procedure, the latent data are generated from a normal approximation to $p(Z|\theta^*, Y)$, while in the adjusted approach the latent data are drawn from $\hat{p}(Z|\theta^*, Y)$, where $\hat{p}$ is the nonparametric estimate of $p(Z|\theta^*, Y)$ based on the observed residuals.

While Schenker and Welsh (1988) have considered the large sample properties of these four imputation methods, further work regarding their small sample/small m properties is warranted.

5.8.5. Nonignorable Nonresponse

In this section, we consider the situation where nonresponse is related to the values of the outcome variable. Rubin (1987b) calls this situation *nonignorable nonresponse.* Rubin (1987b) considers two approaches for handling nonignorable nonresponse: mixture models and selection models. For each of these approaches, we consider how to proceed in the presence or absence of followup data.

5.8.5.1 Mixture Model – Without Followup Data

Consider a response Y and a $d \times 1$ covariate vector X. Rubin (1987b) considers an approach for creating multiple imputations for the missing response values. His idea is to factor the posterior distribution of the missing values and the parameters of the model (to be specified below) into three components. The first component is the predictive distribution of the latent data given the parameters. The second component is the conditional distribution of the nonresponders' parameters given the responders' parameters. The third component is the posterior of the responders' parameters given the observed data. Having specified each of these distributions, Rubin (1987b) applies the method of composition to sample latent data patterns.

The normal mixture model supposes that for the responders, Y_i follows the normal distribution with mean $X_i^T\beta_1 + \alpha_1$ and variance σ_1^2. For the nonresponders, Y_i follows the normal distribution with mean $X_i^T\beta_0 + \alpha_0$ and variance σ_0^2. For simplicity Rubin (1987b) takes $\sigma_1 = \sigma_0 = \sigma$. Hence, under a noninformative prior, the posterior distribution $p(\alpha_1, \beta_1, \sigma | Y)$ factors as the product of an inverse chi-square distribution and a conditional normal distribution. Note that in large samples, this factorization should be approximately correct even if the Y_i's are not normally distributed. The predictive distribution of the latent data, $p(Z_i|\alpha_0, \beta_0, \alpha_1, \beta_1, \sigma, Y)$, is $N(\alpha_0 + X_i^T\beta_0, \sigma^2)$. (Rubin (1987b) assumes that X_i is observed for all units in the sample.) What is the middle distribution linking (α_0, β_0) to (α_1, β_1)?

The conditional distribution of β_0 and α_0 given α_1, β_1 and σ is given by the product of two independent distributions. The conditional marginal of β_0 given α_1, β_1 and σ is a normal distribution, with mean β_1 and variance $C_\beta^2 \beta_1 \beta_1^T$. The conditional marginal of α_0 is specified through the average Y value at $\bar{X}_1$ for the nonresponders in the population, $\eta_0 = \alpha_0 + \bar{X}_1^T \beta_0$. The conditional marginal of η_0 is normal with mean $\eta_1 = \alpha_1 + \bar{X}_1^T \beta_1$ (the average Y value for the responders in the population) and variance $C_\eta^2 \eta_1^2$.

Note that when $C_\eta = C_\beta = 0$, these specifications imply an ignorable nonresponse mechanism. The coefficient C_β specifies the similarity of slopes for responders and nonresponders. Rubin (1987b) refers to C_β as the "prior coefficient of variation" for the nonresponders' regression coefficients. Letting β_{1j} and β_{0j} denote the jth components of β_1 and β_0, respectively, the model supposes that with 95% certainty β_{0j} will lie in the interval $\beta_{1j}(1 \pm 1.96 C_\beta)$. The coefficient C_β is taken for all the covariates. Moreover, it is assumed that there is no bias, i.e. the distribution of β_0 is centered at β_1.

The coefficient C_η specifies the similarity of the expected value of Y for the responders and nonresponders with covariate means equal to the covariate mean of the responders. In other words, one is 95% certain that the expected response for the nonresponders with average covariate value $\bar{X}_1$ is contained in the interval $\eta_1(1 \pm 1.96 C_\eta)$.

In this way, the joint posterior of the missing data Z, η_0, β_0, η_1, β_1 and σ factors as

$$p(Z|\eta_0, \beta_0, \sigma, Y) p(\eta_0, \beta_0 | \eta_1, \beta_1, \sigma, Y) p(\eta_1, \beta_1, \sigma | Y) \ , \qquad (5.8.2)$$

where the components are defined above.

To implement this normal mixture model, Rubin (1987b) applies the method of composition to (5.8.2). In particular, to sample latent data patterns Rubin performs the following three steps:

Step 1. Draw η_1^*, β_1^*, σ_*^2.

a. Draw σ_*^2 from the inverse χ^2 distribution:

$$(n_1 - d) s_1^2 / \chi^2_{(n_1 - d)} \ ,$$

where s_1^2 is the residual mean square for the responders' regression.

b. Draw β_1^* from the conditional normal:

$$N(\hat{\beta}_1, (X_1^T X_1)^{-1} \sigma_*^2) \ ,$$

where $\hat{\beta}_1$ is the least-squares estimate of β_1 and X_1 is the associated design matrix omitting the intercept column.

c. Draw η_1^* from $N(\bar{Y}_1, \sigma_*^2 / n_1)$.

Step 2. Draw β_0^* and η_0^*.

a. Draw β_0^* from

$$N(\beta_1^*, C_\beta^2 \beta_1^* \beta_1^{*T}) \ .$$

b. Draw η_0^* from

$$N(\eta_1^*, C_\eta^2(\eta_1^*)^2) \ .$$

Step 3. Draw Z.
Draw z from

$$N(\eta_0^* + (X_i - \bar{X}_1)^T \beta_0^*, \sigma_*^2) \ .$$

Steps (1)–(3) are repeated m times to form the augmented data sets.

Rubin (1987b) applies this multiple imputation procedure to data from a 1971 survey of high school principals. Rubin found that in this example, the 95% confidence interval for response was insensitive to C_β, but quite sensitive to C_η. This reflects the fact that Y is not well predicted by X. This example illustrates how one may use the normal mixture model to display the sensitivity of the analysis to a range of plausible assumptions regarding the "missingness" mechanism. In general, however, this approach does not yield consistent point estimates with infinitely large sample sizes (Brown, 1990) for a particular model.

5.8.5.2. Mixture Model – With Followup Data

In the presence of followup data, one can use this information to generate the latent data patterns. In particular, for the followups we assume

$$Y = \alpha_0 + X_f^T \beta_0 + \sigma\varepsilon$$

where n_f is the number of followups, $\varepsilon \sim N(0, 1)$ and X_f is the design matrix for the followups. We have the factorization

$$p(Z, \beta_0, \alpha_0, \sigma \,|\, Y) = p(Z \,|\, \beta_0, \alpha_0, \sigma, Y) p(\sigma, \beta_0, \alpha_0 \,|\, Y) \ . \qquad (5.8.3)$$

Under the noninformative prior, $p(\beta_0, \alpha_0, \sigma \,|\, Y)$ factors as $s_f^2(n_f - d)/\chi^2_{(n_f - d)}$ times the conditional normal $N(\hat{\beta}_0, \sigma^2(X_f^T X_f)^{-1})$, where $\hat{\beta}_0$ is the least-squares estimate using the followup data and s_f^2 is the corresponding residual mean square. Note that in large samples, this factorization should be approximately correct even if the Y_i's are not normally distributed. This approach does assume that the followup data are a random sample from the nonresponders. The predictive distribution is $N(\alpha_0 + X_{nf}^T \beta_0, \sigma^2)$, where X_{nf} is the design matrix for the nonfollowups. The predictive distribution can alternatively be approximated using a hot deck type approach on the follow up data. This approach would eliminate the need to specify a normal distribution on the nonfollowup responses.

An application of the method of composition to (5.8.3) yields a latent data pattern:

Step 1. *Draw* α_0^*, β_0^* and σ_*^2 from the inverse chi-square/conditional normal.
Step 2. Draw the latent data for the nonfollowups from $N(\alpha_0^* + X_{nf}^T \beta_0^*, \sigma_*^2)$.

5.8.5.3. Selection Model – Without Followup Data

Greenlees, Reece and Zieschang (1982) introduced the Selection Model for handling nonignorable nonresponse. Greenlees, Reece and Zieschang (1982) assume that the responses follow

$$Y_i = X_i^T \beta + \sigma \varepsilon_i \ ,$$

where $\varepsilon_i \overset{\text{iid}}{\sim} N(0, 1)$. Also assumed in the selection model is that the probability of response depends on Y and other variables V_i: $p(R_i = 1 \mid Y_i, V_i) = 1/[1 + \exp(-\alpha - \gamma Y_i - V_i^T \delta)]$, where $R_i = 1$ if individual i is a responder and 0, otherwise; V_i is a p-dimensional vector of characteristics of individual i; α and γ are scalar parameters; and δ is an $p \times 1$ parameter vector.

For the responders, the contribution to the likelihood is

$$L_i = \frac{1}{1 + \exp(-\alpha - \gamma Y_i - V_i^T \delta)} \frac{1}{\sigma} \phi\left(\frac{Y_i - X_i^T \beta}{\sigma}\right) ,$$

where $\phi(x)$ is the standard normal density. For the nonresponders, the contribution to the likelihood is

$$L_i = \int_{-\infty}^{\infty} \left\{1 - \frac{1}{1 + \exp(-\alpha - \gamma Y - V_i^T \delta)}\right\} \frac{1}{\sigma} \phi\left(\frac{Y - X_i^T \beta}{\sigma}\right) dY \ .$$

The likelihood of the entire sample is given by the product of the L_i's. Greenlees, Reece and Zieschang (1982) suggests the use of the Gauss–Newton algorithm for maximizing the loglikelihood with respect to α, γ, β, δ and σ.

Having obtained the maximum likelihood estimates $\hat{\alpha}$, $\hat{\beta}$, $\hat{\gamma}$, $\hat{\delta}$ and $\hat{\sigma}$, a single imputation is given as:

$$E(Y_i \mid X_i, V_i, R_i = 0) =$$

$$\frac{\displaystyle\int_{-\infty}^{\infty} Y \left\{1 - \frac{1}{1 + \exp(-\hat{\alpha} - \hat{\gamma} Y - V_i^T \hat{\delta})}\right\} \frac{1}{\hat{\sigma}} \phi\left(\frac{Y - X_i^T \hat{\beta}}{\hat{\sigma}}\right) dY}{\displaystyle\int_{-\infty}^{\infty} \left\{1 - \frac{1}{1 + \exp(-\hat{\alpha} - \hat{\gamma} Y - V_i^T \hat{\delta})}\right\} \frac{1}{\hat{\sigma}} \phi\left(\frac{Y - X_i^T \hat{\beta}}{\hat{\sigma}}\right) dY} \ .$$

Greenlees et al. (1982) propose the following rejection/acceptance algorithm for drawing multiple imputations from the conditional predictive distribution of the latent data and this algorithm gives $\hat{\alpha}$, $\hat{\gamma}$, $\hat{\delta}$, $\hat{\beta}$ and $\hat{\sigma}$. Ripley (1987) discusses the theory and motivation for general rejection/acceptance algorithms.

Step 1. Draw e_i from $N(0, 1)$.

Step 2. a. Calculate $Y_i = X_i^T \hat{\beta} + \hat{\sigma} e_i$.

b. $p(R = 0 \mid Y_i, V_i) = 1 - 1/\{1 + \exp(-\hat{\alpha} - \hat{\gamma} Y_i - V_i^T \hat{\delta})\}$.

Step 3. Draw u from uniform $[0, 1]$.

Step 4. Accept Y_i if $p(R = 0 | Y_i, V_i) \geq u$, otherwise reject and return to Step 1.

Greenlees, Reece and Ziechang (1982) apply the selection model to income data from the Current Population Survey. In this case, information from a secondary source allows for a direct test of the method. These authors found that a nonignorable nonresponse model gave more accurate estimates (with regard to mean squared error for the imputed values) than the ignorable nonresponse model.

Rubin (1987b) criticizes the idea of sampling from $p(Z|\hat{\alpha}, \hat{\gamma}, \hat{\beta}, \hat{\delta}, \hat{\sigma}, Y)$, rather than from the product

$$p(Z | \alpha, \gamma, \beta, \delta, \sigma, Y) p(\alpha, \gamma, \beta, \delta, \sigma | Y) \ .$$

Rubin (1987b) shows empirically that the conditional multiple imputation approach underestimates the variability, as would be expected in a first-order-approximation method. Rubin (1987b) suggests approximating $p(\alpha, \gamma, \beta, \delta, \sigma | Y)$ by the matching multivariate normal and then applies the method of composition to sample the latent data patterns. Alternatively, one could use the matching multivariate normal as an importance function and then attach the appropriate weights to the latent data patterns. This idea has not been investigated.

5.8.5.4. Selection Model – With Followup Data

The selection model with followup data is quite similar to the model without followup data: If individual i is a responder or nonfollowup nonresponder, then the contribution to the likelihood is as in the previous subsection. For each followup nonresponder, the contribution is

$$\left[1 - \frac{1}{1 + \exp(-\alpha - \gamma Y_i - V_i^T \delta)}\right] \frac{1}{\sigma} \phi\left(\frac{Y_i - X_i^T \beta}{\sigma}\right) .$$

5.9. Further Importance Sampling Ideas

5.9.1. Sampling from the Predictive Identity

In step (a1) of the data augmentation algorithm, we draw a θ value from the mixture

$$\frac{1}{m} \sum_{j=1}^{m} p(\theta | z_j, Y) \ .$$

Explicitly, we first select a $p(\theta \mid z_i, Y)$ uniformly from the m versions and then sample from this selected augmented posterior $p(\theta \mid z_i, Y)$. In some situations, we may not be able to *directly* sample from the distribution $p(\theta \mid z, Y)$. As such, we need a vehicle to circumvent this problem.

It is straightforward to incorporate importance sampling into the data augmentation algorithm. Given an importance function $I(\theta)$, the predictive identity can be rewritten as

$$p(Z \mid Y) = \int_{\Theta} p(Z \mid \theta, Y) \frac{p(\theta \mid Y)}{I(\theta)} I(\theta)\, d\theta.$$

This suggests that to incorporate importance sampling into data augmentation, replace steps a1 and a2 of the algorithm with

a1′. Generate ϕ from $I(\phi)$.
a2′. Generate z from $p(Z \mid \phi, Y)$, where ϕ is the value generated in (a1′).
a3′. Compute the weight $w = g_i(\phi)/I(\phi)$.

Steps (a1′)–(a3′) are repeated m times. The current approximation to $p(\theta \mid Y)$ is then updated as the weighted mixture of augmented posteriors of θ, given the augmented data from (a2′), i.e.

$$g_{i+1}(\theta) = \sum_{j=1}^{m} w_j p(\theta \mid z_j, Y) \bigg/ \sum_{j=1}^{m} w_j .$$

Wei and Tanner (1990c) illustrate this idea in the context of censored regression data, where it is difficult to directly sample from $p(\theta \mid Y, Z)$. Their idea is to approximate $p(\theta \mid z_j, Y)$ *by* $I_j(\theta)$ and then to sample ϕ from $\sum_{j=1}^{m} w_j I_j(\phi) / \sum_{j=1}^{m} w_j = I(\phi)$ which approximates $g_i(\phi)$.

5.9.2. Sequential Imputation

Kong, Liu and Wong (1992) introduce the method of sequential imputation as an alternative means of simulating latent data patterns $\underset{\sim}{z}(1), \ldots, \underset{\sim}{z}(m)$. Their idea is related to the *SIR*, but their algorithm tends to produce more stable importance weights. In many situations their procedure works well without the need of iteration. Let each latent data pattern $\underset{\sim}{z}(j)$ have n components: $z_1(j), \ldots, z_n(j)$. To draw from $p(z \mid Y)$, sequential imputation imputes the z_t's sequentially and then uses importance sampling weights to avoid iteration. In general, the algorithm starts by drawing z_1^* from $p(z_1 \mid y_1)$ and computing $w_1 = p(y_1)$, where y_1 is the first component of Y. For $t = 2, \ldots, n$, the next two steps are performed sequentially:

a1. Draw $z_t^* \sim p(z_t \mid y_1, z_1^*, \ldots, y_{t-1}, z_{t-1}^*, y_t)$

a2. Compute $p(y_t \mid y_1, z_1^*, \ldots, y_{t-1}, z_{t-1}^*) = p_t$, and let $w_t = w_{t-1} p_t$.

Note that the z_t^*'s are drawn sequentially since each z_t^* is drawn conditioned on the previously imputed missing parts $z_1^*, \ldots, z_{t-1}^*$. Steps (a1) and (a2) are repeated m times to obtain $\underset{\sim}{z}^*(1), \ldots, \underset{\sim}{z}^*(m)$ and $w(1), \cdots, w(m)$, where $z^*(j) = (z_1^*(j), \ldots, z_n^*(j))$ for $j = 1, \ldots, m$ and

$$w(j) = p(y_1) \prod_{t=2}^{n} p(y_t \mid y_1, z_1^*(j), \ldots, y_{t-1}, z_{t-1}^*(j)) .$$

Kong, Liu and Wong (1992) then estimate $p(\theta \mid Y)$ by the weighted mixture

$$\sum_{j=1}^{m} w(j) p(\theta \mid Y, \underset{\sim}{z}^*(j)) \bigg/ \sum_{j=1}^{m} w(j) .$$

From step (a1) we see that $\underset{\sim}{z}^*(j)$ is drawn from

$$p^*(\underset{\sim}{z}^*(j) \mid Y) = p(z_1^*(j) \mid y_1) \prod_{t=2}^{n} p(z_t^*(j) \mid y_1, z_1^*(j), y_2, z_2^*(j), \ldots, y_{t-1}, z_{t-1}^*(j), y_t) ,$$

rather than from $p(Z \mid Y)$ – the correct distribution. However, notice that the ratio $p(\underset{\sim}{z}^*(j) \mid Y)/p^*(\underset{\sim}{z}(j) \mid Y)$ is

$$\frac{p(Y, \underset{\sim}{z}^*(j))}{p(Y)} \frac{p(y_1)}{p(y_1, z_1^*(j))} \prod_{t=2}^{n} \frac{p(y_1, \ldots, y_t, z_1^*(j), \ldots, z_{t-1}^*(j))}{p(y_1, \ldots, y_t, z_1^*(j), \ldots, z_t^*(j))}$$

$$= \frac{p(Y, \underset{\sim}{z}^*(j))}{p(Y)} \frac{p(y_1)}{p(y_1, \ldots, y_n, z_1^*(j), \ldots, z_n^*(j))}$$

$$\times \prod_{t=2}^{n} \frac{p(y_1, \ldots, y_t, z_1^*(j), \ldots, z_{t-1}^*(j))}{p(y_1, \ldots, y_{t-1}, z_1^*(j), \ldots, z_{t-1}^*(j))}$$

$$= \frac{p(y_1)}{p(Y)} \prod_{t=2}^{n} p(y_t \mid y_1, z_1^*(j), \ldots, y_{t-1}, z_{t-1}^*(j)) = \frac{w(j)}{p(Y)} .$$

Since $p(Y)$ does not vary across imputations, the normalized weights $w(j)/\sum_{j=1}^{m} w(j)$ provide the correct adjustment to realize a sample from $p(z \mid Y)$, having actually sampled from $p^*(z \mid Y)$.

Kong, Liu and Wong (1992) illustrate their sequential imputation algorithm on the bivariate normal example of Section 5.1. They realize a comparable plot of the posterior of the correlation coefficient, *without the need to iterate*. In their paper, they also discuss applications to prior sensitivity analysis, case influence, prediction and model selection. Results relating to the efficiency of sequential imputation are presented. Kong, Liu and Wong (1992) also discuss the case where the posterior has to be updated with the arrival of new information.

5.9.3. Calculating the Posterior

Implicit in the calculation

$$\frac{1}{m} \sum_{j=1}^{m} p(\theta \mid Y, z_j)$$

is the availability of the normalizing constant for $p(\theta \mid Y, Z)$. Chen (1992) presents an importance sampling idea which avoids this issue. Note that

$$p(\theta^* \mid Y) = \int_\Theta \int_Z w(\theta \mid Z, Y) \frac{p(\theta^*, Z \mid Y)}{p(\theta, Z \mid Y)}\, p(\theta, Z \mid Y) dZ\, d\theta$$

for an arbitrary conditional density $w(\theta \mid Z, Y)$, since by Fubini's theorem the double integral is equal to

$$\int_Z p(\theta^*, Z \mid Y) \left\{ \int_\Theta w(\theta \mid Z, Y) d\theta \right\} dZ = p(\theta^* \mid Y) \ .$$

Thus, Chen (1992) suggests that we approximate $p(\theta^* \mid Y)$ via

$$\frac{1}{n} \sum_{j=1}^{n} w(\theta_j \mid z_j, Y) \frac{p(\theta^*, z_j \mid Y)}{p(\theta_j, z_j \mid Y)} \ .$$

Note that the normalizing constant for $p(\theta, Z \mid Y)$ is not necessary since it cancels in the ratio $p(\theta^*, Z \mid Y)/p(\theta, Z \mid Y)$. Chen (1992) presents this method for the multicomponent case (see Chapter 6 for a discussion of the Gibbs sampler) and shows that the optimal choice for $w(\theta \mid Z, Y)$ from a minimum variance approach is $p(\theta \mid Z, Y)$. Hence it is important to choose $w(\theta \mid Z, Y)$ to be close in shape to $p(\theta \mid Y, Z)$.

5.10. Sampling in the Context of Multinomial Data

5.10.1. Dirichlet Sampling

In the genetic linkage example, the augmented posterior is a beta distribution. Thus, it is straightforward to sample from the augmented posterior. In more complicated situations, this simplicity may be absent. We now discuss the Dirichlet sampling procedure, which can be used to sample approximately from the posterior distribution of parametric models for multinomial data.

In the genetic linkage example, given the augmented data, the distribution of the last four cell probabilities (p_2, p_3, p_4, p_5) is equal in distribution to that of $(v_2/2, v_3/2, v_4/2, v_5/2)$, where (v_2, v_3, v_4, v_5) has the Dirichlet distribution proportional to

$$v_2^{x_2} v_3^{x_3} v_4^{x_4} v_5^{x_5} \tag{5.10.1}$$

where $\sum_{i=2}^{5} v_i = 1$. Note that it is straightforward to sample from such a Dirichlet distribution. The genetic linkage model, however, specifies that (p_2, p_3, p_4, p_5) must lie on the linear parametric curve

$$C = \left\{ \left(\frac{\theta}{4}, \frac{1}{4} - \frac{\theta}{4}, \frac{1}{4} - \frac{\theta}{4}, \frac{\theta}{4} \right) : \theta \subset [0, 1] \right\} \ .$$

Thus, to sample from the augmented posterior:

a. Observations are drawn from the Dirichlet distribution in (5.10.1).
b. Find the $\hat{\theta}$ that gives cell probabilities $(\hat{p}_2, \hat{p}_3, \hat{p}_4, \hat{p}_5)$ closest to the observed Dirichlet observation (p_2, p_3, p_4, p_5).
c. Accept those points near the parametric curve C e.g. using

$$\left[\sum_{i=2}^{5} (p_i - \hat{p}_i)^2\right]^{1/2} < \varepsilon ,$$

where ε is small. The approximate posterior distribution for θ is obtained by forming the histogram of the corresponding $\hat{\theta}$ values.

