I dedicate this to my brother Florin.
ACKNOWLEDGMENTS

I would like to thank my advisor, Professor Hani Doss, for the invaluable help and guidance with writing this dissertation. I am also thankful to Professors Farid AitSahlia, George Casella, and James Hobert for serving on my supervisory committee and offering me helpful comments. In addition, I owe a debt of gratitude to the Department of Statistics at the University of Florida. I greatly appreciate the chance I have been given to come here and learn Statistics from many exceptional teachers, all while benefiting from the kindness and support of other students and staff.
# TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACKNOWLEDGMENTS</td>
<td>4</td>
</tr>
<tr>
<td>LIST OF TABLES</td>
<td>7</td>
</tr>
<tr>
<td>LIST OF FIGURES</td>
<td>8</td>
</tr>
<tr>
<td>ABSTRACT</td>
<td>9</td>
</tr>
<tr>
<td>CHAPTER</td>
<td></td>
</tr>
<tr>
<td>1 INTRODUCTION</td>
<td>10</td>
</tr>
<tr>
<td>2 ESTIMATION OF BAYES FACTORS AND POSTERIOR EXPECTATIONS</td>
<td>14</td>
</tr>
<tr>
<td>2.1 Estimation of Bayes Factors</td>
<td>17</td>
</tr>
<tr>
<td>2.2 Estimation of Bayes Factors Using Control Variates</td>
<td>19</td>
</tr>
<tr>
<td>2.3 Estimation of Posterior Expectations</td>
<td>21</td>
</tr>
<tr>
<td>2.4 Estimation of Posterior Expectations Using Control Variates</td>
<td>23</td>
</tr>
<tr>
<td>2.5 Estimation of Posterior Expectations Using Control Variates With</td>
<td>25</td>
</tr>
<tr>
<td>Estimated Skeleton Bayes Factors and Expectations</td>
<td></td>
</tr>
<tr>
<td>3 VARIANCE ESTIMATION AND SELECTION OF THE SKELETON POINTS</td>
<td>27</td>
</tr>
<tr>
<td>3.1 Estimation of the Variance</td>
<td>27</td>
</tr>
<tr>
<td>3.2 Selection of the Skeleton Points</td>
<td>32</td>
</tr>
<tr>
<td>4 REVIEW OF PREVIOUS WORK</td>
<td>34</td>
</tr>
<tr>
<td>5 ILLUSTRATION ON VARIABLE SELECTION</td>
<td>38</td>
</tr>
<tr>
<td>5.1 A Markov Chain for Estimating the Posterior Distribution of Model</td>
<td>38</td>
</tr>
<tr>
<td>Parameters</td>
<td></td>
</tr>
<tr>
<td>5.2 Choice of the Hyperparameter</td>
<td>43</td>
</tr>
<tr>
<td>5.3 Examples</td>
<td>47</td>
</tr>
<tr>
<td>5.3.1 U.S. Crime Data</td>
<td>47</td>
</tr>
<tr>
<td>5.3.2 Ozone Data</td>
<td>50</td>
</tr>
<tr>
<td>6 DISCUSSION</td>
<td>55</td>
</tr>
<tr>
<td>APPENDIX</td>
<td></td>
</tr>
<tr>
<td>A PROOF OF RESULTS FROM CHAPTER 1</td>
<td>56</td>
</tr>
<tr>
<td>B DETAILS REGARDING GENERATION OF THE MARKOV CHAIN FROM CHAPTER 5</td>
<td>90</td>
</tr>
</tbody>
</table>
## LIST OF TABLES

<table>
<thead>
<tr>
<th>Table</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>5-1</td>
<td>Posterior inclusion probabilities for the fifteen predictor variables in the U.S. crime data set, under three models. Names of the variables are as in Table 2 of Liang et al. (2008) (but all variables except for the binary variable S have been log transformed).</td>
<td>49</td>
</tr>
<tr>
<td>D-1</td>
<td>The 44 predictors used in the ozone illustration. The symbol “.” represents an interaction.</td>
<td>99</td>
</tr>
</tbody>
</table>
## LIST OF FIGURES

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>5-1</td>
<td>Estimates of Bayes factors for the U.S. crime data. The plots give two different views of the graph of the Bayes factor as a function of $w$ and $g$ when the baseline value of the hyperparameter is given by $w = 0.5$ and $g = 15$. The estimate is (2–13), which uses control variates.</td>
<td>48</td>
</tr>
<tr>
<td>5-2</td>
<td>Estimates of posterior inclusion probabilities for Variables 1 and 6 for the U.S. crime data. The estimate used is (2–16).</td>
<td>50</td>
</tr>
<tr>
<td>5-3</td>
<td>Variance functions for two versions of $\hat{\beta}_{\hat{d}(\hat{d})}$. The left panel is for the estimate based on the skeleton (5–4). The points in this skeleton were shifted to better cover the problematic region near the back of the plot ($g$ small and $w$ large), creating the skeleton (5–5). The maximum variance is then reduced by a factor of 9 (right panel).</td>
<td>51</td>
</tr>
<tr>
<td>5-4</td>
<td>Estimates of Bayes factors for the ozone data. The plots give two different views of the graph of the Bayes factor as a function of $w$ and $g$ when the baseline value of the hyperparameter is given by $w = .2$ and $g = 50$.</td>
<td>52</td>
</tr>
<tr>
<td>5-5</td>
<td>95% confidence intervals of the posterior inclusion probabilities for the 44 predictors in the ozone data when the hyperparameter value is given by $w = .13$ and $g = 75$. A table giving the correspondence between the integers 1–44 and the predictors is given in Appendix D.</td>
<td>54</td>
</tr>
</tbody>
</table>
We consider situations in Bayesian analysis where we have a family of priors \( \nu_h \) on the parameter \( \theta \), where \( h \) varies continuously over a space \( \mathcal{H} \), and we deal with two related problems. The first involves sensitivity analysis and is stated as follows. Suppose we fix a function \( f \) of \( \theta \). How do we efficiently estimate the posterior expectation of \( f(\theta) \) simultaneously for all \( h \) in \( \mathcal{H} \)? The second problem is how do we identify subsets of \( \mathcal{H} \) which give rise to reasonable choices of \( \nu_h \)? We assume that we are able to generate Markov chain samples from the posterior for a finite number of the priors, and we develop a methodology, based on a combination of importance sampling and the use of control variates, for dealing with these two problems. The methodology applies very generally, and we show how it applies in particular to a commonly used model for variable selection in Bayesian linear regression, in which the unknown parameter includes the model and the regression coefficients for the selected model. The prior is a hierarchical prior in which first the model is selected, then the coefficients for this model are chosen, and this prior is indexed by two hyperparameters. These hyperparameters effectively determine whether the selected model will be a large model with many variables, or a parsimonious model with only a few variables, so choosing them is very important. We give two illustrations of our methodology, one on the U.S. crime data of Vandaele and the other on ground level ozone data originally analyzed by Breiman and Friedman.
CHAPTER 1
INTRODUCTION

In the Bayesian paradigm we have a data vector $Y$ with density $p_\theta$ for some unknown $\theta \in \Theta$, and we wish to put a prior density on $\theta$. The available family of prior densities is $\{\nu_h, h \in \mathcal{H}\}$, where $h$ is called a hyperparameter. Typically, the hyperparameter is multivariate and choosing it can be difficult. But this choice is very important and can have a large impact on subsequent inference. There are two issues we wish to consider:

(A) Suppose we fix a quantity of interest, say $f(\theta)$, where $f$ is a function. How do we assess how the posterior expectation of $f(\theta)$ changes as we vary $h$? More generally, how do we assess changes in the posterior distribution of $f(\theta)$ as we vary $h$?

(B) How do we determine if a given subset of $\mathcal{H}$ constitutes a class of reasonable choices?

The first issue is one of sensitivity analysis and the second is one of model selection.

As an example of the kind of problem we wish to deal with, consider the problem of variable selection in Bayesian linear regression. Here, we have a response variable $Y$ and a set of predictors $X_1, \ldots, X_q$, each a vector of length $m$. For every subset $\gamma$ of $\{1, \ldots, q\}$ we have a potential model $M_\gamma$ given by

$$Y = 1_m \beta_0 + X_\gamma \beta_\gamma + \epsilon,$$

where $1_m$ is the vector of $m$ 1’s, $X_\gamma$ is the design matrix whose columns consist of the predictor vectors corresponding to the subset $\gamma$, $\beta_\gamma$ is the vector of coefficients for that subset, and $\epsilon \sim \mathcal{N}_m(0, \sigma^2 I)$. Let $q_\gamma$ denote the number of variables in the subset $\gamma$. The unknown parameter is $\theta = (\gamma, \sigma, \beta_0, \beta_\gamma)$, which includes the indicator of the subset of variables that go into the linear model. A very commonly used prior distribution on $\theta$ is given by a hierarchy in which we first choose the indicator $\gamma$ from
the “independence Bernoulli prior”—each variable goes into the model with a certain probability $w$, independently of all the other variables—and then choose the vector of regression coefficients corresponding to the selected variables. In more detail, the model is described as follows:

$$Y \sim N_m(1, m \beta_0 + X_\gamma \beta_\gamma, \sigma^2 I)$$  \hspace{1cm} (1–1a)

$$(\sigma^2, \beta_0) \sim p(\sigma^2, \beta_0) \propto 1/\sigma^2, \quad \text{and given } \sigma, \beta_\gamma \sim N_{d_\gamma}(0, g \sigma^2 (X'_\gamma X_\gamma)^{-1})$$  \hspace{1cm} (1–1b)

$$\gamma \sim w q_\gamma (1 - w)^{q - q_\gamma}.$$  \hspace{1cm} (1–1c)

The prior on $(\sigma, \beta_0, \beta_\gamma)$ is Zellner’s $g$-prior introduced in Zellner (1986), and is indexed by a hyperparameter $g$. Although this prior is improper, the resulting posterior distribution is proper.

Note that we have used the word “model” in two different ways: (i) a model is a specification of the hyperparameter $h$, and (ii) a model in regression is a list of variables to include. The meaning of the word will always be clear from context.

To summarize, the prior on the parameter $\theta = (\gamma, \sigma, \beta_0, \beta_\gamma)$ is given by the two-level hierarchy (1–1c) and (1–1b), and is indexed by $h = (w, g)$. Loosely speaking, when $w$ is large and $g$ is small, the prior encourages models with many variables and small coefficients, whereas when $w$ is small and $g$ is large, the prior concentrates its mass on parsimonious models with large coefficients. Therefore, the hyperparameter $h = (w, g)$ plays a very important role, and in effect determines the model that will be used to carry out variable selection.

A standard method for approaching model selection involves the use of Bayes factors. For each $h \in \mathcal{H}$, let $m_h(y)$ denote the marginal likelihood of the data under the prior $\nu_h$, that is, $m_h(y) = \int p_\theta(y)\nu_h(\theta)\,d\theta$. We will write $m_h$ instead of $m_h(y)$. The Bayes factor of the model indexed by $h_2$ vs. the model indexed by $h_1$ is defined as the ratio of the marginal likelihoods of the data under the two models, $m_{h_2}/m_{h_1}$, and is denoted throughout by $B(h_2, h_1)$. Bayes factors are widely used as a criterion for comparing
models in Bayesian analyses. For selecting models that are better than others from the family of models indexed by \( h \in \mathcal{H} \), our strategy will be to compute and subsequently compare all the Bayes factors \( B(h, h_*) \), for all \( h \in \mathcal{H} \), and a fixed hyperparameter value \( h_* \). We could then consider as good candidate models those with values of \( h \) that result in the largest Bayes factors.

Suppose now that we fix a particular function \( f \) of the parameter \( \theta \); for instance, in the example, this might be the indicator that variable 1 is included in the regression model. It is of general interest to determine the posterior expectation \( E_h(f(\theta) \mid Y) \) as a function of \( h \) and to determine whether or not \( E_h(f(\theta) \mid Y) \) is very sensitive to the value of \( h \). If it is not, then two individuals using two different hyperparameters will reach approximately the same conclusions and the analysis will not be controversial. On the other hand, if for a function of interest the posterior expectation varies considerably as we change the hyperparameter, then we will want to know which aspects of the hyperparameter (e.g. which components of \( h \)) produce big changes and we may want to see a plot of the posterior expectations as we vary those aspects of the hyperparameter. Except for extremely simple cases, posterior expectations cannot be obtained in closed form, and are typically estimated via Markov chain Monte Carlo (MCMC). It is slow and inefficient to run Markov chains for every hyperparameter value \( h \). Chapter 2 reviews an existing method for estimating \( E_h(f(\theta) \mid Y) \) that bypasses the need to run a separate Markov chain for every \( h \). The method has an analogue for the problem of estimating Bayes factors. Unfortunately, the method has severe limitations, which we also discuss.

The purpose of this work is to introduce a methodology for dealing with the sensitivity analysis and model selection issues discussed above. The basic idea is—not surprisingly—to use Markov chains corresponding to a few values of the hyperparameter in order to estimate \( E_h(f(\theta) \mid Y) \) for all \( h \in \mathcal{H} \) and also the Bayes factors \( B(h, h_*) \) for all \( h \in \mathcal{H} \), and this is done through importance sampling. The difficulty we face is that there
is a severe computational burden caused by the requirement that we handle a very large number of values of $h$.

The main contributions of this work are the development of computationally efficient schemes for estimating large families of posterior expectations and Bayes factors that are based on a combination of MCMC, importance sampling, and the use of control variates, therefore providing an answer to questions (A) and (B) raised earlier. We also provide theory to support the methods we propose. Chapter 2 describes our methodology for estimating Bayes factors and posterior expectations, and gives statements of theoretical results associated with the methodology. In Chapter 3 we discuss estimation of the variance of our estimates. Chapter 4 gives a review of the relevant literature, along with a discussion of how the present work fits in the context of previous related work. In Chapter 5 we return to the problem of variable selection in Bayesian linear regression. There, first we consider a Markov chain algorithm that generates a sequence $(\gamma^{(1)}, \sigma^{(1)}, \beta_0^{(1)}, \beta_\gamma^{(1)}), (\gamma^{(2)}, \sigma^{(2)}, \beta_0^{(2)}, \beta_\gamma^{(2)}), \ldots$, describe theoretical properties of this chain, and show how to implement the methods developed in Chapter 2 to answer questions (A) and (B) posed earlier. We also illustrate our methodology on two data sets. Appendix A contains the proofs of the theorems stated in Chapter 2, Appendix B provides details regarding the generation of our Markov chain and its computational complexity, and Appendix C gives technical details regarding the theoretical properties of the Markov chain.
CHAPTER 2
ESTIMATION OF BAYES FACTORS AND POSTERIOR EXPECTATIONS

Let \( \nu_{h,y} \) denote the posterior density of \( \theta \) given \( Y = y \) when the prior is \( \nu_h \). Suppose we have a sample \( \theta_1, \ldots, \theta_n \) (iid or ergodic Markov chain output) from the posterior density \( \nu_{h_1,y} \) for a fixed \( h_1 \) and we are interested in the posterior expectation

\[
E_h(f(\theta) \mid Y = y) = \int f(\theta) \nu_{h,y}(\theta) \, d\theta
\]

for different values of the hyperparameter \( h \). We may write \( \int f(\theta) \nu_{h,y}(\theta) \, d\theta \) as

\[
\int f(\theta) \frac{p_{\theta}(y) \nu_h(\theta)}{p_{\theta}(y) \nu_{h_1}(\theta)} \nu_{h_1,y}(\theta) \, d\theta = \frac{m_{h_1}}{m_h} \int f(\theta) \frac{\nu_h(\theta)}{\nu_{h_1}(\theta)} \nu_{h_1,y}(\theta) \, d\theta \tag{2–1a}
\]

\[
= \frac{m_{h_1}}{m_h} \int \frac{\nu_h(\theta)}{\nu_{h_1}(\theta)} \frac{\nu_{h_1,y}(\theta) \, d\theta}{\nu_{h_1}(\theta)} \tag{2–1b}
\]

\[
= \frac{\int \frac{\nu_h(\theta)}{\nu_{h_1}(\theta)} \nu_{h_1,y}(\theta) \, d\theta}{\int \frac{\nu_h(\theta)}{\nu_{h_1}(\theta)} \nu_{h_1}(\theta) \, d\theta} \tag{2–1c}
\]

where in (2–1b) we have used the fact that the integral in the denominator is just 1, in order to cancel the unknown constant \( m_{h_1}/m_h \) in (2–1c). The idea to express \( \int f(\theta) \nu_{h,y}(\theta) \, d\theta \) in this way was proposed in a different context by Hastings (1970).

Expression (2–1c) is the ratio of two integrals with respect to \( \nu_{h_1,y} \), each of which may be estimated from the sequence \( \theta_1, \ldots, \theta_n \). We may estimate the numerator and the denominator by

\[
\frac{1}{n} \sum_{i=1}^{n} f(\theta_i) [\nu_h(\theta_i)/\nu_{h_1}(\theta_i)] \quad \text{and} \quad \frac{1}{n} \sum_{i=1}^{n} [\nu_h(\theta_i)/\nu_{h_1}(\theta_i)], \tag{2–2}
\]

respectively. Thus, if we let

\[
W_i^{(h)} = \frac{[\nu_h(\theta_i)/\nu_{h_1}(\theta_i)]}{\sum_{e=1}^{n} [\nu_h(\theta_e)/\nu_{h_1}(\theta_e)]},
\]

then these are weights, and we see that the desired integral may be estimated by the weighted average \( \sum_{i=1}^{n} f(\theta_i) W_i^{(h)} \).
The disappearance of the likelihood function in (2–1a) is very convenient because its computation requires considerable effort in some cases (for example, when we have missing or censored data, the likelihood is a possibly high-dimensional integral). Note that the second average in (2–2) is an estimate of \( m_h/m_{h_1} \), i.e. the Bayes factor \( B(h, h_1) \). Ideally, we would like to use the estimates in (2–2) for multiple values of \( h \) using only a sample from the posterior distribution corresponding to the fixed hyperparameter value \( h_1 \). But, when the prior \( \nu_h \) differs from \( \nu_{h_1} \) greatly, the two estimates in (2–2) are unstable because of the potential that only a few observations will dominate the sums. Their ratio suffers the same defect.

A natural approach for dealing with the instability of these simple estimates is to choose \( k \) hyperparameter values \( h_1, \ldots, h_k \in \mathcal{H} \) and to replace \( \nu_{h_1} \) with a mixture \( \sum_{s=1}^k a_s \nu_{h_s} \), where \( a_s \geq 0 \), for \( s = 1, \ldots, k \), and \( \sum_{s=1}^k a_s = 1 \). For concreteness, consider the estimate of the Bayes factor. To estimate \( B(h, h_1) \) using a sample \( \theta_1, \ldots, \theta_n \) (iid or ergodic Markov chain output) from the posterior mixture \( \nu_y := \sum_{s=1}^k a_s \nu_{h_s,y} \), we are tempted to write

\[
\frac{1}{n} \sum_{i=1}^n \frac{\nu_h(\theta_i)}{\nu_{h_1}(\theta_i)} = \frac{m_h}{m_{h_1}} \frac{1}{n} \sum_{i=1}^n \frac{p_h(y) \nu_h(\theta_i)/m_h}{\sum_{s=1}^k d_s \nu_{h_s}(\theta_i)/m_h} \tag{2–3a}
\]

\[
= B(h, h_1) \frac{1}{n} \sum_{i=1}^n \frac{\nu_{h,y}(\theta_i)}{\sum_{s=1}^k d_s \nu_{h_s,y}(\theta_i)/d_s} \tag{2–3b}
\]

where \( d_s = m_{h_s}/m_{h_1} \), \( s = 1, \ldots, k \). Thus, in order to have \( \nu_y \) in the denominator in (2–3b) (which would imply that the average in (2–3b) converges to 1, so that (2–3b) converges to \( B(h, h_1) \)), we need to start out with \( \sum_{s=1}^k a_s \nu_{h_s}/d_s \) in the denominator of the left side of (2–3a). Unfortunately, this requires the condition that we know the vector \( d = (d_2, \ldots, d_k)' \). Under this condition, if \( \theta_1, \ldots, \theta_n \) are drawn from the mixture \( \nu_{h,y} \), instead of from \( \nu_{h_1,y} \), we may form

\[
\frac{1}{n} \sum_{i=1}^n \frac{\nu_h(\theta_i)}{\sum_{s=1}^k a_s \nu_{h_s}(\theta_i)/d_s} \tag{2–4}
\]
and this quantity converges to
\[ B(h, h_1) \int \frac{\nu_{h,y}(\theta)}{\nu_y(\theta)} \nu_y(\theta) \, d\theta = B(h, h_1). \]

Assuming that for each \( l = 1, \ldots, k \) we have samples \( \theta^{(l)}_i, i = 1, \ldots, n_l \) from the posterior density \( \nu_{h,y} \), then for \( a_s = n_s/n \), the estimate in (2–4) can be written as
\[
\hat{B}(h, h_1, d) = \sum_{l=1}^k \sum_{i=1}^{n_l} \frac{\nu_{h}(\theta^{(l)}_i)}{n_s \nu_{h_s}(\theta^{(l)}_i) / d_s}.
\]
(2–5)

(Note that the combined samples \( \theta^{(l)}_i, i = 1, \ldots, n_l, l = 1, \ldots, k \) form a stratified sample from the mixture distribution \( \nu_{y,y} \).) Doss (2010) shows that under certain regularity conditions the estimate (2–5) is consistent and asymptotically normal.

In virtually all applications, the value of the vector \( d \) is unknown and has to be estimated. Doss (2010) does not deal with the case where \( d \) is unknown. In this Chapter, we assume that \( d \) is estimated via preliminary MCMC runs generated independently of the runs subsequently used to estimate \( B(h, h_1) \). Hence the sampling will consist of the following two stages.

**Stage 1** Generate samples \( \theta^{(l)}_i, i = 1, \ldots, N_l \) from \( \nu_{h,y} \), the posterior density of \( \theta \) given \( Y = y \), assuming that the prior is \( \nu_h \), for each \( l = 1, \ldots, k \), and use these \( N = \sum_{l=1}^k N_l \) observations to form an estimate of \( d \).

**Stage 2** Independently of Stage 1, again generate samples \( \theta^{(l)}_i, i = 1, \ldots, n_l \) from \( \nu_{h,y} \), for each \( l = 1, \ldots, k \), and construct the estimate of the Bayes factor \( B(h, h_1) \) based on this second set of \( n = \sum_{l=1}^k n_l \) observations and the estimate of \( d \) from Stage 1.

From now on, for \( l = 1, \ldots, k \), we use the notations \( A_l \) and \( a_l \) to identify the ratios \( N_l / N \) and \( n_l / n \), respectively.

It is natural to ask why is it necessary to have two steps of sampling, instead of estimating the vector \( d \) and \( B(h, h_1) \) from a single sample. The reason is that we are interested in estimating Bayes factors and posterior expectations for a very large number of values of \( h \), and for each \( h \), the computational time needed is linear in the
total sample size. This fact limits the total sample size and hence the accuracy of the estimates. An increase in accuracy can be achieved essentially for free by estimating \( d \) from long preliminary runs in Stage 1. The cost incurred in Stage 1 is minimal because generating the chains is typically extremely fast, and has to be done only once.

Doss (2010) also developed an improvement of (2–5) that is based on control variates, and showed that this improvement is also consistent and asymptotically normal. Unfortunately, both of these estimates require us to know the vector \( d \) exactly. One may be tempted to believe that using an estimated \( d \) instead of the true \( d \) will not inflate the asymptotic variance—indeed, the literature has errors regarding this point, and this is discussed in Appendix A. Here we provide a careful analysis of the increase in the asymptotic variance that results when we use an estimate of \( d \). A more detailed summary of the main contributions of the present work is as follows.

1. We develop a complete characterization of the asymptotic distribution of both the estimate (2–5) and the improvement that uses control variates for the realistic case where \( d \) is estimated from Stage 1 sampling (Theorems 1 and 2).

2. We develop an analogous theory for the problem of estimating a family of posterior expectations \( E_h(f(\theta) \mid Y = y), h \in \mathcal{H} \) (Theorems 3, 4, and 6).

3. We discuss estimation of the variance, and show how variance estimates can be used to guide selection of the skeleton points \( h_1, \ldots, h_k \).

4. We apply the methodology to the problem of Bayesian variable selection discussed earlier. In particular, we show how our methods enable us to select good values of \( h = (w, g) \) and to also see how the probability that a given variable is included in the regression varies with \((w, g)\).

2.1 Estimation of Bayes Factors

Here, we analyze the asymptotic distributional properties of the estimator that results if in (2–5) we replace \( d \) with an estimate. Geyer (1994) proposes an estimator for \( d \) based on the “reverse logistic regression” method and Theorem 2 therein shows that this estimator is asymptotically normal when the samplers used satisfy certain regularity conditions. This estimator is obtained by maximizing with respect to \( d_2, \ldots, d_k \) the log
quasi-likelihood

\[ l_h(d) = \sum_{l=1}^{k} \sum_{i=1}^{N_l} \log \left( \frac{A_l \nu_{h_l}(\theta_{l}^{(i)0})/d_l}{\sum_{s=1}^{k} A_s \nu_{h_s}(\theta_{l}^{(i)0})/d_s} \right). \tag{2–6} \]

The estimate is the same as the estimates obtained by Gill et al. (1988), Meng and Wong (1996), and Kong et al. (2003). We assume that for all the Markov chains we use a Strong Law of Large Numbers (SLLN) holds for all integrable functions [for sufficient conditions see, e.g., Theorem 2 of Athreya et al. (1996)]. In the next theorem, we show that if \( \hat{d} \) is the estimate produced by Geyer’s (1994) method, or any of the equivalent estimates discussed above, then the estimate of the Bayes factor given by

\[ \hat{B}(h, h_1, \hat{d}) = \sum_{l=1}^{k} \sum_{i=1}^{n_l} \frac{\nu_{h}(\theta_{l}^{(i)})}{\sum_{s=1}^{k} \nu_{h_s}(\theta_{l}^{(i)})/d_s} \] \tag{2–7}

is asymptotically normal if certain regularity conditions are met. In (2–7), \( \hat{d}_1 = 1 \).