This procedure is justified by the following result of Tanner and Wong (1987):

Result. The distribution induced by $p(\theta \mid Y, Z)$ on the curve C is the same as the conditional distribution induced by the Dirichlet distribution in (5.10.1) on C.

In practice, the value of ε is selected by plotting a sequence of estimated posterior distributions of θ corresponding to a sequence of decreasing ε values. This Dirichlet sampling procedure is appropriate when the cell probabilities are linear in θ or if the posterior distribution is relatively concentrated in comparison with the curvature of the parametric surface. Otherwise, the histogram of $\hat{\theta}$ must be multiplied by an adjustment factor.

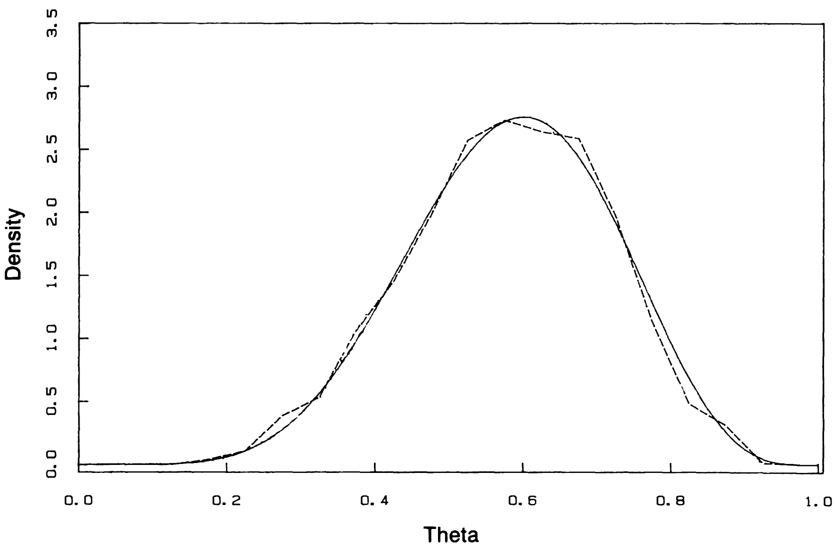


Figure 5.16. Marginal of θ.

Example. *Genetic Linkage (Continued)*
Assume the augmented data is given by (3,2,2,3). Ten thousand observations were drawn from the Dirichlet distribution corresponding to this data vector. For each Dirichlet observation, the value of θ that gives the closest $(\hat{p}_2, \hat{p}_3, \hat{p}_4, \hat{p}_5)$ vector was found using least squares. The histogram of the 3000 accepted values and the true augmented posterior is given in Figure 5.16. The distribution of the accepted $\hat{\theta}$ values appears to follow the true augmented posterior.

5.10.2. Latent Class Analysis

We now turn our attention to the second example of Chapter 1, latent class analysis.

The data presented in that example represent the responses of 3181 participants in the 1972, 1973 and 1974 General Social Surveys, as presented in Haberman (1979). The participants in these surveys were cross-classified by the year of the survey and their responses to each of three questions regarding abortion. Thus, the cell entry n_{abcd} represents the number of subjects who in year $D = d$ give responses a to question A, b to question B, and c to question C. Regarding question A, subjects were asked, "Please tell me whether or not you think it should be possible for a pregnant woman to obtain a legal abortion *if she is married and does not want any more children.*" In question B, the italicized phrase was replaced with "if the family has a very low income and cannot afford any more children," and in question C it was replaced with "if she is not married and does not want to marry the man." For these data, Haberman (1979) considered several models, one which is the traditional latent class model.

In this example, the traditional latent class model assumes that the manifest variables (A, B, C, D) are conditionally independent, given a dichotomous latent variable (X). In other words, if the value of the dichotomous latent variable is known for a given participant, then knowledge of the response to a given question provides no further information regarding the responses to either of the other two questions. For the augmented 3×2^4 table, the latent class model specifies that the expected counts m_{xabcd} follow:

$$\log m_{xabcd} = \lambda + \lambda_x^X + \lambda_a^A + \lambda_b^B + \lambda_c^C + \lambda_d^D + \lambda_{xa}^{XA} + \lambda_{xb}^{XB} + \lambda_{xc}^{XC} + \lambda_{xd}^{XD}$$

where

$$\sum_x \lambda_x^X = \sum_a \lambda_a^A = \cdots = \sum_d \lambda_d^D = \sum_x \lambda_{xa}^{XA} = \sum_a \lambda_{xa}^{XA} = \cdots = \sum_x \lambda_{xd}^{XD} = \sum_d \lambda_{xd}^{XD} = 0 .$$

One parameter of interest associated with this model is the conditional probability of a response a to question A, given that $X = 1$ (which will be denoted as π_{a1}^{AX}). In conjunction with π_{a2}^{AX}, the magnitude of this conditional probability indicates the accuracy of the response a to question A in identifying the latent classification $X = 1$, since the ratio $\pi_{a1}^{AX}/\pi_{a2}^{AX}$ is the likelihood ratio for identifying X based on an observation of A. In the present example, Haberman estimated π_{11}^{AX} to be 0.892.

To obtain the posterior distribution of π_{11}^{AX}, the data augmentation algorithm is implemented as follows. In the initial iteration, the odds of being in the class $X = 2$ (which will be denoted as $\theta_{abcd\cdot}$) is taken to be $\frac{1}{2}$ for all values of a, b, c and d. The unobserved cell counts (n_{abcdx}) are imputed by noticing that conditional on both $\theta_{abcd\cdot}$ and (the observed cell counts) n_{abcd}, the posterior distribution of n_{abcd1} follows a binomial distribution with parameters n_{abcd} and $1/(1 + \theta_{abcd\cdot})$. The posterior distribution of π_{11}^{AX} is then obtained by drawing from the mixture of augmented posterior distributions. In particular, for a given augmented data set, a vector of probabilities $\{P_{abcdx}\}$ is drawn from the Dirichlet distribution $D(n_{11111}, \ldots, n_{22232})$ and some of the observations are discarded using the Euclidean distance criterion, as discussed in Section 5.10.1. The odds of being in the latent class $X = 2$ given that $A = a$, $B = b$, $C = c$ and $D = d$ is updated using the maximum likelihood estimate (under the conditional independence model)

$$\left(\frac{\sum\limits_{b,c,d} p_{abcd2}}{\sum\limits_{b,c,d} p_{abcd1}}\right)\left(\frac{\sum\limits_{a,c,d} p_{abcd2}}{\sum\limits_{a,c,d} p_{abcd1}}\right)\left(\frac{\sum\limits_{a,b,d} p_{abcd2}}{\sum\limits_{a,b,d} p_{abcd1}}\right)\left(\frac{\sum\limits_{a,b,c} p_{abcd2}}{\sum\limits_{a,b,c} p_{abcd1}}\right)\cdot\left(\frac{\sum\limits_{a,b,c,d} p_{abcd1}}{\sum\limits_{a,b,c,d} p_{abcd2}}\right)^3$$

and the algorithm cycles until convergence is achieved. For each augmented

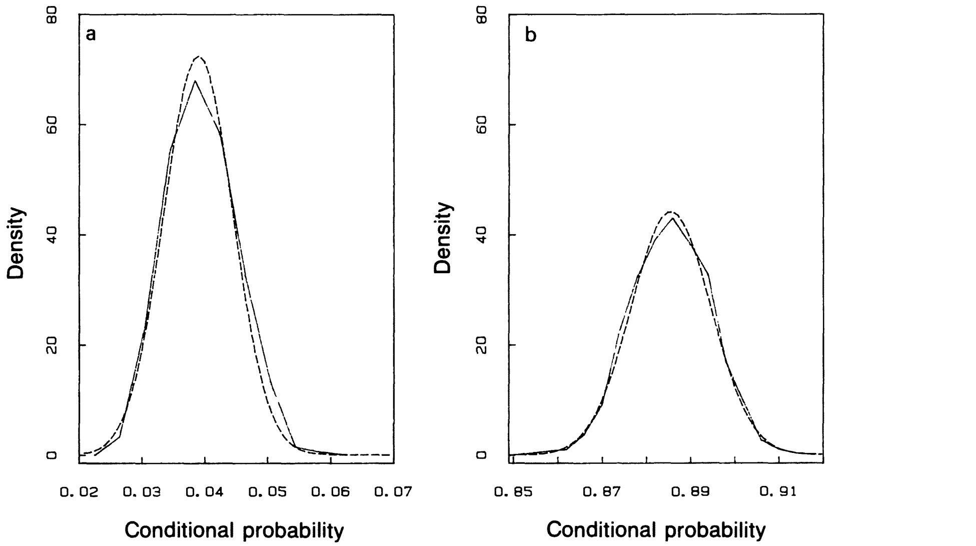


Figure 5.17. Conditional probability (a) left mode, (b) right mode.

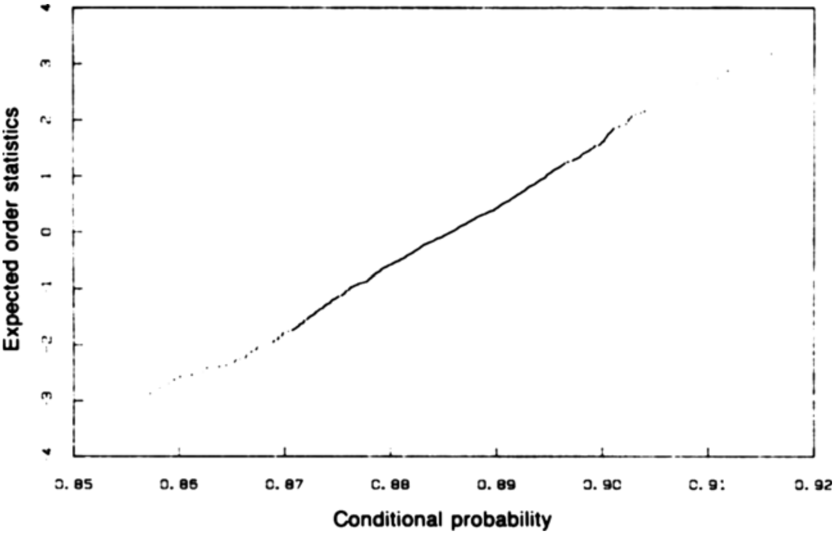


Figure 5.18(a). Rankit plot for right mode.

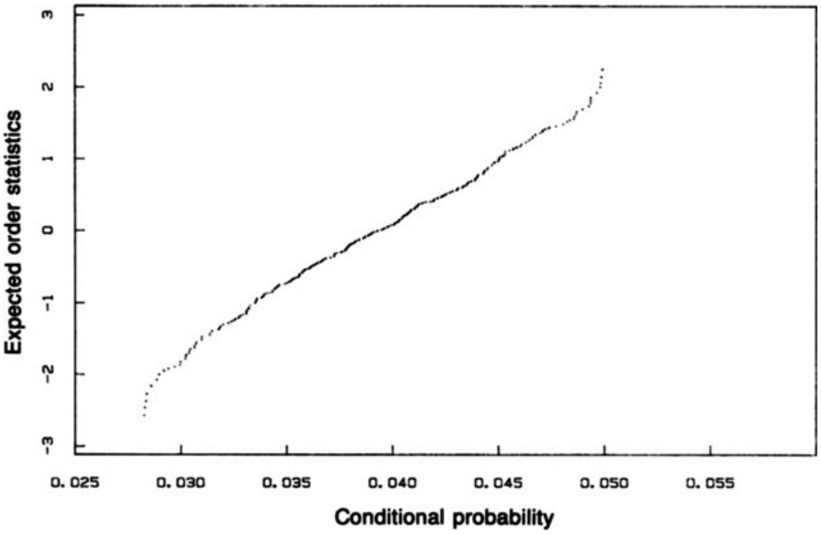


Figure 5.18(b). Rankit plot for left mode.

data set, the conditional probability of interest is calculated from the equation

$$\pi_{11}^{AX} = \frac{\sum\limits_{b,c,d} p_{1bcd1}}{\sum\limits_{a,b,c,d} p_{abcd1}} .$$

In Figure 5.17, the estimated posterior distribution of π_{11}^{AX} is presented, where the values from the 15th to 20th iterations are pooled ($m = 1600$) to form the histogram in Figures 5.17. As can be seen from Figures 5.17, the posterior distribution appears to be bimodal, with one mode occurring at about 0.039 and the other mode occurring at about 0.886. The reason for this bimodality stems from the unidentifiability inherent in the problem. In the latent class model, the data analyst has the choice of identifying a positive attitude toward abortion with the condition that $X = 1$ or with the condition that $X = 2$. The mode occurring at 0.039 occurs if one identifies a positive attitude with $X = 2$; the second mode occurs if a positive attitude is identified with $X = 1$. In this regard, it is important to note that the modes are well separated. Thus, for the present data set, the conditional probability is, in the Bayesian sense, *locally identifiable.*

Conditioning on the identification of a positive attitude toward abortion with $X = 1$, that is, examining the right mode, our point estimate for π_{11}^{AX} is close to the maximum likelihood estimate (0.886 vs. 0.892). (Such an identification is reasonable given the nature of the question.) In addition, there is little evidence of a departure of the normal approximation from the posterior distribution. Comparing the estimated density to the normal curve with matching mean and standard error (0.009), an overall concordance is observed. A similar conclusion is reached by examining the corresponding rankit plot (see Figures 5.18). Regarding the lower mode, some evidence against the normal approximation ($\hat{\mu} = 0.039, \hat{\sigma} = 0.006$) is noted. In particular, the posterior distribution is slightly skewed to the right.

CHAPTER 6

Markov Chain Monte Carlo: The Gibbs Sampler and the Metropolis Algorithm

6.1. Introduction to the Gibbs Sampler

To motivate the Gibbs sampler, we consider a modification of data augmentation which we will refer to as *chained data augmentation.* The Gibbs sampler turns out to be a multivariate extension of chained data augmentation.

6.1.1. Chained Data Augmentation

To begin, consider the two data augmentation equations:

$$\text{Posterior Equation: } p(\theta | Y) = \int_Z p(\theta | Y, Z) p(Z | Y) dZ$$

$$\text{Predictive Equation: } p(Z | Y) = \int_\theta p(Z | Y, \theta) p(\theta | Y) d\theta \ .$$

The following algorithm is motivated by applying the method of composition to each of these equations:

a1. Given z^*, generate θ^* from $p(\theta|z^*, Y)$.
a2. Generate z^* from $p(Z|\theta^*, Y)$ where θ^* is the value obtained in (a1).

The posterior equation suggests that one may draw a value of θ by sampling from $p(\theta|z^*, Y)$ where z^* is the value obtained in (a2). Given the new value for θ, step (a2) may be repeated and the algorithm iterated. It is pointed out that this iterative algorithm is data augmentation with $m = 1$.

We will refer to this version of data augmentation as *chained data augmentation.* Thus, in chained data augmentation we apply the method of composition to both the predictive and posterior equations to realize an observation from $g_i(\theta)$. In data augmentation, we apply the method of composition m times to the predictive equation to yield $z^{(1)}, \ldots, z^{(m)}$ and then apply the Monte Carlo method to the posterior equation to calculate the current approximation to $p(\theta|Y)$.

As noted in Tanner and Wong (1987), the values of θ in chained data augmentation (over the iterations $i = 1, 2, \ldots$) form a Markov Chain with transition function equal to $K(\theta, \phi)$, as defined in Section 5.1. Under the regularity conditions of Section 5.5, this is a Markov Chain with an equilibrium distribution satisfying the fixed point equation (5.1.2) of Section 5.1. By virtue of the results in Section 5.5, we have

$$\theta^{(i)} \xrightarrow{\text{d}} \theta \sim p(\theta|Y) \tag{6.1.1}$$

and

$$\frac{1}{t}\sum_{i=1}^{t} f(\theta^{(i)}) \xrightarrow{\text{a.s.}} E_p(f(\theta)) \tag{6.1.2}$$

for any integrable function f where $\theta^{(i)}$ is the sampled value at iteration i. Thus, (6.1.1) suggests that if the chain is run to equilibrium, one can use the simulated values as a basis for summarizing $p(\theta|Y)$. Since successive $\theta^{(i)}$ are correlated, parallel independent runs of the chain or suitable spacings between realizations will be needed to realize an iid sample from $p(\theta|Y)$. The second result (6.1.2) suggests that the average of a function of interest over values from a chain yields a consistent estimator of its expectation. Geyer (1992) discusses computing the variance of functionals of these correlated $\theta^{(i)}$'s.

EXAMPLE. *Simulation in Hierarchical Models*

Morris (1987) applied chained data augmentation in the context of hierarchical models. In particular, suppose that d population means $\theta = (\theta_1, \ldots, \theta_d)$ are to be estimated having observed d independent normally distributed sample means $Y = (Y_1, \ldots, Y_d)$, where Y_i given θ_i, is iid $N(\theta_i, V_i)$, $i = 1, \ldots, d$, and the $V_i = \text{var}\,(Y_i|\theta_i)$ are known. The conjugate prior distribution is taken for each θ_i, independently, with A unknown and θ_i given A are iid $N(0, A)$, $i = 1, \ldots, d$. The distribution on the hyperparameter A is $cA^{-1-q/2}\exp(-0.5\lambda/A)$, with known $q > 0$ and $\lambda > 0$.

Morris (1987) notes that these choices lead to proper posterior densities. In particular, the distribution $p(\theta|Y, A)$ is normal, with the jth component distributed as

$$N[(1 - B_j)\,Y_j, V_j(1 - B_j)]$$

$j = 1, \ldots, d$ and $B_j = V_j/(V_j + A)$. The distribution $p(A|\theta, Y)$ is a reciprocal chi-squared distribution:

$$\frac{\lambda + \|\theta\|^2}{\chi^2_{d+q}},$$

where $\|\theta\|^2$ denotes the sum of squares. Thus, the chained data augmentation algorithm is given in this case as follows. At iteration i:

1. Sample $\theta_j^{(i)}$ from $N[(1 - B_j^{(i)})Y_j, V_j(1 - B_j^{(i)})]$, with $B_j^{(i)} = V_j/(V_j + A^{(i)})$, $j = 1, \ldots, d$.
2. Sample $A^{(i)}$ from $(\lambda + \|\theta^{(i)}\|^2)/\chi^2_{d+q}$.

Morris (1987) suggests that m chained paths be created to allow for computation of posterior means and posterior distributions. Note that in this case, the hyperparameter A plays the role of the "missing" data.

Example. *Genetic Linkage* (*Continued*)
The application of chained data augmentation to this example is quite straightforward. Given z^*, a chain would be constructed by sampling a value for $\theta(\theta^*)$ from $p(\theta|z^*, Y) = \text{Be}(z^* + X_5 + 1, X_3 + X_4 + 1)$ and then sampling a value for $Z(z^*)$ from $p(Z|\theta^*, Y) = \text{Bi}(125, \theta^*/(\theta^* + 2))$. One then iterates between these two steps to obtain a chain.

6.1.2. Multivariate Chained Data Augmentation – The Gibbs Sampler

We now consider a multivariate extension of the chained data augmentation algorithm, which is known as the systematic scan Gibbs sampler. Given the starting point $(\theta_1^{(0)}, \theta_2^{(0)}, \ldots, \theta_d^{(0)})$, this algorithm iterates the following loop:

a. Sample $\theta_1^{(i+1)}$ from $p(\theta_1|\theta_2^{(i)}, \ldots, \theta_d^{(i)}, Y)$

b. Sample $\theta_2^{(i+1)}$ from $p(\theta_2|\theta_1^{(i+1)}, \theta_3^{(i)}, \ldots, \theta_d^{(i)}, Y)$

$\vdots \qquad \vdots$

d. Sample $\theta_d^{(i+1)}$ from $p(\theta_d|\theta_1^{(i+1)}, \ldots, \theta_{d-1}^{(i+1)}, Y)$.

The vectors $\theta^{(0)}, \theta^{(1)}, \ldots, \theta^{(t)}, \ldots$ are a realization of a Markov chain, with transition probability from θ' to θ,

$$K(\theta', \theta) = p(\theta_1|\theta'_2, \ldots, \theta'_d, Y)p(\theta_2|\theta_1, \theta'_3, \ldots, \theta'_d, Y)p(\theta_3|\theta_1, \theta_2, \theta'_4, \ldots, \theta'_d, Y) \times p(\theta_d|\theta_1, \ldots, \theta_{d-1}, Y) .$$

Chan (1993) Geman and Geman (1984), Liu, Wong and Kong (1991a, b), Tierney (1991) and Schervish and Carlin (1992) present conditions such that:

Result 6.1.1. The joint distribution of $(\theta_1^{(i)}, \ldots, \theta_d^{(i)})$ converges geometrically to $p(\theta_1, \ldots, \theta_d | Y)$, as $i \rightarrow \infty$.

Result 6.1.2. $\frac{1}{t} \sum_{i=1}^{t} f(\theta^{(i)}) \xrightarrow{\text{a.s.}} \int f(\theta) p(\theta | Y) \, d\nu(\theta)$, as $t \rightarrow \infty$.

Applegate, Kannan and Polson (1990), Rosenthal (1992), and Amit (1991) provide bounds for rates of convergence in specific situations.

In many applications of the Gibbs sampler, d simulations are performed per iteration corresponding to the individual d components of the parameter $\theta = (\theta_i, \ldots, \theta_d)$. In other cases the components could be sub vectors of $\underset{\sim}{\theta}$. In this regard, if *highly correlated components* are treated individually, then convergence can be quite slow. If, however, the correlated components are treated as a block (as in data augmentation), this problem may be mitigated. This issue is treated in detail in Liu, Wong and Kong (1991a, b).

Gelman and Rubin (1992) argue that one should monitor the behavior of several chains starting from points sampled from an overdispersed distribution and that one should check both within- and between-chain variation to assess convergence. (Further discussion of this latter idea is presented in Section 6.3.3.) This overdispersed distribution is constructed in three steps. In step 1, one locates the K modes of $p(\theta | Y)$. In step 2, one approximates $p(\theta | Y)$ by a mixture of multivariate t distributions – with each t distribution centered at a mode. Gelman and Rubin (1992) suggest that four degrees of freedom on each t should suffice. The inverse information at each mode (possibly multiplied by a scaling factor) determines the overdispersed approximation to $p(\theta | Y)$. In step 3, they then use the *SIR* method to obtain the starting points for the chains. This last step makes the starting points "more typical of the target distribution" and should speed up the convergence of the algorithm. Gelman and Rubin (1992) suggest that one draw about 1000 points from the mixture of multivariate t distributions, and then draw 10 importance-weighted resamples – using larger samples when more than one major mode exists. It is noted that the first step assumes that one can locate the (dominant) modes, though this may be difficult in practice. It is also noted that the poor man's data augmentation algorithms may provide a useful approximation to $p(\theta | Y)$ from which one can draw starting points for the Gibbs sampler chains.

Geyer (1992) argues that inference can be based on one long run of a Markov chain and presents methods from the time series and operations research literature to quantify the variance of an estimator based on correlated observations. Clearly, several parallel runs or batches within a run can indicate that a single run is not sufficiently long. However, there can never be any guarantee that a run is "long enough". Section 6.3 discusses techniques for monitoring convergence of the Gibbs sampler.

Example. *Genetic Linkage (Continued)*
Gelfand and Smith (1990) consider an extension of the genetic linkage example. They present the data as $Y = (14, 1, 1, 1, 5)$ and the model

$$\left[\left(\frac{\theta}{4} + \frac{1}{8}\right), \frac{\theta}{4}, \frac{\eta}{4}, \frac{\eta}{4} + \frac{3}{8}, \frac{1}{2}(1 - \theta - \eta)\right].$$

They consider the augmented data set

$$X = (X_1, \ldots, X_7) \sim \text{Mult}\left[22; \frac{\theta}{4}, \frac{1}{8}, \frac{\theta}{4}, \frac{\eta}{4}, \frac{\eta}{4}, \frac{3}{8}, \frac{1}{2}(1 - \theta - \eta)\right].$$

Gelfand and Smith let $Z = (X_1, X_5)$. Under the Dirichlet (1, 1, 1) prior for (θ, η), the conditionals are:

$$p(\theta | Y, \eta, Z) = (1 - \eta) Be(X_1 + Y_2 + 1, Y_5 + 1)$$

$$p(\eta | Y, \theta, Z) = (1 - \theta) Be(Y_3 + X_5 + 1, Y_5 + 1)$$

$$p(Z | Y, \theta, \eta) = p(X_1, X_5 | Y, \theta, \eta) = Bi[Y_1, 2\theta(1 + 2\theta)^{-1}] Bi[Y_4, 2\eta(3 + 2\eta)^{-1}].$$

Using Gaussian quadrature techniques, Gelfand and Smith (1991) find that $E(\theta | Y) = 0.520$, $E(\eta | Y) = 0.123$. These authors report the corresponding Tierney–Kadane approximation values of 0.518 and 0.088, respectively. Gelfand and Smith (1991) iterated the Gibbs sampler 10 times and considered 20 replicates of the path. To approximate $E(\eta | Y)$, the authors averaged the η component of the 10th iteration over the 20 replicates. This process was then repeated 5000 times. Gelfand and Smith (1991) calculate

$$E(\eta | Y) = 0.123$$

$$E(\theta^2 | Y) = 0.288$$

$$E(\eta^2 | Y) = 0.022$$

$$\text{var}(\theta | Y) = 0.018$$

$$\text{var}(\eta | Y) = 0.0065$$

These values agree with the exact values based on Gaussian quadrature. Gelfand and Smith (1991) show that by the fourth iteration, the Gibbs sampler correctly computed the area to the left of the 5th, 25th, 50th, 75th and 95th percentage point to two decimal places.

6.1.3. Historical Comments

Geman and Geman (1984) present the Gibbs Sampler in the context of spatial processes involving large numbers of variables e.g. image reconstruction. They consider under which situations the conditional distributions, given "neighborhood" subsets of the variables, uniquely determines the joint distribution. In the situation considered in this chapter, Besag (1974) has shown that if the joint distribution $p(\theta_1, \ldots, \theta_d)$ is positive over its entire domain, then the joint distribution is uniquely determined by the d conditional distributions

$$p(\theta_1|\theta_2, \ldots, \theta_d), \ldots, p(\theta_d|\theta_1, \ldots, \theta_{d-1}) \ .$$

The related Metropolis algorithm is discussed in Section 6.5.

Li(1988) applied the Gibbs sampler in the context of multiple imputation. Li suggests that the complete data be partitioned into $d+1$ parts, $X_0, X_1, \ldots, X_d$, where the observed data is X_0 and $X_1, \ldots, X_d$ is a partition of the missing data. Li assumes that X_i can be sampled from $p(X_i|X_j, j \neq i)$. His algorithm is:

Step 1. Sample $X_1^{(0)}, \ldots, X_d^{(0)}$ from some distribution.
Step 2. a. Sample $X_1^{(j)}$ from

$$p(X_1|X_0, X_2^{(j-1)}, \ldots, X_d^{(j-1)})$$

b. Sample $X_2^{(j)}$ from

$$p(X_2|X_0, X_1^{(j)} X_3^{(j-1)}, \ldots, X_d^{(j-1)})$$

⋮

d. Sample $X_d^{(j)}$ from

$$p(X_d|X_0, X_2^{(j)}, \ldots, X_{d-1}^{(j)}) \ .$$

Step 2 is then cycled until the algorithm converges. Li suggests that multiple paths be considered to check for convergence.

Li (1988) illustrates the method in the context of categorical data, latent varibles and censored life data. Li (1988) provides conditions such that the distribution of $(X_1^{(i)}, \ldots, X_d^{(i)})$ converges to $p(X_1, \ldots, X_d)$ geometrically fast. Like Metropolis et al. (1953) and Geman and Geman (1984), Li represents the process as a Markov chain with the joint posterior distribution as the stationary distribution.

Tanner and Wong (1987) present the data augmentation algorithm which is a two component version of the Gibbs sampler. One of the basic contributions of Tanner and Wong (1987) was to develop the framework in which Bayesian computations can be performed in the context of iterative Monte Carlo algorithms. Moreover, in their rejoinder they sketch a Gibbs sampler approach for handling hierarchical models with t errors. Gelfand and Smith (1990) present a review of data augmentation, the Gibbs sampler and the *SIR* algorithm due to Rubin (1987a). These authors apply the approaches to several examples and provide an initial practical comparison of the methods. The papers by Gelfand et al. (1990), Gelfand and Smith (1991), Carlin, Gelfand and Smith (1992), Carlin and Polson (1991), Gelfand, Smith and Lee (1992), and Carlin, Polson and Stoffer (1992) apply the Gibbs sampler to a variety of important statistical problems. This listing is but a small part of the important and interesting published applications of the Gibbs sampler. Volume 55 (No. 1) of the *Journal of the Royal Statistical Society, B*, presents extensive discussion of Markov Chain Monte Carlo methods.

6.2. Examples

6.2.1. Rat Growth Data

In Chapter 1, data on 30 young rats measured weekly for five weeks were presented. Gelfand et al. (1990) suppose individual straight-line growth curves, i.e.

$$Y_{ij} \sim N(\alpha_i + \beta_i x_{ij}, \sigma^2)$$

where $i = 1, \ldots, 30; j = 1, \ldots, n_i = 5$; and x_{ij} denotes the age in days of the ith rat at measurement j. The slopes and intercepts are modelled as

$$\begin{pmatrix}\alpha_i\\ \beta_i\end{pmatrix} \sim N\left[\begin{pmatrix}\alpha_c\\ \beta_c\end{pmatrix}, \Sigma\right]$$

with conditional independence for $i = 1, \ldots, 30$ assumed. Gelfand et al. (1990) take the prior

$$p(\alpha_c, \beta_c; \Sigma^{-1}; \sigma^2) = p(\alpha_c, \beta_c)\, p(\Sigma^{-1})\, p(\sigma^2)$$

where

$$p(\alpha_c, \beta_c) = N(\eta, C), \qquad p(\Sigma^{-1}) = W[(\rho R)^{-1}, \rho]$$

$$p(\sigma^2) = IG\left(\frac{v_0}{2}, \frac{v_0 \tau_0^2}{2}\right),$$

W is the Wishart distribution and IG is the inverse-gamma distribution. Define $\theta_i = (\alpha_i, \beta_i)^{\mathrm{T}}$, X_i is the appropriate design matrix,

$$\bar{\theta} = \frac{1}{30} \sum_{i=1}^{30} \theta_i, \qquad D_i = \sigma^2 X_i^{\mathrm{T}} X_i + \Sigma^{-1}, \qquad V = (30\Sigma^{-1} + C^{-1})^{-1},$$

$$\mu_c = (\alpha_c, \beta_c)$$

and

$$n = \sum_{i=1}^{30} n_i = 30 \times 5 = 150 \ .$$

The Gibbs sampler is specified by the following distributions

$$p(\theta_i | Y, \mu_c, \Sigma^{-1}, \sigma^2) = N[D_i(\sigma^{-2} X_i^{\mathrm{T}} Y_i + \Sigma^{-1} \mu_c), D_i]$$

$$p(\mu_c | Y, \theta_1, \ldots, \theta_{30}, \Sigma^{-1}, \sigma^2) = N[V(30\Sigma^{-1} \bar{\theta} + C^{-1}\eta), V]$$

$$p(\Sigma^{-1}|Y,\theta_1,\ldots,\theta_{30},\mu_c,\sigma^2) = W\left[\left[\sum_i(\theta_i-\mu_c)(\theta_i-\mu_c)^T+\rho R\right]^{-1},30+\rho\right]$$

$$p(\sigma^2|Y,\theta_1,\ldots,\theta_{30},\mu_c,\Sigma^{-1}) = IG\left\{\frac{n+\nu_0}{2},\frac{1}{2}\left[\sum_i(Y_i-X_i\theta_i)^T(Y_i-X_i\theta_i)+\nu_0\tau_0^2\right]\right\}.$$

For the rat data, Gelfand et al. (1990) take

$$C^{-1}=0,\ \nu_0=0,\ \rho=2 \text{ and } R=\begin{pmatrix}100 & 0\\ 0 & 0.1\end{pmatrix}$$

reflecting the vague prior information. Note that to sample from the inverse gamma distribution, one draws a deviate from the corresponding gamma distribution and forms the reciprocal of this deviate.

For these data, Gelfand et al. (1990) report convergence of the Gibbs sampler in about 35 iterations with $m = 50$. Figures 6.1 present the resulting posterior distribution for α_c and β_c, respectively. The solid line is obtained from the Gibbs sampler, and the dashed line is the normal approximation based on the results of the *EM* algorithm. As can be seen in the plots, the normal approximation underestimates the dispersion of the parameters. An

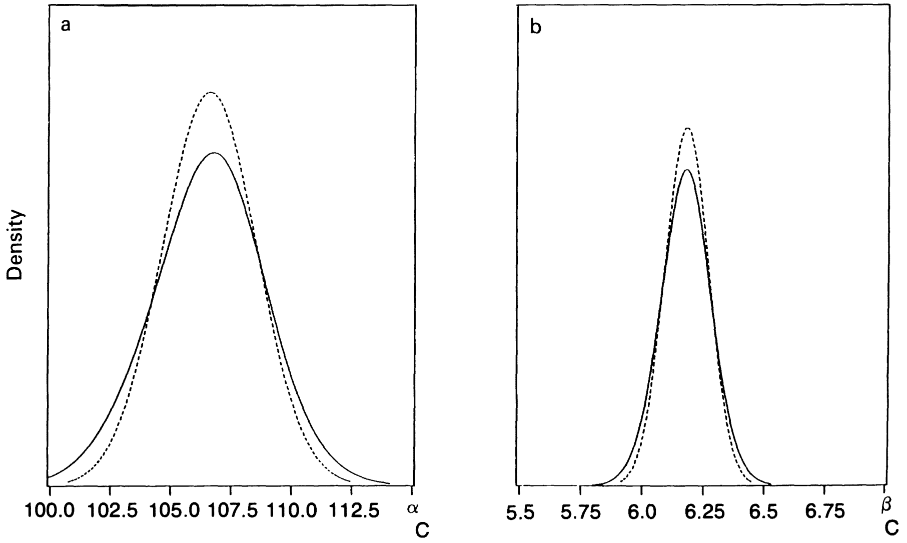


Figure 6.1. (a) Posterior of population intercept. (b) posterior of population slope. solid-Gibbs dashed-EM. Source: Gelfand et al. (1990).

even poorer performance of the normal approximation is noted by Gelfand et al. (1990) for smaller versions of this data set. These lower-dimensional marginals (e.g. $p(\alpha_c | Y)$) were obtained by averaging $p(\alpha_c | \theta, \beta_c, \Sigma^{-1}, \sigma, Y)$ over the sampled values of θ, μ_c, Σ^{-1} and σ (see Section 5.2). Gelfand and Smith (1990) and Liu, Wong and Kong (1991a, b) show that this estimate has a smaller mean-square error than nonparametric estimates (e.g. histogram) based on the realized values. Gelfand and Smith (1990) use a Rao–Blackwell argument to derive their result. Hence, this estimate of the marginal is referred to as the *Rao–Blackwell* estimate. The motivation for this estimate is simply applying the method of Monte Carlo to the integral:

$$p(\theta_1 | Y) = \int p(\theta_1 | \theta_2, \cdots, \theta_d, Y)\, p(\theta_2, \cdots, \theta_d | Y)\, d\theta_2 \cdots d\theta_d.$$

6.2.2. Poisson Process with a Change Point

Carlin, Gelfand and Smith (1992) consider the application of the Gibbs sampler to a Poisson process model with a change point. In particular, Carlin et al. (1992) let $Y_i \sim$ Poisson (θt_i), $i = 1, \ldots, k$, $Y_i \sim$ Poisson (λt_i), $i = k + 1, \ldots, n$ at the first stage. At the second stage, the authors put independent priors on k, θ and λ: k discrete uniform on $\{1, \ldots, n\}$; $\theta \sim G(a_1, b_1)$ and $\lambda \sim G(a_2, b_2)$, where G denotes the Gamma distribution. At the third stage, the authors let $b_1 \sim IG(c_1, d_1)$ independent of $b_2 \sim IG(c_2, d_2)$, where IG denotes the inverse gamma distribution, and suppose that a_1, a_2, c_1, c_2, d_1 and d_2 are known. The conditional distributions are

$$p(\theta | Y, \lambda, b_1, b_2, k) \sim G\left[a_1 + \sum_{i=1}^{k} Y_i, \left(\sum_{i=1}^{k} t_i + b_1^{-1}\right)^{-1}\right]$$

$$p(\lambda | Y, \theta, b_1, b_2, k) \sim G\left[a_2 + \sum_{i=k+1}^{k} Y_i, \left(\sum_{i=k+1}^{n} t_i + b_2^{-1}\right)^{-1}\right]$$

$$p(b_1 | Y, \theta, \lambda, b_2, k) \sim IG[a_1 + c_1, (\theta + d_1^{-1})^{-1}]$$

$$p(b_2 | Y, \theta, \lambda, b_1, k) \sim IG[a_2 + c_2, (\lambda + d_2^{-1})^{-1}]$$

and

$$p(k | Y, \theta, \lambda, b_1, b_2) = \frac{L(Y, k, \theta, \lambda)}{\sum_{k=1}^{n} L(Y, k, \theta, \lambda)}$$

where

$$L(Y, k, \theta, \lambda) = \exp\left[(\lambda - \theta) \sum_{i=1}^{k} t_i\right] (\theta/\lambda)^{\sum_{i=1}^{k} Y_i}.$$

Table 6.1. British coalmining disaster data by year, 1851–1962.

Year	Count	Year	Count	Year	Count	Year	Count
1851	4	1881	2	1911	0	1941	4
1852	5	1882	5	1912	1	1942	2
1853	4	1883	2	1913	1	1943	0
1854	1	1884	2	1914	1	1944	0
1855	0	1885	3	1915	0	1945	0
1856	4	1886	4	1916	1	1946	1
1857	3	1887	2	1917	0	1947	4
1858	4	1888	1	1918	1	1948	0
1859	0	1889	3	1919	0	1949	0
1860	6	1890	2	1920	0	1950	0
1861	3	1891	2	1921	0	1951	1
1862	3	1892	1	1922	2	1952	0
1863	4	1893	1	1923	1	1953	0
1864	0	1894	1	1924	0	1954	0
1865	2	1895	1	1925	0	1955	0
1866	6	1896	3	1926	0	1956	0
1867	3	1897	0	1927	1	1957	1
1868	3	1898	0	1928	1	1958	0
1869	5	1899	1	1929	0	1959	0
1870	4	1900	0	1930	2	1960	1
1871	5	1901	1	1931	3	1961	0
1872	3	1902	1	1932	3	1962	1
1873	1	1903	0	1933	1		
1874	4	1904	0	1934	1		
1875	4	1905	3	1935	2		
1876	1	1906	1	1936	1		
1877	5	1907	0	1937	1		
1878	5	1908	3	1938	1		
1879	3	1909	2	1939	1		
1880	4	1910	2	1940	2		

Carlin et al. (1992) apply this model to the British coalmining disaster data given in Table 6.1.

For these data, Carlin et al. (1992) take $a_1 = a_2 = 0.5$, $c_1 = c_2 = 0$ and $d_1 = d_2 = 1$. Convergence was obtained after 15 iterations with $m = 100$. Figure 6.2 presents $p(k \mid Y)$ for the entire data set (solid line) and for the data set with every fifth year deleted (dashed lines). In both cases, the mode is at $k = 41$. Note that the posterior probability that $k = n$, i.e. there is no change point, is close to 0. Thus, there is strong evidence of a change. Carlin et al. (1992) also present the $p(\theta \mid Y)$ and $p(\lambda \mid Y)$ marginals for these data.

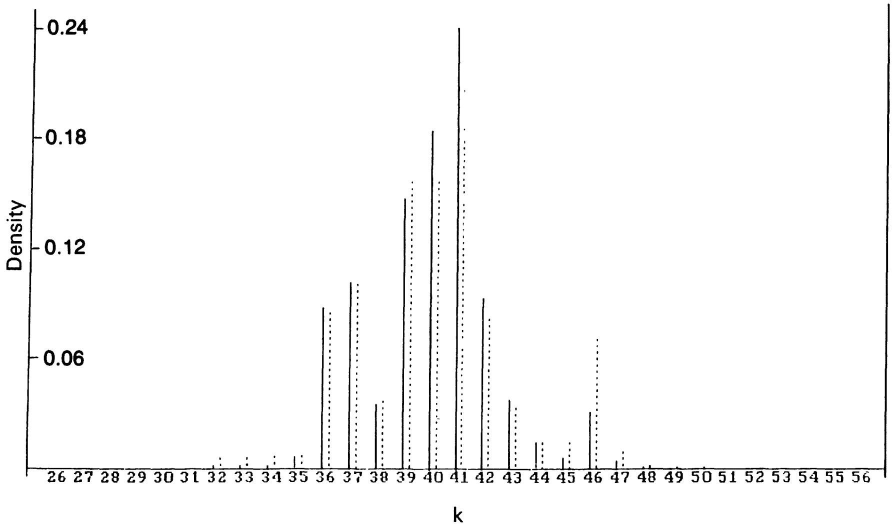


Figure 6.2. Posterior distribution of change point. Source: Carlin et al. (1992).

6.2.3. Generalized Linear Models with Random Effects

Zeger and Karim (1991) consider the generalized linear model with random effects. In particular, the data consist of the response y_{ij} and a $p \times 1$ vector of predictors x_{ij} for observations $j = 1, 2, \ldots, n_i$ within clusters $i = 1, \ldots, I$. Conditional on a random variable b_i, y_{ij} follows an exponential family distribution'

$$f(y_{ij}|b_i) = \exp\{[y_{ij}\theta_{ij} - a(\theta_{ij}) + c(y_{ij})]/\phi\}.$$

This is the well-known generalized linear model (McCullagh and Nelder, 1989). The conditional moments $u_{ij} = E(y_{ij}|b_i) = a'(\theta_{ij})$ and $v_{ij} =$ var $(y_{ij}|b_i) = a''(\theta_{ij})\phi$ satisfy

$$h(u_{ij}) = \eta_{ij} = x_{ij}^T\beta + z_{ij}^T b_i$$

and

$$v_{ij} = g(u_{ij})\phi$$

where h and g are known link and variance functions, respectively; z_{ij} is a $q \times 1$ subset of x_{ij} with random coefficients, β is a vector of regression coefficients and b_i is a $q \times 1$ vector of random effects following a multivariate normal distribution with mean 0 and variance D.

To complete the specification of the model, let $y_i = (y_{i1}, \ldots, y_{in_i})^T$, $X_i = (x_{i1}, \ldots, x_{in_i})^T$, $Z_i = (z_{i1}, \ldots, z_{in_i})^T$, $U_i = (u_{i1}, \ldots, u_{in_i})^T$, $\eta_i = (\eta_{i1}, \ldots, \eta_{in_i})^T$ and $V_i = \text{diag}\,(v_{i1}, \ldots, v_{in_i})$.

Zeger and Karim (1991) apply the Gibbs sampler in this context. To implement the Gibbs sampler, these authors consider three distributions: $p(\beta|b, D, Y)$; $p(D|\beta, b, Y)$ and $p(b|\beta, D, Y)$.

Regarding $p(\beta|b, D, Y)$, Zeger and Karim (1991) show that $p(\beta|b, D, Y) = p(\beta|b, Y)$. Given the values $b^{(K)}$ for the b's, the random-effects model reduces to a generalized linear model with offset $z_{ij}b_i^{(K)}$ for each response. For a flat prior, $p(\beta|b^{(K)}, Y)$ is proportional to the likelihood $\prod_{i,j} f(y_{ij}|b_i^{(K)})$. To sample from the likelihood, Zeger and Karim employ a rejection/acceptance algorithm. (See Section 3.3.3 for a discussion of this algorithm.)

In the context of $p(\beta|b^{(K)}, Y)$, Zeger and Karim (1991) take the majorizing function of the rejection/acceptance algorithm to be $c_1 N(\hat{\beta}^{(K)}, c_2 V_{\hat{\beta}}^{(K)})$, where $\hat{\beta}^{(K)}$ is the maximizer of $p(\beta|b^{(K)}, Y)$ and $V_{\hat{\beta}}^{(K)}$ is the inverse Fisher information. These quantities may be obtained by performing the generalized linear model regression of y_{ij} on x_{ij} with $z_{ij}b_i^{(K)}$ taken as offsets. Zeger and Karim (1991) inflate the variance of the majoring function by taking $c_2 = 2.0$. The quantity c_1 is selected so that the modes of the majorizing density and $p(\beta|b^{(K)}, Y)$ are equal.

Regarding the conditioned mariginal $p(b|\beta^{(K)}, D^{(K)}, Y)$ the authors again use a rejection/acceptance algorithm. In this case, the majorizing function is $cN(A, B)$, where $A = (Z_i^T V_i Z_i + D^{-1})^{-1} \; Z_i^T V_i(y_1^* - X_i\beta)$; $B = c_2(Z_i^T V_i Z_i + D^{-1})^{-1}$; y_i^* is the linear approximation to $h(y_i)$:

$$y_i^* = \eta_i + \left(\frac{\partial u_i}{\partial \eta_i}\right)^{-1} (y_i - u_i)$$

and

$$\eta_i = X_i\beta + Z_i b_i \ .$$

Zeger and Karim (1991) show that $p(D|\beta, b, Y) = p(D|b)$. Assuming that the b_i's are independent normal with mean 0 and variance D and adopting the noninformative prior, the posterior of D^{-1} is a Wishart distribution with parameters

$$S^{(K)} = \sum_{i=1}^{I} b_i^{(K)} b_i^{(K)^T}$$

and I–q + 1 degrees of freedom.