**Theorem 1** Suppose the chains in Stage 2 satisfy conditions A1 and A2 in Doss (2010):

A1 For each \( l = 1, \ldots, k \), the chain \( \{\theta_{l}^{(i)}\}_{i=1}^{\infty} \) is geometrically ergodic.

A2 For each \( l = 1, \ldots, k \), there exists \( \epsilon > 0 \) such that

\[ E \left( \left| \sum_{s=1}^{k} \frac{\nu_{h}(\theta_{l}^{(i)})}{\hat{d}_s \nu_{h_s}(\theta_{l}^{(i)})/d_s} \right|^{2+\epsilon} \right) < \infty. \]

Assume also that the chains in Stage 1 satisfy the conditions in Theorem 2 of Geyer (1994) that imply \( \sqrt{N}(\hat{d} - d) \xrightarrow{d} \mathcal{N}(0, \Sigma) \). In addition, suppose the total sample sizes for the two stages, \( N \) and \( n \), are chosen such that \( n/N \to q \in [0, \infty) \). Then

\[ \sqrt{n}(\hat{B}(h, h_1, \hat{d}) - B(h, h_1)) \xrightarrow{d} \mathcal{N}(0, qc(h)\Sigma c(h) + \tau^2(h)), \]

where \( c(h) \) and \( \tau^2(h) \) are given in equation (A.3) in the Appendix and equation (A.9) in Doss (2010), respectively.

**Remarks**
1. There are two components to the expression for the variance. The first component arises from estimating \( d \), and the second component is the variance that we would have if we had estimated the Bayes factor knowing what \( d \) is. As can be seen from the formula, the first component vanishes if \( q = 0 \), i.e., if the sample size for estimating the parameter \( d \) converges to infinity at a faster rate than does the sample size used to estimate the Bayes factor. In this case the Bayes factor estimator (2–7) using the estimate \( \hat{d} \) has the same asymptotic distribution as the estimator in (2–5) which uses the true value of \( d \). Otherwise, the variance of (2–7) is greater than that of (2–5), and the difference between the variances depends on the magnitude of \( q \).

2. This theorem assumes the sampling is done in two independent stages: Stage 1 samples to estimate \( d \), and Stage 2 samples used together with \( \hat{d} \) to estimate the Bayes factor \( B(h, h_1) \). As a byproduct of our approach, we can get a similar theorem for the situation where both \( d \) and \( B(h, h_1) \) are estimated from the same sample. (However, for the reasons discussed earlier, ordinarily we would not use a single sample.) In more detail, if we impose the same conditions as in the theorem above on samples of total size \( n \) from a single stage, except for the condition that \( n/N \rightarrow q \in [0, \infty) \), then

\[
\sqrt{n} (\hat{B}(h, h_1, \hat{d}) - B(h, h_1)) \xrightarrow{d} \mathcal{N}(0, c(h)\Sigma c(h) + \tau^2(h) + 2c(h)'E'z),
\]

where \( \Sigma \) denotes, as in the statement of Theorem 1, the asymptotic variance of \( \sqrt{n}(\hat{d} - d) \), \( E \) is the matrix given in equation (A.51), and \( z \) is the column vector given in (A.49). A proof of this result is given in Appendix A.

2.2 Estimation of Bayes Factors Using Control Variates

Recall that we have samples \( \theta_i^{(l)} \), \( i = 1, \ldots, n_i \) from \( \nu_{h_l, y} \), \( l = 1, \ldots, k \), with independence across samples (Stage 2 of sampling) and that, based on an independent set of preliminary MCMC runs (Stage 1 of sampling), we have estimated the constants \( d_2, \ldots, d_k \). Also, \( n_i/n = a_l \) and \( n = \sum_{l=1}^{k} n_i \). Let

\[
Y(\theta) = \frac{\nu_h(\theta)}{\sum_{s=1}^{k} \frac{\nu_{h_s}(\theta)}{d_s}}.
\]
Recalling that $\nu_y := \sum_{s=1}^{k} a_s \nu_{h_s,y}$, we have $E_{\nu_y}(Y(\theta)) = B(h, h_1)$, where the subscript $\nu_y$ to the expectation indicates that $\theta \sim \nu_y$. Also, for $j = 2, \ldots, k$, let

$$Z^{(j)}(\theta) = \frac{\nu_h(\theta)/d_j - \nu_{h_1}(\theta)}{\sum_{s=1}^{k} a_s \nu_{h_s}(\theta)/d_s} = \frac{\nu_{h,y}(\theta) - \nu_{h_1,y}(\theta)}{\sum_{s=1}^{k} a_s \nu_{h_s,y}(\theta)}. \tag{2-9}$$

Expression (2–10) shows that $E_{\nu_y}(Z^{(j)}(\theta)) = 0$. This is true even if the priors $\nu_{h_1}$ and $\nu_{h_1}$ are improper, as long as the posteriors $\nu_{h,y}$ and $\nu_{h_1,y}$ are proper, exactly our situation in the Bayesian variable selection example of Chapter 1. On the other hand, the representation (2–9) shows that $Z^{(j)}(\theta)$ is computable if we know the $d_j$’s—it involves the priors and not the posteriors. (A similar remark applies to (2–8).) Therefore, if as in Doss (2010) we define for $l = 1, \ldots, k$, $i = 1, \ldots, n_l$

$$Y_{i,l} = \frac{\nu_h(\theta^{(l)}_i)}{\sum_{s=1}^{k} a_s \nu_{h_s}(\theta^{(l)}_i)/d_s}, \quad Z^{(1)}_{i,l} = 1, \quad Z^{(j)}_{i,l} = \frac{\nu_h(\theta^{(l)}_i)/d_j - \nu_{h_1}(\theta^{(l)}_i)}{\sum_{s=1}^{k} a_s \nu_{h_s}(\theta^{(l)}_i)/d_s}, j = 2, \ldots, k, \tag{2–11}$$

then for any fixed $\beta = (\beta_2, \ldots, \beta_k)$,

$$\hat{\gamma}_{\beta}^d = \frac{1}{n} \sum_{l=1}^{k} \sum_{i=1}^{n_l} (Y_{i,l} - \sum_{j=2}^{k} \beta_j Z^{(j)}_{i,l}), \tag{2–12}$$

is an unbiased estimate of $B(h, h_1)$. The value of $\beta$ that minimizes the variance of $\hat{\gamma}_{\beta}^d$ is unknown. As is commonly done when one uses control variates, we use instead the estimate obtained by doing ordinary linear regression of the response $Y_{i,l}$ on the predictors $Z^{(j)}_{i,l}$, $j = 2, \ldots, k$, and to emphasize that this estimate depends on $d$, we denote it by $\hat{\beta}(d)$. Theorem 1 of Doss (2010) states that the estimator $\hat{B}_{\text{reg}}(h, h_1) = \hat{\gamma}_{\hat{\beta}(d)}^d$, obtained under the assumption that we know the constants $d_2, \ldots, d_k$, has an asymptotically normal distribution. As mentioned earlier, $d_2, \ldots, d_k$ are typically unknown, and must be estimated. Let $\hat{d}_2, \ldots, \hat{d}_k$ be estimates obtained from previous MCMC runs.
and let
\[
\gamma_{\hat{d}(\hat{d})} = \frac{1}{n} \sum_{i=1}^{k} \sum_{l=1}^{n_l} \left( \hat{Y}_{i,l} - \sum_{j=2}^{k} \hat{\beta}_j(\hat{d}) \hat{Z}_{i,l}^{(j)} \right),
\]  
\[(2-13)\]
where \(\hat{Y}_{i,l}\) and \(\hat{Z}_{i,l}^{(j)}\) are like in (2–11), except using \(\hat{d}\) for \(d\), and \(\hat{\beta}(\hat{d})\) is the least squares regression estimator from regressing \(\hat{Y}_{i,l}\) on predictors \(\hat{Z}_{i,l}^{(j)}\), \(j = 2, \ldots, k\). The next theorem gives the asymptotic distribution of this new estimator.

**Theorem 2** Suppose all the conditions from Theorem 1 are satisfied. Moreover, assume that \(R\), the \(k \times k\) matrix defined by
\[
R_{j,j'} = E\left( \sum_{l=1}^{k} a_l Z_{1,l}^{(j)} Z_{1,l}^{(j')} \right), \quad j, j' = 1, \ldots, k,
\]
is nonsingular. Then
\[
\sqrt{n} \left( \gamma_{\hat{d}(\hat{d})} - \mathcal{B}(h, h_1) \right) \xrightarrow{d} \mathcal{N}(0, qw(h)' \Sigma w(h) + \sigma^2(h)).
\]
Expressions for \(w(h)\) and \(\sigma^2(h)\) are given in equation (A.18) to follow and equation (A.7) in Doss (2010), respectively.

### 2.3 Estimation of Posterior Expectations

In this section we give a method for estimating the posterior expectation of a function \(f\) when the prior is \(\nu_h\). Let us denote this quantity by
\[
[f](h) = \int f(\theta) \nu_{h,y}(\theta) d\theta.
\]
Define
\[
Y_{i,l}^{[f]} = \frac{f(\theta_i^{(l)}) \nu_h(\theta_i^{(l)})}{\sum_{s=1}^{k} a_s \nu_h(\theta_i^{(s)}) / d_s} = \frac{f(\theta_i^{(l)}) \nu_h(\theta_i^{(l)}) / m_h}{\sum_{s=1}^{k} a_s \nu_h(\theta_i^{(s)}) / m_h} \cdot \frac{m_h}{m_h_1} = \frac{f(\theta_i^{(l)}) \nu_{h,y}(\theta_i^{(l)})}{\sum_{s=1}^{k} a_s \nu_{h_1,y}(\theta_i^{(s)})} \mathcal{B}(h, h_1).
\]
Assuming a SLLN holds for the Markov chains \(\theta_i^{(l)}\), \(l = 1, \ldots, k\), \(i = 1, \ldots, n_i\), we have
\[
\frac{1}{n_i} \sum_{i=1}^{n_i} Y_{i,l}^{[f]} \xrightarrow{a.s.} \int \frac{f(\theta) \nu_{h,y}(\theta)}{\sum_{s=1}^{k} a_s \nu_{h_1,y}(\theta)} \nu_{h,y}(\theta) d\theta \cdot \mathcal{B}(h, h_1).
\]
Therefore,
\[
\frac{1}{n} \sum_{l=1}^{k} \sum_{i=1}^{n_l} Y_{i,l}^{[r]} = \sum_{l=1}^{k} \sum_{i=1}^{n_l} \frac{1}{n_l} n_l Y_{i,l}^{[r]}
\]
\[
\xrightarrow{a.s.} \int \frac{f(\theta)\nu_{h,y}(\theta)}{\sum_{s=1}^{k} a_s^s \nu_{h_s,y}(\theta)} \sum_{l=1}^{k} a_l \nu_{h_i,y}(\theta) d\theta \cdot B(h, h_1)
\]  
(2–14)

\[
= [\ell](h) \cdot B(h, h_1).
\]

Similarly, we have
\[
\frac{1}{n} \sum_{l=1}^{k} \sum_{i=1}^{n_l} Y_{i,l} \xrightarrow{a.s.} B(h, h_1)
\]
(the \(Y_{i,l}\)'s are defined in (2–11)).

Note that \(Y_{i,l} = Y_{i,l}^{[r]} \) when \(f \equiv 1\). Letting
\[
\hat{\ell}^{[r]}(h, d) = \frac{\sum_{l=1}^{k} \sum_{i=1}^{n_l} Y_{i,l}^{[r]}}{\sum_{i=1}^{k} \sum_{l=1}^{n_l} Y_{i,l}},
\]  
(2–15)

we see that \(\hat{\ell}^{[r]}(h, d) \xrightarrow{a.s.} \ell^{[r]}(h)\). Replacing the unknown \(d\) with an estimate \(\hat{d}\) obtained from Stage 1 sampling, we form the estimator
\[
\hat{\ell}^{[r]}(h, \hat{d}) = \frac{\sum_{l=1}^{k} \sum_{i=1}^{n_l} f(\theta_{i}^{(l)}) \nu_{h}(\theta_{i}^{(l)})}{\sum_{s=1}^{k} \sum_{i=1}^{n_l} \nu_{h_s}(\theta_{i}^{(l)}) / \hat{a}_s}.
\]  
(2–16)

It is the asymptotic behavior of this estimator that we are concerned with in the following theorem.

**Theorem 3** Suppose the conditions stated in Theorem 1 are satisfied and, in addition, for each \(l = 1, \ldots, k\), there exists an \(\epsilon > 0\) such that
\[
E\left(\left| Y_{1,l}^{[r]} \right|^{2+\epsilon}\right) < \infty.
\]  
(2–17)

Then
\[
\sqrt{n}(\hat{\ell}^{[r]}(h, \hat{d}) - \ell^{[r]}(h)) \xrightarrow{d} \mathcal{N}(0, q \nu(h) \Sigma \nu(h) + \rho(h)).
\]
Expressions for \( v(h) \) and \( \rho(h) \) are given in equations (A.22) and (A.21), respectively, in the Appendix.

2.4 Estimation of Posterior Expectations Using Control Variates

We assume in this section that the values of the vector \( d \) and the posterior expectations \( I[f_i(h)] \), for \( j = 1, \ldots, k \), are available to us. In reality, these quantities are seldom known, and the next section deals with the case when they are estimated based on previous MCMC runs. Recall that the integral we want to estimate is

\[
I[f](h) = \int f(\theta)\nu_{h,y}(\theta)\,d\theta.
\]

In (2–14) we established that \( (1/n) \sum_{l=1}^{k} \sum_{i=1}^{n_i} Y_{i,l}^{[r]} \) is a strongly consistent estimator of \( I[f](h) \cdot B(h, h_1) \). Define

\[
Z_{i,l}^{[r][0]} = 1, \quad Z_{i,l}^{[r][j]} = \frac{f(\theta_i^{(l)})\nu_h(\theta_i^{(l)})/d_j}{\sum_{s=1}^{k} a_s\nu_h(\theta_i^{(s)})/d_s} - I[f](h_j), \quad j = 1, \ldots, k,
\]

and let

\[
Z^{[r][j]}(\theta) = \frac{f(\theta)\nu_h(\theta)/d_j}{\sum_{s=1}^{k} a_s\nu_h(\theta)/d_s} - I[f](h_j), \quad j = 1, \ldots, k.
\]

With \( \nu_{h,y} \) denoting the mixture distribution \( \sum_{s=1}^{k} a_s\nu_{h,s,y} \), it can be easily checked that

\[
E_{\nu_{h,y}}(Z^{[r][j]}(\theta)) = 0, \quad \text{for } j = 1, \ldots, k,
\]

so we can use the \( Z^{[r][j]} \)'s as control variates to reduce the variance of the original estimator \( (1/n) \sum_{l=1}^{k} \sum_{i=1}^{n_i} Y_{i,l}^{[r]} \). Doing so gives the estimator

\[
\frac{1}{n} \sum_{l=1}^{k} \sum_{i=1}^{n_i} \left( Y_{i,l}^{[r]} - \sum_{j=1}^{k} \beta_j^{[r]} Z_{i,l}^{[r][j]} \right),
\]

where \( \beta_j^{[r]} \)'s denote the least squares estimates resulting from the regression of \( Y_{i,l}^{[r]} \) on predictors \( Z_{i,l}^{[r][j]} \). The Bayes factor \( B(h, h_1) \) will be estimated as before in Section 2.2, using the estimator \( (1/n) \sum_{l=1}^{k} \sum_{i=1}^{n_i} Y_{i,l} \) and the \( k - 1 \) control variates \( Z^{(j)} \), for \( j = 2, \ldots, k \). The ratio of these two control variate adjusted estimators provides us with an
improved estimator for the posterior expectation $I^{[r]}(h)$, which is given by

$$
\hat{I}_{\hat{\beta}, \hat{\beta}^{[r]}} = \frac{\sum_{l=1}^{k} \sum_{i=1}^{n_l} \left( Y_{i,l} - \sum_{j=1}^{k} \hat{\beta}_j^{[r]} Z_{i,l}^{[r](j)} \right)}{\sum_{l=1}^{k} \sum_{i=1}^{n_l} \left( Y_{i,l} - \sum_{j=1}^{k} \hat{\beta}_j Z_{i,l}^{(j)} \right)}. \quad (2-18)
$$

**Theorem 4** Suppose conditions A1 and A2 stated in Theorem 1 are satisfied and the matrix $R$ defined in Theorem 2 is nonsingular. Also, suppose that

A3 for each $l = 1, \ldots, k$, there exists $\epsilon > 0$ such that $E \left( Y_{1,l}^{[r]} \right)^{2+\epsilon} < \infty$;

A4 for each $l = 1, \ldots, k$, there exists $\epsilon > 0$ such that $E_{\nu_h}(|f^{2+\epsilon}(\theta)|) < \infty$;

A5 for each $l = 1, \ldots, k$,

$$
E_{\nu_{h,l}} \left( \frac{\nu_h(\theta)}{\sum_{s=1}^{k} d_s \nu_h(\theta) / d_s} \right) < \infty;
$$

A6 the $(k + 1) \times (k + 1)$ matrix $R^{[r]}$ defined by

$$
R_{j+1,j'+1}^{[r]} = E \left( \sum_{l=1}^{k} a_l Z_{1,l}^{[r](j)} Z_{1,l}^{[r](j')} \right), \quad j, j' = 0, \ldots, k,
$$

is nonsingular.

Then

$$
\sqrt{n} \left( \hat{I}_{\hat{\beta}, \hat{\beta}^{[r]}} - I^{[r]}(h) \right) \xrightarrow{d} N(0, r(h)),
$$

where $r(h)$ is given in equation (A.33) of the Appendix.

**Remarks**

1. If $h = h_j$ for some $j \in \{1, \ldots, k\}$, our estimator of posterior expectation $\hat{I}_{\hat{\beta}, \hat{\beta}^{[r]}}$ given above in (2–18) has zero variance. To see why, note that in this case the response $Y^{[r]}$ can be written as

$$
Y^{[r]} = d_j I^{[r]}(h_j) + d_j Z^{[r](j)},
$$

so there is no noise in the regression of $Y^{[r]}$ on predictors $Z^{[r](j)}$’s, and as a consequence, the numerator of this estimator is constant (specifically, $nd_j I^{[r]}(h_j)$). Through similar arguments, the denominator was shown to be constant ($nd_j$) in Doss (2010). Hence, for $h = h_j$, $\hat{I}_{\hat{\beta}, \hat{\beta}^{[r]}}$ is a perfect estimator of $I^{[r]}(h_j)$.

2. Theorem 4 pertains to the case where $d$ and the posterior expectations $I^{[r]}(h_j), j = 1, \ldots, k$ are known. There do exist some situations where this is the case. For example, in the hierarchical model for Bayesian linear regression discussed in
Chapter 1, for each \(j = 1, \ldots, k\), the marginal likelihood \(m_h\) is a sum of \(2^q\) terms, and if \(q\) is relatively small, these marginal likelihoods are computable, so the vector \(d\) is available. Likewise, for some functions \(f\), the posterior expectations \(I^r(h)\) can be numerically obtained; see Section 3 of George and Foster (2000). So it is possible to calculate \(d_I\) for \(I^r(h)\), \(j = 1, \ldots, k\) for skeleton points \(h_1, \ldots, h_k\), and the method described in this section enables us to efficiently estimate the family \(I^r(h), h \in \mathcal{H}\).

2.5 Estimation of Posterior Expectations Using Control Variates With Estimated Skeleton Bayes Factors and Expectations

This section undertakes estimation of the posterior expectation \(I^r(h) = \int f(\theta) \nu_h(y, \theta) d\theta\) in the case where the quantities \(d\) and \(I^r(h_j)\)’s are unknown and estimated based on previous MCMC runs. Let

\[
e = (I^r(h_1), \ldots, I^r(h_k))^\top \quad \text{and} \quad \hat{e} = \left(\frac{1}{N_1} \sum_{i=1}^{N_1} f(\theta_i^{(1)}), \ldots, \frac{1}{N_k} \sum_{i=1}^{N_k} f(\theta_i^{(k)})\right)^\top,
\]

i.e. \(e\) is the vector of true expectations, which had been assumed known in Theorem 4, and \(\hat{e}\) is its natural estimate based on the samples in Stage 1. To account for the fact that the responses \(Y_{i,l}^r\) and covariates \(Z_{i,l}^{r(j)}\) used in the previous section are now unknown, as they involve the unknown \(d\) and \(e\), we need to consider new responses and covariates based on estimates \(\hat{d}\) and \(\hat{e}\) obtained from Stage 1 samples. Define

\[
\hat{Y}_{i,l}^r = \frac{f(\theta_i^{(l)}) \nu_h(\theta_i^{(l)})}{\sum_{s=1}^k a_s \nu_h(\theta_i^{(l)}) / \hat{d}_s} \quad \text{and} \quad \hat{Z}_{i,l}^{r(j)} = \frac{f(\theta_i^{(l)}) \nu_h(\theta_i^{(l)}) / \hat{d}_j}{\sum_{s=1}^k a_s \nu_h(\theta_i^{(l)}) / \hat{d}_s} - \hat{d}_j, \quad j = 1, \ldots, k.
\]

Hence a new control variate adjusted estimator for \(I^r(h)\) corresponding to the estimator (2–18) of the previous section, which assumed knowledge of \(d\) and \(I^r(h_j)\)’s, is

\[
\hat{\gamma}_{\alpha(\hat{d}), \beta(\hat{d}), \hat{r}(\hat{d})} = \frac{\hat{Y}_{i,l}^r - \sum_{j=1}^k [\beta^r(\hat{d})]_{j,l} \hat{Z}_{i,l}^{r(j)}}{\sum_{l=1}^{\eta_l} \sum_{j=1}^k [\beta^r(\hat{d})]_{j,l} \hat{Z}_{i,l}^{r(j)}} - \frac{\sum_{l=1}^{\eta_l} \sum_{j=1}^k [\beta^r(\hat{d})]_{j,l} \hat{Z}_{i,l}^{r(j)}}{\sum_{l=1}^{\eta_l} \sum_{j=1}^k [\beta^r(\hat{d})]_{j,l} \hat{Z}_{i,l}^{r(j)}} + \frac{\hat{Z}_{i,l}^{r(j)}}{\sum_{l=1}^{\eta_l} \sum_{j=1}^k [\beta^r(\hat{d})]_{j,l} \hat{Z}_{i,l}^{r(j)}} + \frac{\hat{Z}_{i,l}^{r(j)}}{\sum_{l=1}^{\eta_l} \sum_{j=1}^k [\beta^r(\hat{d})]_{j,l} \hat{Z}_{i,l}^{r(j)}}
\]

where \(\beta^r(\hat{d})\) is the least squares estimate resulting from the regression of \(\hat{Y}_{i,l}^r\) on predictors \(\hat{Z}_{i,l}^{r(j)}\). Theorem 6 establishes the asymptotic normality of this estimator.

But before stating this theorem, we give an auxiliary theorem which shows the joint
asymptotic normality of
\[ \sqrt{N} \left( \begin{pmatrix} \hat{d} \\ \hat{e} \end{pmatrix} - \begin{pmatrix} d \\ e \end{pmatrix} \right). \]

Let us first define separable and inseparable Monte Carlo samples as introduced by Geyer (1994). The Monte Carlo sample \( \{\theta^{(l)}_i\}_{i=1}^\infty, \, l = 1, \ldots, k \) is said to be separable if there are disjoints subsets \( L \) and \( M \) of \( \{1, \ldots, k\} \) such that for each \( \theta \) in the sample and each \( l \in L \) and \( m \in M \) either \( \nu_{h,\gamma}(\theta) \) or \( \nu_{h_m,\gamma}(\theta) \) are zero. A Monte Carlo sample that is not separable is said to be inseparable.

**Theorem 5** Assume that the Monte Carlo sample from Stage 1, \( \{\theta^{(l)}_i\}_{i=1}^\infty, \, l = 1, \ldots, k \), is inseparable, and the following conditions hold:

- **B1** for each \( l = 1, \ldots, k \), the chain \( \{\theta^{(l)}_i\}_{i=1}^\infty \) is geometrically ergodic
- **B2** for each \( l = 1, \ldots, k \), there exists \( \epsilon > 0 \) such that \( E_{\nu_{h,\gamma}}(|f|^{2+\epsilon}(\theta)) < \infty \).

Then
\[ \sqrt{N} \left( \begin{pmatrix} \hat{d} \\ \hat{e} \end{pmatrix} - \begin{pmatrix} d \\ e \end{pmatrix} \right) \xrightarrow{d} \mathcal{N}(0, V), \tag{2-19} \]

where \( V \) is given in equation (A.48) in the Appendix.

**Theorem 6** If the conditions stated in Theorem 4 and (2–19) hold, then
\[ \sqrt{n} \left( \begin{pmatrix} \hat{d} \\ \hat{e} \end{pmatrix}_{\bar{\beta}(\bar{d}), \bar{\beta}(\bar{d})} - \bar{f}(h) \right) \xrightarrow{d} \mathcal{N}(0, \psi(h)), \]

with \( \psi(h) \) given in equation (A.57) in the Appendix.
CHAPTER 3
VARIANCE ESTIMATION AND SELECTION OF THE SKELETON POINTS

Estimation of the variance of our estimates is important for several reasons. In addition to the usual need for providing error margins for our point estimates, variance estimates are of great help in selecting the skeleton points.

3.1 Estimation of the Variance

There are two approaches one can use to estimate the variance of any of our estimates. For the sake of concreteness, consider $\hat{B}(h, h_1, \hat{d})$, whose asymptotic variance is the expression $\kappa^2(h) = qc(h)\Sigma c(h) + \tau^2(h)$ (see Theorem 1).