Zeger and Karim (1991) apply the algorithm to fit a logistic-normal random-effects model to infectious disease data on 250 Indonesian children. The children were examined up to six times for the presence of respiratory infection. There were a total of 1200 responses. The covariates considered are: age in months (centered at 36); presence/absence of xerophthalmia; cosine and sine terms for the annual cycle; gender; height for age as a percentage of the National Center for Health Statistics (NCHS) standard (centered at 90%) and presence of stunting, defined as being below 85% in height for age. The intercept is taken as a random effect with a normal distribution.

Table 6.2. Posterior characteristics.

Explanatory variable	Mean	Mode	St. dev.	90% C.I.	
Intercept	− 2.74	− 2.72	0.23	− 3.10,	− 2.34
Age (months)	− 0.04	− 0.04	0.01	− 0.05,	− 0.02
Xerophthalmia (0 – no; 1 – yes)	0.64	0.64	0.51	− 0.31,	1.34
Sex (0 – male, 1 – female)	− 0.61	− 0.57	0.18	− 0.91,	− 0.33
Seasonal cosine	− 0.17	− 0.16	0.17	− 0.45,	0.13
Seasonal sine	− 0.45	− 0.44	0.26	− 0.89,	− 0.03
Height for age (% of NCHS standard)	− 0.05	− 0.05	0.03	− 0.10,	− 0.01
Stunted (< 85% height for age; 0 – no, 1 – yes)	0.18	0.24	0.47	− 0.63,	0.91
D	0.80	0.78	0.40	0.25,	1.54

Table 6.2 taken from Zeger and Karim (1991) lists the estimated posterior mean, mode, standard deviation and 90% confidence interval for each parameter. These statistics were computed from 1000 simulated values obtained after the algorithm converged. As can be seen from Table 6.2, respiratory infection appears to be related to age, gender, season and height for age. Further analysis, including a presentation of the marginal densities, is given in Zeger and Karim (1991).

6.3. Assessing Convergence of the Chain

6.3.1. The Gibbs Stopper

The basic idea behind the Gibbs stopper is to assign the weight w to the vector $\theta = (\theta_1, \ldots, \theta_d)$, which has been drawn from the current approximation to the joint density g_i via

$$w(\theta) = \frac{q(\theta_1, \ldots, \theta_d | Y)}{g_i(\theta_1, \ldots, \theta_d)},$$

where $q(\theta_1, \ldots, \theta_d | Y)$ is proportional to the posterior density $p(\theta_1, \ldots, \theta_d | Y)$. As g_i converges toward $p(\theta_1, \ldots, \theta_d | Y)$, the distribution of the weights (associated with values of θ drawn from g_i) should converge toward a spike distribution. This observation has been found useful in assessing convergence of the Gibbs sampler, as well as in transforming a sample from g_i into a sample from the exact distribution; see Ritter and Tanner (1992). Historically, the idea of using importance weights to monitor convergence of the data augmentation algorithm was first presented in the rejoinder of Tanner and Wong (1987) and illustrated in Wei and Tanner (1990), as discussed in Section 5.4.

To write down the functional form for g_i for the Gibbs sampler, we introduce notation following Schervish and Carlin (1992). Let $p^{(i)}(\theta) = p(\theta_i | \theta_1, \ldots, \theta_{i-1}, \theta_{i+1}, \ldots, \theta_d, Y)$. For two vectors θ and θ', define for each $i < d$, $\theta^{(i')} = (\theta_1, \ldots, \theta_i, \theta'_{i+1}, \ldots, \theta'_d)$ and $\theta^{(d')} = \theta$. As noted in Schervish and Carlin (1992), if g_i is the joint density of the observations sampled at iteration i, then the joint density of the observations sampled at the next iteration (g_{i+1}) is given by

$$\int K(\theta', \theta) g_i(\theta') \, d\lambda(\theta'), \qquad K(\theta', \theta) = \prod_{i=1}^{d} p^{(i)}(\theta^{(i')}) \tag{6.3.1}$$

[see also Tanner and Wong (1987) and Liu, Wong and Kong (1991a, b)]. One may approximate the integral in (6.1.1) via the method of Monte Carlo. In particular, given the observations $\theta^1, \theta^2, \ldots, \theta^m$, use the Monte Carlo sum

$$\frac{1}{m} \sum_{j=1}^{m} K(\theta^j, \theta) \tag{6.3.2}$$

to approximate $g_{i+1}(\theta)$. Ritter and Tanner (1992) suggest using θ values from independent chains. One may also use *successive* θ values from one chain to construct the Monte Carlo sum (6.3.2). Note that construction of (6.3.2.) requires the normalizing constants (or good approximations to the normalizing constants) for the conditional distributions. Also note that we are examining, through $p(\theta_k | \theta_1, \ldots, \theta_{k-1}, \theta_{k+1}, \ldots, \theta_d, Y)$, the first component of each θ vector along with components 2–d of the other $m - 1$ θ vectors, the first and second components of each θ vector along with components 3–d of the other $m - 1$ θ vectors, etc, thereby expanding the coverage of the parameter space. The effort in constructing (6.3.2) will yield useful information regarding the state of the Markov chain vis-a-vis the equilibrium distribution. An illustration of the potential of this approach is given below.

To illustrate this convergence diagnostic, consider the "Witch's Hat" distribution presented in Matthews (1991). The posterior under consideration is proportional to a mixture of a multivariate normal distribution and uniform distribution on the open d-dimensional hypercube $(0, 1)^d$ C:

$$(1 - \delta)\left(\frac{1}{\sqrt{2\pi\sigma}}\right)^d e^{-\frac{1}{2}\sum_{i=1}^{d}\left(\frac{y_i - \theta_i}{\sigma}\right)^2} + \delta I_{(Y \varepsilon C)} \; . \tag{6.3.3}$$

We have chosen $\delta = 10^{-11}$, $\sigma = 0.03$, $d = 9$ and $Y = (0.9, 0.9, 0.9, 0.9, 0.9, 0.9, 0.9, 0.9, 0.9)$. A cursory examination of the posterior reveals a spike centered at Y, with a flat "brim" extending out to the boundary of the unit hypercube. We proceed under the assumption that the posterior at hand is analytically complicated, thereby not allowing for an easy recognition of the location of or number of spikes. Clearly, in a situation where determining the number and location of the modes is straightforward, one would focus attention around these points, possibly along the lines suggested by Gelman and Rubin (1992).

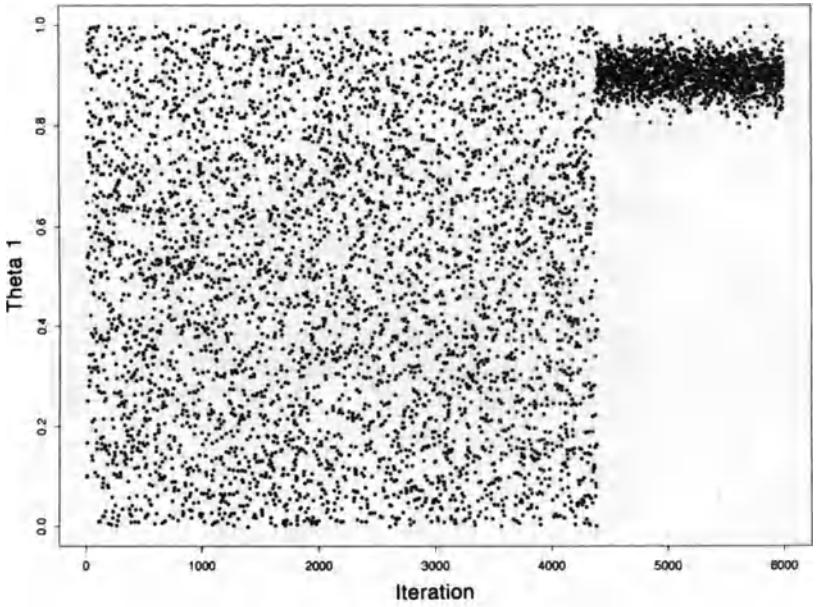


Figure 6.3. θ_1 vs. iteration.

Figure 6.3 presents a history of the θ_1 marginal, starting from the point (0.1, 0.1, 0.1, 0.1, 0.1, 0.1, 0.1, 0.1, 0.1). The plots for the other marginals are quite similar. As can be seen from the plot, the Markov chain wanders about the hypercube until iteration 4400 or so, at which point it locates the mode. A run shorter than 4400 iterations would not have detected the spike.

Figure 6.4 presents a plot of $w(\theta^{(i)})$ vs. iteration i, $i = 1, 270$. The weights in this plot are based on the first 270 successive (9-dimensional) points in the chain. The 270 weights in the plot were standardized to have mean zero and unit standard deviation. As can be seen in the plot, all 270 weights are equal, with the exception of two weights.

The reason for the outlying weights is easily explained. The conditional densities $p(\theta_k|\theta_1, \ldots, \theta_{k-1}, \theta_{k+1}, \ldots, \theta_d, Y)$ used in computing K in (6.3.1) are proportional to (6.3.3) for specified θ_k and $0 < \theta_i < 1$. When θ_i is "far" from y_i, $i \neq k$, $p(\theta_k|\theta_1, \ldots, \theta_{k-1}, \theta_{k+1}, \ldots, \theta_d, Y)$ is virtually equal to unity, independent of the value of θ_k. When θ_i is "near" y_i, $i \neq k$, $p(\theta_k|\theta_1, \ldots, \theta_{k-1}, \theta_{k+1}, \ldots, \theta_d, Y)$ (as a function of θ_k) follows the normal curve. The outlying weight points noted in Figure 6.2 stem from the fact that in three of the 270 products in (6.3.2), which are averaged to compute g_{i+1}, one of the terms $p(\theta_k|\theta_1, \ldots, \theta_{k-1}, \theta_{k+1}, \ldots, \theta_d, Y)$ has seven of the eight θ_i's within three σ's of 0.9, $i \neq k$, with θ_k "far" from 0.9. (For future reference call these components θ_i^*, $i \neq$ k.) This term is nearly zero, thereby causing the

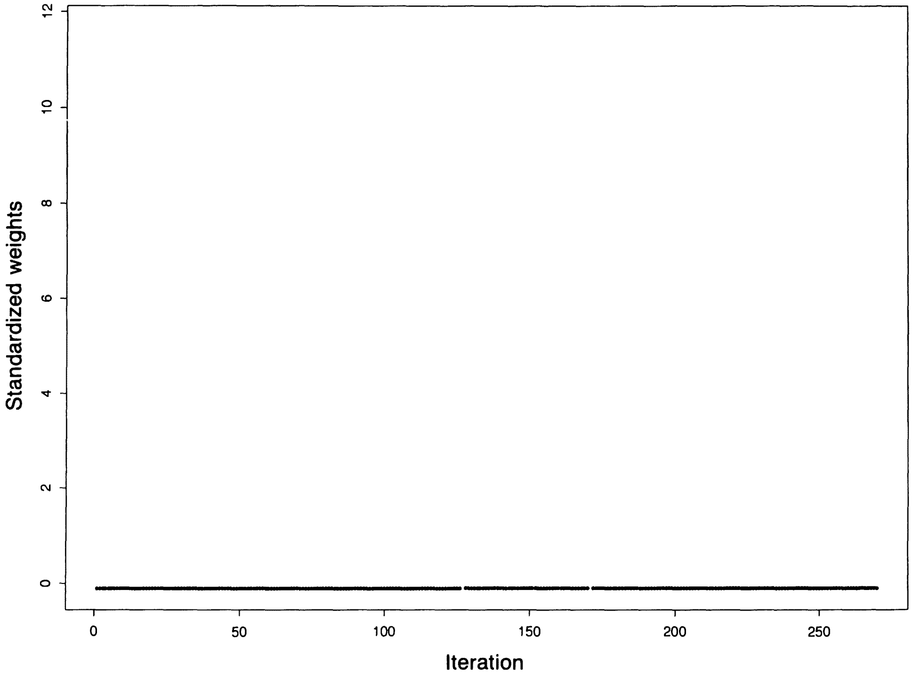


Figure 6.4. Weights vs. iteration.

entire product of terms in K to be small. Thus, rather than averaging 270 products equal to unity to compute g_{i+1}, the average was diminished by the three near-zero products, leading to the outlying weight.

As a followup to the investigation of the cause of the two outliers, the maximizer (θ_k^*) of $p(\theta_k|\theta_1,\ldots,\theta_{k-1},\theta_{k+1},\ldots,\theta_d, Y)$ was located, where the θ_i's were set equal to the θ_i^* $(i \neq k)$ values identified in the previous paragraph. A chain was started from this point. Figure 6.5 presents a plot of the history of the θ_1 marginal for this path, the other marginals are similar. As indicated in Figure 6.5, the chain moved immediately into the neighborhood of the spike.

In this example, we see how a careful examination of the outlying weights helps to locate the spike much earlier than the 4400 iterations required by the original chain. Of course, this Gibbs stopper technique is not infallible. Had we considered only the first 100 points in the Markov chain, we would have missed the outliers. Similarly, one would expect that smaller values of σ and higher values of d would require more terms in (6.3.2) – though both of these modifications would probably increase the run time of the chain as well.

Roberts (1992) essentially suggests to calculate the average of the Gibbs stopper weights (his Λ_n). He formally shows that the expectation of Λ_n goes to unity as the chain moves to the equilibrium distribution. Roberts (1992) also calculates the asymptotic variance of Λ_n (as the number of iterations goes to infinity).

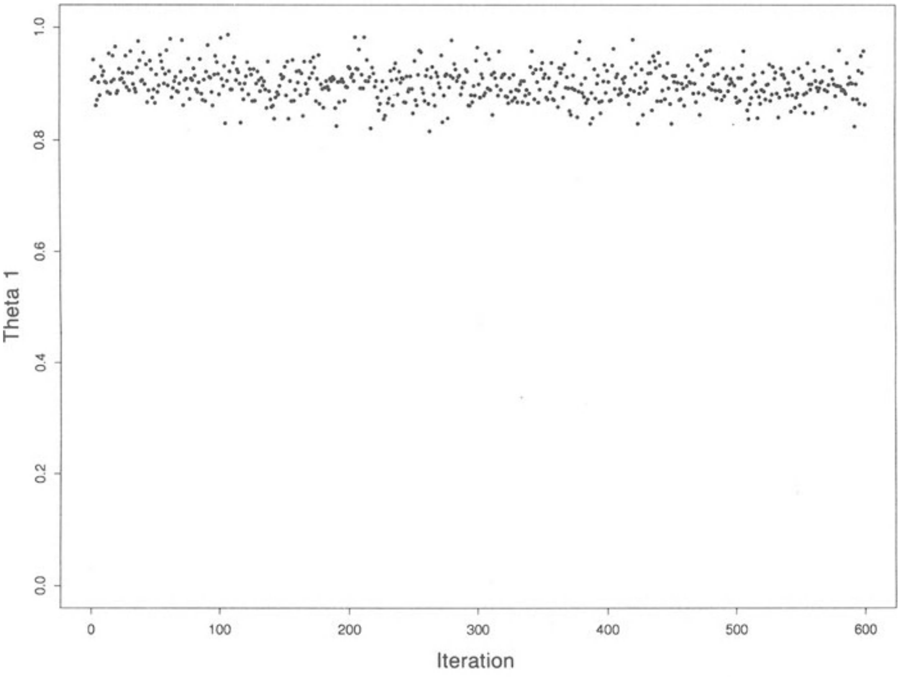


Figure 6.5. θ_1 vs. Iteration.

Additional weight plots based on nonoverlapping segments (iterations 271–810, 811–1850, 1851–4010, etc.) of the original Markov chain are presented in Figures 6.6–6.9. One does not see in the first three plots of this series a degeneration of the distribution of the weights about a spike, as would be expected if the chain was in equilibrium – providing further indication to the data analyst of an anomaly regarding convergence of the chain. These plots clearly highlight the slowly mixing nature of the chain – a valuable diagnostic for the data analyst. The final plot in the series highlights the fact that the equilibrium distribution has two distinct domains of attraction – the "brim" and the "spike". See Section 6.4.3 for further illustration of Gibbs stopper methods.

6.3.2. Control/Variates

Liu and Liu (1993) have suggested using the following control variates when m chains have been generated:

$$U^{(i,j,t)} = \frac{p(\theta^{(j,t)}|Y)}{p(\theta^{(i,t)}|Y)} \frac{K(\theta^{(j,t-1)}, \theta^{(i,t)})}{K(\theta^{(j,t-1)}, \theta^{(j,t)})}$$

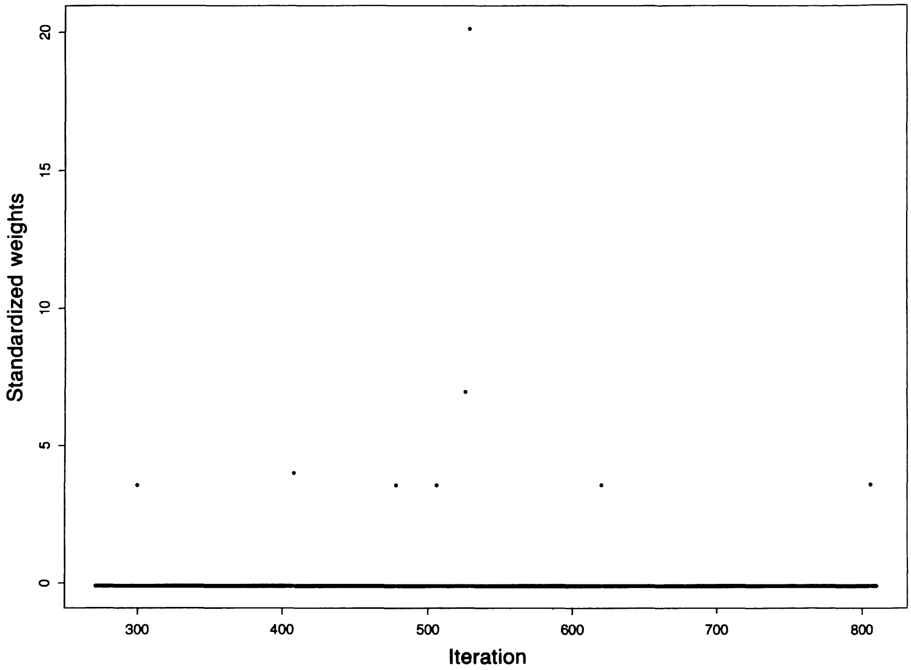
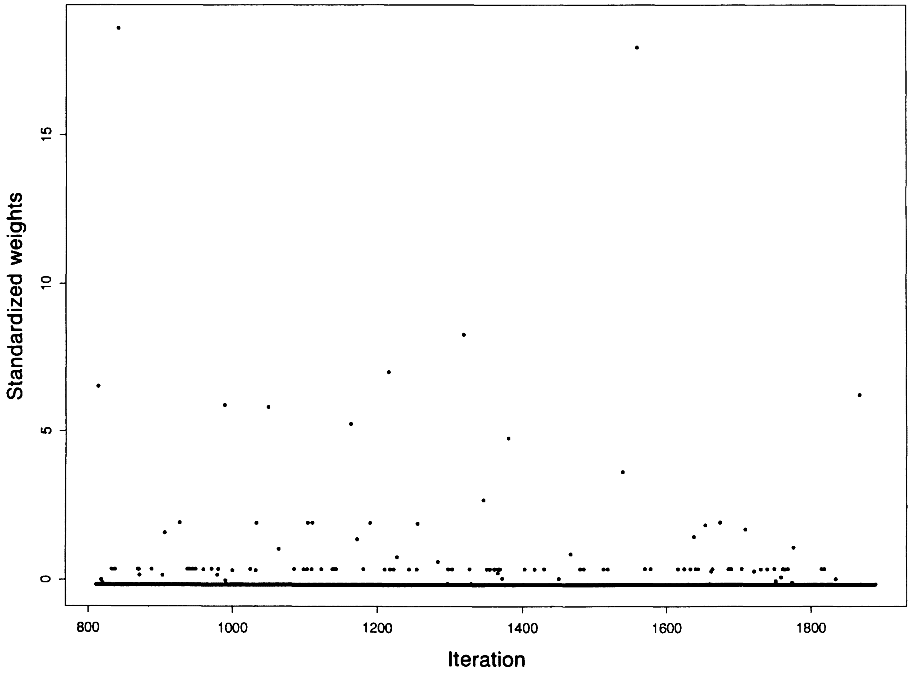


Figure 6.6. Weights vs. Iteration.



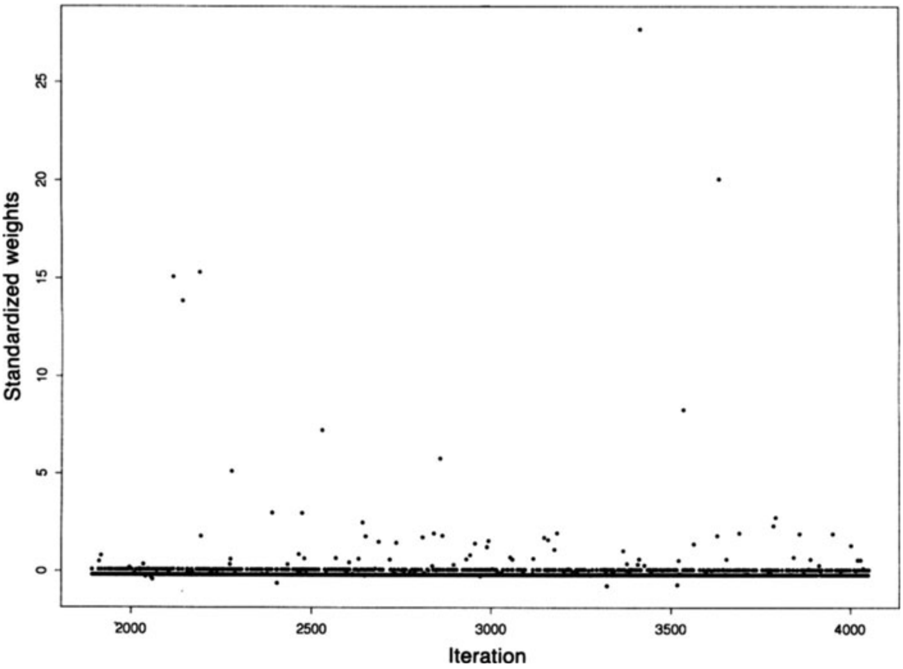
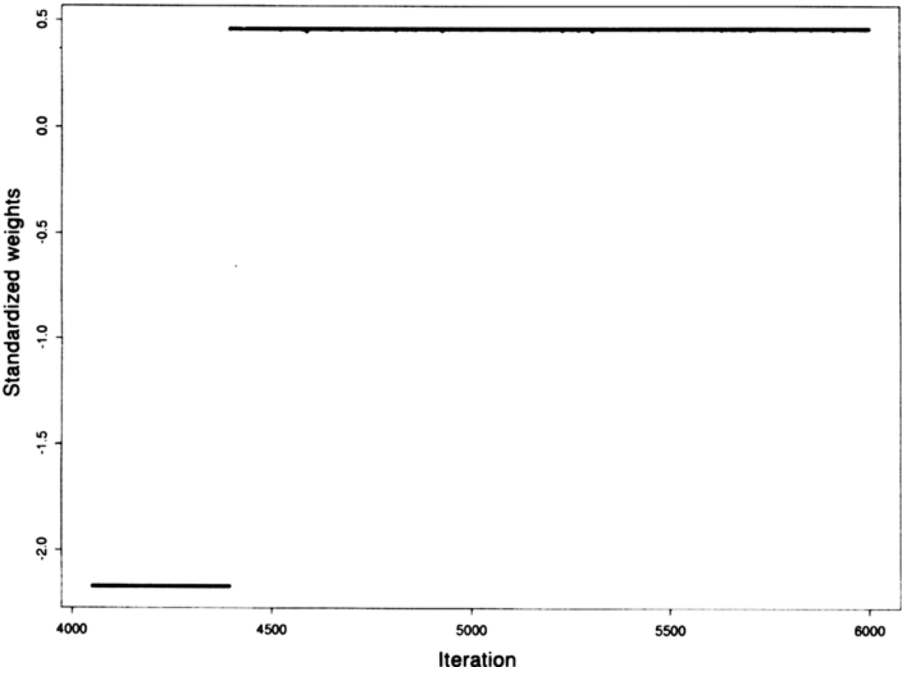


Figure 6.8. Weights vs. Iteration.



where $\theta^{(j,t)}$ is the tth value in chain j, $\theta^{(i,t)})$ is the corresponding value in chain i, for $i \neq j$, $i, j = 1, \ldots, m$ and $\mathrm{K}(\cdot,\cdot)$ is defined in 6.3.1. They prove that the expected value of $U^{(i,j,t})$ is equal to

$$\mathrm{var}_p\left(\frac{g_t(\theta)}{p(\theta \mid Y)}\right) + 1 \ ,$$

where g_t is the current approximation to the posterior. Hence, each $U^{(i,j,t)}$ is an unbiased estimate of the distance between $g_t(\theta)$ and $p(\theta \mid Y)$. In their paper, Liu and Liu (1993) plot the average of the logarithm of $U^{(i,j,t)}$ (over the $m(m-1)$ pairs of chains) vs. t. As $g_t(\theta)$ convergences to $p(\theta \mid Y)$, the resulting curve should fluctuate about a line parallel to the x-axis.

6.3.3. Alternative Methods

Other approaches for assessing convergence have been proposed. Gelman and Rubin (1992) consider the ratio of between-chain (across m chains) variation to within-chain (up to iteration n) variation. If this ratio is high, then there may be a reason to believe that further iteration will be of value. Formally, they compute:

$$\sqrt{\hat{\mathrm{R}}} = \sqrt{\left(\frac{n-1}{n} + \frac{(m+1)}{mn}\frac{B}{W}\right)\frac{df}{df-2}}$$

where

$B/n =$ the variance between the m chain means, $\bar{x}_i.$, each based on n values of x, $B/n = \Sigma_{i=1}^{m}(\bar{x}_{i.} - \bar{x}_{..})^2/(m-1)$,

$W =$ the average of the m within-chain variances, s_i^2, each based on $n-1$ degrees of freedom, i.e. $W = \Sigma_{i=1}^{m} s_i^2/m$, and

$$df = 2\,\hat{V}^2/\widehat{\mathrm{var}}(\hat{V})$$

where

$$\hat{V} = \frac{n-1}{n} W + \left(\frac{m+1}{\mathrm{mn}}\right) B$$

and

$$\widehat{\mathrm{var}}(\hat{V})\left(\frac{n-1}{n}\right)^2 \frac{1}{m}\widehat{\mathrm{var}}(s_i^2) + \left(\frac{m+1}{mn}\right)^2 \frac{2}{m-1} B^2 + 2\frac{(m-1)(n-1)}{mn^2} \cdot \frac{n}{m}$$
$$\left[\widehat{\mathrm{cov}}(s_i^2, \bar{x}_{i.}^2) - 2\bar{x}_{..}\,\widehat{\mathrm{cov}}(s_i^2, \bar{x}_{i.})\right]$$

and the estimated variances and covariances are obtained from the m sample values of $\bar{x}_{i.}$ and s_i^2. When $\hat{\mathrm{R}} \approx 1$ for all scalar estimands of interest Gelman and Rubin (1992) conclude that one can terminate iteration.

The idea of comparing between-chain variation to within-chain variation was presented in Tanner and Wong (1987) in the context of data augmentation. This approach may not work well if the chains are trapped within the same subregion. Applications of this methodology are given in Gelman and Rubin (1992).

Raftery and Lewis (1992) discuss the problem how many times to iterate the Gibbs sampler. In particular, suppose one wishes to estimate a particular quartile of the posterior distribution $p(U \le \mu | Y)$ to within $\pm r$ with probability s, where U is a function of the parameter vector θ. Raftery and Lewis (1992) present formulas (for the single chain case) for the number of initial M iterations that are to be discarded, and for a further N iterations of which every kth is stored.

Suppose one calculates U_t for each iteration t and then forms $Z_t = \delta(U_t \le \mu)$, where $\delta(\cdot)$ *is the indicator function.* $\{Z_t\}$ is a binary 0–1 process. Consider the new process $\{Z_t^{(k)}\}$, where $Z_t^{(k)} = Z_{1+(t-1)k}$ and let

$$\begin{pmatrix} 1-\alpha, & \alpha \\ \beta, & 1-\beta \end{pmatrix}$$

be the transition matrix for $(Z_t^{(k)})$. The number of "burn-in" iterations to be discarded is equal to mk, where

$$m = \frac{\log\left(\dfrac{\varepsilon(\alpha+\beta)}{\max(\alpha, \beta)}\right)}{\log(1-\alpha-\beta)}$$

and ε is the required distance between $p(z_m^{(k)} = i | z_0{}^{(k)} = j)$ and the equilibrium probability. The formula for N is kn, where

$$n = \frac{\alpha\beta(2-\alpha-\beta)}{(\alpha+\beta)^3}\left[\frac{\Phi\left(\dfrac{1}{2}(1+s)\right)}{r}\right]^2$$

and $\Phi(\cdot)$ is the standard normal cdf. To implement this idea one needs an initial run to estimate α and β. A poor estimate of α and β may result if the chain is trapped in a subregion of the parameter space.

6.4. The Griddy Gibbs Sampler

In many practical situations the distribution $p(\theta_i | \theta_1, \ldots \theta_{i-1}, \theta_{i+1}, \ldots \theta_d, Y)$ is univariate. When it is difficult to directly sample from $p(\theta_i | \theta_1, \ldots \theta_{i-1}, \theta_{i+1}, \ldots \theta_d, Y)$, one may form a simple approximation to the inverse cdf based on the evaluation of $p(\theta_i | \theta_1, \ldots \theta_{i-1}, \theta_{i+1}, \ldots \theta_d, Y)$ on a grid of points. Ritter and Tanner (1992) consider the following algorithm:

Step 1. Evaluate $p(\theta_i|\theta_1, \ldots \theta_{i-1}, \theta_{i+1}, \ldots \theta_d, Y)$ at $\theta_i = \phi_1, \phi_2, \ldots, \phi_n$, to obtain $w_1, w_2, \ldots, w_n$.

Step 2. Use $w_1, w_2, \ldots, w_n$ to obtain an approximation to the inverse cdf of $p(\theta_i|\theta_1, \ldots \theta_{i-1}, \theta_{i+1}, \ldots \theta_d, Y)$.

Step 3. Sample a uniform (0,1) deviate and transform the observation via the approximate inverse cdf.

Remark 6.4.1. The function $p(\theta_i|\theta_1, \ldots \theta_{i-1}, \theta_{i+1}, \ldots \theta_d, Y)$ need only be known up to a proportionality constant, since the normalization can be obtained from the $w_1, w_2, \ldots, w_n$ directly.

Remark 6.4.2. The grid points need not be uniformly spaced. In fact, good grids put more points in neighborhoods of high mass and fewer points in neighborhoods of low mass. One approach to address this goal is to construct the grid so that the mass under the current approximation to the marginal distribution between successive grid points is approximately constant. This point is discussed in more detail below.

Remark 6.4.3. The number of points in the grid need not be constant over the iterations of the Gibbs sampler. At early iterations, n may be small. As the algorithm iterates toward the joint distribution, n can be increased. In practice, when the algorithm appears to have stabilized with $n = n^*$ grid points, at the next iteration one may use a finer grid e.g. $n = 2n^*$ or $4n^*$. The Gibbs stopper may also be used to regulate n. See Section 6.4.3.

Remark 6.4.4. Simple approximations to the inverse cdf are:

a. Piecewise constant corresponding to a discrete distribution for $\phi_1, \ldots, \phi_n$ with probabilities $p(\phi_i) = w_i / \sum_{j=1}^{n} w_j$.

b. Piecewise linear corresponding to a piecewise uniform distribution on the interval $[a_i, a_{i+1}]$, $i = 1, \ldots, n$, with $\phi_i \in [a_i, a_{i+1}]$ and density $f_i = w_i / \sum_{j=1}^{n} \omega_j$, where $\omega_i = w_i(a_{i+1} - a_i)$. Typically, ϕ_i is centered in the interval $[a_i, a_{i+1}]$.

More sophisticated approximations may be based on piecewise quadratic interpolation or higher-order splines. In general, when the conditional marginal is easy to evaluate, one may wish to use a simpler approximation to the inverse cdf and use a finer grid. When the conditional marginal is more difficult to evaluate, one may wish to use a coarser grid, but a more clever approximation.

Remark 6.4.5. Over an unbounded interval it is important to compare w_1 and w_n to $M = \max(w_1, \ldots, w_n)$. If w_1 is greater than $f \cdot M (0 < f < 1)$, then the grid must be augmented with points to the left of ϕ_1. If w_n is greater than $f \cdot M$, then the grid must be augmented with points to the right of ϕ_n. In practice, it is sufficient to force the grid to grow if at either end of the grid the density is more than 10% of the maximum value. If, however, interest centers on the tails of the distribution, then f should be smaller than 0.1. We will refer to this device as the *grid-grower*.

Remark 6.4.6. Several alternative methods are available for handling the situation where it is difficult to directly sample from the conditionals. These methods tend to rely on importance sampling or acceptance/rejection approaches. Wei and Tanner (1990c) presented an importance sampling approach see Section 5.9.1. Carlin and Gelfand (1991) and Zeger and Karim (1991) presented rejection/acceptance modifications to the Gibbs sampler. (See Section 6.2.3 for an approach of Zeger and Karim). Also of note is the work by Gilks and Wild (1992), who used tangent and secant approximations above and below the log-posterior to develop an acceptance/rejection scheme. Acceptance/rejection methods are exact in the sense that they produce samples from the required distribution. Possible drawbacks of these methods include a low acceptance rate, the need to know the normalizing constant of the conditional distribution, and possible restrictions on the distribution (e.g. log-concavity). As illustrated in Section 6.4.3, nonlinear regression problems may lead to conditionals that are not log-concave and may in fact be multimodal. Moreover, in nonlinear regression the conditionals, typically, are known only up to a multiplicative constant. Both importance sampling and rejection/acceptance algorithms require a higher degree of programming sophistication on the part of the data analyst and thereby detract from the conceptual simplicity and appeal of the Gibbs sampler. Specification of an importance sampling function that is easy to sample from, yet provides a "good match" to the density of interest, may require the specification of several tuning constants. In the context of rejection/acceptance, specification of tuning constants is required to ensure that the importance function, modulo a multiplicative constant, dominates the density of interest *everywhere* (or at least over a region of high content; e.g. 95%).

The griddy Gibbs sampler preserves the conceptual and implementational simplicity of the Gibbs sampler. Whereas acceptance/rejection algorithms may require elaborate code, the griddy Gibbs sampler in its simplest form can be implemented in 30 to 50 lines of Fortran code, not including the subroutine that computes the posterior. Enhancements to this simple form (e.g. grid growing) can be added in a highly modular form. This method is sufficiently flexible to allow the analyst freedom to vary the specification of the model without the need to limit attention to log-concave or unimodal posteriors. An interesting alternative approach based on the Metropolis algorithm is discussed in Section 6.5.3.

6.4.1. Example

For the genetic linkage model with data (14, 0, 1, 5), the distribution of the latent data (x_2) given θ and $Y, p(x_2|\theta, Y)$, is the binomial distribution with parameters 14 and $\theta/(\theta + 2)$. The augmented posterior, $p(\theta|x_2, Y)$, is (under a flat prior) proportional to

$$\theta^{x_2+x_5}(1 - \theta)^{x_3+x_4} \ . \tag{6.4.1}$$

To implement the Gibbs sampler, one would draw x_2^* from $p(x_2|\theta^*, Y)$, where θ^* is the current value for θ. The next step would be to sample a new θ^* from $p(\theta|x_2^*, Y)$. These two steps would then be iterated. Independent replicates of this path could be constructed to realize an independent sample of θ values from $p(\theta|Y)$. For the purpose of this example, the fact that the augmented posterior is a beta distribution is ignored. We only assume that one can calculate (6.4.1) at any value of θ. Let m denote the number of such paths.

The griddy Gibbs sampler based on a linear approximation of the inverse cdf was initially implemented with a five-point grid on [0, 1]. The algorithm

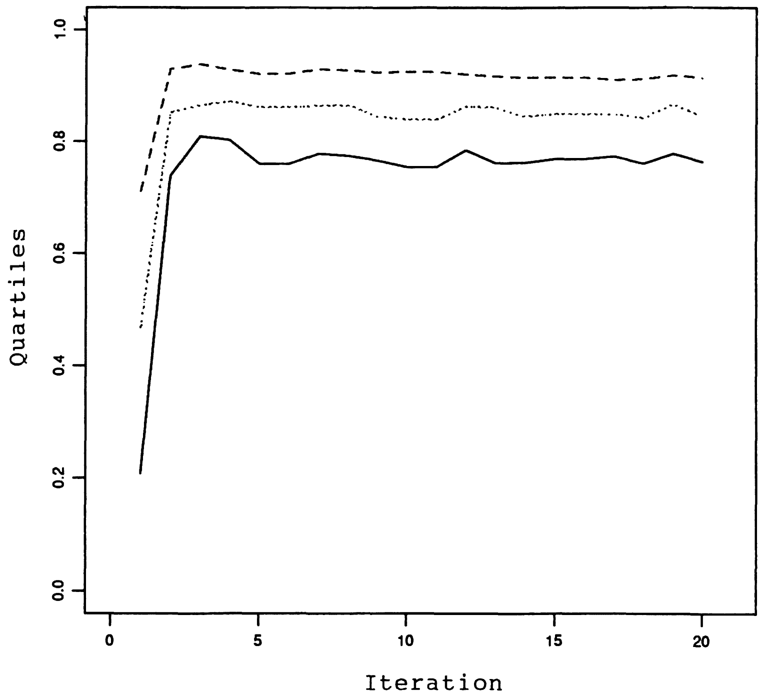


Figure 6.10. Sample quartiles vs. iteration. n = 5, 10, 20, 40 for 5 iterations each.

was allowed to run for five iterations, at which point the number of grid points was increased to ten. The algorithm continued for five more iterations, at which point the number of grid points was increased to 20. The final five iterations used 40 grid points.

Figure 6.10 summarizes the history of this process by plotting the upper, middle and lower quartile for θ vs. iteration number. (Three hundred independent griddy Gibbs sampler paths were created to obtain the independent θ values.) As can be seen from Figure 6.10 the median appears to have

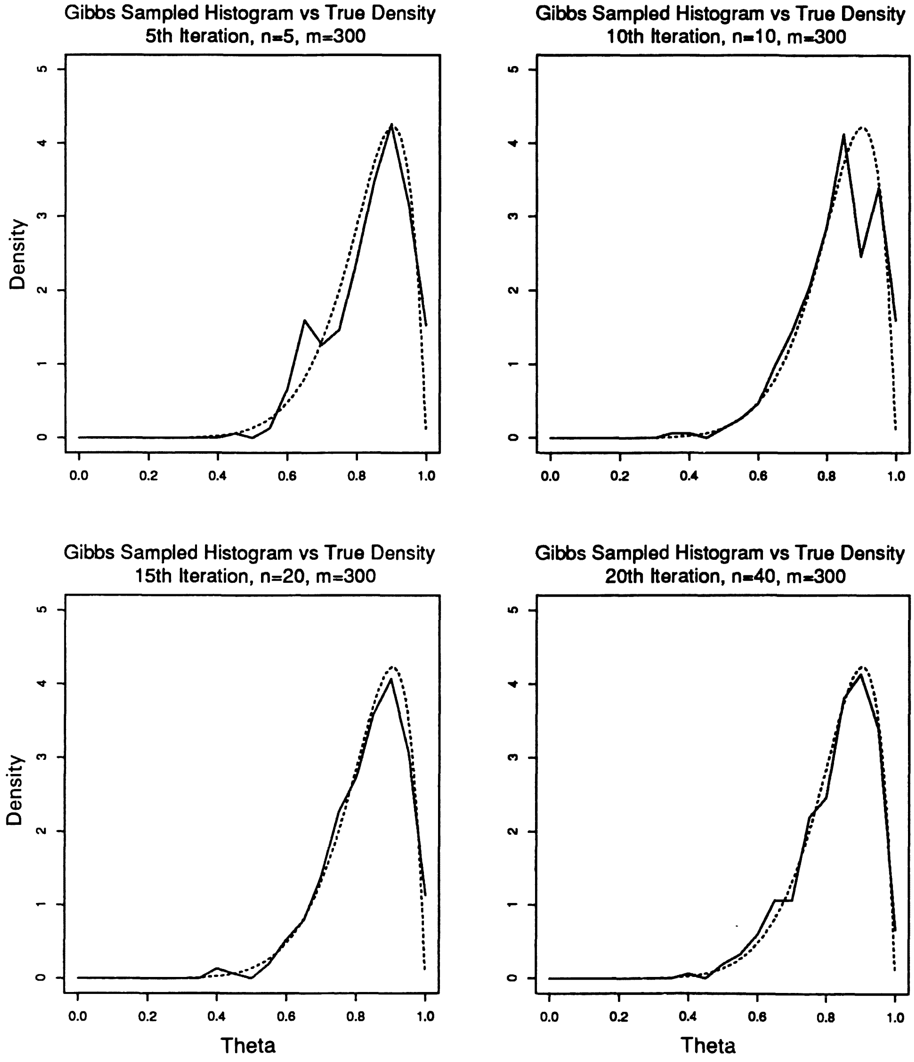


Figure 6.11. Gibbs sampler histogram and true density for iterations 5, 10, 15 and 20.

stabilized quite quickly with a small number of grid points. However, the lower tail does require further iteration and a finer mesh to stabilize.

As can be seen in Figure 6.11, the histograms of the θ values after 15 and 20 iterations (with $n = 20$ and 40, respectively) are quite similar and both are congruent with the true posterior. Thus, a grid with 20 equally spaced points appears to capture the shape of $p(\theta|x_2, Y)$.

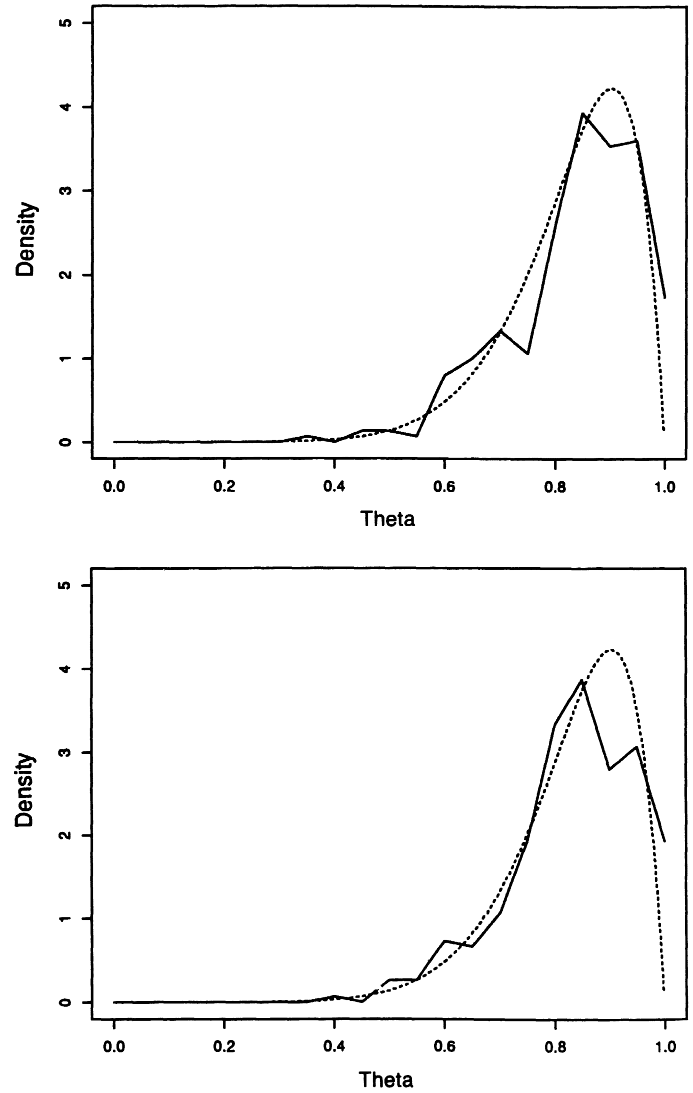


Figure 6.12. Gibbs sampler histogram and true density for $n = 5$ and $n = 10$.

Figure 6.12 presents two density plots after 30 iterations using a five-point and a ten-point grid, respectively, for all iterations. As one might expect, an increase in iterations with too coarse a grid does not improve the density estimate.

6.4.2. Adaptive Grid/Grid Grower

In the previous section, a grid with uniform gap size was used at each iteration. Clearly, a grid which has more points in a region of high mass and fewer points in a region of less mass is preferable.

One approach to address this goal is to make use of the empirical cdf ($\hat{F}$) of the X_i values (obtained from the m replicates of the griddy Gibbs path) from the current iteration. The new n grid points $\phi_1, \ldots, \phi_n$ can be obtained by computing the empirical quantiles $\hat{F}^{-1}(\xi_i)$ for preselected values $0 < \xi_1 < \cdots < \xi_n < 1$. This grid is referred to as the *adaptive grid.*

Remark 6.4.7. Usually, the ξ_i are chosen to be equidistant. This would form a grid which is dense in regions of high mass and sparse in regions of low mass.

Remark 6.4.8. It is not efficient to modify the grid at each iteration. Rather, one can branch to this adaptive-grid method every j iterations. In the example below, $j = 5$.

Remark 6.4.9. One may use the adaptive grid idea to increase the number of points, as well as to specify where to place the points, i.e. n may be some multiple (other than one) of the number of grid points in the previous iteration.

Remark 6.4.10. The adaptive grid may be used in conjunction with the grid-grower introduced above.