**Spectral Methods** If $X_0, X_1, X_2, \ldots$ is a Markov chain and $f$ is a function, the asymptotic variance of $(1/n) \sum_{i=0}^{n-1} f(X_i)$ (when it exists) is the infinite series

$$\text{Var}(f(X_0)) + 2 \sum_{j=1}^{\infty} \text{Cov}(f(X_0), f(X_j))$$

where the variances and covariances are calculated under the assumption that $X_0$ has the stationary distribution. Spectral methods involve estimating an initial segment of the series, using techniques from time series; see Geyer (1992) for a review. Our problem is more complicated because we are dealing with multiple chains. In our situation, the term $\tau^2(h)$ may be estimated through spectral methods, and this is done in a straightforward manner. We now give technical details regarding the consistency of this method. The quantity $\tau^2(h)$ is given by $\tau^2(h) = \sum_{l=1}^{k} \tilde{a}_l \tilde{\tau}_l^2(h)$, where $\tilde{\tau}_l^2(h)$ is the asymptotic variance of

$$
\frac{1}{n_l} \sum_{i=1}^{n_l} \frac{\nu_{h_l}(\theta_i^{(l)})}{\sum_{s=1}^{k} \tilde{a}_s \nu_{h_s}(\theta_i^{(l)})/d_s}.
$$

(See equation (A.9) of Doss (2010).) Because for each $l$ we will be estimating $\tilde{\tau}_l^2(h)$ by the asymptotic variance of

$$
\frac{1}{n_l} \sum_{i=1}^{n_l} \frac{\nu_{h_l}(\theta_i^{(l)})}{\sum_{s=1}^{k} \tilde{a}_s \nu_{h_s}(\theta_i^{(l)})/d_s}.
$$
where \( \hat{d} \) is formed from Stage 1 runs, it is necessary to consider the quantity \( \tau^2(h, u) \), defined as the asymptotic variance of

\[
\tau^2(h, u) = \frac{1}{n} \sum_{i=1}^{n} \frac{\nu_h(\theta^{(i)}_i)}{\sum_{s=1}^{k} a_s \nu_s(\theta^{(i)}_s)/u_s},
\]

where \( u = (u_1, u_2, \ldots, u_k)' \). After defining

\[
f_u(\theta) = \frac{\nu_h(\theta)}{\sum_{s=1}^{k} a_s \nu_s(\theta)/u_s},
\]

we get

\[
\tau^2(h, u) = \text{Var}(f_u(\theta^{(1)}_1)) + 2 \sum_{g=1}^{\infty} \text{Cov}(f_u(\theta^{(1)}_1), f_u(\theta^{(1)}_{1+j})).
\]

We now proceed to establish continuity of \( \tau^2(h, u) \) in \( u \), and to do this we will show that for each \( l = 1, \ldots, k \), \( \tau^2(h, u) \) is continuous in \( u \). For the rest of this discussion expectations and variances are taken with respect to \( \nu_{h,y} \), and we drop \( l \) from the notation. Let \( u^{(n)} \) be any sequence of vectors such that \( u^{(n)} \to d \). Then trivially \( f_{u^{(n)}}(\theta) \to f_u(\theta) \) for all \( \theta \), and letting \( \epsilon = \min\{d_1, \ldots, d_k\} \), there exists a positive integer \( n(\epsilon) \) such that \( \|u^{(n)} - d\| \leq \epsilon \) for all \( n \geq n(\epsilon) \). Consequently,

\[
f_{u^{(n)}}(\theta) \leq f_{2d}(\theta) = 2f_d(\theta) \quad \text{for all } \theta \text{ and all } n \geq n(\epsilon) \quad (3–2)
\]

and we can apply the Lebesgue Dominated Convergence Theorem twice to conclude that

\[
\text{Var}(f_{u^{(n)}}(\theta_1)) = E(f_{u^{(n)}}^2(\theta_1)) - [E(f_{u^{(n)}}(\theta_1))]^2
\]

converges to

\[
E(f_{d}^2(\theta_1)) - [E(f_{d}(\theta_1))]^2 = \text{Var}(f_{d}(\theta_1)).
\]

Note that condition A2 guarantees that the dominating function in (3–2) has finite expectation. Similarly, for each of the covariance terms,

\[
\text{Cov}(f_{u^{(n)}}(\theta_1), f_{u^{(n)}}(\theta_{1+j})) \to \text{Cov}(f_{d}(\theta_1), f_{d}(\theta_{1+j})).
\]
If we define
\[ c_u(j) = \text{Cov}(f_u(\theta_1), f_u(\theta_1+j)), \quad j = 1, 2, \ldots, \]
then \( \sum_{j=1}^{\infty} c_u(j) \) is absolutely convergent. This is because under geometric ergodicity, the so-called strong mixing coefficients \( \alpha(j) \) decrease to 0 exponentially fast (a definition of strong mixing is given on p. 349 of Ibragimov (1962)), and
\[ \text{Cov}(f_u(\theta_1), f_u(\theta_1+j)) \leq [\alpha(j)]^\beta [E(|f_u(\theta_1)|^{2+\epsilon})]^{2/(2+\epsilon)}, \tag{3–3} \]
for some \( \beta > 0 \). See Theorem 18.5.3 of Ibragimov and Linnik (1971) or Lemma 7.7 of Chapter 7 in Durrett (1991). Since \( c_{u^n}(j) \to c_u(j) \) for each \( j \), (3–2) and (3–3) enable us to again apply Dominated Convergence to conclude that \( \sum_{j=1}^{\infty} c_{u^n}(j) \to \sum_{j=1}^{\infty} c_u(j) \), and this proves that \( \tau^2(h, u) \) is continuous in \( u \).

Let \( g(u) \) be the spectral density at 0 of the series \( f_u(\theta_i) \). Note that \( g(u) \) is equal to \( \tau^2(h, u) \), except for a normalizing constant. Under strong mixing (implied by geometric ergodicity), standard spectral density estimates \( \hat{g}(u) \) are consistent, and bounds on the discrepancy \( |\hat{g}(u) - g(u)| \) depend on the mixing rate and bounds on the moments of the function \( f_u(\theta) \) (Rosenblatt 1984). By (3–2), the rate is uniform as long as \( \|u - d\| \) is small, and the condition that \( \|\hat{d} - d\| \) is small is guaranteed if the Stage 1 sample size \( N \) is large.

Geyer (1994) gives an expression for \( \Sigma \) involving infinite series of the form (3–1), and this enables estimation of \( \Sigma \) by spectral methods. Now, \( c(h) \) is a vector each of whose components is an integral with respect to the posterior \( \nu_{h,y} \) (see (A.3)). The estimate derived in Section 2.3 (see (2–16)) is designed precisely to estimate such posterior expectations. Combining, we arrive at an overall estimate of \( \kappa^2(h) \), and the asymptotic variances of our other estimates are handled similarly.

**Methods Based on Regeneration** The cleanest approach to estimating asymptotic variances is based on regeneration. Let \( X_0, X_1, X_2, \ldots \) be a Markov chain on the measurable space \( (\mathcal{X}, \mathcal{B}) \), let \( K(x, A) \) be the Markov transition distribution, and assume...
that \( \pi \) is a stationary probability distribution for the chain. Suppose that for each \( x \),
\[ K(x, \cdot) \]
has density \( k(x, \cdot) \) with respect to a dominating measure \( \mu \). Regeneration methods require the existence of a function \( s: \mathcal{X} \rightarrow [0, 1) \), whose expectation with respect to \( \pi \) is strictly positive, and a probability density \( d \) with respect to \( \mu \), such that
\[ k(\cdot, \cdot) \]
satisfies
\[ k(x, x') \geq s(x)d(x') \quad \text{for all } x, x' \in \mathcal{X}. \tag{3–4} \]
This is called a minorization condition and, as we describe below, it can be used to introduce regenerations into the Markov chain driven by \( k \). These regenerations are the key to constructing a simple, consistent estimator of the variance in the central limit theorem. Define
\[ r(x, x') = \frac{k(x, x') - s(x)d(x')}{1 - s(x)}. \]
Note that, for fixed \( x \in \mathcal{X} \), \( r(x, x') \) is a density function in \( x' \). We may therefore write
\[ k(x, x') = s(x)d(x') + (1 - s(x))r(x, x'), \]
which gives a representation of \( k(x, \cdot) \) as a mixture of two densities, \( d(\cdot) \) and \( r(x, \cdot) \).
This provides an alternative method of simulating from \( k \). Suppose that the current state of the chain is \( X_n \). We generate \( \delta_n \sim \text{Bernoulli}(s(X_n)) \). If \( \delta_n = 1 \), we draw \( X_{n+1} \sim d \); otherwise, we draw \( X_{n+1} \sim r(X_n, \cdot) \). Note that, if \( \delta_n = 1 \), the next state of the chain is drawn from \( d \), which does not depend on the current state. Hence, the chain “forgets” the current state and we have a regeneration. To be more specific, suppose we start the Markov chain with \( X_0 \sim d \) and then use the method described above to simulate the chain. Each time \( \delta_n = 1 \), we have \( X_{n+1} \sim d \) and the process stochastically restarts itself; that is, the process regenerates. Even though the description above involves generating observations from \( r \), there are clever tricks that enable the user to bypass generating from \( r \), and to obtain the sequence \((X_n, \delta_n)\) in a way that requires only generating directly from \( k \); see, e.g., Tan and Hobert (2009).
Here is how the regenerative method is used to get valid asymptotic standard errors. Suppose we wish to approximate the posterior expectation of some function $f(X)$. Suppose further that the Markov chain is to be run for $R$ regenerations (or tours); that is, we begin by drawing the starting value from $d$ and we stop the simulation the $R$th time that a $\delta_n = 1$. Let $0 = \tau_0 < \tau_1 < \tau_2 < \cdots < \tau_R$ be the (random) regeneration times, i.e. $\tau_t = \min\{n > \tau_{t-1} : \delta_{n-1} = 1\}$ for $t \in \{1, 2, \ldots, R\}$. The total length of the simulation, $\tau_R$, is random. Let $N_1, N_2, \ldots, N_R$ be the lengths of the tours, i.e. $N_t = \tau_t - \tau_{t-1}$, and define $S_t = \sum_{n=\tau_{t-1}}^{\tau_t-1} f(X_n), t = 1, \ldots, R$. Note that the $(N_t, S_t)$ pairs are iid, and a strongly consistent estimator of $E_{\pi}(f(X))$ is $\bar{T}_{\tau_R} = \bar{S}/\bar{N} = (1/\tau_R) \sum_{n=0}^{\tau_R-1} f(X_n)$, where $\bar{S} = (1/R) \sum_{t=1}^{R} S_t$ and $\bar{N} = (1/R) \sum_{t=1}^{R} N_t$, and the asymptotic variance of $\bar{T}_{\tau_R}$ may be estimated very simply by $\sum_{t=1}^{R} (S_t - \bar{T}_{\tau_R} N_t)^2/(R\bar{N}^2)$. Moment and ergodicity conditions that guarantee strong consistency of this variance estimator are given in Hobert et al. (2002). This method has recently been applied successfully in a number of problems involving continuous state spaces; see, e.g., Tan and Hobert (2009) and the references therein, and we use the method in the illustration in Chapter 5.

In our framework of multiple chains, one might think that we need to identify a sequence of times $0 = \tau_0 < \tau_1 < \tau_2 < \cdots < \tau_R$ at which all the chains regenerate. This is not the case, and we need only identify, for each chain, a sequence of regeneration times for that chain. Since the overall estimate is essentially a function of averages involving the $k$ chains, its asymptotic variance is a function of the asymptotic variances of averages formed from the individual chains.

Consider the function $B(h, h_1); h \in \mathcal{H}$, and an estimator, such as $\hat{B}(h, h_1, \hat{d})$ (for the rest of this discussion, we will denote these by $B(h)$ and $\hat{B}(h)$, for brevity). It is of interest to provide a confidence band (region, if $h$ is multidimensional) for $B(h)$ that is valid simultaneously for all $h \in \mathcal{H}$. A closely related problem is to produce a confidence interval for $\arg\max_{h \in \mathcal{H}} B(h)$. The traditional way of forming confidence bands that are valid globally is to proceed as follows:
1 Establish a functional central limit theorem that says that $n^{1/2}(\hat{B}(h) - B(h))$ converges in distribution to a Gaussian process $W(h); h \in \mathcal{H}$.

2 Find the distribution of $\sup_{h \in \mathcal{H}} |W(h)|$.

If $s_\alpha$ is the $(1 - \alpha)$-quantile of the distribution of this supremum, then the band $\hat{B}(h) \pm s_\alpha/n^{1/2}$ has asymptotic coverage probability equal to $1 - \alpha$. The value $s_\alpha$ is typically too difficult to compute analytically, but can be obtained by simulation [see, e.g. Burr and Doss (1993) among many others]. The maximal inequalities needed to establish functional central limit theorems typically require an iid structure, and for this reason we believe that the regeneration method offers the best hope for establishing such theorems.

### 3.2 Selection of the Skeleton Points

The asymptotic variances of any of our estimates depend on the choice of the points $h_1, \ldots, h_k$. For concreteness, consider $\hat{B}(h, h_1, \hat{d})$, and to emphasize this dependence, let $V(h, h_1, \ldots, h_k)$ denote the asymptotic variance of $\hat{B}(h, h_1, \hat{d})$. For fixed $h_1, \ldots, h_k$, identifying the set of $h$'s for which $V(h, h_1, \ldots, h_k)$ is finite is typically a feasible problem. For instance, Doss (1994) considered the pump data example discussed in Tierney (1994), for which the hyperparameter $h$ has dimension 3, and determined this set for the case $k = 1$. He showed that one can go as far away from $h_1$ as one wants in certain directions, but in other directions the range is limited. (The calculation can be extended to any $k$.) Suppose now that we fix a range $\mathcal{H}$ over which $h$ is to vary. Typically, we will want more than just a positioning of $h_1, \ldots, h_k$ that guarantee that $V(h, h_1, \ldots, h_k)$ is finite for all $h \in \mathcal{H}$, and we will face the problem below.

**Design Problem** Find the values of $h_1, \ldots, h_k$ that minimize $\max_{h \in \mathcal{H}} V(h, h_1, \ldots, h_k)$.

Unfortunately, except for extremely simple cases, it is not possible to calculate $V(h, h_1, \ldots, h_k)$ analytically (even if $k = 1$, $V(h, h_1)$ is an infinite sum each of whose terms depends on the Markov transition distribution in a complicated way), and maximizing it over $h \in \mathcal{H}$ would present additional difficulties. Furthermore, even if
we were able to calculate $\max_{h \in \mathcal{H}} V(h, h_1, \ldots, h_k)$, the design problem would involve the minimization of a function of $k \times \dim(\mathcal{H})$ variables, and in general, solving the design problem is hopeless.

In our experience, we have found that the following method works reasonably well. Having specified the range $\mathcal{H}$, we select trial values $h_1, \ldots, h_k$ and plot the estimated variance as a function of $h$, using one of the methods described above. If we find a region in $\mathcal{H}$ where this variance is unacceptably large, we “cover” this region by moving some $h_i$’s closer to the region, or by simply adding new $h_i$’s in that region, which increases $k$. This is illustrated in the example in Chapter 5.
CHAPTER 4
REVIEW OF PREVIOUS WORK

Vardi (1985) introduced the following \( k \)-sample model for biased sampling. There is an unknown distribution function \( F \), which we wish to estimate. For each weight function \( w_l, l = 1, \ldots, k \), we have a sample \( X_{i1}, \ldots, X_{i_n} \overset{\text{iid}}{\sim} F_l \), where

\[
F_l(x) = \frac{1}{W_l} \int_{-\infty}^{x} w_l(s) dF(s). \tag{4–1}
\]

In (4–1), \( W_l = \int_{-\infty}^{\infty} w_l(s) dF(s) \). The weight functions \( w_1, \ldots, w_k \) are known, but the normalizing constants \( W_1, \ldots, W_k \) are not. Vardi (1985) was interested in conditions that guarantee that a nonparametric maximum likelihood estimator (NPMLE) exists and is unique, and he gave the form of the NPMLE. (The conditions for existence and uniqueness involve issues regarding the supports of the \( F_l \)'s and do not concern us in the present paper.)

To estimate \( F \), a preliminary step is to estimate the vector \( (W_1, \ldots, W_k) \). Vardi (1985) and Gill et al. (1988) show that \( W \) may be estimated by the solution to the system of \( k \) equations

\[
W_l = \int \frac{w_l(y)}{\sum_{j=1}^{k} a_j w_j(y)/W_j} d\check{F}_n(y), \quad l = 1, \ldots, k, \tag{4–2}
\]

where \( a_j \equiv n_j/n, n = \sum_{j=1}^{k} n_j \), and \( \check{F}_n \) is the empirical distribution function that gives mass \( 1/n \) to each of the \( X_{il} \). Actually, the solution to (4–2) is not unique: it is trivial to see that if the vector \( W \) solves (4–2), then so does \( \alpha W \), for any \( \alpha \). However, it turns out that knowing \( W \) only up to a multiplicative constant is all that is needed, and to avoid non-identifiability issues, we define the vector \( V = (W_2/W_1, \ldots, W_k/W_1) \).

Gill et al. (1988) show that if \( \hat{W} \) is any solution to (4–2), and \( \hat{V} \) is defined by \( \hat{V} = (\hat{W}_2/\hat{W}_1, \ldots, \hat{W}_k/\hat{W}_1) \), then \( n^{1/2}(\hat{V} - V) \) is asymptotically normal (Proposition 2.3 in Gill et al. (1988)). Once an estimate of \( W \) is formed, it is relatively easy to form an estimate \( \hat{F}_n \) of \( F \), and consequently of integrals of the form \( \int h dF \). Gill et al. (1988)
obtain functional weak convergence results of the sort \( n^{1/2} \left( \int h \, d\tilde{F}_n - \int h \, dF \right) \overset{d}{\to} Z(h) \),
where \( Z \) is a mean-0 Gaussian process indexed by \( h \in H \), where \( H \) is a large class of square integrable functions.

It is not difficult to see that our setup is the same as that considered in Vardi (1985) and Gill et al. (1988): their \( F \) corresponds to our \( \nu_{h,y} \); their \( w_l \) to \( \nu_{h}/\nu_{h} \); \( F_l \) to \( \nu_{h,y} \); \( W_l \) to \( m_{h_i}/m_{h} \); and \( V \) to \( d \). But there are major differences between our framework and theirs. They deal with iid samples, and so can use empirical process theory, whereas we deal with Markov chains, for which such a theory is not available. In their framework, the samples arise from some experiment, and they are seeking optimal estimates given data that is given to them. In contrast, our samples are obtained by Monte Carlo, so we have control over design issues. In particular, we are concerned with computational efficiency, in addition to statistical efficiency; hence our interest in the two-stage sampling method for preliminary estimation of \( d \) and for enabling the use of control variates.

Geyer (1994) also deals with the setup in Vardi (1985) and Gill et al. (1988), i.e. the \( k \)-sample model for biased sampling, and he also considers the problem of estimating \( d \). As mentioned in Section 2.1, his estimator is obtained by maximizing (2–6), and the solution is numerically identical to the solution to the system (4–2). However, he considers the situation where each of the \( k \) samples are Markov chains, as opposed to iid samples, and assuming that the chains satisfy certain mixing conditions, he obtains a central limit theorem for \( n^{1/2}(\hat{d} - d) \). Naturally, the variance of the limiting distribution is different from the variance obtained in Gill et al. (1988), and is typically larger.

In Section 7 of their paper Meng and Wong (1996) consider the situation where for each \( l = 1, \ldots, k \), we have an iid sample from the density \( f_l = q_l/m_l \), where the functions \( q_1, \ldots, q_k \) are known, but the normalizing constants \( m_1, \ldots, m_k \) are not, and we wish to estimate the vector \( (m_2/m_1, \ldots, m_k/m_1) \). Without going into detail, we mention that they develop a family of “bridge functions” and show that, in the iid setting, the optimal bridge function gives rise to an estimate identical to that of Geyer (1994). They obtain their
estimate through an iterative scheme which is fast and stable [Meng and Wong (1996, p. 849)] and this is the computational method we use in the present paper.

Owen and Zhou (2000) consider the problem of estimating an integral of the form $I = \int h(x)f(x)\,dx$, where $f$ is a probability density that is completely known (as opposed to known up to a normalizing constant) and $h$ is a known function. They wish to estimate $I$ through importance sampling. They assume they can generate sequences $X_{l1}, \ldots, X_{ln_l} \overset{iid}{\sim} p_l$, $l = 1, \ldots, k$, where the $p_l$’s are completely known densities. The doubly indexed sequence $X_{li}$, $i = 1, \ldots, n_l$, $l = 1, \ldots, k$ forms a (stratified) sample from the mixture density $p = \sum_{l=1}^{k} a_l p_l$, where $a_l = n_l / \sum_{l=1}^{k} n_l$, so one can carry out importance sampling with respect to this mixture. They point out that since the $p_l$’s are completely known, they can form the functions $H_j(x) = [p_j(x)/p(x)] - 1$, $j = 1, \ldots, k$, and these satisfy $E_{p}(H_j(X)) = 0$, where the subscript indicates that the expectation is taken with respect to the mixture density $p$. Therefore, these $k$ functions can be used as control variates. What we do in Chapter 2 is similar, except that we are working with densities whose functional form is known, but whose normalizing constants are not.

Kong et al. (2003) also consider the $k$-sample model for biased sampling, but have a different perspective, and we describe their work in the notation of the present paper. They assume that there are probability measures $Q_1, \ldots, Q_k$, with densities $q_1/m_1, \ldots, q_k/m_k$, respectively, relative to some dominating measure $\mu$, and for each $l = 1, \ldots, k$, we have an iid sample $\{X_{li}\}_{i=1}^{n_l}$ from $Q_l$. Here, the $q_l$’s are known, but the $m_l$’s are not. Their objective is to estimate all possible ratios $m_l/m_j$, $l, j \in \{1, \ldots, k\}$ or, equivalently, the vector $d = (1, m_2/m_1, \ldots, m_k/m_1)$. In their highly unorthodox approach, Kong et al. (2003) obtain the maximum likelihood estimate $\hat{\mu}$ of the dominating measure itself ($\hat{\mu}$ is given up to an overall multiplicative constant). They can then estimate the ratios $m_l/m_j$, since the normalizing constants are known functions of $\mu$ (i.e. $m_r = \int q_r(x)\,d\mu(x)$, and $q_r$ is known). They show that the resulting estimate of $d$ is obtained
by solving the system

\[ d_r = \sum_{i=1}^{k} \sum_{s=1}^{l_s} \sum_{j=1}^{n_i} \frac{q_r(X_{ij})}{n_s q_s(X_{is})/d_s}, \quad r = 1, \ldots, k, \]  

(4–3)

which is easily seen to be identical to the system (4–2) of Gill et al. (1988).

Tan (2004) shows how control variates can be incorporated in the likelihood framework of Kong et al. (2003). When there are \( r \) functions \( H_j, j = 1, \ldots, r \) for which we know that \( \int H_j \, d\mu = 0 \), the parameter space is restricted to the set of all sigma-finite measures satisfying these \( r \) constraints. For the case where \( X_{ii}, i = 1, \ldots, n_i \) are iid for each \( l = 1, \ldots, k \), he obtains the maximum likelihood estimate of \( \mu \) in this reduced parameter space, and therefore of corresponding estimates of \( d \) and \( m_h/m_{h_1} \), and shows that this approach gives estimates that are asymptotically equivalent to estimates that use control variates via regression. He also obtains results on asymptotic normality of his estimators that are valid when we have the iid structure.

The estimates of \( d \) in Gill et al. (1988), Geyer (1994), Meng and Wong (1996), and Kong et al. (2003) are all equivalent. Theorem 1 of Tan (2004) establishes asymptotic optimality of this estimate under the iid assumption. When the samples are Markov chain draws, the asymptotically optimal estimate is essentially impossible to obtain (Romero 2003). But the estimate derived under the iid assumption can still be used in the Markov chain setting if one can develop asymptotic results that are valid in the Markov chain case, and this is done by Geyer (1994), whose results we use in all our theorems.
CHAPTER 5
ILLUSTRATION ON VARIABLE SELECTION

There exist many classes of problems in Bayesian analysis in which the sensitivity analysis and model selection issues discussed earlier arise; see Chapter 6. Here we give an application involving the hierarchical prior used in variable selection in the Bayesian linear regression model discussed in Chapter 1. This chapter consists of three parts. First we discuss an MCMC algorithm for this model and state some of its theoretical properties; then we discuss the literature on selection of the hyperparameter \( h \); and finally we present two detailed illustrations of our methodology.

5.1 A Markov Chain for Estimating the Posterior Distribution of Model Parameters

The design of MCMC algorithms for estimating the posterior distribution of \( \theta \) under (1–1) revolves around the generation of the indicator variable \( \gamma \). We now briefly review the algorithms for running a Markov chain on \( \gamma \) that are proposed in the literature, and the main issues of implementation of these algorithms. Raftery et al. (1997) and Madigan and York (1995) discuss the following Metropolis-Hastings algorithm for generating a sequence \( \gamma^{(1)}, \gamma^{(2)}, \ldots \). If the current state is \( \gamma \), a new state \( \gamma^* \) is formed by selecting at random a coordinate \( j \), setting \( \gamma^*_j = 1 - \gamma_j \), and \( \gamma^*_k = \gamma_k \) for \( k \neq j \). The proposal \( \gamma^* \) is then accepted or rejected with the Metropolis-Hastings acceptance probability \( \min\{p(\gamma^* | Y) / p(\gamma | Y) , 1\} \). Madigan and York (1995) call this algorithm MC^3. Clyde et al. (1996) propose a modification of this algorithm in which we do not select a component at random and update it, but instead sequentially update all components. They call this the “Hybrid Algorithm.” (Strictly speaking, this is a Metropolized Gibbs sampler, and is not actually a Metropolis-Hastings algorithm.) Smith and Kohn (1996) propose a Gibbs sampler which simply cycles through the coordinates \( \gamma_j \) one at a time. George and McCulloch (1997) show that when compared with MC^3, the Gibbs sampler algorithm gives estimates with smaller standard error, and is also slightly faster, at least in several simulation studies they conducted.
Kohn et al. (2001) consider Metropolized Gibbs algorithms which are the same as the Hybrid Algorithm of Clyde et al. (1996), except that at coordinate $j$, instead of deterministically proposing to go from $\gamma_j$ to $\gamma_j^* = 1 - \gamma_j$, the proposed value $\gamma_j^*$ is equal to $1 - \gamma_j$ with probability depending on the current state $\gamma$. Kohn et al. (2001) describe two such algorithms, and show that these are more computationally efficient than the Gibbs sampler in situations where on average $q_\gamma$ is small, i.e. the models are sparse. They also conduct a detailed simulation study of one of their sampling schemes (their “SS(2)”) which suggests that, while the scheme produces estimates whose standard errors are a bit larger than those produced by the Gibbs sampler, this disadvantage is more than outweighed by its computational efficiency.