1. Every i_0 iterations compute $\hat{F}$ and derive a grid of size n from it, as described above. This grid is called the *global grid.*
2. For the first of the d conditional distributions, augment the global grid, if necessary, using the grid-grower. The corresponding grid will be called the *local grid.* Use this local grid for the next d conditional distribution and augment, if necessary, using the grid-grower.

If the global grid is far from the hump of the specific marginal conditional distribution, then the grid-grower will correct the problem at the expense of function evaluations. Having captured the important part of the density, the grid-grower would tend not to be used.

To illustrate the adaptive grid, consider the genetic linkage example. In this case, we let the griddy Gibbs sampler run for ten iterations, with $n = 10$ at

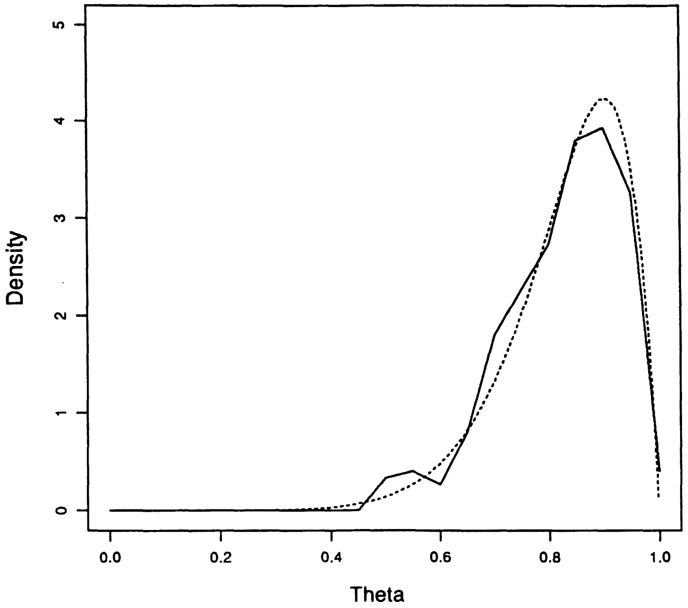


Figure 6.13. Gibbs sampled histogram vs. true density. 10th iteration, $n = 10$, $m = 300$, adaptive grid.

each iteration. The adaptive-grid routine was called every five iterations. As in the earlier discussion of this example, a linear approximation to the empirical inverse cdf of the conditional was used. Every five iterations, the grid was adapted to the current marginal cdf of θ. We chose $\xi_i = 1/2n + (i - 1)/n$.

Figure 6.13 presents the density plot of the 300 θ values at the final iteration, along with the true posterior distribution. A comparison of the bottom panel of the previous figure with Figure 6.13 reveals an improved plot with fewer marginal conditional function calls required.

To illustrate the grid-grower in combination with the adaptive grid, Figure 6.14 is presented. This figure presents the density of 500 θ values after 20 iterations having started with a grid on [0.0, 0.2]. Note that the correct mode is near 0.9. The adaptive-grid routine was called every two iterations. Within two iterations the grid-grower captured the general shape of the density and moved the grid to [0.5, 1.0].

6.4.3. Nonlinear Regression

As a second example, we consider the biochemical oxygen demand (BOD) data from Marske (1967) which was analyzed by Bates and Watts (1988). To determine the biochemical oxygen demand, a sample of stream water was

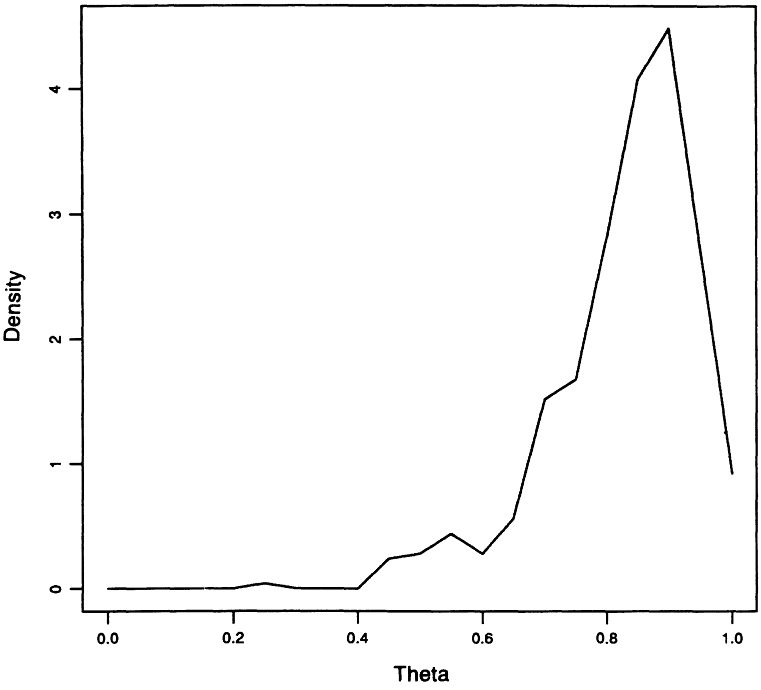


Figure 6.14. 500 θ Values, 20 Iterations.

taken, injected with soluble organic matter, inorganic nutrients and dissolved oxygen, and subdivided into BOD bottles. Each bottle was inoculated with a mixed culture of microorganisms, sealed and incubated at constant temperature. The bottles were opened periodically and analyzed for dissolved oxygen concentration from which the BOD was calculated in milligrams per liter. The values shown in Table 6.3 are the averages of two analyses on each bottle.

Table 6.3. Biochemical Oxygen Demand vs. Time.

Time (days)	BOD (mg/I)
1	8.3
2	10.3
3	19.0
4	16.0
5	15.6
7	19.8

Source: Marske, 1967

A nonlinear model was derived based on exponential decay with a fixed rate constant (Bates and Watts, 1988) as:

$$y_i = \theta_1[1 - \exp(-\theta_2 x_i)] + \varepsilon_i$$

where y is the biochemical oxygen demand at time x. For independent normal errors with constant variance σ^2, the likelihood for the BOD problem is:

$$L(\theta_1,\theta_2,\sigma^2|Y) = \exp\left[-n\log\sigma - \frac{1}{2}\frac{S(\theta_1,\theta_2)}{\sigma^2}\right]$$

where $S(\theta_1,\theta_2)$ is the sum of squared residuals.

Assuming a flat prior and integrating out σ^2 results in the following (improper) posterior:

$$p(\theta_1,\theta_2|Y) \propto [S(\theta_1,\theta_2)]^{-[n/2-1]} \ .$$

Figure 6.15 presents the likelihood contours. Contours above the approximate 93% F level are open with respect to θ_2. Note that in this region the model is insensitive to changes in θ_2. Moreover, for values of θ_2 less than 0, θ_1 is negative, thus realizes a fold in the 99.9% contour.

For the purpose of the present analysis, a flat (strictly positive) prior for (θ_1,θ_2) is adopted over the region $[-20,50] \times [-2,6]$. Outside this region the prior for (θ_1,θ_2) is zero. This amounts to using the restriction that $p(\theta_1,\theta_2|Y)$ be normalised on the domain $[-20,50] \times [-2,6]$ to integrate to unity. Under this restriction the posterior is bimodal, with the main mode at (19,0.5) and a minor mode at the lower-left corner.

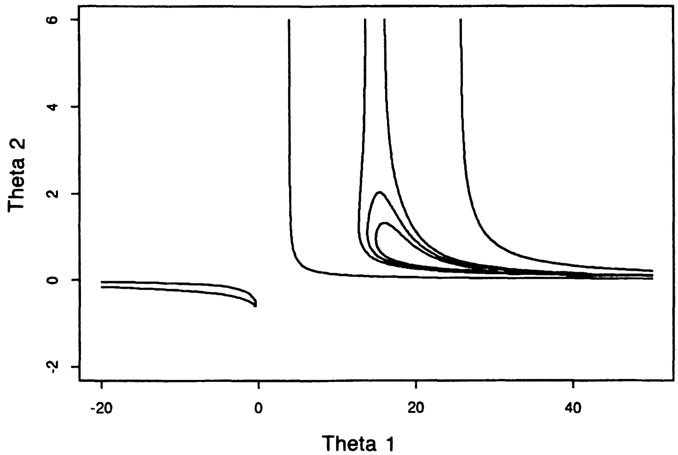


Figure 6.15. 80%, 90%, 95% and 99.9%. Normal likelihood contours for θ_1, and θ_2 labelled by approximate frequency coverage using the F statistic.

We illustrate in Figure 6.16 the griddy Gibbs sampler with adaptive grids. Grids of 20 points were used spanning the ranges [– 20,50] and [– 2,6]. The algorithm was initialized with a uniform sample on the domain and 500 paths of griddy Gibbs sampler were run. Figure 6.16 presents the scatter plot of (θ_1,θ_2) for selected iterations as well as the quartiles of θ_1 and θ_2 as functions of the number of iterations. Note that the quartiles appear to stabilize quite early.

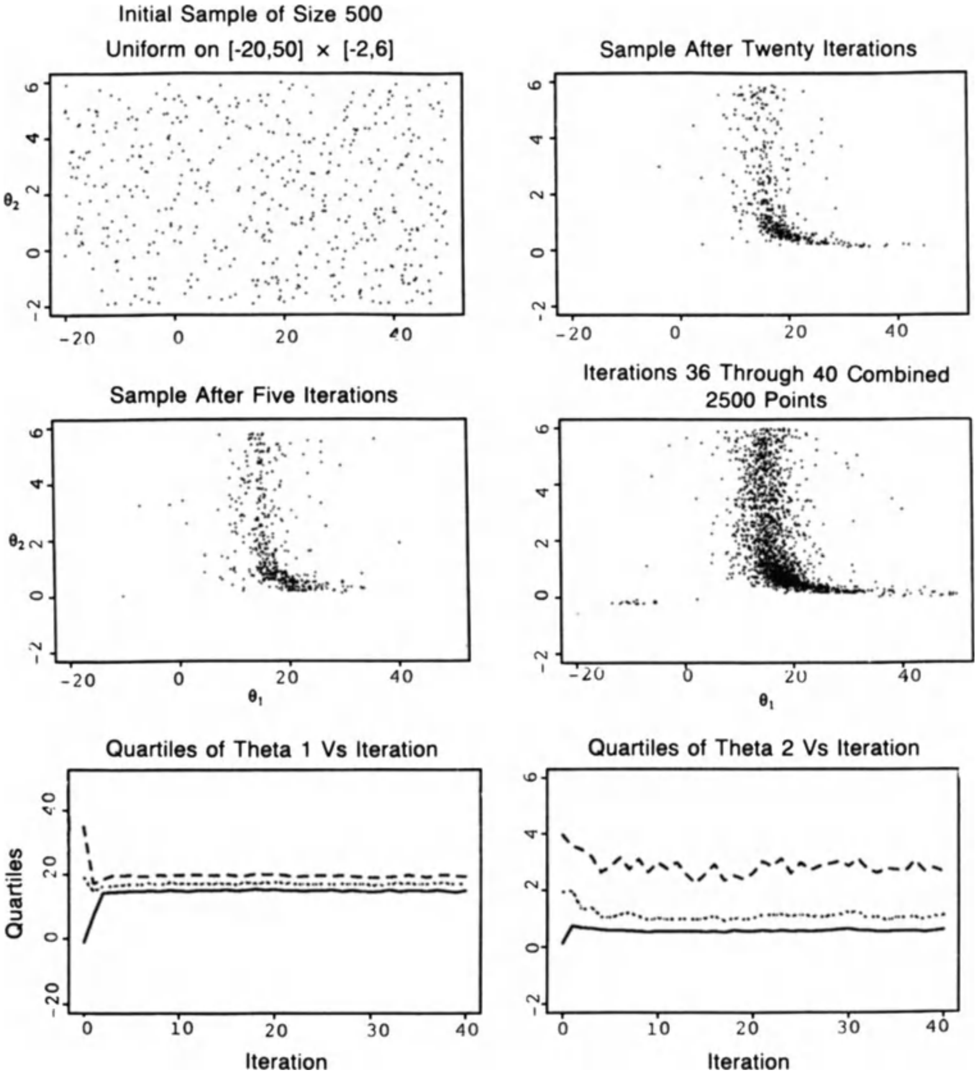


Figure 6.16. BOD example, flexible grid with 20 points.

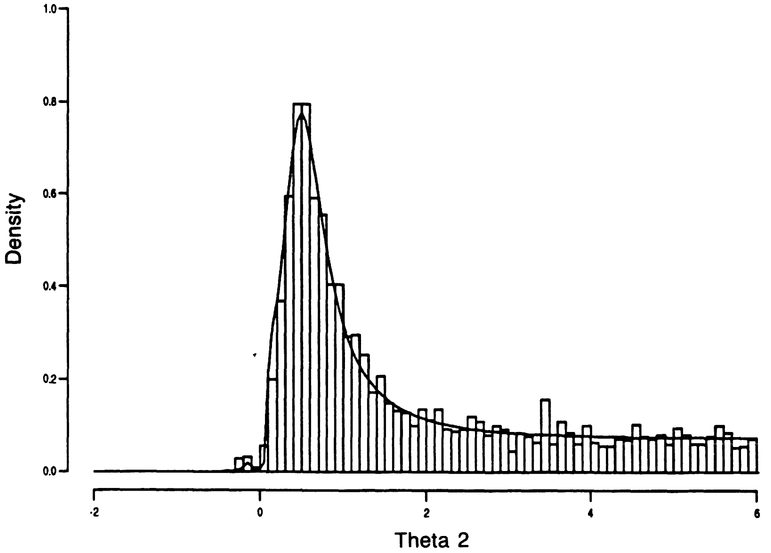


Figure 6.17. θ_2 Marginal. Solid: Integrated; Bars: Gibbs sample.

Figure 6.17 compares the θ_2 marginals obtained (1) by numerical integration and (2) by forming the histogram of the θ_2 values of iterations 36–40 combined. From this plot, it can be seen that the griddy Gibbs sampler has recovered this marginal.

Finally, Figure 6.18 presents the Gibbs stopper weights for iterations 1, 5, 35 and 40. As can be seen from these plots, a major component of the decrease in variation occurred by the fifth iteration. Beyond the fifth iteration, one may wish to increase the number of points in the grid or the number of chains to improve the precision of the estimate. Had all four distributions been similar in shape to the top plot, one would tune the number of points in the grid or the number of chains to push the approximation toward the true joint distribution.

6.4.4. Cox Model

As a third example, we consider data taken from Glasser (1965). The response is the survival time in days for subjects with primary lung tumors. Two covariates are considered: the age of the patient and the performance status. Of the 16 cases, six are right-censored. We model these data using the Cox

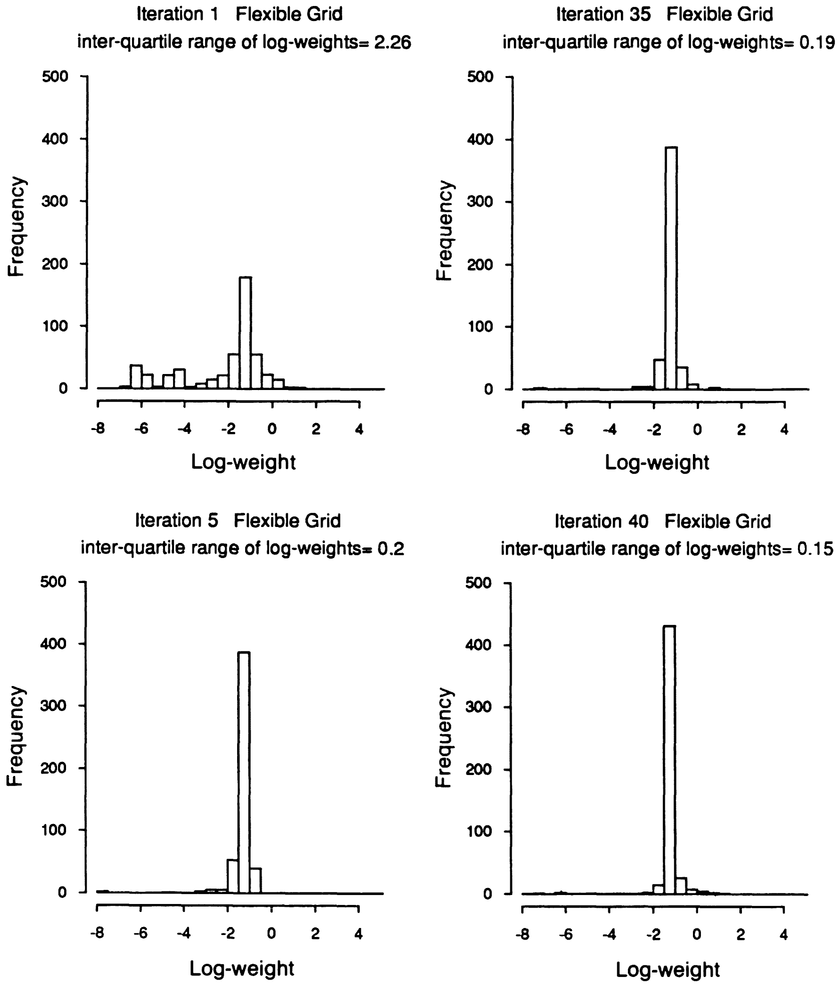


Figure 6.18. Stopper weights for flexible grid with $n = 20$.

proportional hazards model (Lawless, 1982, Chapter 7). The partial likelihood is used as the basis of inference for this model. For the Glasser data, the mode of the partial likelihood is located at -0.217 (performance) and at -0.005 (age), with corresponding estimated standard errors of 0.162 and 0.042, respectively. The results after 20 iterations of the griddy Gibbs sampler (applied to the partial likelihood with a flat prior) are quite similar. The average parameter value based on 300 independent chains is -0.208 for performance and -0.005 for age, and the associated standard deviations are 0.162 and 0.043, respectively. The estimated survivor function corresponding to a performance status of 7 and an age of 51, at 141 days, is 0.625 and 0.641, respectively, for the modal partial likelihood and the griddy Gibbs approaches. The estimated standard error based on the griddy Gibbs approach is 0.085. The griddy Gibbs point estimate for the survivor function (at $t = 141$) was obtained by averaging the survivor function $S(t|\mathbf{x}) = S_0(t)^{\exp(\mathbf{x}\boldsymbol{\beta})}$ over the 300 independent values of the parameter vector ($\boldsymbol{\beta}$), where $\mathbf{x}$ is the known covariate vector. (The dependency of the baseline survivor function ($S_0(t)$) on the parameter vector is given by Lawless (1982, page 361).) The associated estimated standard error was obtained by computing the standard deviation of these 300 survivor function values. Clayton (1991) and Sinha (1993) apply data augmentation and Gibbs sampler algorithms to the related frailty models.

6.5 The Metropolis Algorithm

The Metropolis algorithm (Metropolis et al. 1953) was developed to investigate the equilibrium properties of large systems of particles such as electrons in an atom. This algorithm has been extensively used in the statistical physics literature (Hammersley and Handscomb, 1964) to simulate complex systems. Hastings (1970) suggests a generalization of the Metropolis algorithm. He illustrates how to use this algorithm to simulate Poisson and normal deviates, as well as random orthogonal matrices. In this section, we show how one can use the Metropolis algorithm to construct a Markov chain with equilibrium distribution $\pi(x)$. However, before presenting the algorithm, we briefly review some elementary ideas related to discrete-space Markov chains.

6.5.1. Elements of Discrete-space Markov Chains

Assume we have a set of states, $S = \{s_1, s_2, \ldots, s_d\}$. At time 0, the process begins in one of these states, the process moves from state s_i to state s_j at unit time intervals. For a Markov chain, the status of the chain at time $n + 1$ depends only on its status at time n. Hence, the transition probability p_{ij} that the state moves from s_i to s_j may be organized in a $d \times d$ matrix, where $p_{ij} \geqslant 0$ and $\sum\limits_j p_{ij} = 1$ for all rows i.

As an illustration, we consider an example from Snell (1988). In the Land of Oz there are three kinds of weather: rain, nice and snow. The transition probabilities are:

$$P = \begin{array}{c} \\ R \\ N \\ S \end{array} \left(\begin{array}{ccc} R & N & S \\ 0.5 & 0.25 & 0.25 \\ 0.5 & 0 & 0.5 \\ 0.25 & 0.25 & 0.5 \end{array} \right).$$

Suppose on day 0 the weather is nice (N) and let, for example, $p(X_i = S)$ be the probability of sunny weather on day i. The probabilities of the three states (R, N, S) on day 1 would then follow from the transition matrix to be

$$(p(X_1 = R), p(X_1 = N), p(X_1 = S)) = (w_1^{(1)}, w_2^{(1)}, w_3^{(1)}) = w^{(1)} = (0.5, 0, 0.5) \ .$$

What are the corresponding probabilities for day 2? In general, we compute the probabilities for day $n + 1$ by noting that

$$\begin{aligned} w_j^{(n+1)} &= p(X_{n+1} = s_j) \\ &= \sum_k p(X_n = s_k, X_{n+1} = s_j) \\ &= \sum_k p(X_n = s_k) p(X_{n+1} = s_j \mid X_n = s_k) \\ &= \sum_k w_k^{(n)} p_{kj} \ . \end{aligned}$$

In matrix notation, we have $w^{(n+1)} = w^{(n)} P = w^{(0)} P^n$. In this way, starting from a nice day, i.e. $p(X_0 = N) = 1$, we have:

$$\begin{aligned} w^{(1)} &= (0.5, 0, 0.5) \\ w^{(2)} &= (0.375, 0.25, 0.375) \\ w^{(3)} &= (0.406, 0.188, 0.406) \\ w^{(4)} &= (0.398, 0.203, 0.398) \\ w^{(5)} &= (0.4, 0.2, 0.4) \\ \mathrm{w}^{(6)} &= (0.4, 0.2, 0.4) \ . \end{aligned}$$

Similarly, if $p(X_0 = R) = 1$ (i.e. we started on a rainy day) or $p(X_0 = S) = 1$ (i.e. day 0 was snowy) we would have $w^{(6)} = (0.4, 0.2, 0.4)$. Thus, if we were to simulate the behavior of this chain, we would find in the long run that 40% of the days would be in state R, 20% in state N and 40% in state S, irrespective of the starting point of the chain. The vector $(0.4, 0.2, 0.4)$ is called the *equilibrium distribution.*

More generally, a Markov chain with transition matrix P will have an equilibrium distribution π iff $\pi = \pi P$. This chain is a *reversible chain* iff $\pi_i p_{ij} = \pi_j p_{ji}$ for all $i \neq j$. Note that $\pi_i p_{ij} = \pi_j p_{ji}$ implies $\pi = \pi P$, since

$$\pi p_j = \sum_i \pi_i p_{ij} = \sum_i \pi_j p_{ji} = \pi_j \sum_i p_{ji} = \pi_j \quad .$$

In this way, to sample from the equilibrium distribution π, we run a Markov chain with transition matrix P satisfying $\pi_i p_{ij} = \pi_j p_{ji}$ until the chain appears to have settled down to equilibrium. One of the contributions of Metropolis et al. (1953) is in detailing a general way of constructing P, to realize a chain with equilibrium distribution π.