All the algorithms mentioned above require, in one way or another, the calculation of $p(\gamma^* | Y) / p(\gamma | Y)$. Because of the conjugate nature of model (1–1), the marginal likelihood of model $\gamma$ is available in closed form, and therefore $p(\gamma | Y)$ is available up to a normalizing constant. We have

$$p(\gamma | Y) \propto (1 + g)^{-q_\gamma/2} S^{-(m-1)} \left[ 1 + g(1 - R_\gamma^2) \right]^{-(m-1)/2} \left( \frac{1}{w} \right)^{q_\gamma},$$

(5–1)

where $S^2 = \sum_{j=1}^m (Y_j - \bar{Y})^2$ and $R_\gamma^2$ is the coefficient of determination of model $\gamma$.

As is standard for model (1–1), we assume that the columns of the design matrix are centered, and in this case, $R_\gamma^2 = Y'X_\gamma (X_\gamma'X_\gamma)^{-1}X_\gamma'Y / S^2$. The main computational burden in obtaining (5–1) is the calculation of $R_\gamma^2$, which is time-consuming if $q_\gamma$ is large. Smith and Kohn (1996) note that, when $\gamma^*$ and $\gamma$ differ in only one component, $R_\gamma^2$ can be obtained rapidly from $R_\gamma^2$. We return to this point in Appendix B.

In our situation, we need to generate a Markov chain on $\theta$, because the Bayes factor estimates given in Chapter 2 require samples from the posterior distribution of $\theta$. The algorithm we use in the present paper is based on the Gibbs sampler on $\gamma$ introduced in Smith and Kohn (1996) (although the computational implementation we use is different from theirs), followed by three steps to generate $\sigma$, $\beta_0$, and $\beta_\gamma$. In a bit more detail, let
Suppose the current state is \((\gamma^{(i)}, \sigma^{(i)}, \beta_0^{(i)}, \beta_{\gamma}^{(i)})\). We proceed as follows.

1. We update \(\gamma^{(i)}\) to \(\gamma^{(i+1)}\) using \(V(\gamma^{(i)}, \cdot)\). The generation of \(\gamma^{(i+1)}\) does not involve \((\sigma^{(i)}, \beta_0^{(i)}, \beta_{\gamma}^{(i)})\).

2. We generate \(\sigma^{(i+1)}\) from the conditional distribution of \(\sigma\) given \(\gamma = \gamma^{(i+1)}\) and the data.

3. We generate \(\beta_{\gamma}^{(i+1)}\) from the conditional distribution of \(\beta_{\gamma}\) given \(\gamma = \gamma^{(i+1)}, \sigma = \sigma^{(i+1)}\), and the data.

4. We generate \(\beta_0^{(i+1)}\) from the conditional distribution of \(\beta_0\) given \(\gamma = \gamma^{(i+1)}, \sigma = \sigma^{(i+1)}, \beta_{\gamma} = \beta_{\gamma}^{(i+1)}\), and the data.

The details describing the distributions involved and the computations needed are given in Appendix B. The algorithm above gives a sequence \(\theta^{(1)}, \theta^{(2)}, \ldots\), and it is easy to see that this sequence is a Markov chain.

As Markov chains on the \(\gamma\) sequence, the relative performance of the Gibbs sampler vs. SS(2) depends, in part, on \(m, q, h\), and the data set itself, and neither algorithm is uniformly superior to the other. In principle, in Step 1 of our algorithm we can use any Markov transition function that generates a chain on \(\gamma\), including SS(2). We chose to work with the Gibbs sampler because it is easier to develop a regeneration scheme for this chain than for the other chains.

The output of the chain can be used in several ways. An obvious way is to use the highest posterior probability model (HPM). Unfortunately, when \(q\) is bigger than around 20, the number of models, \(2^q\), is very large, and it may happen that no single model has appreciable probability, and in any case, it is very difficult or impossible to identify the HPM from the Markov chain output. Barbieri and Berger (2004) argue in favor of the median probability model (MPM), which is defined to be the model that includes all variables \(j\) for which the marginal inclusion probability \(P(\gamma_j = 1 \mid Y) \geq 1/2\). We mention
here the Bayesian Adaptive Sampling method of Clyde et al. (2009), which gives an algorithm for providing samples without replacement from the set of models. Under certain conditions, the algorithm has the feature that these are perfect samples without replacement; it then enables an efficient search for the HPM.

**Uniform Ergodicity**

Let \( \Theta = \{0, 1\}^q \times (0, \infty) \times \mathbb{R}^{q+1} \), let \( \nu \) be the (prior) distribution of \( \theta \) specified by (1–1b) and (1–1c), and let \( \nu_y \) be the posterior distribution of \( \theta \) given \( Y = y \). (For the remainder of this section the subscript \( h \) is suppressed since we are dealing with a single specification of this hyperparameter.) Let \( K \) denote the Markov transition function for the Markov chain on \( \theta \) described in the beginning of this chapter, i.e. \( K(\theta_0, \cdot) \) is the distribution of \( \theta_1 \) given that the current state is \( \theta_0 \), and let \( K^n(\theta_0, \cdot) \) denote the corresponding \( n \)-step Markov transition function. Harris ergodicity of the chain is the condition that \( \|K^n(\theta, \cdot) - \nu_y(\cdot)\| \to 0 \) for all \( \theta \in \Theta \), where \( \| \cdot \| \) denotes supremum over all Borel subsets of \( \Theta \). This condition is guaranteed by the so-called “usual regularity conditions,” namely that the chain has an invariant probability measure, is irreducible, aperiodic, and Harris recurrent; see, e.g., Theorem 13.0.1 of Meyn and Tweedie (1993). These usual regularity conditions are typically easy to check; in the present context, they are implied for example if the Markov transition function has a density (with respect to the product of counting measure on \( \{0, 1\}^q \) and Lebesgue measure on \( (0, \infty) \times \mathbb{R}^{q+1} \)) which is everywhere positive, which is the case in our situation. Uniform ergodicity is the far stronger condition that there exist constants \( c \in [0, 1) \) and \( M > 0 \) such that for any \( n \in \mathbb{N} \),

\[
\|K^n(\theta, \cdot) - \nu_y(\cdot)\| \leq Mc^n \quad \text{for all } \theta.
\]

**Proposition 1** The chain driven by \( K \) is uniformly ergodic.

The proof of Proposition 1 is given in Appendix C. Let \( \theta_0, \theta_1, \ldots \) be a Markov chain driven by \( K \), let \( l \) be a real-valued function of \( \theta \) (for example \( l(\theta) = l(\gamma_1 = 1) \), the indicator that variable 1 is in the model), and suppose we wish to form confidence
intervals for the posterior expectation of $l(\theta)$. Suppose that $E(P(\theta)) < \infty$. Then since the chain is uniformly ergodic, Corollary 4.2 of Cogburn (1972) implies that, with $\text{Var}(l(\theta_0))$ and $\text{Cov}(l(\theta_0), l(\theta_j))$ calculated under the assumption that $\theta_0$ has the stationary distribution, the series

$$\kappa^2 = \text{Var}(l(\theta_0)) + 2 \sum_{j=1}^{\infty} \text{Cov}(l(\theta_0), l(\theta_j))$$

(5–2)

converges absolutely, and if $\kappa^2 > 0$, then with $\theta_0$ having an arbitrary distribution, the estimate $\bar{l}_n = (1/n) \sum_{i=0}^{n-1} l(\theta_i)$ satisfies

$$n^{1/2}(\bar{l}_n - E[l(\theta)] \xrightarrow{d} \mathcal{N}(0, \kappa^2) \quad \text{as } n \to \infty.$$ 

The Markov chain driven by $K$ is also regenerative, and in Appendix C we give an explicit minorization condition that can be used to introduce regenerations into the chain. Functions that run the chain and implement the regeneration scheme are provided in the R package bvslr, available from http://www.stat.ufl.edu/~ebuta/BVSLR.

In Chapters 1 and 2, $\nu_h$ and $\nu_{h,y}$ refer to the prior and posterior densities, and all estimates in Chapter 2 involve ratios of these prior densities. In the Bayesian linear regression model that we are considering here, the priors $\nu_h$ on $(\gamma, \sigma, \beta_0, \beta_\gamma)$ are actually probability measures on $\{0, 1\}^q \times (0, \infty) \times \mathbb{R}^{q+1}$, which in fact are not absolutely continuous with respect to the product of counting measure on $\{0, 1\}^q$ and Lebesgue measure on $(0, \infty) \times \mathbb{R}^{q+1}$. For $h_1 = (w_1, g_1)$ and $h_2 = (w_2, g_2)$, the Radon-Nikodym derivative of $\nu_{h_1}$ with respect to $\nu_{h_2}$ is given by

$$\left[\frac{d\nu_{h_1}}{d\nu_{h_2}}\right](\gamma, \sigma, \beta_0, \beta_\gamma) = \left(\frac{w_1}{w_2}\right)^{q_\gamma} \left(\frac{1 - w_1}{1 - w_2}\right)^{q - q_\gamma} \times \phi_{q_\gamma}(\beta_\gamma; 0, g_1 \sigma^2(X'_\gamma X_\gamma)^{-1})$$

(5–3)

$$\phi_{q_\gamma}(\beta_\gamma; 0, g_2 \sigma^2(X'_\gamma X_\gamma)^{-1}).$$

where $\phi_{q_\gamma}(u; a, V)$ is the density of the $q_\gamma$-dimensional normal distribution with mean $a$ and covariance $V$, evaluated at $u$ (Doss (2007)). It is immediate that all formulas in Chapter 2 remain valid if ratios of the form $\nu_h(\theta)/\nu_{h_1}(\theta)$ (see, e.g., equation (2–2)) are replaced by the Radon-Nikodym derivative $[d\nu_h/d\nu_{h_1}](\theta)$. Fortunately, evaluation of (5–3)
requires neither matrix inversion nor calculation of a determinant, so can be done very quickly. Note that in view of (5–3), it is not enough to have Markov chains running on the $\gamma$’s and we need Markov chains running on the $\theta$’s (or at least $(\gamma, \sigma, \beta_\gamma)$).

5.2 Choice of the Hyperparameter

As mentioned earlier, regarding $w$, the proposals in the literature are quite simple: either $w$ is fixed at $1/2$, or a beta prior is put on $w$. The discussion below focuses primarily on $g$, for which there is an extensive literature, and we now summarize the portion of this literature that is directly relevant to the present work. Broadly speaking, recommendations regarding $g$ can be divided into three categories:

Data-Independent Choices  In the simple case where the setup is given by (1–1) but without (1–1c), i.e. the true model $\gamma$ is assumed known, the posterior distribution of $\beta$ given $\sigma$ is $\mathcal{N}((g/(g + 1))\hat{\beta}_\gamma, (g/(g + 1))\sigma^2(X_\gamma'X_\gamma)^{-1})$, where $\hat{\beta}_\gamma$ is the usual least squares estimate of $\beta$. If $q$ is fixed and $m \to \infty$, under standard conditions $X_\gamma'X_\gamma/m \to \Sigma$, where $\Sigma$ is a positive definite matrix; therefore if $g$ is fixed, this distribution is approximately a point mass at $(g/(g + 1))\hat{\beta}_\gamma$, so the posterior is not even consistent, and we see that a necessary condition for consistency is that $g \to \infty$. Data-independent choices of $g$ include Kass and Wasserman’s (1995) recommendation of $g = m$, and Fernandez et al.’s (2001) recommendation of $g = \max(m, q^2)$, following up on Foster and George’s (1994) earlier recommendation of $g = q^2$.

Liang et al. (2008) argue that, in general, data-independent choices of $g$ have the following undesirable property, referred to as the “Information Paradox.” When the data give overwhelming evidence in favor of model $\gamma$ (e.g. $\|\hat{\beta}_\gamma\| \to \infty$), then using $\gamma_0$ to denote the null model (i.e. the model that includes only the intercept), the ratio of posterior probabilities $p(\gamma | Y)/p(\gamma_0 | Y)$ does not tend to infinity.
Empirical Bayes (EB) Methods In global EB procedures, an estimate of \( g \) common for all models is derived from its marginal likelihood; see George and Foster (2000). In local EB, an estimate of \( g \) is derived for each model; see Hansen and Yu (2001). Unfortunately, the EB method is in general computationally demanding because the likelihood is a sum over all \( 2^q \) models \( \gamma \), so it is practically feasible only for relatively small values of \( q \). Liang et al. (2008) show that the EB method is consistent in the frequentist sense: if \( \gamma_* \) is the true model, then if \( g \) is chosen via the EB method, the posterior probability \( P(\gamma = \gamma_* | Y) \) converges to 1 as \( m \to \infty \). See Theorem 3 of Liang et al. (2008) for a precise statement. (This result refers only to the case where \( w \) is fixed at 1/2, and only \( g \) is estimated.) Liang et al. (2008) propose an EM algorithm for estimating \( g \) in the global EB setting. In their algorithm, the model indicator and \( \sigma \) are treated as missing data. While their approach is certainly useful, there are some problems associated with it. Each step in the EM algorithm involves a sum of \( 2^q \) terms. Unless \( q \) is relatively small, complete enumeration is not possible, and Liang et al. (2008) propose summing only over the most significant terms. However, determining which terms these are may be very difficult in some problems. Also, the EM algorithm gives a single point estimate. What we do is different: we estimate the Bayes factor for all \( g \) (and \( w \)). This enables us in particular to estimate the maximizing values; but it also allows us to rule out large regions of the hyperparameter space. Additionally, our method allows us to carry out sensitivity analysis. We also mention very briefly that if we are interested only in the maximizing values, then the method proposed in the present paper can be used to form a stochastic search algorithm. The basic requirement for such algorithms is that we know the gradient \( \partial B(h, h_1) / \partial h \). But the same methodology used to estimate \( B(h, h_1) \) can also be used to estimate its gradient. For example, in the simple estimate (2–7), we just replace \( \nu_h(\theta_i^{(0)}) \) by \( \partial \nu_h(\theta_i^{(0)}) / \partial h \).
**Fully Bayes (FB) Methods** The most common prior on \( g \) is the Zellner and Siow (1980) prior, an inverse-gamma which results in a multivariate Cauchy prior for \( \beta \). The family of “hyper-\( g \)” priors is introduced by Cui and George (2008) and developed further by Liang et al. (2008), who show that these have several desirable properties. In particular, they do not suffer from the information paradox, and they exhibit important consistency properties.

Both the EB methods and FB methods have their own advantages and disadvantages. Cui and George (2008) give evidence that EB methods outperform FB methods. This is based on extensive simulation studies in cases where numerical methods are feasible. Also, FB methods require one to specify hyperparameters of the prior on the hyperparameter \( h \), and different choices lead to different inferences. Additionally, in EB methods, one uses a model with a single value of \( h \), and the resulting inference is more parsimonious and interpretable.

On the other hand, as with many likelihood-based methods, special care needs to be taken when the maximizing value is at the boundary. When we use the EB method, if the maximizing value of \( w \) is 0 or 1, the posterior assigns probability one to the null model or full model (model that includes all variables), respectively. This is similar to the very simple situation in which we have \( X \sim \text{binomial}(n, p) \): if we observe \( X = 0 \), then not only is the maximum likelihood estimate of \( p \) equal to 0, but the associated standard error estimate is also 0, and the naive Wald-type confidence interval for \( p \) is the singleton \( \{0\} \). Of course in this simple case there exist modifications to the maximum likelihood estimate \( \hat{p} = X/n \) which yield procedures that do not give rise to this degeneracy. How to develop corresponding modifications to the maximum likelihood estimate of the Bernoulli parameter \( w \) in the present context is a problem that is much more difficult, but certainly worthy of investigation.

Scott and Berger (2010) consider the same model for variable selection that we consider here, i.e. model (1–1), but with a Zellner-Siow prior on \( g \), and the remaining
parameter, $w$, estimated by maximum likelihood. They show that if the null model has the largest marginal likelihood, then the MLE of $w$ is 0 and if the full model has the largest marginal likelihood, then the MLE of $w$ is 1. Each of these gives rise to the degeneracy discussed above. Their result is not true in our setup, in which we do not put a prior on $g$, but rather estimate both $w$ and $g$ by maximum likelihood. To see this, consider a very simple example, in which $Y = (2, 1, 9, 5)'$ and

$$X = \begin{pmatrix} 1 & 3 \\ 5 & 3 \\ 8 & 7 \\ 8 & 10.5 \end{pmatrix}.$$  

We have $R^2_{\gamma = (1,1)} = 0.52$, $R^2_{\gamma = (1,0)} = 0.51$, $R^2_{\gamma = (0,1)} = 0.40$, and $R^2_{\gamma = (0,0)} = 0$. Now

$$P(Y | \gamma, g) = c(Y) \frac{(1 + g)^{(3-q_2)/2}}{(1 + g(1 - R^2_\gamma))^{3/2}},$$

where $c(Y)$ does not depend on $g$ or $\gamma$. Therefore,

$$(\hat{g}, \hat{w}) = \arg\max_{(g, w)} \sum_{\gamma} w^q (1 - w)^{q - q_2} \frac{(1 + g)^{(3-q_2)/2}}{(1 + g(1 - R^2_\gamma))^{3/2}} \approx (.5, .2).$$

From equation (38) of Scott and Berger (2010) we know that under the Zellner-Siow null prior, we have

$$\frac{P(Y | \gamma)}{P(Y | \gamma = (0, 0))} = \int_0^\infty \frac{(1 + g)^{(3-q_2)/2}}{(1 + g(1 - R^2_\gamma))^{3/2}} \left(\frac{2}{\pi}\right)^{1/2} g^{-3/2} \exp(-2/g) dg$$

$$= \begin{cases} .72 < 1 & \text{for } \gamma = (1, 0) \\ .58 < 1 & \text{for } \gamma = (0, 1) \\ .31 < 1 & \text{for } \gamma = (1, 1) \end{cases}$$
and hence the null model has the strictly largest marginal likelihood among all models. Lemma 4.1 of Scott and Berger (2010) implies that, with a Zellner-Siow prior on \( g \), \( \hat{\nu} = 0 \), while in our setup, the same data give \( \hat{\nu} > 0 \).

5.3 Examples

We illustrate our methods on two examples. The first is the U.S. crime data of Vandaele (1978), which can be found in the R library MASS under the name UScrime. We use this data set because it has been studied in several papers already so we can compare our results with previous analyses, and also because the number of variables is small enough to enable a closed-form calculation of the marginal likelihood \( m_h \), so we can compare our estimates with the gold standard. The second data set is the ozone data originally analyzed by Breiman and Friedman (1985). We use this data set because it involves 44 variables, even though only a few of those are important, and we wanted to show how our methodology handles a data set with this character.

5.3.1 U.S. Crime Data

The data set gives, for each of \( m = 47 \) U.S. states, the crime rate, defined as number of offenses per 100,000 individuals (the response variable), and \( q = 15 \) predictors measuring different characteristics of the population, such as average number of years of schooling, average income, unemployment rate, etc.

To be consistent with what is done in the literature, we applied a log transformation to all variables, except the indicator variable. We took the baseline hyperparameter to be \( h_1 = (w_1, g_1) = (0.5, 15) \), and our goal was to estimate \( B(h, h_1) \) for the 924 values of \( h \) obtained when \( w \) ranges from 0.1 to 0.91 by increments of 0.03, and \( g \) ranges from 4 to 100 by increments of 3. We used (2–13) and this estimate was based on 16 chains each of length 10,000, corresponding to the skeleton grid of hyperparameter values

\[
(w, g) \in \{0.3, 0.5, 0.6, 0.8\} \times \{15, 50, 100, 225\}
\]

(5–4)
for the Stage 1 samples, and 16 new chains, each of length 1000, corresponding to the same hyperparameter values, for the Stage 2 samples. The plots in Figure 5-1 give graphs of the estimate (2–13) as \( w \) and \( g \) vary, from two different angles. These indicate that values for \( w \) around 0.65 and for \( g \) around 20 seem appropriate, while values of \( w \) less than 0.3 and values of \( g \) greater than 60 should be avoided. A side calculation showed that, interestingly, for \( g = \max\{m, q^2\} (= 225) \), the estimate of \( B((w, g), (.65, 20)) \) is less than 0.008 regardless of the value of \( w \), so this choice should not be used for this data set. With the long chains used and the estimate that uses control variates, the Bayes factor estimates in Figure 5-1 are extremely accurate—root mean squared errors are less than 0.04 uniformly over the entire domain of the plot and considerably less in the convex hull of the skeleton grid (our calculation of the root mean squared errors used the closed-form expression for the Bayes factors based on complete enumeration). The figure took about a half hour to generate on an Intel 2.8 GHz Q9550 running Linux. (The accuracy we obtained is overkill and the figure can be created in a few minutes if we use more typical Markov chain lengths.)

![Figure 5-1](image-url)

Figure 5-1. Estimates of Bayes factors for the U.S. crime data. The plots give two different views of the graph of the Bayes factor as a function of \( w \) and \( g \) when the baseline value of the hyperparameter is given by \( w = 0.5 \) and \( g = 15 \). The estimate is (2–13), which uses control variates.
Table 5-1 gives the posterior inclusion probabilities for each of the fifteen predictors, i.e. \( P(\gamma_i = 1 | y) \) for \( i = 1, \ldots, 15 \), under several models. Line 2 gives the inclusion probabilities when we use model (1–1) with the values \( w = .65 \) and \( g = 20 \), which are the values at which the graph in Figure 5-1 attains its maximum. Line 4 gives the inclusion probabilities when the hyper-\( g \) prior “HG3” in Liang et al. (2008) is used. As can be seen, the inclusion probabilities we obtained under the EB model are comparable to, but somewhat larger than, the probabilities when the HG3 prior is used. This is not surprising since our model allows \( w \) to be chosen, and the data-driven choice gives a value (.65) greater than the value \( w = .5 \) used in Liang et al. (2008). (Table 2 of Liang et al. (2008) gives a comparison of posterior inclusion probabilities for a total of ten models taken from the literature.) Line 3 of Table 5-1 gives the inclusion probabilities under model (1–1) when we use \( w = .5 \) and the value of \( g \) that maximizes the likelihood with \( w \) constrained to be .5. It is interesting to note that the inclusion probabilities are then strikingly close to those under the HG3 model.

Table 5-1. Posterior inclusion probabilities for the fifteen predictor variables in the U.S. crime data set, under three models. Names of the variables are as in Table 2 of Liang et al. (2008) (but all variables except for the binary variable S have been log transformed).

<table>
<thead>
<tr>
<th></th>
<th>Age</th>
<th>S</th>
<th>Ed</th>
<th>Ex0</th>
<th>Ex1</th>
<th>LF</th>
<th>M</th>
<th>N</th>
<th>NW</th>
<th>U1</th>
<th>U2</th>
<th>W</th>
<th>X</th>
<th>Prison</th>
<th>Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>EB (20,.65)</td>
<td>.93</td>
<td>.39</td>
<td>.99</td>
<td>.70</td>
<td>.51</td>
<td>.34</td>
<td>.35</td>
<td>.52</td>
<td>.83</td>
<td>.40</td>
<td>.76</td>
<td>.55</td>
<td>1.00</td>
<td>.96</td>
<td>.55</td>
</tr>
<tr>
<td>EB (20,.5)</td>
<td>.85</td>
<td>.29</td>
<td>.97</td>
<td>.67</td>
<td>.45</td>
<td>.22</td>
<td>.22</td>
<td>.38</td>
<td>.70</td>
<td>.27</td>
<td>.62</td>
<td>.38</td>
<td>.99</td>
<td>.90</td>
<td>.39</td>
</tr>
<tr>
<td>HG3</td>
<td>.84</td>
<td>.29</td>
<td>.97</td>
<td>.66</td>
<td>.47</td>
<td>.23</td>
<td>.23</td>
<td>.39</td>
<td>.69</td>
<td>.27</td>
<td>.61</td>
<td>.38</td>
<td>.99</td>
<td>.89</td>
<td>.38</td>
</tr>
</tbody>
</table>

Figure 5-2 gives plots of the posterior inclusion probabilities for Variables 1 and 6, as \( w \) and \( g \) vary. The literature recommends various choices for \( g \) [in particular \( g = m \) in Kass and Wasserman (1995), \( g = q^2 \) in Foster and George (1994), \( g = \max(m, q^2) \) in Fernandez et al. (2001)], and posterior inclusion probabilities for all these choices combined with any choice of \( w \) can be read directly from the figure. The extent to which these probabilities change with the choice of \( g \) is quite striking.
Selection of the skeleton points was discussed at the end of Chapter 3, and we now return to this issue. Consider the Bayes factor estimate based on the skeleton (5–4), which was chosen in an ad-hoc manner. The left panel in Figure 5-3 gives a plot of the variance of this estimate, as a function of $h$. As can be seen from the plot, the variance is greatest in the region where $g$ is small and $w$ is large. We changed the skeleton from (5–4) to

$$(w, g) \in \{.5, .7, .8, .9\} \times \{10, 15, 50, 100\}$$

and reran the algorithm. The variance for the estimate based on (5–5) is given by the right panel of Figure 5-3, from which we see that the maximum variance has been reduced by a factor of about 9.

### 5.3.2 Ozone Data

This data set was originally analyzed in Breiman and Friedman (1985), was used in many papers since, and was recently analyzed in a Bayesian framework by Casella and Moreno (2006) and Liang et al. (2008). The data consist of daily measurements of ozone concentration and eight meteorological quantities in the Los Angeles basin for 330 days of 1976. The response variable is the daily ozone concentration, and we follow Liang et al. (2008) in considering 44 possible predictors: the eight meteorological...
measurements, their squares, and their two-way interactions. Liang et al. (2008) give a review of the literature on priors for the hyperparameter \( g \) and advocate the hyper-\( g \) priors. They compare 10 variable selection techniques (including three hyper-\( g \) priors) on this data set by using a cross-validation procedure: the data set is randomly split in two halves, one of which (the training sample) is used for selecting the model (for the Bayesian methods this is the highest probability model), while the other (the validation sample) is used for measuring the predictive accuracy. The predictive accuracy of method \( j \) is measured through the square-root of the mean squared prediction error (RMSE) of the selected model \( \gamma_j \), defined by 

\[
\text{RMSE}(\gamma_j) = \left( n_V^{-1} \sum_{i \in V} (Y_i - \hat{Y}_i)^2 \right)^{1/2}.
\]

Here, \( V \) is the validation set, \( n_V \) is its size, and \( \hat{Y}_i \) is the fitted value of observation \( i \) under model \( \gamma_j \). Liang et al. (2008) point out the curious fact that the RMSE’s of the 10 methods are all very close (they range from 4.4 to 4.6), but the selected models differ greatly in the number of variables selected, which range from 3 to 18.
We investigated the performance of our methodology using a split of the data into training and validation sample identical to the one used by Liang et al. (2008). We took the baseline hyperparameter to be the pair $h_1 = (w_1, g_1) = (.2, 50)$ and the skeleton grid of hyperparameters to consist of the 16 pairs

$$(w, g) \in \{.1, .2, .3, .5\} \times \{15, 50, 100, 150\}.$$  

To identify the value of $h$ that maximizes the Bayes factor $B(h, h_1)$, we estimated this quantity for a grid of the 750 values of $h$ obtained when $w$ ranges from .01 to .5 by increments of .02, and $g$ ranges from 5 to 150 by increments of 5. These estimates were based on 16 chains each of length 10,000, corresponding to the skeleton grid of hyperparameter values for the Stage 1 samples, and 16 new chains, each of length 1000, corresponding to the same hyperparameter values, for the Stage 2 samples. Figure 5-4 gives a plot of these estimates of $B(h, h_1)$ as a function of $w$ and $g$. The standard error is less than .014 over the entire range of the plot.

![Figure 5-4](image)

Figure 5-4. Estimates of Bayes factors for the ozone data. The plots give two different views of the graph of the Bayes factor as a function of $w$ and $g$ when the baseline value of the hyperparameter is given by $w = .2$ and $g = 50$.

The value of $h$ at which the maximum $B(h, h_1)$ is attained is $h = (.13, 75)$. We ran a new chain of length 100,000 corresponding to this value of $h$, and based on it we estimated the highest probability model to be the model containing the 4 variables
dpdg, ibt, vh.ibh, and humid.ibt (see Appendix D for a description of these variables). This model yields an out-of-sample RMSE of 4.5. Since the empirical Bayes choice of $\hat{w}$ is relatively small ($\hat{w} = .13$), it is not surprising that the highest probability model includes only 4 variables—fewer than in any of the hyper-$g$ models recommended by Liang et al. (2008), which all include at least 6 variables. But it is interesting to note that nevertheless, this model gives an RMSE that is essentially the same as the RMSE of any of the other models.

We applied the regeneration algorithm described in Appendix B to the chain corresponding to the hyperparameter $h = (.13, 75)$ deemed optimal by our previous analysis. We ran the chain until $R = 3000$ regenerations occurred, which took 85,000 iterations. From the output, we obtained estimates of the posterior inclusion probabilities for every one of the 44 predictors, and formed the corresponding 95% confidence intervals, using the regeneration method discussed in Chapter 3. These are displayed in Figure 5-5.

Our choice of $R$ was arbitrary, but this choice should ultimately be based on the degree of accuracy one desires for the estimates of the quantities of interest. We considered our choice to be satisfactory for this particular analysis since the confidence intervals for the posterior inclusion probabilities for the 44 predictors have margins of error of at most 1%. Note that our chain regenerates relatively often with the average length of a tour ($\overline{N}$) being about 28. Mykland et al. (1995) recommend that one check that the coefficient of variation $CV(N) = (\text{Var}(N))^{1/2}/E(N)$ of the average tour length is less than .1 before deeming $\kappa^2$ to be estimated properly by $\hat{\kappa}^2$. Their criterion seems to be met here since the strongly consistent estimator $\hat{CV}(\overline{N}) = (\sum_{t=1}^{R}(N_t - \overline{N})^2/(R \overline{N})^2)^{1/2}$ equals .02.
Figure 5-5. 95% confidence intervals of the posterior inclusion probabilities for the 44 predictors in the ozone data when the hyperparameter value is given by $w = .13$ and $g = 75$. A table giving the correspondence between the integers 1–44 and the predictors is given in Appendix D.
CHAPTER 6
DISCUSSION

The following fact is obvious, but it may be worthwhile to state it explicitly. If $h_1$ is fixed, maximizing $B(h, h_1)$ and maximizing the marginal likelihood $m_h$ are equivalent. Thus, the value of $h$ that maximizes $m_h$ is by definition the empirical Bayes method. Thus, the development in Chapter 2 can be used to implement empirical Bayes methods.

Our methodology for dealing with the sensitivity analysis and model selection problems discussed in Chapter 1 can be applied to many classes of Bayesian models. In addition to the usual parametric models, we mention also Bayesian nonparametric models involving mixtures of Dirichlet processes (Antoniak (1974)), in which one of the hyperparameters is the so-called total mass parameter—very briefly, this hyperparameter controls the extent to which the nonparametric model differs from a purely parametric model. (Among the many papers that use such models, we mention in particular Burr and Doss (2005), who give a more detailed discussion of the role of the total mass parameter.) The approach developed in Sections 2.1 and 2.2 can be used to select this parameter.

When the dimension of $h$ is low, it will be possible to plot $B(h, h_1)$, or at least plot it as $h$ varies along some of its dimensions. Empirical Bayes methods are notoriously difficult to implement when the dimension of the hyperparameter $h$ is high. In this case, it is possible to use the methods developed in Sections 2.1 and 2.2 to enable approaches based on stochastic search algorithms. These require the calculation of the gradient $\partial B(h, h_1)/\partial h$. We note that the same methodology used to estimate $B(h, h_1)$ can also be used to estimate its gradient. For example, in (2–7), $\nu_k(\theta_i^{(l)})$ is simply replaced by $\partial \nu_k(\theta_i^{(l)})/\partial h$.  

55
APPENDIX A
PROOF OF RESULTS FROM CHAPTER 1

Proof of Theorem 1

We begin by writing
\[ \sqrt{n}(\hat{B}(h, h_1, \hat{d}) - B(h, h_1)) = \sqrt{n}(\hat{B}(h, h_1, \hat{d}) - \hat{B}(h, h_1, d)) + \sqrt{n}(\hat{B}(h, h_1, d) - B(h, h_1)). \]  
(A.1)

The second term on the right side of (A.1) involves randomness coming only from the second stage of sampling. This term was analyzed by Doss (2010), who showed that it is asymptotically normal, with mean 0 and variance \( \tau^2(h) \). The first term ostensibly involves randomness from both Stage 1 and Stage 2 sampling. However, as will emerge from our proof, the randomness from Stage 2 is of lower order, and effectively all the randomness is from Stage 1. This randomness is non-negligible. We mention here the often-cited work of Geyer (1994) (whose nice results we use in the present paper). In the context of a setup very similar to ours, his Theorem 4 states that using an estimated \( d \) and using the true \( d \) results in the same asymptotic variance. From our proof (refer also to Remark 2 of Section 2.1), we see that this statement is not correct.

To analyze the first term on the right side of (A.1), we define the function \( F(u) = \hat{B}(h, h_1, u) \), where \( u = (u_2, \ldots, u_k)' \) is a real vector with \( u_l > 0 \), \( l = 2, \ldots, k \). Then, by the Taylor series expansion of \( F \) about \( d \), we get
\[ \sqrt{n}(\hat{B}(h, h_1, \hat{d}) - \hat{B}(h, h_1, d)) = \sqrt{n}(F(\hat{d}) - F(d)) \]
\[ = \sqrt{n}\nabla F(\hat{d})'(\hat{d} - d) + \frac{\sqrt{n}}{2} (\hat{d} - d)' \nabla^2 F(d^*)(\hat{d} - d), \]  
(A.2)

where \( d^* \) is between \( d \) and \( \hat{d} \).

First, we show that the gradient \( \nabla F(d) = (\partial F(d)/\partial d_2, \ldots, \partial F(d)/\partial d_k)' \) converges almost surely to a finite constant. For \( j = 2, \ldots, k \), the \( (j - 1)^{th} \) component of this vector converges almost surely since, with the SLLN assumed to hold for the Markov chains
used, we have

\[
[\nabla F(d)]_{j-1} = \sum_{i=1}^{k} \sum_{l=1}^{n_j} \frac{n_i \nu_h(\theta_j^{(i)}) \nu_h(\theta_i^{(l)})}{d_j^2 \left( \sum_{s=1}^{k} n_s \nu_h(\theta_j^{(i)})/d_s \right)^2} \]

\[
= \sum_{i=1}^{k} \frac{1}{n_i} \sum_{l=1}^{n_j} \frac{a_j a_i \nu_h(\theta_j^{(i)}) \nu_h(\theta_i^{(l)})}{d_j^2 \left( \sum_{s=1}^{k} a_s \nu_h(\theta_j^{(i)})/d_s \right)^2}
\]

a.s. \[
= \frac{1}{d^2} \sum_{i=1}^{k} a_i \int \frac{a_j \nu_h(\theta) \nu_h(\theta)}{(\sum_{s=1}^{k} a_s \nu_h(\theta)/d_s)^2} \nu_{h,y}(\theta) d\theta
\]

\[
= \frac{1}{d_j^2} \int \frac{m_h}{m_{h_1}} \cdot \frac{a_j \nu_h(\theta)}{(\sum_{s=1}^{k} a_s \nu_h(\theta)/d_s)^2} \cdot \left( \sum_{i=1}^{k} a_i \nu_h(\theta)/d_i \right) \cdot \nu_{h,y}(\theta) d\theta
\]

\[
= \frac{B(h, h_1)}{d^2} \int \frac{a_j \nu_h(\theta)}{\sum_{s=1}^{k} a_s \nu_h(\theta)/d_s} \cdot \nu_{h,y}(\theta) d\theta := [c(h)]_{j-1}.
\]

(A.3)

The last integral is clearly finite, and the last equality in (A.3) indicates that \( c(h) \) denotes the constant vector to which \( \nabla F(d) \) converges.

Next, we show that the random Hessian matrix \( \nabla^2 F(d^*) \) of second-order derivatives of \( F \) evaluated at \( d^* \) is bounded in probability. To this end, it suffices to show that each element of this matrix, say \( \left[ \nabla^2 F(d^*) \right]_{t-1,j-1} \), where \( t, j \in \{2, \ldots, k\} \), is \( O_p(1) \). Since

\[ \|d^* - d\| \leq \|\hat{d} - d\| \overset{p}{\to} 0 \], it follows that \( d^* \overset{p}{\to} d \).

Let \( \epsilon \in (0, \min(d_2, \ldots, d_k)) \). Then we have \( P(\|d^* - d\| \leq \epsilon) \to 1 \). We now show that, on the set \( \{\|d^* - d\| \leq \epsilon\} \), \( \nabla^2 F(d^*) \) is bounded in probability. Let

\[ I = I(\|d^* - d\| \leq \epsilon). \]
For \( t \neq j \), we have

\[
|[\nabla^2 F(d^*)]_{t-1,j-1}| \cdot I = \sum_{i=1}^{k} \frac{2}{n_i} \sum_{i=1}^{n_i} \frac{a_j a_i \nu_h(\theta^{(i)}) \nu_h(\theta^{(i)})}{d_j^3 (\sum_{s=1}^{k} a_s \nu_h(\theta^{(i)})/d_s)^3} \cdot I
\]

\[
\leq \sum_{i=1}^{k} \frac{2}{n_i} \sum_{i=1}^{n_i} \frac{a_j a_i \nu_h(\theta^{(i)}) \nu_h(\theta^{(i)})}{(d_j - \epsilon)^2 (d_t - \epsilon)^2 \left[ \sum_{s=1}^{k} a_s \nu_h(\theta^{(i)})/(d_s + \epsilon) \right]^3}
\]

\[
as \frac{2}{(d_j - \epsilon)^2 (d_t - \epsilon)^2} \sum_{i=1}^{k} \int \left\{ \frac{a_j a_i \nu_h(\theta) \nu_h(\theta) \nu_h(\theta)}{\left[ \sum_{s=1}^{k} a_s \nu_h(\theta)/(d_s + \epsilon) \right]^3} \right\} \nu_h(\theta) d\theta.
\]

Note that the expression inside the braces in (A.4) is clearly bounded above by a constant, so expression (A.4) is finite. Similarly, for \( t = j \),

\[
|\nabla^2 F(d^*)|_{t-1,j-1} \cdot I
\]

\[
\leq \sum_{i=1}^{k} \frac{2}{n_i} \sum_{i=1}^{n_i} \frac{a_j a_i \nu_h(\theta^{(i)}) \nu_h(\theta^{(i)})}{d_j^3 (\sum_{s=1}^{k} a_s \nu_h(\theta^{(i)})/d_s)^3}
\]

\[
\leq \sum_{i=1}^{k} \frac{2}{n_i} \sum_{i=1}^{n_i} \frac{a_j a_i \nu_h(\theta^{(i)}) \nu_h(\theta^{(i)})}{(d_j - \epsilon)^3 \left[ \sum_{s=1}^{k} a_s \nu_h(\theta^{(i)})/(d_s + \epsilon) \right]^2}
\]

\[
\leq \sum_{i=1}^{k} \frac{2}{n_i} \sum_{i=1}^{n_i} \frac{a_j a_i \nu_h(\theta^{(i)}) \nu_h(\theta^{(i)})}{(d_j - \epsilon)^3 \left[ \sum_{s=1}^{k} a_s \nu_h(\theta^{(i)})/(d_s + \epsilon) \right]^2}
\]

\[
as \frac{2}{(d_j - \epsilon)^3} \sum_{i=1}^{k} B(h, h_i) \int \left\{ \frac{a_j a_i \nu_h(\theta) \nu_h(\theta)}{\left[ \sum_{s=1}^{k} a_s \nu_h(\theta)/(d_s + \epsilon) \right]^2} \right\} \nu_h(\theta) d\theta.
\]

Again, this limit is a finite constant by the same reasoning we used earlier. Since \( P(\|d^* - d\| \leq \epsilon) \to 1 \), it follows that \( \nabla^2 F(d^*) \) is bounded in probability. Now, by
combining (A.1) and (A.2), we obtain
\[
\sqrt{n} \left( \bar{B}(h, h_1, \hat{d}) - B(h, h_1) \right) = \sqrt{n} \frac{\nabla F(d)'}{N} \sqrt{n} (\hat{d} - d) + \frac{1}{2\sqrt{n}} \sqrt{n} \left( \nabla^2 F(d^*) \right) [\sqrt{n} (\hat{d} - d)]
\]
\[+ \sqrt{n} \left( \bar{B}(h, h_1, d) - B(h, h_1) \right) = \sqrt{q} c(h)' \sqrt{n} \left( \hat{d} - d \right) + \sqrt{n} \left( \bar{B}(h, h_1, d) - B(h, h_1) \right) + o_p(1), \tag{A.5} \]
where the last line follows from the previously established fact that \( \nabla F(d) \xrightarrow{\text{a.s.}} c(h) \), and the assumptions of Theorem 1 that \( \sqrt{n}/N \to \sqrt{q} \) and that \( \sqrt{N} (\hat{d} - d) \) converges in distribution (hence is \( O_p(1) \)). Because the two sampling stages (for estimating \( d \) and \( B(h, h_1) \)) are assumed to be independent, using the assumption that \( \sqrt{N} (\hat{d} - d) \xrightarrow{d} \mathcal{N}(0, \Sigma) \) in conjunction with the result \( \sqrt{n} \left( \bar{B}(h, h_1, d) - B(h, h_1) \right) \xrightarrow{d} \mathcal{N}(0, \tau^2(h)) \) established in Theorem 1 of Doss (2010) under conditions A1 and A2, we conclude that
\[
\sqrt{n} \left( \bar{B}(h, h_1, \hat{d}) - B(h, h_1) \right) \xrightarrow{d} \mathcal{N}(0, q c(h)' \Sigma c(h) + \tau^2(h)). \tag{\text{□}} \]

Proof of Theorem 2

We begin by writing
\[
\sqrt{n} \left( \hat{\beta}^d_{\hat{d}(d)} - B(h, h_1) \right) = \sqrt{n} \left( \hat{\beta}^d_{\hat{d}(d)} - \hat{\beta}^d_{\hat{\beta}(d)} \right) + \sqrt{n} \left( \hat{\beta}^d_{\hat{\beta}(d)} - B(h, h_1) \right), \tag{A.6} \]
where the second term on the right side of (A.6) was analyzed by Doss (2010) who showed that it is asymptotically normal, with mean 0 and variance \( \sigma^2(h) \). Our plan is to show that \( \hat{\beta}(d) \) and \( \hat{\beta}(\hat{d}) \) converge in probability to the same limit, which we denote \( \beta_{\lim} \).

We then expand the first term on the right side of (A.6) by writing
\[
\sqrt{n} \left( \hat{\beta}^d_{\hat{d}(d)} - \hat{\beta}^d_{\hat{\beta}(d)} \right) = \sqrt{n} \left( \hat{\beta}^d_{\hat{d}(d)} - \hat{\beta}^d_{\hat{\beta}(d)} \right) + \sqrt{n} \left( \hat{\beta}^d_{\hat{\beta}(d)} - \hat{\beta}^d_{\hat{\beta}(d)} \right) + \sqrt{n} \left( \hat{\beta}^d_{\hat{\beta}(d)} - \hat{\beta}^d_{\hat{\beta}(d)} \right) + \sqrt{n} \left( \hat{\beta}^d_{\hat{\beta}(d)} - \hat{\beta}^d_{\hat{\beta}(d)} \right). \tag{A.7} \]
Our proof is organized as follows:

- We note that the third term on the right side of (A.7) was shown to converge to 0 in probability by Doss (2010).
- We will show the first term on the right side of (A.7) also converges to 0 in probability.
- The second term on the right side of (A.7) involves randomness from both Stage 1 and Stage 2. However, we will show that the randomness from Stage 2 is asymptotically negligible, and that this term is asymptotically equivalent to an expression of the form \( w(h)'(\hat{d} - d) \), where \( w(h) \) is a deterministic vector. This will show that the second term is asymptotically normal.

Now we prove that the first term on the right side of (A.7) is \( o_p(1) \), and to do this we begin by showing that \( \hat{\beta}(d) \) and \( \hat{\beta}(\hat{d}) \) converge in probability to the same limit. Let \( Z \) be the \( n \times k \) matrix whose transpose is

\[
Z' = \begin{pmatrix}
1 & \ldots & 1 & 1 & \ldots & 1 & \ldots & 1 \\
Z_{1,1}^{(2)} & \ldots & Z_{n,1}^{(2)} & Z_{1,2}^{(2)} & \ldots & Z_{n,2}^{(2)} & \ldots & Z_{1,k}^{(2)} & \ldots & Z_{n,k}^{(2)} \\
\vdots & \ddots & \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots \\
Z_{1,1}^{(k)} & \ldots & Z_{n,1}^{(k)} & Z_{1,2}^{(k)} & \ldots & Z_{n,2}^{(k)} & \ldots & Z_{1,k}^{(k)} & \ldots & Z_{n,k}^{(k)}
\end{pmatrix} \quad (A.8)
\]

and let \( Y \) be the vector

\[
Y = (Y_{1,1}, \ldots, Y_{n,1}, Y_{1,2}, \ldots, Y_{n,2}, \ldots, Y_{1,k}, \ldots, Y_{n,k})'.
\] (A.9)

Let \( \hat{Z} \) be the \( n \times k \) matrix corresponding to \( Z \) when we replace \( d \) by \( \hat{d} \). Similarly, \( \hat{Y} \) is like \( Y \), but using \( \hat{d} \) for \( d \).

For fixed \( j, j' \in \{2, \ldots, k\} \), consider the function

\[
G(u) = \frac{1}{n} \sum_{l=1}^{k} \sum_{i=1}^{n_l} \frac{\nu_{h_l}(\theta_i^{(l)})}{u_l} - \nu_{h_1}(\theta_i^{(1)}) \cdot \frac{\nu_{h_j}(\theta_i^{(j)})}{u_j} - \nu_{h_1}(\theta_i^{(1)}) \cdot \frac{\nu_{h_j}(\theta_i^{(j)})}{u_j} \cdot \frac{1}{\sum_{s=1}^{k} a_s \nu_{h_s}(\theta_i^{(s)}) / u_s}, \quad (A.10)
\]

where \( u = (u_2, \ldots, u_k)' \) and \( u_l > 0 \), for \( l = 2, \ldots, k \). (On the right side of (A.10), \( u_1 \) is taken to be 1.) Note that setting \( u = d \) gives \( G(d) = \frac{1}{n} \sum_{l=1}^{k} \sum_{i=1}^{n_l} Z_{i,l}^{(l)} Z_{i,l}' \). By the Mean Value
Theorem, we know that there exists a $d^*$ between $d$ and $\hat{d}$ such that

$$G(\hat{d}) = G(d) + \nabla G(d^*)(\hat{d} - d) = R_{j,j'} + \nabla G(d^*)(\hat{d} - d) + o_p(1).$$

Note that the last equality above comes from applying the SLLN. Next we show that $\nabla G(d^*) = O_p(1)$. We have three cases for $t = 2, \ldots, k$.

**Case 1: $t \notin \{j, j'\}$.** We have

$$\left| [\nabla G(d^*)]_{t-1} \right| \cdot I \leq \sum_{i=1}^{k} 2a_i \left[ \sum_{n=1}^{n_i} \frac{\left[ \nu_{h_i}(\theta_i) / d_i - \nu_{h_i}(\theta_i) \right]}{d_i^2} \sum_{s=1}^{k} \frac{a_i \nu_{h_i}(\theta_i)}{s} \right] \bigg\| \sum_{n=1}^{n_i} \frac{\left[ \nu_{h_i}(\theta_i) / d_i - \nu_{h_i}(\theta_i) \right]}{d_i^2} \sum_{s=1}^{k} \frac{a_i \nu_{h_i}(\theta_i)}{s} \bigg\|^2$$

The term inside the inner sum is bounded, so we can conclude that $[\nabla G(d^*)]_{t-1}$ is bounded in probability, as it is bounded by a $O_p(1)$ term on $I$.

**Case 2: $j \neq j'$, $t \in \{j, j'\}$, say $t = j$.** We have

$$[\nabla G(d^*)]_{j-1} = \sum_{i=1}^{k} a_i \left[ \sum_{n=1}^{n_i} \frac{2(\nu_{h_i}(\theta_i) / d_i - \nu_{h_i}(\theta_i)) \left( \nu_{h_i}(\theta_i) / d_i - \nu_{h_i}(\theta_i) \right)}{d_i^2} a_i \nu_{h_i}(\theta_i) \right]$$

and this is bounded in probability.