6.5.2. Metropolis' Method

We first present the idea in the discrete case. The continuous case is taken up later. Let $Q = \{q_{ij}\}$ be a specified symmetric transition matrix. At a given step, randomly draw state s_j from the ith row of Q. With known probability α_{ij} move from s_i to s_j, otherwise remain at step s_i. This construction defines a Markov chain with transition matrix $p_{ij} = \alpha_{ij} q_{ij}$ $(i \neq j)$ and $p_{ii} = 1 - \Sigma_{j \neq i} p_{ij}$. Following Metropolis et al. (1953), we let

$$\alpha_{ij} = \begin{cases} 1 & \text{if } \pi_j/\pi_i \geqslant 1 \\ \pi_j/\pi_i & \text{if } \pi_j/\pi_i \leqslant 1 \end{cases} .$$

This chain is reversible, since

$$\begin{aligned} \pi_i p_{ij} &= \pi_i \min\left\{1, \frac{\pi_j}{\pi_i}\right\} q_{ij} \\ &= \min\{\pi_i, \pi_j\} q_{ij} \\ &= \min\{\pi_i, \pi_j\} q_{ji} \\ &= \pi_j p_{ji} . \end{aligned}$$

The equilibrium distribution will be unique if Q is irreducible (Ripley, 1987). A sufficient condition for convergence (if π is not constant) is being able to move from any state to any other under Q (Ripley, 1987).

Barker (1965) takes $\alpha_{ij} = \pi_j/(\pi_i + \pi_j)$. Note that the resulting chain is reversible, since $\pi_i p_{ij} = \pi_i \pi_j/(\pi_i + \pi_j) q_{ij} = \pi_j p_{ji}$.

In the continuous case, where π is a density with respect to a measure μ and $f(x, y)$ is a symmetric transition probability function, [i.e. $f(x, y) = f(y, x)$], then the Metropolis algorithm is given by:

a. If the chain is currently at $X_n = x$, then generate a candidate value y^* for next location X_{n+1} from $f(x, y)$.
b. With probability

$$\alpha(x, y^*) = \min\left\{\frac{\pi(y^*)}{\pi(x)}, 1\right\}$$

accept the candidate value and move the chain to $X_{n+1} = y^*$. Otherwise, reject and let $X_{n+1} = x$. Thus, the Metropolis algorithm yields a series of *dependent* realizations forming a Markov chain with π as its equilibrium distribution. A key observation is that the Metropolis algorithm only requires that π be defined up to the normalizing constant, since the constant drops out in the ratio $\pi(y^*)/\pi(x)$.

Tierney (1991) presents a number of suggestions for $f(x, y)$. If $f(x, y) = f(y - x)$, then the chain is driven by a random-walk process. Possible candidates for f are the multivariate normal, multivariate t or split

t (with a small degrees of freedom). In situations where the multivariate normal or multivariate t is used to generate candidate values, one would center the normal or the t at the current state of the chain x, with the variance–covariance matrix possibly equal to some multiple of the inverse information at the posterior mode. Müller (1993) discusses the choice of scale issue in detail. Besides presenting a range of hybrid strategies by cycling/mixing different chains, Tierney (1991) presents formal conditions for convergence, rates of convergence and limiting behavior of averages. Gelfand (1992) presents an analogue to the Gibbs stopper for the Metropolis algorithm.

A generalization of the Metropolis algorithm due to Hastings (1970) takes

$$\alpha(x\ y) = \begin{cases} \min\left\{\dfrac{\pi(y)q(y,x)}{\pi(x)q(x,y)}, 1\right\} & \text{if } \pi(x)q(x,y) > 0 \\ 1 & \text{if } \pi(x)q(x,y) = 0 \end{cases}$$

where $q(x, y)$ is an *arbitrary* transition probability function. Note that if q is symmetric, i.e. $q(x, y) = q(y, x)$, as would be the case in using a multivariate normal or multivariate t to drive the algorithm, then the Hastings algorithm reduces to the Metropolis algorithm. Hastings (1970) considers the case where $q(x, y) = q(y)$, which is closely related to importance sampling. Tierney (1991) calls these *independence chains.*

EXAMPLE. *Genetic Linkage (Continued)*

We illustrate the Metropolis algorithm with the very simple genetic linkage model presented in Section 4.1. In the notation of this section we have:

$$\pi(\theta) = (2+\theta)^{125}(1-\theta)^{38}\theta^{34} \propto p(\theta \mid Y)\ .$$

Note that the normalizing constant for $p(\theta \mid Y)$ is not required for implementing the Metropolis algorithm. Also note that the Metropolis algorithm works with the observed posterior, rather than with $p(\theta \mid Z, Y)$ and with $p(Z \mid \theta, Y)$ – as would be the case with data augmentation (Gibbs sampler).

For this example, candidate values were generated from the uniform distribution on (0, 1) i.e. $f(x, y) = 1$ on the interval (0, 1). Use of the normal or t distribution to drive the algorithm would lead to candidate values outside the support of $\pi(\theta)$ – various fixes could be developed for this problem, though. One thousand independent chains were generated and each chain was iterated 1000 times. Figure 6.19 presents the true distribution along with the estimated distribution (based on a histogram of the final value from each of the 1000 chains). One thousand iterations was probably not necessary in this case. However, given the simplicity of the problem the computation was not burdensome. Figure 6.20 presents the results for the observed data set (14, 0, 1, 5), yielding the posterior proportional to $(2+\theta)^{14}(1-\theta)\theta^5$.

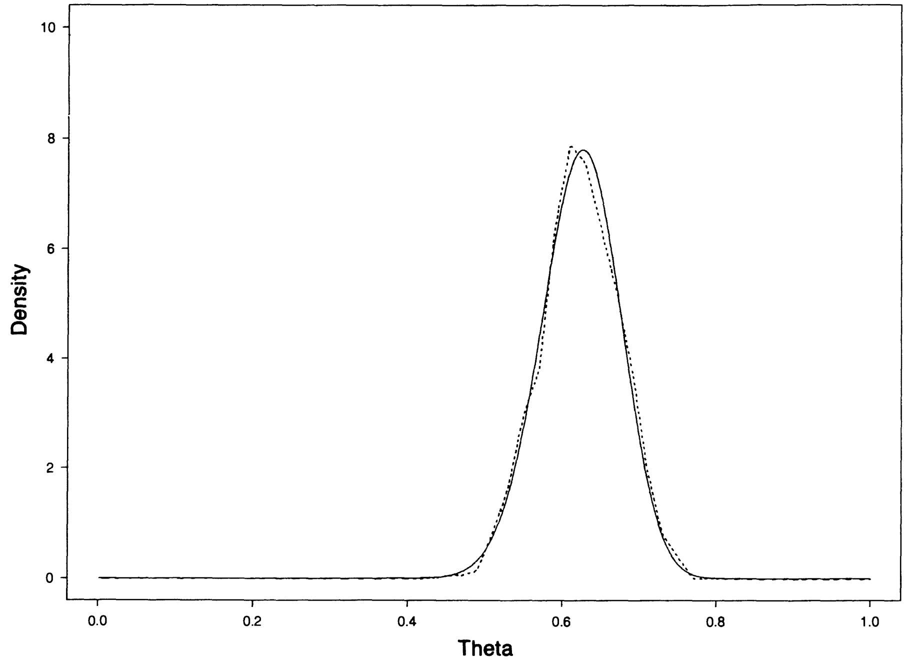


Figure 6.19. Posterior of θ solid: true; dashed: Metropolis.

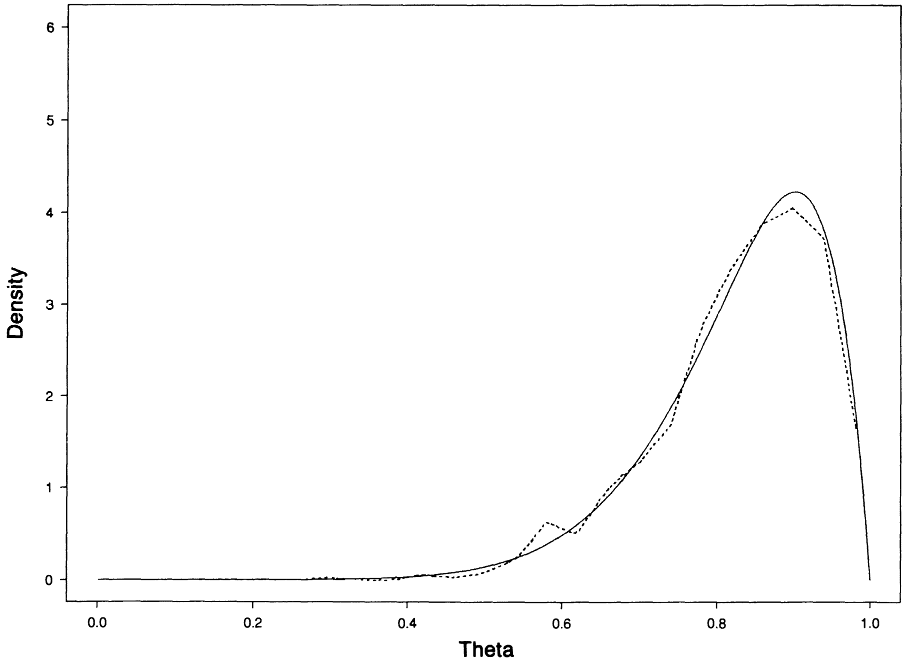


Figure 6.20. Posterior of θ solid: true; dashed: Metropolis.

6.5.3. Metropolis Subchains

Müller (1993) suggests embedding a Metropolis random-walk subchain in the Gibbs sampler in the situation where it is difficult to sample directly from a conditional distribution, e.g. $p(\theta_1 | \theta_2, \ldots, \theta_d, Y)$. In particular, given the values from iteration i, $\theta_2^{(i)}, \theta_3^{(i)}, \ldots, \theta_d^{(i)}$, Müller (1993) lets, $\pi(\rho) = p(\rho | \theta_2^{(i)}, \theta_3^{(i)}, \ldots, \theta_d^{(i)}, Y)$. The quantity $\theta_1^{(i+1)}$ is then the Tth value in the Metropolis chain:

a. Generate y^* from the distribution $f(y - x)$, where x is the current state of the chain.
b. With probability $\alpha(x, y^*) = \min(1, \pi(y^*)/\pi(x))$ accept y^* as the new state; otherwise the chain remains at x.

This algorithm, in contrast to rejection/acceptance algorithms, does not require the construction of a function which dominates $\pi(\rho)$ everywhere. However, it does require specification of the driving function $f(y - x)$. Like the griddy Gibbs sampler of section 6.4, this algorithm relies on repeated evaluation of the conditional $p(\rho | \theta_2, \ldots, \theta_d, Y)$, modulo the normalizing constant. Moreover, this Metropolis' subchain idea is iterative and, as such, requires some assessment of whether the subchain has reached equilibrium. Tierney (1991) suggests using the grid-based approximation of $\pi(\rho)$ described in Section 6.4 to drive the Metropolis subchain. Gelman (1992) suggests other ways to construct these within-Gibbs subchains.

6.6. Conditional Inference via the Gibbs Sampler

6.6.1. Introduction

Several approaches are available for inference in the presence of nuisance parameters. This book has focused on the Bayesian/likelihoodist approach where the nuisance parameters are eliminated by integration and inference is based on a marginal likelihood or marginal posterior distribution.

A frequentist approach bases inference on the conditional distribution of the sufficient statistics corresponding to the parameters of interest, given the sufficient statistics corresponding to the remaining parameters fixed at their observed values. This device yields a distribution which is independent of the nuisance parameters in the exponential family case. As noted in Cox and Snell (1989), this is the standard Neyman–Pearson approach to the elimination of nuisance parameters. Andersen (1980) surveys the use of conditional likelihood functions to facilitate inference. The program of considering the exact permutation distribution of the conditional sufficient statistics in the case of logistic regression has been suggested by Cox (1971) and implemented in a computationally feasible manner by Bayer and Cox (1979), Tritchler (1984) and Hirji, Mehta and Patel (1987). In the contingency-table case, Patefield (1981) presents an algorithm for generating $r \times c$ contigency tables, conditional on row and column totals of the observed tables, to allow for a Monte Carlo conditional test of the model of independence. Mehta and Patel (1983, 1986a, b) present sophisticated network algorithms to perform exact calculations. In typical applications, however, hypothesis testing is generally performed by comparing the likelihood ratio, Wald or score statistic to the appropriate asymptotic distribution under the null; usually, a chi-square approximation is used. Confidence regions are typically constructed by inverting these approximate tests. It is well known that for small, highly stratified or unbalanced data sets, these asymptotic methods may be misleading.

Kolassa and Tanner (1992) propose a Gibbs–Skovgaard algorithm for the construction of a Markov chain whose equilibrium distribution $\tilde{F}$ approximates the distribution F of the sufficient statistics of interest, conditional on the sufficient statistics associated with the nuisance parameters. They approximate the univariate conditional distribution for a sufficient statistic corresponding to a canonical parameter of interest, conditional on all other sufficient statistics, using a saddlepoint approximation derived by Skovgaard (1987). The Gibbs sampler is then used to sample from this Markov chain with equilibrium distribution $\tilde{F}$. To generate a single chain realization, one draws an observation from the (Skovgaard) approximate distribution of each sufficient statistic conditional on all other sufficient statistics. Following the Gibbs scheme, one systematically scans all sufficient statistics of interest. Under certain regularity conditions, the resulting chain is ergodic and the equilibrium distribution is an approximation to the joint distribution of interest.

Their aim is to sample accurately from the multivariate conditional distribution of sufficient statistics corresponding to the null hypothesis. No considerations of power arise in their work; rather, Kolassa and Tanner (1992) present a method of calculating p values that are closer to those arising from the exact sampling distribution of the test statistic chosen than those arising from asymptotic methods.

Before proceeding, other uses of Markov chain methods in the frequentist context are mentioned. Gelfand and Carlin (1991) and Geyer and Thompson (1992) illustrate how these methods can be used to locate the modes of the likelihood (or posterior). Besag and Clifford (1989) discuss the application of Markov chain methods to the problem of significance testing. *Assuming* one can devise an ergodic Markov chain whose equilibrium distribution is the same as the distribution of the sufficient statistic and whose transition probabilities are easy to sample from, Besag and Clifford (1989, 1991) suggest methods for computing p-values.

6.6.2. Skovgaard's Approximation

Skovgaard (1987) uses saddlepoint methods to derive approximate tail probabilities for conditional distributions.

Let

$$K_X(\beta) = \log[E(\exp(\beta^T X))]$$

be the cumulant generating function of X, a sum of n independent random vectors. Denote by $\hat{\beta}$ the multivariate saddlepoint, which is determined by the saddlepoint equation:

$$K'_X(\hat{\beta}) = x \ ,$$

where K'_X is the vector of derivatives of K_X and x is the observed value of X. Let $\hat{\gamma}$ be a vector of length d such that

$$K_{X_j}(\hat{\gamma}) = x_j \quad \text{for } j = 2, \ldots, d \ ,$$

$$\hat{\gamma}_1 = 0 \ ,$$

where K_{X_j} denotes the derivative of K_X with respect to component j of its argument. Skovgaard (1987) approximates the conditional cumulative distribution function as

$$\Phi(\sqrt{n}\hat{w}) - \phi(\sqrt{n}\hat{w})\left(\frac{\sqrt{|K''_{X^{-1}}(\hat{\gamma})|}}{\hat{\beta}_1\sqrt{n\,|K''_X(\hat{\beta})|}} - \frac{1}{\sqrt{n}\hat{\omega}}\right) \tag{6.6.1}$$

where $\hat{\beta}_1$ is the first component of $\hat{\beta}$,

$$\hat{\omega} = \operatorname{sgn}(\hat{\beta}_1)\sqrt{2[\hat{\beta}^{\mathrm{T}}x - K_X(\hat{\beta})] - 2[\hat{\gamma}^{\mathrm{T}}x - K_X(\hat{\gamma}])},$$

$K''_{X^{-1}}$ is the $(d-1)\times(d-1)$ submatrix of the matrix of second derivatives of K_X, corresponding to all components of β and X except the first, and Φ and ϕ are the normal distribution function and density, respectively.

Also of interest are inversion techniques for lattice distributions. Skovgaard (1987) derives a counterpart of (6.6.1) in the lattice case:

$$\Phi(\sqrt{n}\hat{\omega}) - \phi(\sqrt{n}\hat{\omega})\left(\frac{\sqrt{|K''_{X^{-1}}(\hat{\gamma})|}}{2\sinh(\frac{1}{2}\hat{\beta}_1)\sqrt{n\,|K''_X(\hat{\beta})|}} - \frac{1}{\sqrt{n}\hat{\omega}}\right) \tag{6.6.2}$$

and x_1 is corrected for continuity when calculating $\hat{\beta}$. That is, if possible values for X_1 and Δ units apart, $\hat{\beta}$ solves $K'_X(\hat{\beta}) = \tilde{x}$ where $\tilde{x}_j = x_j$ if $j \neq 1$ and $\tilde{x}_1 = x_1 - \frac{1}{2}\Delta$. Skovgaard illustrates this formula in the context of the hypergeometric distribution, which has applications to testing independence in 2×2 tables.

6.6.3. Two-way Contingency Tables

Inference on a subset of parameters in a regression model in the presence of other parameters is often difficult, especially in the case of nonlinear regression models where the device of orthogonality is not available. Cox and Snell (1989) suggest conditioning on the sufficient statistics corresponding to regression parameters not of immediate interest. The situation in this section exhibits approximate tests of the hypothesis of independence in contingency tables. This situation represents a special case of Poisson regression, in which an $r \times c$ table of independent Poisson random variables is observed. One assumes a log-linear model for the means with row and column effects, and row × column interactions. The test of independence is equivalent to testing whether these interactions are simultaneously zero. Row and column total means are nuisance parameters; hence, we condition on the observed row and column totals, to test the hypothesis of independence. Conditioning on row and column totals leaves random $(r-1) \times (c-1)$ sufficient statistics. These sufficient statistics not conditioned on may be chosen to be table entries in all but the last row and column. Denote these as $T_1, \ldots, T_{(r-1)(c-1)}$. The idea in Kolassa and Tanner (1992) is to generate a Markov chain whose equilibrium distribution is approximately the joint distribution of $T_1, \ldots, T_{(r-1)(c-1)}$, conditional on the row and column totals. They use the Gibbs sampler to reduce this to the problem of sampling from a distribution approximating that of T_i conditional on $T_j \; \forall_j \neq i$ and the row and column totals.

Kolassa and Tanner (1992) apply the Gibbs–Skovgaard algorithm to test the model of independence in an $r \times c$ contingency table $\mathbf{X} = (X_{ij})$, where X_{ij} are independent Poisson random variables with means $\alpha + \beta_i + \gamma_j + \delta_{ij}$. As presented here, this model is overparameterized. Constrain these parameters so that $\beta_r = \gamma_c = 0$ and $\delta_{ij} = 0$ if $i = r$ and $j = c$. The rc sufficient statistics for the saturated model can be chosen to be the raw table entries $\mathbf{X}$. However, an alternative selection of sufficient statistics will make implementing the Gibbs–Skovgaard algorithm more convenient. This algorithm requires sampling conditional on the canonical sufficient statistics associated with main, row and column effects α, β_i and γ_j. Hence, a collection $\mathbf{T}$ of sufficient statistics must be found to include these canonical sufficient statistics. These sufficient statistics are the grand total, the $r-1$ row totals for all but the last row, and the $c-1$ column totals for all but the last column. Canonical sufficient statistics for the parameters δ_{ij} are the $(r-1) \times (c-1)$ table entries in all but the last row and column. Let $\mathbf{T}$ be this vector of sufficient statistics.

The Gibbs sampler is used to sample from the Markov chain whose equilibrium distribution is approximately the joint distribution of these $(r-1) \times (c-1)$ table entries $(T_1, \ldots, T_{(r-1)(c-1)})$ conditional on the other $r + c - 1$ sufficient statistics. One could proceed by drawing random observations uniform on (0, 1), and inverting the conditional cumulative distribution function approximation (6.6.2) to get a random observation from a distribution approximating the true conditional distribution. In this case (6.6.2) would be calculated with $K_{\mathbf{T}}$ in place of K_X. Kolassa and Tanner (1992), however, show how this may be simplified, first by showing how the desired conditional distribution may be generated by a derived 2×2 table rather than by the original larger $r \times c$ table, and then by presenting an alternative to the approximation (6.6.2) not requiring the joint cumulant generating function of a table observation and the marginals, but only the joint cumulant generating function of raw table entries.

Kolassa and Tanner (1992) prove a theorem which allows one to reduce the problem of sampling from the conditional distribution of an entry in this $r \times c$ table to the problem of sampling from a hypergeometric distribution arising from a 2^2 table. More formally, their theorem states that the distribution of one of these first $(r-1) \times (c-1)$ components of **T**, the entry in row i and column j of the table **X**, conditional on all other sufficient statistics, is the same as that if X_{ij} has a hypergeometric distribution generated by the derived 2×2 table

$$\begin{array}{ccc} X_{ij} & x_{ic} & x_{ij} + x_{ic} \\ x_{rj} & x_{rc} & x_{rj} + x_{rc} \\ & & \\ x_{ij} + x_{rj} & x_{ic} + x_{rc} & \end{array} \qquad (6.6.3)$$

conditional on row and column totals for this reduced table.

The idea in Kolassa and Tanner (1992) is to sample approximately from this conditional distribution by drawing an observation from a uniform distribution on (0, 1) and inverting an approximation to the conditional cumulative distribution function arising in (6.6.3) at this random observation. They show that the Skovgaard (1987) approximation in this case can be reduced to:

$$P(X_{ij} \leqslant x_{ij} \mid R, C, N) = \Phi(\hat{\omega}) - \phi(\hat{\omega}) \left\{ \left(\frac{x_{i.} x_{.j} x_{r.} x_{.c}}{x_{..} \tilde{x}_{ij} \tilde{x}_{ic} \tilde{x}_{rj} \tilde{x}_{rc}} \right)^{1/2} \times \right.$$
$$\left. [2 \sinh(\rho)]^{-1} - \hat{\omega}^{-1} \right\}$$

where $\rho = \log[\tilde{x}_{ij}\tilde{x}_{rc}/(\tilde{x}_{rj}\tilde{x}_{ic})]$ is the log of the observed continuity-corrected odds ratio, $R = x_{ij} + x_{ic}$, $C = x_{ij} + x_{rj}$, $N = x_{ij} + x_{ic} + x_{rj} + x_{rc}$, $\tilde{x}_{ij} = x_{ij} - \frac{1}{2}$, $\tilde{x}_{rj} = x_{rj} + \frac{1}{2}$, $\tilde{x}_{ic} = x_{ic} + \frac{1}{2}$, $\tilde{x}_{rc} = x_{rc} - \frac{1}{2}$, and

$$\hat{\omega} = \operatorname{sgn}(\rho) \left\{ 2 \left[\sum_{\substack{i=i,r \\ m=j,c}} \tilde{x}_{lm} \log(\tilde{\mathrm{x}}_{lm}) - \sum_{l=i,r} \mathrm{x}_{l.} \log(x_{l.}) \right. \right.$$
$$\left. \left. - \sum_{m=j,c} x_{.m} \log(\mathrm{x}_{.m}) + \mathrm{x}_{..} \log(x_{..}) \right] \right\}^{1/2}.$$

In this case, · denotes summation over the reduced table: $x_{l.} = x_{lj} + x_{lc}$, $x_{.m} = x_{im} + x_{rm}$, and $x_{..} = x_{ij} + x_{ic} + x_{rj} + x_{rc}$.