**Case 3: $t = j = j'$.** We have

$$[\nabla G(d^*)]_{j-1} = \sum_{i=1}^{k} a_i \left[ \sum_{n=1}^{n_i} \frac{\nu_{h_i}(\theta_i) / d_i - \nu_{h_i}(\theta_i)}{d_i^2} \sum_{s=1}^{k} \frac{a_i \nu_{h_i}(\theta_i)}{s} \right]$$

and again this is bounded in probability.
Therefore

\[ G(\tilde{d}) = R_{j,j'} + \nabla G(d^*)(\tilde{d} - d) + o_p(1) = R_{j,j'} + O_p(1) + o_p(1) \xrightarrow{p} R_{j,j'}. \]

Similar arguments extend to the case \( j = 1 \) or \( j' = 1 \). By the fact that \( R \) is assumed invertible, we have

\[ n(\tilde{Z}^\prime \tilde{Z})^{-1} \xrightarrow{p} R^{-1}. \]  \hspace{1cm} (A.11)

In a similar way, it can be shown that

\[ \tilde{Z}\tilde{Y}/n \xrightarrow{p} v, \]  \hspace{1cm} (A.12)

where \( v \) is the same limit vector to which \( Z'Y/n \) has been proved to converge in Doss (2010). Combining (A.11) and (A.12) we have

\[
(\hat{\beta}_0(\tilde{d}), \hat{\beta}(\tilde{d})) = [n(\tilde{Z}^\prime \tilde{Z})^{-1}][\tilde{Z}\tilde{Y}/n] \xrightarrow{p} (\beta_{0,\lim}, \beta_{\lim}) = R^{-1}v.
\]

Let \( e(j,l) = E(Z^{(j)}_{i,l}) \). We now have

\[
\sqrt{n}(\tilde{\beta}(\tilde{d}) - \tilde{\beta}_{\lim}) = \sum_{j=2}^{k} (\beta_{j,\lim} - \beta_j(\tilde{d})) \left( \sum_{l=1}^{k} a_l n^{1/2} \sum_{i=1}^{n_j} (\tilde{Z}_{i,l}^{(j)} - e(j,l)) n_l \right)
\]

\[ = \sum_{j=2}^{k} o_p(1) \left( \sum_{l=1}^{k} a_l n^{1/2} \sum_{i=1}^{n_j} (\tilde{Z}_{i,l}^{(j)} - e(j,l)) n_l \right). \]  \hspace{1cm} (A.13)

To show that (A.13) converges to 0 in probability it suffices to show that for each \( l \) and \( j \)

\[ n_l^{1/2} \sum_{i=1}^{n_j} \left( \frac{\tilde{Z}_{i,l}^{(j)} - e(j,l)}{n_l} \right) = o_p(1). \]  \hspace{1cm} (A.14)

For fixed \( j \in \{2, \ldots, k\} \) and \( l \in \{1, \ldots, k\} \), define

\[ H(u) = n_l^{-1/2} \sum_{i=1}^{n_j} u_{j}(\theta_i^{(l)}/u_j - u_{j}(\theta_i^{(l)}) \sum_{s=1}^{k} a_s u_{s}(\theta_i^{(l)})/u_s. \]
for $u = (u_2, \ldots, u_k)'$ with $u_i > 0$, $l = 2, \ldots, k$, $u_1 = 1$. Note that $H(d) = n_i^{-1/2} \sum_{i=1}^{n_i} Z_{i,l}^{(j)}$. To see why (A.14) is true we begin by writing

$$n_i^{1/2} \sum_{i=1}^{n_i} \left( \frac{Z_{i,l}^{(j)} - e(j, l)}{n_i} \right) = n_i^{1/2} \sum_{i=1}^{n_i} \left( \frac{Z_{i,l}^{(j)} - Z_{i,l}^{(j)}}{n_i} \right) + n_i^{1/2} \sum_{i=1}^{n_i} \left( \frac{Z_{i,l}^{(j)} - e(j, l)}{n_i} \right)$$

$$= H(\hat{d}) - H(d) + O_p(1). \quad (A.15)$$

Note that the fact that $n_i^{1/2} \sum_{i=1}^{n_i} ([Z_{i,l}^{(j)} - e(j, l)]/n_i) = O_p(1)$, which was used to establish the second equality in (A.15), is proved in Doss (2010). Now, applying the Mean Value Theorem to the function $H$, we know that there exists a point $d^*$ between $d$ and $\hat{d}$ such that (A.15) becomes

$$n_i^{1/2} \sum_{i=1}^{n_i} \left( \frac{Z_{i,l}^{(j)} - e(j, l)}{n_i} \right) = \nabla H(d^*)(\hat{d} - d) + O_p(1)$$

$$= \sqrt{n_i^{1/2} \nabla H(d^*)' \sqrt{n_i^{1/2} \nabla H(d^*)} + O_p(1), \quad (A.16)$$

so that the right side of (A.16) is $O_p(1)$. To see this last assertion, note that the $(t - 1)^{th}$ element of the gradient of $H$, $[\nabla H(d)]_{t-1}$, is given by

$$\begin{align*}
&\begin{cases}
\frac{1}{n_i^{1/2}} \sum_{i=1}^{n_i} \frac{(\nu_h(\theta_i^{(l)})/d_j - \nu_h(\theta_i^{(l)})) a_t \nu_h(\theta_i^{(l)})}{d^2 \left( \sum_{s=1}^{k} a_s \nu_h(\theta_i^{(l)})/d_s \right)^2} & \text{if } t \neq j, \\
\frac{1}{n_i^{1/2}} \sum_{i=1}^{n_i} \frac{-\nu_h(\theta_i^{(l)})}{d^2 \sum_{s=1}^{k} a_s \nu_h(\theta_i^{(l)})/d_s} + \sum_{i=1}^{n_i} \frac{(\nu_h(\theta_i^{(l)})/d_j - \nu_h(\theta_i^{(l)})) a_j \nu_h(\theta_i^{(l)})}{d^2 \left( \sum_{s=1}^{k} a_s \nu_h(\theta_i^{(l)})/d_s \right)^2} & \text{if } t = j.
\end{cases}
\end{align*}$$
Let $\epsilon \in (0, \min(d_2, \ldots, d_k))$. Then $P(\|d^* - d\| \leq \epsilon) \to 1$. For $t \neq j$ we have

$$|n_i^{-1/2}[\nabla H(d^*)]_{i-1}| \cdot I \leq n_i^{-1} \sum_{i=1}^{n} \frac{|\nu_h(\theta_i^{(l)}) / d_j^* - \nu_h(\theta_i^{(l)})| a_t \nu_h(\theta_i^{(l)})}{d_j^2 (\sum_{s=1}^{k} a_s \nu_h(\theta_i^{(l)}) / d_j^*)^2}$$

$$\leq n_i^{-1} \sum_{i=1}^{n} \nu_h(\theta_i^{(l)}) a_t \nu_h(\theta_i^{(l)}) + n_i^{-1} \sum_{i=1}^{n} \frac{\nu_h(\theta_i^{(l)}) a_t \nu_h(\theta_i^{(l)})}{d_j^2 (\sum_{s=1}^{k} a_s \nu_h(\theta_i^{(l)}) / d_j^*)^2}$$

$$\leq n_i^{-1} \sum_{i=1}^{n} \nu_h(\theta_i^{(l)}) a_t \nu_h(\theta_i^{(l)}) + n_i^{-1} \sum_{i=1}^{n} \frac{\nu_h(\theta_i^{(l)}) a_t \nu_h(\theta_i^{(l)})}{(d_j^2 (d_j - \epsilon) (\sum_{s=1}^{k} a_s \nu_h(\theta_i^{(l)}) / (d_j^* + \epsilon))^2}$$

$$+ n_i^{-1} \sum_{i=1}^{n} \nu_h(\theta_i^{(l)}) a_t \nu_h(\theta_i^{(l)}) = O_p(1) + O_p(1) = O_p(1).$$

Similarly,

$$|n_i^{-1/2}[\nabla H(d^*)]_{i-1}| \cdot I \leq \frac{1}{n_i} \sum_{i=1}^{n} \frac{\nu_h(\theta_i^{(l)})}{(d_j^2 (d_j - \epsilon) (\sum_{s=1}^{k} a_s \nu_h(\theta_i^{(l)}) / (d_j^* + \epsilon))^2}$$

$$+ \frac{1}{n_i} \sum_{i=1}^{n} \frac{\nu_h(\theta_i^{(l)}) a_t \nu_h(\theta_i^{(l)})}{(d_j - \epsilon)^3 (\sum_{s=1}^{k} a_s \nu_h(\theta_i^{(l)}) / (d_j^* + \epsilon))^2}$$

$$+ \frac{1}{n_i} \sum_{i=1}^{n} \frac{\nu_h(\theta_i^{(l)}) a_t \nu_h(\theta_i^{(l)})}{(d_j - \epsilon)^2 (\sum_{s=1}^{k} a_s \nu_h(\theta_i^{(l)}) / (d_j^* + \epsilon))^2},$$

and the right side of this inequality is $O_p(1)$, as it is the sum of three $O_p(1)$ terms.

So (A.16) now implies that

$$n_i^{1/2} \sum_{i=1}^{n} \left( \frac{\tilde{Z}_{i,l}^{(j)} - e(j, l)}{n_i} \right) = \sqrt{\frac{1}{n}} \sqrt{O_p(1) O_p(1) + O_p(1) = O_p(1).}$$

We now consider $\sqrt{n}(\tilde{\gamma}_{\beta_{lm}} - \gamma_{\beta_{lm}})$, the middle term in (A.7). Define

$$K(u) = \frac{1}{n} \sum_{l=1}^{k} \sum_{i=1}^{n} \left( \frac{\nu_h(\theta_i^{(l)})}{\sum_{s=1}^{k} a_s \nu_h(\theta_i^{(l)}) / u_s} - \sum_{j=2}^{k} \beta_{j, \lim} \frac{\nu_h(\theta_i^{(l)}) / u_j - \nu_h(\theta_i^{(l)})}{\sum_{s=1}^{k} a_s \nu_h(\theta_i^{(l)}) / u_s} \right),$$

64
where $u = (u_2, \ldots, u_k)'$, and $u_l > 0$ for $l = 2, \ldots, k$. By Taylor series expansion, we have
\[
\sqrt{n} (\tilde{\theta}^d - \tilde{\theta}^{d^*}) = \sqrt{n} \nabla K(d)'(\hat{d} - d) + \sqrt{n} \frac{1}{2} (\hat{d} - d)' \nabla^2 K(d^*)(\hat{d} - d), \tag{A.17}
\]
where $d^*$ is between $\hat{d}$ and $d$. We now focus our attention on $\nabla K(d)$. For $t = 2, \ldots, k$ we have
\[
[\nabla K(d)]_{t-1} = \frac{1}{n} \sum_{l=1}^{k} \sum_{i=1}^{n_t} \left[ \begin{array}{l} \nu_h(\theta_j^{(t)}) a_t \nu_h(\theta_j^{(t)}) \\ d_t^2 (\sum_{s=1}^{k} a_s \nu_h(\theta_i^{(t)}) / d_s)^2 \\ - \sum_{j=2}^{k} \beta_{j, \text{lim}} \left( \nu_h(\theta_j^{(t)}) / d_j - \nu_h(\theta_i^{(t)}) \right) a_t \nu_h(\theta_j^{(t)}) \\ d_t^2 (\sum_{s=1}^{k} a_s \nu_h(\theta_i^{(t)}) / d_s)^2 \\ + \beta_{t, \text{lim}} \frac{\nu_h(\theta_i^{(t)})}{d_t^2} \sum_{s=1}^{k} a_s \nu_h(\theta_i^{(t)}) / d_s \\ - \beta_{t, \text{lim}} \frac{\nu_h(\theta_i^{(t)}) / d_t - \nu_h(\theta_i^{(t)}) a_t \nu_h(\theta_i^{(t)})}{d_t^2 (\sum_{s=1}^{k} a_s \nu_h(\theta_i^{(t)}) / d_s)^2} \right] \\
\cdot \nabla h(y) d \theta \\
- \sum_{j=2}^{k} \beta_{j, \text{lim}} \int \frac{a_t \nu_h(\theta)}{d_t^2} \sum_{s=1}^{k} a_s \nu_h(\theta) / d_s \cdot \nu_h(y) d \theta \\
+ \sum_{j=2}^{k} \beta_{j, \text{lim}} \int \frac{a_t \nu_h(\theta)}{d_t^2} \sum_{s=1}^{k} a_s \nu_h(\theta) / d_s \cdot \nu_h(y) d \theta \\
- \beta_{t, \text{lim}} \int \frac{a_t \nu_h(\theta)}{d_t^2} \sum_{s=1}^{k} a_s \nu_h(\theta) / d_s \cdot \nu_h(y) d \theta \\
+ \beta_{t, \text{lim}} \int \frac{a_t \nu_h(\theta)}{d_t^2} \sum_{s=1}^{k} a_s \nu_h(\theta) / d_s \cdot \nu_h(y) d \theta \\
= \frac{B(h, h_1)}{d_t^2} \int \frac{a_t \nu_h(\theta)}{d_t^2} \sum_{s=1}^{k} a_s \nu_h(\theta) / d_s \cdot \nu_h(y) d \theta \\
- \sum_{j=2}^{k} \beta_{j, \text{lim}} \int \frac{a_t \nu_h(\theta)}{d_t^2} \sum_{s=1}^{k} a_s \nu_h(\theta) / d_s \cdot \nu_h(y) d \theta \\
+ \sum_{j=2}^{k} \beta_{j, \text{lim}} \int \frac{a_t \nu_h(\theta)}{d_t^2} \sum_{s=1}^{k} a_s \nu_h(\theta) / d_s \cdot \nu_h(y) d \theta + \beta_{t, \text{lim}} \frac{1}{d_t} \\
:= [v(h)]_{t-1}. \tag{A.18}
\]
where the notation in (A.18) indicates that \( w(h) \) denotes the finite vector limit to which \( \nabla K(d) \) converges. We now deal with the Hessian matrix \( \nabla^2 K(d^*) \). For \( t \neq u \) we have

\[
[n \nabla^2 K(d^*)]_{n-1,u-1} = \frac{1}{n} \sum_{i=1}^{k} \sum_{j=1}^{n_i} \left[ \frac{2 \nu_h(\theta_i^{(l)}) a_t \nu_{hs}(\theta_i^{(l)}) a_u \nu_{hs}(\theta_i^{(l)})}{d_t^2 d_u^2 (\sum_{s=1}^{k} a_s \nu_{hs}(\theta_i^{(l)})/d_s)^3} \right. \\
- \sum_{j=2}^{k} \sum_{j \neq t}^{n} \beta_{j,lim} \frac{2 [\nu_h(\theta_i^{(l)})/d_t^* - \nu_h(\theta_i^{(l)})] a_t \nu_{hs}(\theta_i^{(l)}) a_u \nu_{hs}(\theta_i^{(l)})}{d_t^2 d_u^2 (\sum_{s=1}^{k} a_s \nu_{hs}(\theta_i^{(l)})/d_s)^3} \\
+ \beta_{u,lim} \frac{\nu_h(\theta_i^{(l)}) a_t \nu_{hs}(\theta_i^{(l)})}{d_t^2 d_u^2 (\sum_{s=1}^{k} a_s \nu_{hs}(\theta_i^{(l)})/d_s)^2} \\
- \beta_{u,lim} \frac{2 (\nu_h(\theta_i^{(l)})/d_u^* - \nu_h(\theta_i^{(l)}) a_t \nu_{hs}(\theta_i^{(l)}) a_u \nu_{hs}(\theta_i^{(l)})}{d_t^2 d_u^2 (\sum_{s=1}^{k} a_s \nu_{hs}(\theta_i^{(l)})/d_s)^3} \\
+ \beta_{t,lim} \frac{\nu_h(\theta_i^{(l)}) a_t \nu_{hs}(\theta_i^{(l)})}{d_t^2 d_u^2 (\sum_{s=1}^{k} a_s \nu_{hs}(\theta_i^{(l)})/d_s)^2} \\
- \beta_{t,lim} \frac{2 (\nu_h(\theta_i^{(l)})/d_t - \nu_h(\theta_i^{(l)}) a_t \nu_{hs}(\theta_i^{(l)}) a_u \nu_{hs}(\theta_i^{(l)})}{d_t^2 d_u^2 (\sum_{s=1}^{k} a_s \nu_{hs}(\theta_i^{(l)})/d_s)^3} \right],
\]

and as before, it can be shown that this is bounded in probability. Similarly, we can show that the diagonal terms of \( \nabla^2 K(d^*) \) are also bounded in probability. Therefore, using the fact that \( \nabla^2 K(d^*) \) is bounded in probability, we can now rewrite (A.17) as

\[
\sqrt{n}(\tilde{\eta} - \tilde{\eta}) = \sqrt{\frac{n}{N}} w(h)' \sqrt{N}(\tilde{\eta} - h) + \sqrt{\frac{n}{N^2 \sqrt{N}}} \sqrt{N}(\tilde{\eta} - d) O_p(1) \sqrt{N}(\tilde{\eta} - d) \\
= \sqrt{q w(h)'} \sqrt{N}(\tilde{\eta} - d) + o_p(1).
\]

Together with (A.6), this gives

\[
\sqrt{n}(\tilde{\eta} - B(h, h_1) = \sqrt{q w(h)'} \sqrt{N}(\tilde{\eta} - d) + \sqrt{n}(\tilde{\eta} - B(h, h_1) + o_p(1) \\
\xrightarrow{d} \mathcal{N}(0, q w(h)') \Sigma w(h) + \sigma^2(h)),
\]

by the independence of the two sampling stages, the assumption that \( \sqrt{N}(\tilde{\eta} - d) \) is asymptotically normal with mean 0 and variance \( \Sigma \), and the result from Doss (2010) that \( \sqrt{n}(\tilde{\eta} - B(h, h_1)) \) is asymptotically normal with mean 0 and variance \( \sigma^2(h) \).
Proof of Theorem 3

First, we note that

$$\sqrt{n}(\hat{\ell}[r](h, \hat{d}) - \ell[r](h)) = \sqrt{n}(\hat{\ell}[r](h, \hat{d}) - \ell[r](h, d)) + \sqrt{n}(\ell[r](h, d) - \ell[r](h)). \quad (A.19)$$

We begin by analyzing the second term on the right side of (A.19), which only involves randomness from the second stage of sampling, and show that it is asymptotically normal. As for the first term, a closer examination reveals that it is also asymptotically normal, with all its randomness coming from Stage 1. The asymptotic normality of the sum of these two terms then follows immediately from the independence of the two stages of sampling.

Note that $\sum_{i=1}^{k} a_i E(Y_{1,i}) = \ell[r](h) \cdot B(h, h_1)$, and in particular, when $f \equiv 1$, this gives $\sum_{i=1}^{k} a_i E(Y_{1,i}) = B(h, h_1)$. Also, we have

$$n^{1/2} \left( \frac{1}{n} \sum_{i=1}^{k} \sum_{i=1}^{n} Y_{i,i} - \ell[r](h) \cdot B(h, h_1) \right) = n^{1/2} \left( \frac{1}{n} \sum_{i=1}^{k} \sum_{i=1}^{n} Y_{i,i} - \sum_{i=1}^{k} a_i E(Y_{1,i}) \right)$$

$$= \sum_{i=1}^{k} a_i^{1/2} \cdot \frac{1}{n^{1/2}} \sum_{i=1}^{n} \left[ \begin{bmatrix} Y_{i,i} \\ Y_{i,1} \end{bmatrix} - \begin{bmatrix} E(Y_{1,i}) \\ E(Y_{1,1}) \end{bmatrix} \right]. \quad (A.20)$$

By condition (2–17), assumption A2 of Theorem 1, and the assumed geometric ergodicity and independence of the $k$ Markov chains used, the vector in (A.20) converges in distribution to a normal random vector with mean 0 and covariance matrix $\Gamma(h) = \sum_{i=1}^{k} a_i \Gamma_i(h)$, where

$$\Gamma_i(h) = \begin{pmatrix} \gamma_{11} & \gamma_{12} \\ \gamma_{21} & \gamma_{22} \end{pmatrix},$$
\[ \gamma_{11} = \text{Var}(Y_{1,l}^r) + 2 \sum_{g=1}^{\infty} \text{Cov}(Y_{1,l}^r, Y_{1+g,l}^r), \]
\[ \gamma_{12} = \gamma_{21} = \text{Cov}(Y_{1,l}^r, Y_{1,l}^r) + \sum_{g=1}^{\infty} [\text{Cov}(Y_{1,l}^r, Y_{1+g,l}^r) + \text{Cov}(Y_{1,l}^r, Y_{1+g,l}^r)], \]
\[ \gamma_{22} = \text{Var}(Y_{1,l}^r) + 2 \sum_{g=1}^{\infty} \text{Cov}(Y_{1,l}^r, Y_{1+g,l}^r). \]

Since \( \tilde{\gamma}^r(h, d) \) is given by the ratio (2–15), in view of (A.20), its asymptotic distribution may be obtained by applying the delta method to the function \( g(u, v) = u/v \). This gives
\[ \sqrt{n}(\tilde{\gamma}^r(h, d) - \gamma^r(h)) \xrightarrow{d} N(0, \rho(h)), \]
where
\[ \rho(h) = \nabla g(\gamma^r(h)B(h, h_1), B(h, h_1))' \cdot \Gamma(h) \cdot \nabla g(\gamma^r(h)B(h, h_1), B(h, h_1)), \quad (A.21) \]

with \( \nabla g(u, v) = (1/v, -u/v^2)' \).

We now consider the first term on the right side of (A.19). Define
\[ L(u) = \sum_{l=1}^{k} \sum_{i=1}^{n_l} \frac{f(\theta_{l}^{(i)})\nu_0(\theta_{l}^{(i)})}{\sum_{s=1}^{k} \partial_s \nu_0(\theta_{l}^{(i)})/u_s} \sum_{l=1}^{k} \sum_{i=1}^{n_l} \frac{\nu_0(\theta_{l}^{(i)})}{\sum_{s=1}^{k} \partial_s \nu_0(\theta_{l}^{(i)})/u_s} \]for \( u = (u_2, \ldots, u_k)' \) with \( u_i > 0 \) for \( l = 2, \ldots, k \). Then
\[ L(d) = \tilde{\gamma}^r(h, d) = \sum_{l=1}^{k} \sum_{i=1}^{n_l} \frac{Y_{l,i}^r}{\sum_{l=1}^{k} \sum_{i=1}^{n_l} Y_{l,i}^r}, \]
and \( \sqrt{n}(\tilde{\gamma}^r(h, d) - \gamma^r(h, d)) = \sqrt{n}(L(\hat{d}) - L(d)) \). Now, by the Taylor series expansion of \( L \) about \( d \) we get
\[ \sqrt{n}(\tilde{\gamma}^r(h, \hat{d}) - \gamma^r(h, d)) = \sqrt{n}\nabla L(d)'(\hat{d} - d) + \frac{\sqrt{n}}{2}(\hat{d} - d)'\nabla^2 L(d^*)(\hat{d} - d), \]
where \( d^* \) is between \( d \) and \( \hat{d} \). First, we show that the gradient \( \nabla L(d) \) converges almost surely to a finite constant vector by proving that each one of its components,
\[ [L(d)]_{j-1}, j = 2, \ldots, k, \text{ converges almost surely. We have} \]
\[
[L(d)]_{j-1} = \sum_{i=1}^{k} \sum_{l=1}^{n_i} \frac{a_i f(\theta_i^{(l)}) \nu_h(\theta_i^{(l)}) \nu_h(\theta_i^{(l)})}{d_j^2 \left( \sum_{s=1}^{k} a_s \nu_h(\theta_i^{(l)}) / d_s \right)^2}
\]
\[
\sum_{i=1}^{k} \sum_{l=1}^{n_i} \frac{\nu_h(\theta_i^{(l)})}{\sum_{s=1}^{k} a_s \nu_h(\theta_i^{(l)}) / d_s}
\]
\[
= \frac{B(h, h_1)}{d_j^2} \int \frac{a_i f(\theta) \nu_h(\theta)}{\sum_{s=1}^{k} a_s \nu_h(\theta) / d_s} \cdot \nu_{h,y}(\theta) d\theta
\]
\[
= \frac{B(h, h_1) \cdot B(h, h_1) \cdot \frac{a_i \nu_h(\theta)}{d_j^2} \int \frac{\nu_{h,y}(\theta) d\theta}{\sum_{s=1}^{k} a_s \nu_h(\theta) / d_s}}{B(h, h_1)^2}
\]
\[
= \frac{1}{d_j^2} \int \frac{a_i f(\theta) \nu_h(\theta)}{\sum_{s=1}^{k} a_s \nu_h(\theta) / d_s} \cdot \nu_{h,y}(\theta) d\theta - \frac{|\theta|}{d_j^2} \int \frac{a_i \nu_h(\theta)}{d_j^2} \int \frac{\nu_{h,y}(\theta) d\theta}{\sum_{s=1}^{k} a_s \nu_h(\theta) / d_s} \cdot \nu_{h,y}(\theta) d\theta,
\]
\[
:= [\nu(h)]_{j-1}, j = 2, \ldots, k. \quad (A.22)
\]

As in the proof of Theorem 1, it can be shown that each element of the second-derivative matrix \(\nabla^2 L(d^*)\) is \(O_p(1)\). Now, we can rewrite (A.19) as
\[
\sqrt{n}(|\theta| L(h, \hat{d}) - |\theta| L(h)) = \sqrt{\frac{n}{N} \nabla L(d)'} \sqrt{\hat{N}}(\hat{d} - d)
\]
\[
+ \frac{1}{2 \sqrt{N}} \sqrt{\frac{n}{N} \left[ \sqrt{N}(\hat{d} - d) \right] \nabla^2 L(d^*) \left[ \sqrt{N}(\hat{d} - d) \right]}
\]
\[
+ \sqrt{n}(|\theta| L(d, h) - |\theta| L(h))
\]
\[
= \sqrt{n} q \nu(h) \sqrt{\hat{N}}(\hat{d} - d) + \sqrt{n}(|\theta| L(h, d) - |\theta| L(h)) + o_p(1).
\]

Since the two sampling stages are assumed to be independent, we conclude that
\[
\sqrt{n}(|\theta| L(h, \hat{d}) - |\theta| L(h)) \xrightarrow{d} \mathcal{N}(0, q \nu(h) \Sigma \nu(h) + \rho(h)). \quad \Box
\]
Proof of Theorem 4

Here $Z$ and $Y$ represent the matrix and vector, respectively, previously defined in (A.8) and (A.9). In addition, let $Z^{[r]}$ denote the $n \times (k + 1)$ matrix with transpose

$$(Z^{[r]})' = \begin{pmatrix} 1 & \ldots & 1 & 1 \ldots 1 & \ldots & 1 \ldots 1 \\ Z_{1,1}^{[r]} & \ldots & Z_{n,1}^{[r]} & Z_{1,2}^{[r]} & \ldots & Z_{n,2}^{[r]} & \ldots & Z_{1,k}^{[r]} & \ldots & Z_{n,k}^{[r]} \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \ddots & \vdots & \ddots & \vdots \\ Z_{1,1}^{[r]} & \ldots & Z_{n,1}^{[r]} & Z_{1,2}^{[r]} & \ldots & Z_{n,2}^{[r]} & \ldots & Z_{1,k}^{[r]} & \ldots & Z_{n,k}^{[r]} \end{pmatrix} \quad (A.23)$$

and let $Y^{[r]}$ be the vector

$$Y^{[r]} = (Y_{1,1}^{[r]}, \ldots, Y_{n,1}^{[r]}, Y_{1,2}^{[r]}, \ldots, Y_{n,2}^{[r]}, \ldots, Y_{1,k}^{[r]}, \ldots, Y_{n,k}^{[r]})'.$$

We know from Doss (2010) that the least squares estimate when $Y$ is regressed on $Z$, denoted by $(\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k) := (\hat{\beta}_0, \hat{\beta})$, converges almost surely to $(\beta_{0,\lim}, \beta_{1,\lim}) = R^{-1}v$. In a similar way, we will show here that the least squares estimate when $Y^{[r]}$ is regressed on $Z^{[r]}$, $(\hat{\beta}_0^{[r]}, \hat{\beta}_1^{[r]}, \ldots, \hat{\beta}_k^{[r]}) = (\hat{\beta}_0^{[r]}, \hat{\beta}_1^{[r]}, \ldots, \hat{\beta}_k^{[r]})$, converges almost surely to a vector $(\beta_{0,\lim}^{[r]}, \beta_{1,\lim}^{[r]})$.

Note that, under the assumption that $[(Z^{[r]})'Z^{[r]}]^{-1}$ exists,

$$(\beta_{0,\lim}^{[r]}, \beta_{1,\lim}^{[r]}) = n[(Z^{[r]})'Z^{[r]}]^{-1} \frac{(Z^{[r]})'Y^{[r]}}{n}.$$  

Since A4 is satisfied, we have

$$\frac{1}{n} \sum_{i=1}^{n} \sum_{l=1}^{n_j} Z_{i,l}^{[r]} Z_{i,l}^{[r]'} = \sum_{i=1}^{n} \frac{1}{n} \sum_{l=1}^{n_j} Z_{i,l}^{[r]} Z_{i,l}^{[r]'} \xrightarrow{a.s.} R_{j+1,j'+1}^{[r]}, \quad j, j' = 0, \ldots, k,$$

and hence $(Z^{[r]})'Z^{[r]}/n \xrightarrow{a.s.} R^{[r]}$. Therefore by A6, with probability one $(Z^{[r]})'Z^{[r]}$ is nonsingular for large $n$, and furthermore

$$n[(Z^{[r]})'Z^{[r]}]^{-1} \xrightarrow{a.s.} (R^{[r]})^{-1}. \quad (A.25)$$
By condition A5, we also have
\[
\frac{(Z^{[r]})' Y^{[r]}}{n} = \left( \frac{1}{n} \sum_{i=1}^{n} Z_{i,1}^{[r](0)} Y_{i,1}^{[r]} \right) \ldots \left( \frac{1}{n} \sum_{i=1}^{n} Z_{i,k}^{[r](k)} Y_{i,k}^{[r]} \right) + \sum_{i=1}^{k} a_i E \left( Z_{i,1}^{[r](0)} Y_{i,1}^{[r]} \right) \ldots + \sum_{i=1}^{k} a_i E \left( Z_{i,k}^{[r](k)} Y_{i,k}^{[r]} \right). \tag{A.26}
\]

Let \( v^{[r]} = (v_0^{[r]}, \ldots, v_k^{[r]})' \) denote the vector on the right side of (A.26). Combining (A.25) and (A.26) we get
\[
(\beta_0^{[r]}, \beta_i^{[r]}) \xrightarrow{a.s.} (\beta_{0,\text{lim}}^{[r]}, \beta_{i,\text{lim}}^{[r]}) = (R^{[r]})^{-1} v^{[r]}.	ag{A.27}
\]

Let
\[
\hat{\beta}_{i,\text{lim}}^{[r]} = \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{k} \left( \beta_{i,j}^{[r]} - \left( \frac{1}{n} \sum_{i=1}^{n} U_{i,j}^{[r]} \right) \right) = \frac{1}{n} \sum_{i=1}^{n} \left( \frac{1}{n} \sum_{i=1}^{n} U_{i,j}^{[r]} \right),
\]

where
\[
U_{i,j}^{[r]} = \begin{pmatrix} U_{i,j}^{[r](1)} \\ U_{i,j}^{[r](2)} \end{pmatrix} = \begin{pmatrix} Y_{i,j}^{[r]} - \beta_{j,\text{lim}}^{[r]} Z_{i,j}^{[r](j)} \\ Y_{i,j}^{[r]} - \sum_{j=1}^{k} \beta_{j,\text{lim}}^{[r]} Z_{i,j}^{[r](j)} \end{pmatrix}.
\]

Also, let \( \mu_{i}^{[r]} = E(U_{i,j}^{[r]}) \).

Now since A2, A3, and A4 hold, for each \( i = 1, \ldots, k \) we have
\[
n_i^{1/2} \left( \frac{1}{n} \sum_{i=1}^{n} U_{i,j}^{[r]} - \mu_{i}^{[r]} \right) \xrightarrow{d} \mathcal{N}(0, \Sigma_{i}^{[r]}),
\]

where
\[
\Sigma_{i}^{[r]} = \begin{pmatrix} \sigma_{i,11}^{2} & \sigma_{i,12} \\ \sigma_{i,21} & \sigma_{i,22}^{2} \end{pmatrix},
\]

with
\[
\sigma_{i,11}^{2} = \text{Var}(U_{1,j}^{[r](1)}) + 2 \sum_{g=1}^{\infty} \text{Cov}(U_{1,j}^{[r](1)}, U_{1+j,g}^{[r](1)}),
\]
\[
\sigma_{i,12} = \text{Cov}(U_{1,j}^{[r](1)}, U_{1,j}^{[r](2)}) + \sum_{g=1}^{\infty} \left[ \text{Cov}(U_{1,j}^{[r](1)}, U_{1+j,g}^{[r](2)}) + \text{Cov}(U_{1,j}^{[r](2)}, U_{1+j,g}^{[r](1)}) \right],
\]
\[
\sigma_{i,22}^{2} = \text{Var}(U_{1,j}^{[r](2)}) + 2 \sum_{g=1}^{\infty} \text{Cov}(U_{1,j}^{[r](2)}, U_{1+j,g}^{[r](2)}).
\]
By the assumed independence of the $k$ Markov chains, we have

$$n^{1/2} \left( \hat{J}_{\beta_{\lim}, \beta_{\lim}} - \sum_{i=1}^{k} a_i \mu_i^{[r]} \right) \xrightarrow{d} \mathcal{N}(0, \Sigma^{[r]}), \quad (A.28)$$

where

$$\Sigma^{[r]} = \sum_{i=1}^{k} a_i \Sigma_i^{[r]}. \quad (A.29)$$

We now show that

$$\sum_{i=1}^{k} a_i \mu_i^{[r]} = \left( [^{[r]}(h)B(h, h_1) \right), \quad (A.30)$$

and to do this we write

$$\sum_{i=1}^{k} a_i \mu_i^{[r]} = \sum_{i=1}^{k} a_i E(U_{1,i}^{[r]})$$

$$= \sum_{i=1}^{k} a_i \left( E(Y_{1,i}^{[r]}) - \sum_{j=1}^{k} \beta_j^{[r]} E(Z_{1,i}^{[j]}) \right)$$

$$= \sum_{i=1}^{k} a_i \left( Y_{1,i}^{[r]} - \sum_{j=1}^{k} \beta_j^{[r]} \sum_{i=1}^{k} a_i E(Z_{1,i}^{[j]}) \right)$$

$$= \sum_{i=1}^{k} a_i E(Y_{1,i}) - \sum_{j=2}^{k} \beta_j^{[r]} \sum_{i=1}^{k} a_i E(Z_{1,i}^{[j]})$$

$$= \left( \sum_{i=1}^{k} a_i E(Y_{1,i}^{[r]}) \right)$$

$$= \left( \sum_{i=1}^{k} a_i E(Y_{1,i}) \right)$$

$$= \left( [^{[r]}(h)B(h, h_1) \right).$$

The next-to-last equality being a consequence of the readily verifiable fact that

$$\sum_{i=1}^{k} a_i E(Z_{1,i}^{[j]}) = 0 \quad \text{and} \quad \sum_{i=1}^{k} a_i E(Z_{1,i}^{[j]}) = 0 \quad \text{for} \quad j = 2, \ldots, k. \quad (A.31)$$

From (A.30) and (A.28) we conclude that

$$n^{1/2} \left[ \frac{1}{n} \left( \hat{J}_{\beta_{\lim}, \beta_{\lim}} - \left( [^{[r]}(h)B(h, h_1) \right) \right) \right] \xrightarrow{d} \mathcal{N}(0, \Sigma^{[r]}).$$
Consider now the difference

\[ n^{1/2} \left( \hat{J}_{\hat{\beta}, \hat{\beta}^l} | \hat{\beta}_{\hat{\beta}, \hat{\beta}^l} \right) = \left( n^{1/2} \sum_{i=1}^{N} \left( \beta_{i, \hat{\beta}} - \hat{\beta} \right) \left( \frac{1}{n} \sum_{i=1}^{n} \sum_{i=1}^{n} Z_{i,j} \right) \right) \]

where the last equality follows from (A.31). By the assumption that the chains are geometrically ergodic (condition A1), the boundedness of \( Z_{i,j} \)'s, and the moment condition imposed on \( f \) in A4, we know that \( n^{1/2} \sum_{i=1}^{n} \left( \frac{Z_{i,j} - E(Z_{i,j})}{n} \right) \) and \( n^{1/2} \sum_{i=1}^{n} \left( \frac{Z_{i,j} - E(Z_{i,j})}{n} \right) \) are asymptotically normal, hence \( O_p(1) \). This fact, combined with (A.27) and the corresponding result for \( (\hat{\beta}_0, \hat{\beta}) \), yields

\[ n^{1/2} \left( \hat{J}_{\hat{\beta}, \hat{\beta}^l} | \hat{\beta}_{\hat{\beta}, \hat{\beta}^l} \right) = o_p(1). \]

Hence we can conclude that

\[ n^{1/2} \left( \hat{J}_{\hat{\beta}, \hat{\beta}^l} | \hat{\beta}_{\hat{\beta}, \hat{\beta}^l} \right) \rightarrow \mathcal{N}(0, \Sigma^{[f]}). \]  

(A.32)

Now applying the delta method with the function \( g(u, v) = u/v \) we have

\[ n^{1/2} \left( \sum_{i=1}^{N} \left( \frac{Y_{i,j} - \sum_{i=1}^{n} \hat{\beta} \cdot Z_{i,j} \cdot \hat{\beta}}{n} \right) - f^{[f]}(h) \right) \rightarrow \mathcal{N}(0, r(h)), \]

i.e.

\[ n^{1/2} \left( \hat{J}_{\hat{\beta}, \hat{\beta}^l} - f^{[f]}(h) \right) \rightarrow \mathcal{N}(0, r(h)), \]

where

\[ r(h) = \nabla g \left( f^{[f]}(h) B(h, h_1), B(h, h_1) \right) \cdot \Sigma^{[f]} \cdot \nabla g \left( f^{[f]}(h) B(h, h_1), B(h, h_1) \right), \]  

(A.33)

with \( \nabla g(u, v) = (1/v, -u/v^2)' \) and \( \Sigma^{[f]} \) as in (A.29).
Proof of Theorem 5

We begin by reviewing some related notation and results established by Geyer (1994). Recall that $N_j$ denotes the length of the $j^{th}$ chain in Stage 1 samples, $N = \sum_{j=1}^{k} N_j$, and $A_j = N_j / N$. Using the notation

$$\eta_j = -\log m_j + \log(A_j), \quad \text{for } j = 1, \ldots, k,$$

Geyer’s (1994) reverse logistic regression estimator $\hat{\eta} = (\hat{\eta}_1, \ldots, \hat{\eta}_k)$ for the unknown vector $\eta$ is obtained by maximizing the log quasi-likelihood

$$l_N(\eta) = \sum_{l=1}^{k} \sum_{i=1}^{N_l} \log(p_l(\theta_l^{(i)} , \eta)),$$

where

$$p_l(\theta, \eta) = \frac{\nu_{h_l}(\theta)e^{\eta_l}}{\sum_{s=1}^{k} \nu_{h_s}(\theta)e^{\eta_s}}, \quad \text{for } l = 1, \ldots, k.$$  

(A.35)

Theorem 1 of Geyer (1994) states that this maximizer is unique up to an additive constant if the Monte Carlo sample is inseparable. Geyer (1994) also proves that, under certain conditions, $\sqrt{N}(\hat{\eta} - \eta_0)$ is asymptotically normal, where $\eta_0$ is defined by

$$[\eta_0]_j = \eta_j - \frac{1}{k} \sum_{s=1}^{k} \eta_s, \quad j = 1, \ldots, k.$$  

Our proof is structured as follows. First, we extend Geyer’s (1994) proof in order to show that the $2k$-dimensional vector

$$\sqrt{N} \left( \begin{pmatrix} \hat{\eta} \\ \hat{e} \end{pmatrix} - \begin{pmatrix} \eta_0 \\ e \end{pmatrix} \right) =: \begin{pmatrix} U^{(1)} \\ \vdots \\ U^{(2k)} \end{pmatrix} =: U$$

is asymptotically normal. Then, by getting back to the $d$ notation through a transformation, we show that our vector of interest

$$\sqrt{N} \left( \begin{pmatrix} \hat{d} \\ \hat{e} \end{pmatrix} - \begin{pmatrix} d \\ e \end{pmatrix} \right)$$

74
is also asymptotically normal.

To carry out the first step, we will express each $U^{(j)}$, $j = 1, \ldots, k$, as the sum of a linear combination of standardized averages of functions of the $\theta^{(1)0}_i$’s and a $o_p(1)$ quantity. We will also need the central limit theorem to hold for these averages. Hence, for each $j = 1, \ldots, k$, we plan to find constants $\alpha^{(j)}_1, \ldots, \alpha^{(j)}_k$ and functions $\phi^{(j)}_1, \ldots, \phi^{(j)}_k$, which satisfy the conditions

$$
E_{\nu_{i,j}} \left( \phi^{(j)}_l(\theta) \right) = 0 \quad \text{and} \quad E_{\nu_{i,j}} \left( |\phi^{(j)}_l(\theta)|^{2+\epsilon} \right) < \infty, \quad l = 1, \ldots, k, \tag{A.37a}
$$

$$
U^{(j)} = \alpha^{(j)}_1 \frac{1}{\sqrt{N}} \sum_{i=1}^{N_j} \phi^{(j)}_1(\theta^{(1)0}_i) + \cdots + \alpha^{(j)}_k \frac{1}{\sqrt{N_k}} \sum_{i=1}^{N_k} \phi^{(j)}_k(\theta^{(k)0}_i) + o_p(1) \tag{A.37b}
$$

for some $\epsilon > 0$. Note that conditions (A.37a) and B1 yield central limit theorems for the averages in the linear combination above.

For $U^{(k+1)}, \ldots, U^{(2k)}$, condition (A.37) is clearly satisfied since

$$
U^{(j+k)} = \sqrt{N}(\hat{\theta}_j - e_j) = \frac{1}{\sqrt{A_j}} \frac{1}{\sqrt{N_j}} \sum_{i=1}^{N_j} \left( f(\theta^{(j)0}_i) - e_j \right), \quad \text{for} \quad j = 1, \ldots, k,
$$

and the moment conditions in (A.37a) hold (see B2 in the statement of this theorem).

Next, we show that condition (A.37) also holds for the first $k$ components of $U$. In the proof of his Theorem 2, Geyer (1994) defines the matrix $B_N$ via

$$
- \frac{1}{N} (\nabla l_N(\hat{\eta}_N) - \nabla l_N(\eta_0)) = B_N(\hat{\eta}_N - \eta_0), \tag{A.38}
$$

where $l_N$ was defined in (A.34), and establishes that $B_N \xrightarrow{a.s.} B$, where $B$ is given by equation (19) in Geyer (1994). He also shows that, with $u$ being the $k$-dimensional column vector of 1’s,

$$
\begin{pmatrix}
B_N \\
u'
\end{pmatrix} \frac{1}{\sqrt{N}} \nabla l_N(\eta_0) = \begin{pmatrix}
\frac{1}{\sqrt{N}} \nabla l_N(\eta_0) \\
0
\end{pmatrix}. \tag{A.39}
$$
[See equation (31) in Geyer (1994).] Note that, by applying the Mean Value Theorem to \( \nabla l_N(\eta) \), \( B_N \) defined in (A.38) can also be expressed as

\[
B_N = -\frac{1}{N} \nabla^2 l_N(\eta^*),
\]

for some \( \eta^* \) between \( \hat{\eta}_N \) and \( \eta_0 \). Hence, with \( p_r, r = 1, \ldots, k \) defined as in (A.35), the elements of \( B_N \) are given by

\[
[B_N]_{r,s} = -\frac{1}{N} \sum_{l=1}^{k} \sum_{i=1}^{N_l} p_r(\theta_i^{(l)0}, \eta^*) [1 - p_r(\theta_i^{(l)0}, \eta^*)] p_s(\theta_i^{(l)0}, \eta^*), \quad r \neq s,
\]

\[
[B_N]_{r,r} = \frac{1}{N} \sum_{l=1}^{k} \sum_{i=1}^{N_l} p_r(\theta_i^{(l)0}, \eta^*) [1 - p_r(\theta_i^{(l)0}, \eta^*)], \quad r = 1, \ldots, k,
\]

which makes it easy to verify that \( B_N u = 0 \). Combining this with equation (A.39), it can be shown that

\[
\sqrt{N}(\hat{\eta}_N - \eta_0) = B_N^+ \frac{1}{\sqrt{N}} \nabla l_N(\eta_0), \quad (A.40)
\]

where

\[
B_N^+ = (B_N + \frac{1}{k} uu')^{-1} - \frac{1}{k} uu'
\]

is the Moore-Penrose inverse of \( B_N \). Furthermore, letting \( B^+ \) denote the Moore-Penrose inverse of \( B \), we can alternatively write the equality in (A.40) as

\[
\sqrt{N}(\hat{\eta}_N - \eta_0) = (B_N^+ - B^+ + B^+) \frac{1}{\sqrt{N}} \nabla l_N(\eta_0) \quad (A.41)
\]

\[
= (B_N^+ - B^+) \frac{1}{\sqrt{N}} \nabla l_N(\eta_0) + B^+ \frac{1}{\sqrt{N}} \nabla l_N(\eta_0).
\]

Now, using the result \( B_N \xrightarrow{\text{a.s.}} B \) established by Geyer (1994), we can easily deduce that

\[
B_N^+ \xrightarrow{\text{a.s.}} B^+, \quad (A.42)
\]
by writing

\[ B_N^+ = \left( B_N + \frac{1}{k} uu' \right)^{-1} - \frac{1}{k} uu' \]

\[ \xrightarrow{\text{a.s.}} \left( B + \frac{1}{k} uu' \right)^{-1} - \frac{1}{k} uu' = B^+, \]

where the last equality comes from Geyer (1994).

Next, we establish asymptotic normality of \( \nabla l_N(\eta_0) / \sqrt{N} \). Since the gradient \( \nabla l_N(\eta_0) \) is the vector whose \( r \)th element is given by

\[ \frac{\partial l_N(\eta_0)}{\partial \eta_r} = N_r - \sum_{l=1}^{k} \sum_{i=1}^{N_j} p_r(\theta_l^{(i)} | \eta_0) , \]

we can see that

\[ \frac{1}{\sqrt{N}} \frac{\partial l_N(\eta_0)}{\partial \eta_r} = \frac{1}{\sqrt{N}} \left( N_r - \sum_{l=1}^{k} \sum_{i=1}^{N_j} p_r(\theta_l^{(i)} | \eta_0) \right) \]

\[ = \sqrt{A_r} \frac{1}{\sqrt{N_i}} \sum_{i=1}^{N_r} \left( 1 - p_r(\theta_i^{(r)} | \eta_0) \right) - \sum_{l=1}^{k} \sqrt{A_l} \frac{1}{\sqrt{N_l}} \sum_{i=1}^{N_l} p_r(\theta_l^{(l)} | \eta_0) , \]

\[ = \sqrt{A_r} \frac{1}{\sqrt{N_i}} \sum_{i=1}^{N_r} \left( 1 - p_r(\theta_i^{(r)} | \eta_0) \right) - \left[ 1 - E \left( p_r(\theta_1^{(r)} | \eta_0) \right) \right] \]

\[ - \sum_{i=1}^{k} \sqrt{A_l} \frac{1}{\sqrt{N_l}} \sum_{i=1}^{N_l} \left[ p_r(\theta_l^{(l)} | \eta_0) - E \left( p_r(\theta_l^{(l)} | \eta_0) \right) \right] \]

\[ = - \sum_{i=1}^{k} \sqrt{A_l} \frac{1}{\sqrt{N_l}} \sum_{i=1}^{N_l} \left[ p_r(\theta_l^{(l)} | \eta_0) - E \left( p_r(\theta_l^{(l)} | \eta_0) \right) \right] , \]

which is a linear combination of the form given in (A.37b) and, because \( 0 \leq p_r(\theta, \eta) \leq 1 \) for all \( \theta \) and \( \eta \), condition (A.37a) is also satisfied. Note that we are allowed to insert the
expectations in the next-to-last equality because

\[-\sqrt{A_r} \frac{1}{\sqrt{N_r}} N_r [1 - E(p_r(\theta^{(r)}_1, \eta_0))] + \sum_{l=1 \atop l \neq r}^k \sqrt{A_l} \frac{1}{\sqrt{N_l}} N_l E(p_r(\theta^{(r)}_1, \eta_0))\]

\[= -\sqrt{N A_r} \left(1 - \int \frac{\nu_h(\theta)e^{\eta_r}}{\sum_{s=1}^k \nu_h(\theta)e^{\eta_s}} \nu_{h,y}(\theta) d\theta\right) + \sqrt{N} \sum_{l=1 \atop l \neq r}^k A_l E(p_r(\theta^{(l)}_1, \eta_0))\]

\[= -\sqrt{N A_r} \sum_{l=1 \atop l \neq r}^k \int \frac{\nu_h(\theta)e^{\eta_r}}{\sum_{s=1}^k \nu_h(\theta)e^{\eta_s}} \nu_{h,y}(\theta) d\theta + \sqrt{N} \sum_{l=1 \atop l \neq r}^k A_l E(p_r(\theta^{(l)}_1, \eta_0))\]

\[= -\sqrt{N A_r} \sum_{l=1 \atop l \neq r}^k \int \frac{m_{h_l}e^{\eta_r}}{m_{h_r}} \nu_{h,y}(\theta) d\theta + \sqrt{N} \sum_{l=1 \atop l \neq r}^k A_l E(p_r(\theta^{(l)}_1, \eta_0))\]

\[= 0.\]

The asymptotic normality of \(\nabla I_N(\eta_0)/\sqrt{N}\) now follows from the Cramér-Wold device.

In view of this convergence in distribution and the convergence result in (A.42), (A.41) gives

\[\sqrt{N}(\hat{\eta}_N - \eta_0) = o_p(1) + B^+ \frac{1}{\sqrt{N}} \nabla I_N(\eta_0) = B^+ \frac{1}{\sqrt{N}} \nabla I_N(\eta_0) + o_p(1).\]

Therefore, we can now easily see that condition (A.37) is also satisfied by the first \(k\) components of \(U\), i.e. \(\sqrt{N}(\hat{\eta}_N - \eta_0)\) because, as we have shown in (A.43), every element of \(\nabla I_N(\eta_0)/\sqrt{N}\) is a linear combination of the form (A.37).
Now that we have shown that

\[
U = \begin{pmatrix}
\alpha_1^{(1)} \frac{1}{\sqrt{N_1}} \sum_{i=1}^{N_1} \phi_1^{(1)}(\theta_1^{(1)}_i) + \cdots + \alpha_k^{(1)} \frac{1}{\sqrt{N_k}} \sum_{i=1}^{N_k} \phi_k^{(1)}(\theta_1^{(1)}_i) \\
\vdots \\
\alpha_1^{(k+1)} \frac{1}{\sqrt{N_1}} \sum_{i=1}^{N_1} \phi_1^{(k+1)}(\theta_1^{(1)}_i) + \cdots + \alpha_k^{(k+1)} \frac{1}{\sqrt{N_k}} \sum_{i=1}^{N_k} \phi_k^{(k+1)}(\theta_1^{(1)}_i) \\
\vdots \\
\alpha_1^{(2k)} \frac{1}{\sqrt{N_1}} \sum_{i=1}^{N_1} \phi_1^{(2k)}(\theta_1^{(1)}_i) \\
\end{pmatrix} + o_p(1),
\]

we can prove that \( U \) is asymptotically normal by using the Cramér-Wold device. Let us denote the asymptotic variance of \( U \) by \( S \). Then

\[
S = \begin{pmatrix}
S^{(1)} & S^{(2)} \\
S^{(2)} & S^{(3)}
\end{pmatrix},
\]

where

\[
S^{(1)} = B^+ C B^+,
\]  \hspace{1cm} (A.44)

with

\[
C_{rs} = \sum_{l=1}^{k} A_l \text{Cov}(p_r(\theta_1^{(l)}_0, \eta_0), p_s(\theta_1^{(l)}_0, \eta_0))
\]

\[+ \sum_{l=1}^{k} A_l \sum_{g=1}^{\infty} \text{Cov}(p_r(\theta_1^{(l)}_0, \eta_0), p_s(\theta_1^{(l)}_0, \eta_0)) + \text{Cov}(p_r(\theta_1^{(l)}_0, \eta_0), p_s(\theta_1^{(l)}_0, \eta_0))],
\]

for \( r = 1, \ldots, k \), and \( s = 1, \ldots, k \),

\[
S^{(3)}_{rr} = \frac{1}{A_r} \left[ \text{Var}(f(\theta_1^{(r)}_0)) + 2 \sum_{g=1}^{\infty} \text{Cov}(f(\theta_1^{(r)}_0), f(\theta_1^{(r)}_{1+g})) \right],
\]  \hspace{1cm} (A.45)

\[
S^{(3)}_{rs} = 0 \text{ when } r \neq s, \quad r = 1, \ldots, k, \quad s = 1, \ldots, k,
\]

and

\[
S^{(2)} = B^+ D,
\]  \hspace{1cm} (A.46)
with

\[
D_{rs} = - \text{Cov}(p_r(\theta^{(s)}_1, \eta_0), f(\theta^{(s)}_1)) \nonumber \\
- \sum_{g=1}^{\infty} [\text{Cov}(p_r(\theta^{(s)}_1, \eta_0), f(\theta^{(s)}_{1+g})) + \text{Cov}(p_r(\theta^{(s)}_{1+g}, \eta_0), f(\theta^{(s)}_1))],
\]

for \( r = 1, \ldots, k, \) and \( s = 1, \ldots, k. \) Now, having established the convergence result

\[
U = \sqrt{N} \left( \begin{pmatrix} \hat{\eta} \\ \hat{e} \end{pmatrix} - \begin{pmatrix} \eta_0 \\ e \end{pmatrix} \right) \xrightarrow{d} \mathcal{N}(0, S), \tag{A.47}
\]

consider the function \( g : \mathbb{R}^{2k} \to \mathbb{R}^{2k-1} \) given by

\[
g \left( \begin{pmatrix} \eta \\ e \end{pmatrix} \right) = \begin{pmatrix} e^{\eta_1 - \eta_2 A_2 / A_1} \\ e^{\eta_1 - \eta_3 A_3 / A_1} \\ \vdots \\ e^{\eta_1 - \eta_k A_k / A_1} \\ e \end{pmatrix},
\]

where \( \eta \) and \( e \) are \( k \)-dimensional vectors. The delta method applied to (A.47) with \( g \) as the transformation gives

\[
\sqrt{N} \left( \begin{pmatrix} \hat{d}_2 \\ \vdots \\ \hat{d}_k \\ \hat{e} \end{pmatrix} - \begin{pmatrix} d_2 \\ \vdots \\ d_k \\ e \end{pmatrix} \right) \xrightarrow{d} \mathcal{N}(0, V),
\]

where

\[
V = \nabla g \begin{pmatrix} \eta_0 \\ e \end{pmatrix} S \nabla g \begin{pmatrix} \eta_0 \\ e \end{pmatrix}, \tag{A.48}
\]
with

\[
\nabla g \left( \begin{array}{c} \eta \\ e \end{array} \right) = \begin{pmatrix}
  e^{m-n_2}A_2/A_1 & e^{m-n_3}A_3/A_1 & \ldots & e^{m-n_k}A_k/A_1 & 0 & 0 & \ldots & 0 \\
  -e^{m-n_2}A_2/A_1 & 0 & \ldots & 0 & 0 & 0 & \ldots & 0 \\
  0 & -e^{m-n_3}A_3/A_1 & \ldots & 0 & 0 & 0 & \ldots & 0 \\
  \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\
  0 & 0 & \ldots & -e^{m-n_k}A_k/A_1 & 0 & 0 & \ldots & 0 \\
  0 & 0 & \ldots & 0 & 1 & 0 & \ldots & 0 \\
  0 & 0 & \ldots & 0 & 0 & 1 & \ldots & 0 \\
  0 & 0 & \ldots & 0 & 0 & 0 & \ldots & 1 
\end{pmatrix},
\]

and \( S \) given by (A.44), (A.46), and (A.45).