This equation provides an approximation to the distributions of each of the sufficient statistics representing individual table counts conditional on all other sufficient statistics. Random observations from a Markov chain are generated by cycling through the Gibbs sampling procedure and at each stage drawing a random sample from the cumulative distribution function given above. Cycling through the Gibbs sampling scheme many times will give a multivariate sample whose distribution is close to the equilibrium distribution, which in turn is an approximation to the joint distribution of the sufficient statistics representing table counts conditional on row and column totals. Kolassa and Tanner (1992) also consider the case of three-way tables and of logistic regression.

6.6.4. Example

This example is a 4 × 4 contingency table of responses to the questionnaire item "Sex is fun for me and my partner (a) never or occasionally, (b) fairly often, (c) very often, (d) almost always" for 91 married couples from the Tucson metropolitan area, as reported by Hout, et al. (1987) and cited by Agresti (1990) (see Table 6.4):

Table 6.4. Sexual Response Data

	Wife's Response			
Husband's Response	Never or ocassionally	Fairly often	Very often	Almost always
Never or ocassionally	7	7	2	3
Fairly often	2	8	3	7
Very often	1	5	4	9
Almost always	2	8	9	14

These data were chosen to assess the quality of the Markov chain algorithm in a situation in which usual asymptotic approximations may be inappropriate. Again, the hypothesis of independence was tested by generating random tables using the Gibbs sampler and Skovgaard's approximation as described above. The statistic for the likelihood ratio test was calculated for each simulated table.

Kolassa and Tanner (1992) simulated 50,000 independent Markov chains for 30 iterations. For each integer k between 1 and 30, they estimated the p-value after iteration k generated by the test statistic by calculating the test statistic for the observed table, and for each of the 50,000 tables represented by the state of the chain at time k. They report as the estimated p-value the proportion of sample tables with a test-statistic value as high or higher than the observed value. Convergence was assessed by monitoring these p-values across iterations; in this example note that after fewer than 10 iterations the

Table 6.5. Results for Sexual Response Data

Test Method	Likelihood ratio
StatXact Monte Carlo *p* value	0.111
Gibbs–Skovgaard *p* value	0.107
Asymptotic *p* value	0.078

estimated *p*-values become stable. The *p*-value based on the final 9000 iterations of one chain of length 10,000 was 0.109. These results were compared (see Table 6.5) with those obtained from the StatXact implementation of the Monte Carlo algorithm due to Patefield (1981), using 50,000 samples. The results are given in Table 6.5.

The Gibbs–Skovgaard algorithm represents an improvement over the asymptotic value, in that the Gibbs–Skovgaard p value lies closer to the true p value than does the p value derived from the χ^2 approximation. It is not argued that the Gibbs–Skovgaard algorithm is the preferred algorithm for testing the model of independence in the $r \times c$ contingency table. The goal is to illustrate the accuracy of this general algorithm in a situation where the exact (up to Monte Carlo error) result is known. This work provides a *general algorithm* to handle small sample inference in the exponential family.

References

Agresti, A. (1990). *Categorical Data Analysis.* New York: Wiley Interscience.

Aitkin, M. (1980). A note on regression analysis of censored data. *Technometrics* **27**, 161–163.

Amit, Y. (1991). On rates of convergence of stochastic relaxation for Gaussian and non-Gaussian distributions. *Journal of Multivariate Analysis* **38**, 82–99.

Andersen, E. B. (1980). *Discrete Statistical Models.* Amsterdam: North-Holland.

Applegate, D., Kannan, R., and Polson, N. (1990). Random polynomial time algorithms for sampling from joint distributions. Technical report, School of Computer Science, Carnegie Mellon University.

Barker (1965). Monte Carlo calculations of radial distribution functions for a proton–electron plasma. *Australian Journal of Physics* **18**, 119–133.

Bates, D. M., and Watts, D. G. (1988). *Nonlinear Regression Analysis and Its Applications.* New York: Wiley.

Bayer, L., and Cox, C. (1979). Algorithm AS142. *Applied Statistics* **28**, 319–24.

Besag, J., and Clifford, P. (1989). Generalized Monte Carlo significance tests. *Biometrika* **76**, 633–642.

Besag, J., and Clifford, P. (1991). Sequential Monte Carlo *p*-values. *Biometrika* **78**, 301–304.

Besag, J. (1974). Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society B* **36**, 192–326.

Bickel, P. J., and Doksum, K. A. (1977). *Mathematical Statistics.* Oakland: Holden-Day.

Bishop, Y. M. M., Fienberg, S., and Holland, P. (1975). *Discrete Multivariate Analysis.* Cambridge: MIT Press.

Box, G. E. P., and Tiao, G. (1973). *Bayesian Inference in Statistical Analysis.* Reading: Addison–Wesley.

Brown, C. H. (1990). Asymptotic properties of estimators with nonignorable missing data. Technical Report, Department of Biostatistics, Johns Hopkins.

Buonaccorsi, J. P., and Gatsonis, C. A. (1988). Bayesian inference for ratios of coefficients in a linear model. *Biometrics* **44**, 87–102.

Carlin, B. P., and Gelfand, A. E. (1991). An iterative Monte Carlo method for nonconjugate Bayesian analysis. *Statistics and Computing* **1**, 119–128.

Carlin, B. P., Gelfand, A. E., and Smith, A. F. M. (1992). Hierarchical Bayesian analysis of change point problems. *Journal of the Royal Statistical Society C* **41**, 389–405.

Carlin, B. P., and Polson, N. G. (1991). Inference for nonconjugate Bayesian models using the Gibbs sampler. *Canadian Journal of Statistics* **19**, 399–405.

Carlin, B. P., Polson, N. G., and Stoffer, D. S. (1992). A Monte Carlo approach to nonnormal and nonlinear state space modelling. *Journal of the American Statistical Association* **87**, 493–500.

Chan, K. S. (1993) Asymptotic behavior of the Gibbs sampler, *Journal of the American Statistical Association* **88**, 320–326.

Chen, M. H. (1992). Importance weighted marginal Bayesian posterior density estimation. Technical report, Department of Statistics, Purdue University.

Chib, S. (1992). Bayes inference in the tobit censored regression model. *Journal of Econometrics* **51**, 79–99.

Clayton, D. G. (1991). A Monte Carlo method for Bayesian inference in frailty models. *Biometrics* **47**, 467–485.

Cochran, W. G. (1977). *Sampling Techniques*. New York: Wiley.

Cook, D., and Weisberg, S. (1982). *Residuals and Influence in Regression*. London: Chapman and Hall.

Cox, D. R., and Hinkley, D. V. (1974). *Theoretical Statistics*. London: Chapman & Hall.

Cox, D. R. (1971). *Analysis of Binary Data*. London: Chapman and Hall.

Cox, D. R., and Snell, E. J. (1989). *Analysis of Binary Data*. London: Chapman and Hall.

Davis, P., and Rabinowitz, P. (1984). *Methods of Numerical Integration*. New York: Academic Press.

Dempster, A. P., Laird, N., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society* B **39**, 1–38.

Efron, B., and Hinkley, D. V. (1978). Assessing the accuracy of the maximum likelihood estimator: observed versus expected Fisher information. *Biometrika* **65**, 457–482.

Gaver, D. P., and O'Muircheartaigh, I. G. (1987). Robust empirical Bayes analysis of event rates. *Technometrics* **29**, 1–15.

Gelfand, A. E., and Carlin, B. P. (1991). Maximum likelihood estimation for constrained or missing data model. Research Report 91-002, Division of Biostatistics, University of Minnesota.

Gelfand, A. E. (1992). Discussion to the papers of Gelman and Rubin and of Geyer. *Statistical Science*, **4**, 486–487.

Gelfand, A. E., Hills, S. E., Racine-Poon, A., and Smith, A. F. M. (1990). Illustration of Bayesian inference in normal data models using Gibbs sampling. *Journal of the American Statistical Association* **85**, 972–985.

Gelfand, A. E., and Smith, A. F. M. (1990). Sampling based approaches to calculating marginal densities. *Journal of the American Statistical Association* **85**, 398–409.

Gelfand, A. E., and Smith, A. F. M. (1991). Gibbs sampling for marginal posterior expectations. *Communications in Statistics A* **20**, 1747–1766.

Gelfand, A. E., Smith, A. F. M., and Lee, T. M. (1992). Bayesian analysis of constrained parameter and truncated data problems. *Journal of the American Statistical Association* **87**, 523–532.

Gelman, A. (1992). Iterative and non-iterative simulation algorithms. In: *Computing Science and Statistics: Proceedings of the 24th Symposium on the Interface*. Interface Foundation of North America, Fairfax VA.

Gelman, A., and Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*. **4**, 457–472.

Geman, S., and Geman, D. (1984). Stochastic relaxation, Gibbs distribution and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **6**, 721–741.

Geyer, C. (1992). Practical Markov chain Monte Carlo. *Statistical Science.* **4**, 473–482.

Geyer, C. J., and Thompson, E. A. (1992). Constrained Monte Carlo maximum likelihood for dependent data (with discussion) *Journal of the Royal Statistical Society Series B*, **54**, 657–699.

Geweke, J. (1989). Bayesian inference in econometric models using Monte Carlo integrations. *Econometrica* **24**, 1317–1339.

Gilks, W. R., and Wild, P. (1992). Adaptive rejection sampling for Gibbs sampling. *Journal of the Royal Statistical Society* C **41**, 337–348.

Glasser, M. (1965). Regression analysis with dependent variable censored. *Biometrics* **21**, 300–307.

Glynn, R. J., Laird, N. M., and Rubin, D. B. (1986). Mixing modelling versus selection modelling with nonignorable nonresponse. In: H. Wainer (ed.), *Drawing Inferences from Self-Selected Samples.* New York: Springer.

Goodman, L. A. (1974). Exploratory latent structure analysis using both identifiable and unidentifiable models. *Biometrika* **61**, 215–231.

Greenlees, J. S., Reece, W. S., and Zieschang, K. Y. (1982). Imputation of missing values. *Journal of the American Statistical Association* **77**, 251–261.

Haberman, S. J. (1979). *Analysis of Qualitative Data.* New York: Academic Press.

Hammersley, J. M., and Handscomb, D. C. (1979). *Monte Carlo Methods.* London: Chapman and Hall.

Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika* **57**, 97–109.

Hirji, K. F., Mehta, C. R., and Patel, N. R. (1987). Computing distributions for exact logistic regression. *Journal of the American Statistical Association* **82**, 1110–1117.

Hout, M., Duncan, O. D., and Sobel, M. E. (1987). Association and heterogeneity: structural models of similarities and differences. *Sociological Methodology* **17**, 145–184.

Jamshidian, M., and Jennrich, R. (1993). Conjugate gradient acceleration of the EM algorithm. *Journal of the American Statistical Association* **88**, 221–228.

Kass, R. E., Tierney, L., and Kadane, J. B. (1989). Approximate methods for assessing influence and sensitivity in Bayesian analysis. *Biometrika* **76**, 663–674.

Kong, A., Liu, J., and Wong, W. H. (1992). Sequential imputations and Bayesian missing data problems. Technical Report, Dept of Statistics, University of Chicago.

Kolassa, J. E., and Tanner, M. A. (1992). *Approximate conditional inference via the Gibbs sampler.* Technical Report, Department of Biostatistics, University of Rochester.

Lawless, J. F. (1982). *Statistical Models and Methods for Lifetime Data.* New York: Wiley.

Leonard, T., Hsu, J. S. J., and Tsui, K. W. (1989). Bayesian marginal inference. *Journal of the American Statistical Association* **84**, 1051–1058.

Little, R. J. A., and Rubin, D. B. (1983). On jointly estimating parameters and missing data. *The American Statistician* **37**, 218–220.

Little, R. J. A., and Rubin, D. B. (1987). *Statistical Analysis With Missing Data.* New York: Wiley.

Li, K. H. (1988). Imputation using Markov chains. *Journal of Statistical Computation and Simulation* **30**, 57–79.

Liu, C., and Liu, J. (1993). Comment on Markov chain Monte Carlo. *Journal of the Royal Statistical Society B*, **55**, 82–83.

Liu, J., Wong, W. H., and Kong, A. (1991a). Correlation structure and convergence rate of the Gibbs sampler (I): applications to the comparison of estimators and augmentation scheme. Technical Report, Department of Statistics, University of Chicago.

Liu, J., Wong, W. H., and Kong, A. (1991b). Correlation structure and convergence-rate of the Gibbs sampler (II): applications to various scans. Technical Report, Department of Statistics, University of Chicago.

Louis, T. A. (1982). Finding observed information using the EM algorithm. *Journal of the Royal Statistical Society* B **44**, 98–130.

Marske, D. (1967). Biochemical oxygen demand data interpretation using sums of squares surface. M.Sc. Thesis, University of Wisconsin.

Matthews, P. (1991) A slowly mixing Markov chain with implications for Gibbs sampling. Technical Report, Department of Mathematics and Statistics, University of Maryland.

McCullagh, P., and Nelder, J. (1989). *Generalized Linear Models.* London: Chapman and Hall.

Mehta, C., and Patel, N. (1983). A network algorithm for performing Fisher's exact test in $r \times c$ contingency tables. *Journal of the American Statistical Association* **78**, 427–434.

Mehta, C., and Patel, N. (1986a). A hybrid algorithm for Fisher's exact test on unordered $r \times c$ contingency tables. *Communications in Statistics* **15**, 387–403.

Mehta, C., and Patel, N. (1986b). FEXACT: a fortran subroutine for Fisher's exact test on unordered $r \times c$ contingency tables. *ACM Transactions on Mathematical Software* **12**, 295–302.

Mehta, C., and Patel, N. (1987). Computing distributions for exact logistic regression. *Journal of the American Statistical Association* **82**, 1110–1117.

Meilijson, I. (1989) A fast improvement to the EM algorithm on its own terms. *Journal of the Royal Statistical Society.* B **51**, 127–138.

Mendenhall, W. M., Parsons, J. T., Stringer, S. P., Cassissi, N. J., and Million, R. R. (1989). T2 oral tongue carcinoma treated with radiotherapy: analysis of local control and complications. Radiotherapy and Oncology **16**, 275–282.

Meng, X. L., and Rubin, D. B. (1991) Using EM to obtain asymptotic variance–covariance matrices. *Journal of the American Statistical Association* **86**, 899–909.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., and Teller, E. (1953). Equations of state calculations by fast computing machines. *Journal of Chemical Physics* **21**, 1087–1091.

Miller, R. (1980). *Survival Analysis.* New York: Wiley.

Miller, R. G., and Halpern, J. W. (1982). Regression with censored data. *Biometrika* **69**, 521–531.

Morris, C. (1983). Parametric empirical Bayes inference: theory and applications (with discussion. *Journal of the American Statistical Association* **78**, 47–65.

Morris, C. (1987). Comment on "The calculation of posterior distributions by data augmentation" *Journal of the American Statistical Association* **82**, 542–543.

Mosteller, F., and Wallace, D. L. (1964). *Inference and Disputed Authorship: The Federalist.* Reading: Addison–Wesley.

Müller, P. (1993). A generic approach to posterior integration and Gibbs sampling. *Journal of the American Statistical Association*, to appear.

Naylor, J. C., and Smith, A. F. M. (1982). Applications of a method for the efficient computation of posterior distributions. *Applied Statistics* **31**, 214–225.

Odell, P. L., and Feiveson, A. H. (1966). A numerical procedure to generate a sample covariance matrix. *Journal of the American Statistical Association* **61**, 199–203.

Orchard, T., and Woodbury, M. A. (1972). A missing information principle: theory and applications. *Proceedings of the 6th Berkeley Symposium on Mathematical Statistics*, Vol. 1, pp. 697–715.

Patefield, W. M. (1981). An efficient method of generating $r \times c$ tables with given row and column totals (Algorithm AS 159). *Applied Statistics* **30**, 91–97.

Raftery, A. E., and Lewis, S. M. (1992) How many iterations in the Gibbs sampler?. In: J. M. Bernardo, J. O. Bergen, A. P. Dawid and A. F. M. Smith (eds.) *Bayesian Statistics.* Oxford University Press.

Rao, C. R. (1973). *Linear Statistical Inference and Applications.* New York: Wiley.

Ripley, B. (1987). *Stochastic Simulation.* New York: Wiley.

Ritter, C., and Tanner, M. A. (1992) The Gibbs stopper and the griddy Gibbs sampler. *Journal of the American Statistical Association.* **87**, 861–868.

Rosenthal, J. S. (1991). Rates of convergence for data augmentation on finite sample spaces. Tech. Report, Department of Mathematics, Harvard.

Rosenthal, J. (1992). Rates of convergence for Gibbs sampling for variance component models. Technical Report, Department of Mathematics, Harvard University.

Roberts, G. O. (1992). Convergence diagnostics of the Gibbs sampler. In: J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith (eds.), Bayesian Statistics. Oxford University Press.

Rubenstein, R. (1981). *Simulation and the Monte Carlo Method.* New York: Wiley.

Rubin, D. B. (1987a), Comment on "The calculation of posterior distributions by data augmentation", by M. A. Tanner and W. H. Wong. *Journal of the American Statistical Association* **82**, 543–546.

Rubin, D. B. (1987b). *Multiple Imputation for Nonresponse in Surveys.* New York: Wiley.

Schenker, N., and Welsh, A. H. (1988). Asymptotic results for multiple imputation. *Annals of Statistics* **16**, 1550–1566.

Schervish, M. J., and Carlin, B. P. (1992). On the convergence of successive substitution sampling. *Journal of Computational and Graphical Statistics* **1**, 111–127.

Schmee, J., and Hahn, G. J. (1979). A simple method for regression analysis with censored data. *Technometrics* **21**, 417–432.

Sinha, D. (1993). Semiparametric Bayesian analysis of multiple event time data. *Journal of the American Statistical Association*, to appear.

Sinha, D., Tanner, M. A., and Hall, W. J. (1992). Maximization of the marginal likelihood of grouped survival data. Technical Report, Department of Biostatistics, University of Rochester.

Skovgaard, I. M. (1987). Saddlepoint expansions for conditional distributions. *Journal of Applied Probability* **24**, 875–887.

Smith, C. A. B. (1977). Comment on "Maximum Likelihood from Incomplete Data Via the EM Algorithm", by A. P. Dempster, N. M. Laird and D. B. Rubin, *Journal of the Royal Statistical Society B* **39**, 24–25.

Smith, A. F. M., Skene, A. M., Shaw, J. E. H., Naylor, J. C., and Dransfield, M. (1985). The implementation of the Bayesian paradigm. *Communications in Statistics* **14**, 1079–1102.

Smith, A. F. M., and Gelfand, A. E. (1992). Bayesian statistics without tears. *American Statistician* **46**, 84–88.

Snell, J. L. (1988). *Introduction to Probability.* New York: Random House.

Tanner, M. A. (1991). *Tools for Statistical Inference.* New York: Springer.

Tanner, M. A., and Wong, W. H. (1987). The calculation of posterior distributions by data augmentation. *Journal of the American Statistical Association* **82**, 528–540.

Thompson, P. A., and Miller, R. B. (1986). Sampling the future: a Bayesian approach to forecasting. *Journal of Business and Economic Statistics* **4**, 427–436.

Tiao, G., and Fienberg, S. (1969). Bayesian estimation of latent roots and vectors with special reference to the bivariate normal distribution. *Biometrika* **56**, 97–104.

Tierney, L., and Kadane, J. B. (1986). Accurate approximations for posterior moments and marginal densities. *Journal of the American Statistical Association* **81**, 82–86.

Tierney, L., Kass, R. E., and Kadane, J. B. (1989). Approximate marginal densities of nonlinear functions. *Biometrika* **76**, 425–434.

Tierney, L. (1991). Markov chains for exploring posterior distributions. Technical Report, School of Statistics, University of Minnesota.

Tritchler, D. (1984). An algorithm for exact logistic regression. *Journal of the American Statistical Association.* **79**, 709–711.

Turnbull, B. W., Brown, B. W., and Hu, M. (1974). Survivorship analysis of heart transplant data. *Journal of the American Statistical Association* **69**, 74–80.

Wasserman, L. A. (1989). A robust Bayesian interpretation of likelihood regions *Annals of Statistics* **17**, 1387–1393.

Wei, G. C. G., and Tanner, M. A. (1990a). A Monte Carlo implementation of the EM algorithm and the poor man's data augmentation algorithm. *Journal of the American Statistical Association* **85**, 699–704.

Wei, G. C. G., and Tanner, M. A. (1990b). Calculating the content and boundary of the HPD region via data augmentation. *Biometrika* **77**, 649–652.

Wei, G. C. G., and Tanner, M. A. (1990c). Posterior computations for censored regression data. *Journal of the American Statistical Association* **85**, 829–839.

Wong, W. H., and Li, B. (1992). Laplace expansion for posterior densities of nonlinear function of parameters, *Biometrika* **79**, 393–398.

Wu, C. F. J. (1983). On the convergence properties of the EM algorithm. *Annals of Statistics* **11**, 95–103.

Zeger, S., and Karim, R. (1991). Generalized linear models with random effects. *Journal of the American Statistical Association* **86**, 79–86.

Zellner, A., and Rossi, P. E. (1984). Bayesian analysis dichotomous quantal response models. *Journal of Econometrics* **25**, 365–393.

Subject Index

Springer Series in Statistics

(continued from p. ii)

Reinsel: Elements of multivariate Time Series Analysis.
Reiss: A Course on Point Processes
Reiss: Approximate Distributions of Order Statistics: With Applications to Nonparametric Statistics.
Ross: Nonlinear Estimation.
Sachs: Applied Statistics: A Handbok of Techniques, 2nd edition.
Salsburg: The Use of Restricted Significance Tests in Clinical Trials.
Sarndal/Swensson/Wretman: Model Assisted Survey Sampling.
Seneta: Non-Negative Matrices and Markov Chains, 2nd edition.
Shedler: Regeneration and Networks of Queues.
Siegmund: Sequential Analysis: Tests and Confidence Intervals.
Tanner: Tools for Statistical Inference: Methods for the Exploration of Posterior Distributions and Likelihood Functions.
Todorovic: An Introduction to Stochastic Processes and Their Applications.
Tong: The Multivariate Normal Distribution.
Vapnik: Estimation of Dependences Based on Empirical Data.
West/Harrison: Bayesian Forecasting and Dynamic Models.
Wolter: Introduction to Variance Estimation.
Yaglom: Correlation Theory of Stationary and Related Random Functions I: Basic Results.
Yaglom: Correlation Theory of Stationary and Related Random Functions II: Supplementary Notes and References.