Proof of Remark 2 to Theorem 1

Following the lines of the proof of Theorem 1 with \( q = 1 \), we get as in (A.5) that

\[
\sqrt{n} \left( \hat{B}(h, h_1, \hat{d}) - B(h, h_1) \right) = c(h)' \sqrt{n} (\hat{d} - d) + \sqrt{n} (\hat{B}(h, h_1, d) - B(h, h_1)) + o_p(1),
\]

where \( c(h) \) is the constant column vector given in (A.3). This decomposition can be rewritten as

\[
\sqrt{n} (\hat{B}(h, h_1, \hat{d}) - B(h, h_1)) = (c(h)', 1) \begin{pmatrix}
  \sqrt{n} (\hat{d} - d) \\
  \sqrt{n} (\hat{B}(h, h_1, d) - B(h, h_1))
\end{pmatrix} + o_p(1).
\]

Now note that in order to establish the asymptotic normality of \( \sqrt{n} (\hat{B}(h, h_1, \hat{d}) - B(h, h_1)) \), it is enough to show that

\[
\begin{pmatrix}
  \sqrt{n} (\hat{d} - d) \\
  \sqrt{n} (\hat{B}(h, h_1, d) - B(h, h_1))
\end{pmatrix}
\]

is asymptotically normal. Using the \( \eta \)-notation introduced in the proof of Theorem 5, let

\[
T = \sqrt{n} \left( \begin{pmatrix}
  \hat{\eta} \\
  \hat{B}(h, h_1, d)
\end{pmatrix} - \begin{pmatrix}
  \eta_0 \\
  B(h, h_1)
\end{pmatrix} \right).
\]

81
As was done for $U$ in the proof of Theorem 5, we can write $T$ as

$$
T = \left( \begin{array}{c}
\alpha_1^{(1)} \frac{1}{\sqrt{n_1}} \sum_{i=1}^{n_1} \phi_1^{(1)}(\theta_1^{(1)}) + \cdots + \alpha_k^{(1)} \frac{1}{\sqrt{n_k}} \sum_{i=1}^{n_k} \phi_1^{(1)}(\theta_1^{(k)}) \\
\vdots \\
\alpha_1^{(k)} \frac{1}{\sqrt{n_1}} \sum_{i=1}^{n_1} \phi_1^{(k)}(\theta_1^{(1)}) + \cdots + \alpha_k^{(k)} \frac{1}{\sqrt{n_k}} \sum_{i=1}^{n_k} \phi_1^{(k)}(\theta_1^{(k)}) \\
\sum_{i=1}^{k} a_1^{1/2} \frac{1}{\sqrt{n_1}} \sum_{i=1}^{n_1} (Y_{i,1} - E(Y_{1,1}))
\end{array} \right) + o_p(1),
$$

where the first $k$ components are the same as the first $k$ components of the vector $U$ in (A.36), and $Y_{i,1}$ is given in (2–11). By applying the Cramér-Wold device, we can conclude that the vector $T$ converges in distribution to a normal random variable with mean 0 and variance

$$
Z = \left( \begin{array}{c}
S^{(1)} \\
z \\
z' \\
\tau^2(h)
\end{array} \right),
$$

in which $S^{(1)}$ is the $k \times k$ matrix given in (A.44), $z$ is the $k \times 1$ vector given by

$$
z = B^+ y \quad \text{(A.49)}
$$

where

$$
y_r = \sum_{l=1}^{k} A_l \text{Cov}(p_r(\theta_1^{(l)}), Y_{1,l}) + \sum_{l=1}^{k} A_l \sum_{g=1}^{\infty} \text{Cov}(p_r(\theta_1^{(l)}), Y_{1+g,l}) + \text{Cov}(p_r(\theta_1^{(l)}), Y_{1,l}),
$$

for $r = 1, \ldots, k$, with $B^+$ as in Theorem 5, and as in (A.9) of Doss (2010)

$$
\tau^2(h) = \sum_{l=1}^{k} a_l \left[ \text{Var}(Y_{1,l}) + 2 \sum_{g=1}^{\infty} \text{Cov}(Y_{1,l}, Y_{1+g,l}) \right].
$$
Now define the function \( g : \mathbb{R}^{k+1} \to \mathbb{R}^k \) by

\[
g(\eta, b) = \left( e^{\frac{n-m}{n} A_2/A_1}, e^{\frac{n-m}{n} A_3/A_1}, \ldots, e^{\frac{n-m}{n} A_k/A_1}, b \right),
\]

where \( \eta \) is a \( k \)-dimensional vector and \( b \) is a real number. Applying the delta method to the previously established result that \( T \xrightarrow{d} \mathcal{N}(0, Z) \), we get

\[
\sqrt{n} \left[ \left( \hat{d} - d \right) - \left( \hat{B}(h, h_1) - B(h, h_1) \right) \right] \xrightarrow{d} \mathcal{N} \left( 0, \nabla g \left( B(h, h_1) \right) \right) Z \nabla g \left( B(h, h_1) \right),
\]

where

\[
\nabla g \left( \eta \right) = \begin{pmatrix} E & 0 \\ 0' & 1 \end{pmatrix}, \tag{A.50}
\]

with

\[
E = \begin{pmatrix} e^{\frac{n-m}{n} A_2/A_1} & e^{\frac{n-m}{n} A_3/A_1} & \ldots & e^{\frac{n-m}{n} A_k/A_1} \\ -e^{\frac{n-m}{n} A_2/A_1} & 0 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & -e^{\frac{n-m}{n} A_3/A_1} & \ldots & 0 \\ 0 & 0 & \ldots & -e^{\frac{n-m}{n} A_k/A_1} \end{pmatrix}, \tag{A.51}
\]

and \( 0 \) in (A.50) representing the column vector of \( k \) zeros.

Hence, we know that \( \sqrt{n}(\hat{B}(h, h_1, \hat{d}) - B(h, h_1)) \) has an asymptotically normal distribution with mean \( 0 \) and variance

\[
c(h)' \Sigma c(h) + \tau^2(h) + 2c(h)'E'z,
\]

where \( \Sigma \) denotes, as in the statement of Theorem 1, the asymptotic variance of \( \sqrt{n}(\hat{d} - d) \), \( E \) is given in (A.51), and \( z \) in (A.49).
Proof of Theorem 6

Let

\[ \hat{J}_{d,e}^{(d,e)} = \frac{1}{n} \sum_{i=1}^{n} \left( \frac{Y^r_{i,l}}{Y_{i,l}} - \left( \sum_{j=1}^{k} \beta^r_{j} Z^r_{i,j} - \sum_{j=2}^{k} \beta^r_{j} Z^r_{i,j} \right) \right) \]

where the superscripts \( d, e \) indicate the values of \( d \) and \( e \) used when computing \( Y \)'s and \( Z \)'s, while the subscripts indicate the coefficients of \( Z \)'s. With \( \mu^r_{[l]} \) as in the proof of Theorem 4 we now write

\[ \sqrt{n} \left( \hat{J}_{d,e}^{(d,e)} - \sum_{l=1}^{k} a_{l} \mu^r_{[l]} \right) = \sqrt{n} \left( \hat{J}_{d,e}^{(d,e)} - \sum_{l=1}^{k} a_{l} \mu^r_{[l]} \right) \]

(A.52)

Note that the second quantity on the right side of (A.52), which involves only known \( d \) and \( e \), was shown to be asymptotically normal with mean 0 and variance \( \Sigma^r \) in the proof of Theorem 4 [see (A.32)]. Now let us expand the first term on the right side of (A.52) by writing

\[ \sqrt{n} \left( \hat{J}_{d,e}^{(d,e)} - \sum_{l=1}^{k} a_{l} \mu^r_{[l]} \right) = \sqrt{n} \left( \hat{J}_{d,e}^{(d,e)} - \sum_{l=1}^{k} a_{l} \mu^r_{[l]} \right) + \sqrt{n} \left( \hat{J}_{d,e}^{(d,e)} - \sum_{l=1}^{k} a_{l} \mu^r_{[l]} \right) \]

(A.53)

We next proceed as follows:

1. We note that the third term on the right side of (A.53) was shown to converge to 0 in probability in the proof of Theorem 4.
2. We show that the first term on the right side of (A.53) also converges to 0 in probability.
3. We show that the second term on the right side of (A.53) is asymptotically normal.

To deal with the second step, as in the proof of Theorem 1, first we show that \( \hat{\beta}^r_{[l]}(d) \) and \( \hat{\beta}^r_{[l]}(\hat{d}) \) converge in probability to the same limit, which we denoted in the proof of
Theorem 4 by $\beta^{[r]}_{\text{lim}}$. For fixed $j, j' \in \{1, \ldots, k\}$, consider the function

$$G(u, v) = \frac{1}{n} \sum_{l=1}^{n} \sum_{i=1}^{n} \left[ \frac{f(\theta_i^{(l)}) \nu_h(\theta_i^{(l)}) / u_j - v_j}{\sum_{s=1}^{k} a_s \nu_h(\theta_i^{(l)}) / u_s} \right] \cdot \left[ \frac{f(\theta_i^{(l)}) \nu_h(\theta_i^{(l)}) / u_{j'} - v_{j'}}{\sum_{s=1}^{k} a_s \nu_h(\theta_i^{(l)}) / u_s} \right],$$

where $u = (u_2, \ldots, u_k)'$ with $u_l > 0$, for $l = 2, \ldots, k$, and $v = (v_1, \ldots, v_k)'$. Note that setting $u = d$ and $v = e$ gives

$$G(d, e) = \frac{1}{n} \sum_{l=1}^{n} \sum_{i=1}^{n} Z_{i,l}^{[r]}(f) Z_{i,l}^{[r](\nu)}.$$

By the Mean Value Theorem, we know that there exists a $(d^*, e^*)$ between $(\hat{d}, \hat{e})$ and $(d, e)$ such that

$$G(\hat{d}, \hat{e}) = G(d, e) + \nabla G(d^*, e^*)' \left( \begin{pmatrix} \hat{d} \\ \hat{e} \end{pmatrix} - \begin{pmatrix} d \\ e \end{pmatrix} \right) = R_{j+1, j'+1}^{[r]} + \nabla G(d^*, e^*)' \left( \begin{pmatrix} \hat{d} \\ \hat{e} \end{pmatrix} - \begin{pmatrix} d \\ e \end{pmatrix} \right) + o_p(1).$$

As in previous proofs, with some calculations we can show that $\nabla G(d^*, e^*) = O_p(1)$. Therefore $G(\hat{d}, \hat{e}) \xrightarrow{p} R_{j+1, j'+1}^{[r]}$, and since $R^{[r]}$ is assumed invertible, we have

$$n \left[ \left( Z^{[r]} \right)^T Z^{[r]} \right]^{-1} \xrightarrow{p} (R^{[r]})^{-1},$$

where $Z^{[r]}$ is obtained from the matrix $Z^{[r]}$ in (A.23) by replacing $d$ and $e$ with $\hat{d}$ and $\hat{e}$.

The same reasoning extends to the case where $j = 0$ or $j' = 0$. In a similar way, if we let $\hat{Y}^{[r]}$ denote the vector obtained from $Y^{[r]}$ in (A.24) by replacing $d$ with $\hat{d}$ and we recall that $v^{[r]}$ was defined to be the vector on the right side of (A.26), it can be proved that

$$\frac{Z^{[r]} \hat{Y}^{[r]}}{n} \xrightarrow{p} v^{[r]}.$$
which together with the previous result implies that $\beta_{\mathbf{r}1}(d)$ and $\beta_{\mathbf{r}1}(\hat{d})$ converge in probability to the same limit. Also,

$$\sqrt{n} \left( \hat{\beta}_{\mathbf{r}1}(\hat{d}, \theta) - \hat{\beta}_{\mathbf{r}1}(\hat{d}) \right) = \left( n^{1/2} \sum_{j=1}^{k} (\hat{\beta}_{\mathbf{r}1, \mathbf{j}, \mathbf{r}1}(\hat{d}) - \hat{\beta}_{\mathbf{j}, \mathbf{r}1}(\hat{d})) + \frac{1}{n} \sum_{i=1}^{p} \sum_{l=1}^{k} \hat{\beta}_{\mathbf{r}1, \mathbf{j}, \mathbf{r}1}(\hat{d}) \right)$$

From the proof of Theorem 2 we already know that the second component of this last vector, denoted therein by $\sqrt{n} (\hat{\beta}_{\mathbf{r}1}(\hat{d}) - \hat{\beta}_{\mathbf{r}1}(\hat{d}))$, is $o_p(1)$. In an analogous manner, it can be shown that the first component is $o_p(1)$. Thus the whole vector is $o_p(1)$.

As for the middle term of the right side of (A.53), if we define

$$K_{\mathbf{r}1}(u, \nu) = \frac{1}{n} \sum_{i=1}^{p} \sum_{l=1}^{k} \left( \frac{f(\theta_i^{(l)})}{u_i} - \sum_{j=1}^{k} \hat{\beta}_{\mathbf{r}1, \mathbf{j}, \mathbf{r}1}(\hat{d}) \right) + \frac{1}{n} \sum_{i=1}^{p} \sum_{l=1}^{k} \hat{\beta}_{\mathbf{j}, \mathbf{r}1}(\hat{d}) \right)$$

where $u = (u_2, \ldots, u_k)'$, with $u_i > 0$ for $i = 2, \ldots, k$, and $\nu = (\nu_1, \nu_2, \ldots, \nu_k)'$, then

$$\sqrt{n} \left( \hat{\beta}_{\mathbf{r}1}(\hat{d}, \theta) - \hat{\beta}_{\mathbf{r}1}(\hat{d}, \theta) \right) = \left( \sqrt{n} (K_{\mathbf{r}1}(\hat{d}, \theta) - K_{\mathbf{r}1}(d, \theta)) \right)$$

with $K$ defined as in the proof of Theorem 2. From this same proof, we know that

$$\sqrt{n} (K(\hat{d}) - K(d)) = \sqrt{n} (\nu(h) \sqrt{N}(\hat{d} - d) + o_p(1).$$

We will now show that, similarly,

$$\sqrt{n} (K_{\mathbf{r}1}(\hat{d}, \theta) - K_{\mathbf{r}1}(d, \theta)) = \sqrt{n} (\nu(h) \sqrt{N}(\hat{d} - d) + o_p(1).$$

86
with \( w^{[r]}(h) \) defined by (A.56) below. By Taylor series expansion, we get

\[
\sqrt{n}(K^{[r]}(\hat{d}, \hat{e}) - K^{[r]}(d, e)) = \sqrt{n} \nabla K^{[r]}(d, e) \left( \frac{\hat{d} - d}{\hat{e} - e} \right) + \sqrt{n} \frac{1}{2} \left( \frac{\hat{d} - d}{\hat{e} - e} \right) \nabla^2 K^{[r]}(d^*, e^*) \left( \frac{\hat{d} - d}{\hat{e} - e} \right),
\]

where \((d^*, e^*)\) is between \((\hat{d}, \hat{e})\) and \((d, e)\).

Below we compute the gradient \( \nabla K^{[r]}(d, e) \) and show that it converges almost surely to a vector \( w^{[r]}(h) \). We have

\[
\frac{\partial K^{[r]}}{\partial u_t}(d, e) = \frac{1}{n} \sum_{i=1}^{k} \sum_{s=1}^{n} \left[ \frac{f(\theta^{(i)}_s)\nu_h(\theta^{(i)}_s)\alpha_t\nu_h(\theta^{(i)}_s)}{d^2_t \left( \sum_{s=1}^{k} \alpha_s\nu_h(\theta^{(i)}_s)/d_s \right)^2} - \sum_{j=1}^{k} \beta_{t, \lim}^{[r]} \frac{f(\theta^{(i)}_s)\nu_h(\theta^{(i)}_s)\alpha_t\nu_h(\theta^{(i)}_s)}{d^2_t \left( \sum_{s=1}^{k} \alpha_s\nu_h(\theta^{(i)}_s)/d_s \right)^2} \right]
\]

\[
\text{a.s.} \quad \frac{B(h, h_t)}{d^2_t} \int \frac{f(\theta)\alpha_t\nu_h(\theta)}{\sum_{s=1}^{k} \alpha_s\nu_h(\theta)/d_s} \cdot \nu_{h, y}(\theta) d\theta
\]

\[
- \sum_{j=1}^{k} \beta_{t, \lim}^{[r]} \int \frac{f(\theta)\alpha_t\nu_h(\theta)}{d^2_t \sum_{s=1}^{k} \alpha_s\nu_h(\theta)/d_s} \cdot \nu_{h, y}(\theta) d\theta + \beta_{t, \lim}^{[r]} \frac{f^{[r]}(h_t)}{d_t}
\]

\[
= \frac{B(h, h_t)}{d^2_t} \int \frac{f(\theta)\alpha_t\nu_h(\theta)}{\sum_{s=1}^{k} \alpha_s\nu_h(\theta)/d_s} \cdot \nu_{h, y}(\theta) d\theta
\]

\[
- \sum_{j=1}^{k} \beta_{t, \lim}^{[r]} \int \frac{f(\theta)\alpha_t\nu_h(\theta)}{d^2_t \sum_{s=1}^{k} \alpha_s\nu_h(\theta)/d_s} \cdot \nu_{h, y}(\theta) d\theta + \beta_{t, \lim}^{[r]} \frac{f^{[r]}(h_t)}{d_t}
\]

\[
= w^{[r]}_{t-1}(h), \quad \text{for} \ t = 2, \ldots, k, \quad \text{(A.56a)}
\]
and
\[
\frac{\partial K^{[r]}_{t}}{\partial v} (d, e) = \beta^{[r]}_{t, \lim} := w^{[r]}_{k-1+ t}(h), \quad \text{for } t = 1, \ldots, k. \tag{A.56b}
\]

Proceeding as we did in the proof of Theorem 2 when we showed that \(\nabla^2 K(d^*) = O_p(1)\), we can show here that \(\nabla^2 K^{[r]}(d^*, e^*)\) is bounded in probability. Hence
\[
\sqrt{n} (K^{[r]}(\hat{d}, \hat{e}) - K^{[r]}(d, e)) = \sqrt{q} w^{[r]}(h)' \sqrt{N} \left( \begin{pmatrix} \hat{d} \\ \hat{e} \end{pmatrix} - \begin{pmatrix} d \\ e \end{pmatrix} \right) + o_p(1),
\]
and together with (A.55) and (A.54) this implies that
\[
\sqrt{n} \left( \frac{\partial^2 K^{[r]}(d, e)}{\partial \beta \partial \beta} - \frac{\partial^2 K^{[r]}(d, e)}{\partial \mu \partial \mu} \right) = \left( \sqrt{q} w^{[r]}(h)' \sqrt{N} \left( \begin{pmatrix} \hat{d} \\ \hat{e} \end{pmatrix} - \begin{pmatrix} d \\ e \end{pmatrix} \right) + o_p(1) \right) \\
\sqrt{q} w^{[r]}(h)' \sqrt{N} (\hat{d} - d) + o_p(1)
\]
where \(w_0(h)\) is the column-vector obtained from \(w(h)\) by concatenating \(k\) zeros at its end. Now returning to (A.52) and (A.53) we get
\[
\sqrt{n} \left( \frac{\partial^2 K^{[r]}(d, e)}{\partial \beta \partial \mu} - \sum_{l=1}^{k} a_l \mu^{[r]}_l \right) = \sqrt{q} \left( \begin{pmatrix} w^{[r]}(h)' \\ w_0(h)' \end{pmatrix} \right) \sqrt{N} \left( \begin{pmatrix} \hat{d} \\ \hat{e} \end{pmatrix} - \begin{pmatrix} d \\ e \end{pmatrix} \right) \\
+ \sqrt{n} \left( \frac{\partial^2 K^{[r]}(d, e)}{\partial \beta \partial \beta} - \sum_{l=1}^{k} a_l \mu^{[r]}_l \right) + o_p(1)
\]
\[
\xrightarrow{d} \mathcal{N} \left( 0, q \left( \begin{pmatrix} w^{[r]}(h)' \\ w_0(h)' \end{pmatrix} V (w^{[r]}(h), w_0(h)) + \Sigma^{[r]} \right) \right).
\]

We can now apply the delta method with the function \(g(u, \nu) = u/\nu\) to get our result
\[
\sqrt{n} \left( \frac{\partial^2 K^{[r]}(d, e)}{\partial \beta \partial \mu} - j^{[r]}(h) \right) \xrightarrow{d} \mathcal{N}(0, \psi(h)),
\]
88
where

\[
\psi(h) = \nabla g(I[f(h)]B(h, h_1), B(h, h_1))' \cdot \left( q \begin{pmatrix} w^{[r]}(h)' \\ w_0(h)'
\end{pmatrix} V(w^{[r]}(h), w_0(h)) + \Sigma^{[r]} \right)
\cdot \nabla g(I[f(h)]B(h, h_1), B(h, h_1)).
\]  

(A.57)

with \(\nabla g(u, v) = (1/v, -u/v^2)'\).

\hfill \Box
APPENDIX B
DETAILS REGARDING GENERATION OF THE MARKOV CHAIN FROM CHAPTER 5

To generate a Markov chain of length $n$ on $\theta = (\gamma, \sigma, \beta_0, \beta_\gamma)$ for a fixed choice of the hyperparameter $h = (w, g)$, we use the following sampling scheme. First, we pick an arbitrary value for $\gamma^{(0)}$. Then we draw $\sigma^{2(0)}, \beta_0^{(0)},$ and $\beta_\gamma^{(0)}$ as indicated in Steps 2–4 below (with $i = 0$). To generate the rest of the chain, we iterate through Steps 1–4 described below for each $i = 1, \ldots, n - 1$.

**Step 1** In this stage we generate the binary vector $\gamma^{(i)}$ by using a Gibbs sampler on $\gamma = (\gamma_1, \gamma_2, \ldots, \gamma_q)$. Thus, we first generate $\gamma_1^{(i)} | \gamma_{\neq 1}^{(i)}, \ldots, \gamma_q^{(i-1)}, Y$ according to the following Bernoulli distribution:

$$p(\gamma_1 | \gamma_{\neq 1}, Y) \propto p(\gamma | Y) \propto (1 + g)^{-q/y/2} S^{-(m-1)} [1 + g(1 - R^2_\gamma)]^{-(m-1)/2} \left( \frac{w}{1-w} \right)^q,$$

where, recall that $S^2 = \sum_{j=1}^m (Y_j - \overline{Y})^2$, and $R^2_\gamma$ is the coefficient of determination of model $\gamma$; see (5–1). Similarly, generate $\gamma_2^{(i)}$ from $p(\gamma_2^{(i)} | \gamma_1^{(i)}, \gamma_3^{(i-1)}, \ldots, \gamma_q^{(i-1)}, Y)$, and so on for $\gamma_3^{(i)}, \ldots, \gamma_q^{(i)}$. (This Gibbs sampler is not identical to that of Smith and Kohn (1996) in that in our model the prior on $\beta_0$ is a flat prior, whereas Smith and Kohn (1996) use a proper prior on $\beta_0$.)

**Step 2** Generate $\sigma^{2(i)} | \gamma^{(i)}, Y$ according to the density

$$p(\sigma^2 | \gamma, Y) \propto p(Y | \gamma, \sigma^2) p(\sigma^2) \propto \int p(Y | \gamma, \sigma^2, \beta_0, \beta_\gamma) p(\beta_0, \beta_\gamma | \gamma, \sigma^2) d\beta_0 d\beta_\gamma \cdot p(\sigma^2) \propto (\sigma^2)^{-(m+1)/2} \exp \left\{ -\frac{S^2}{2\sigma^2(g+1)} \cdot \left[ 1 + g(1 - R^2_\gamma) \right] \right\},$$

an inverse gamma density.
To see why the last relationship statement is true, we first consider the integral with respect to \( \beta_0 \). We have

\[
\int p(Y \mid \gamma, \sigma^2, \beta_0, \beta_\gamma)p(\beta_0) \, d\beta_0
\]

\[
\propto \int (\sigma^2)^{-m/2} \exp\left[ -\frac{1}{2\sigma^2}(Y - 1_m \beta_0 - X_\gamma \beta_\gamma)'(Y - 1_m \beta_0 - X_\gamma \beta_\gamma) \right] \, d\beta_0
\]

\[
\propto (\sigma^2)^{-m/2} \exp\left[ -\frac{1}{2\sigma^2}(Y - X_\gamma \beta_\gamma)'(Y - X_\gamma \beta_\gamma) \right] \int \exp\left[ -\frac{m}{2\sigma^2}(\beta_0^2 - 2\beta_0 \bar{Y}) \right] \, d\beta_0
\]

\[
\propto (\sigma^2)^{-m/2} \exp\left[ -\frac{1}{2\sigma^2}(Y - X_\gamma \beta_\gamma)'(Y - X_\gamma \beta_\gamma) \right] \exp\left( \frac{m\bar{Y}^2}{2\sigma^2} \right).
\]

So we may now write

\[
p(\sigma^2 \mid \gamma, Y)
\]

\[
\propto (\sigma^2)^{-m/2} \exp\left( \frac{m\bar{Y}^2}{2\sigma^2} \right) \int \exp\left[ -\frac{1}{2\sigma^2}(Y - X_\gamma \beta_\gamma)'(Y - X_\gamma \beta_\gamma) \right] p(\beta_\gamma \mid \gamma, \sigma^2) \, d\beta_\gamma
\]

\[
\propto (\sigma^2)^{-m+1+q_\gamma/2} \exp\left( -\frac{S^2}{2\sigma^2} \right) \int \exp\left[ -\frac{1}{2\sigma^2}(g + 1)\beta_\gamma X_\gamma^T X_\gamma \beta_\gamma + \frac{1}{2\sigma^2} Y'X_\gamma \beta_\gamma \right] d\beta_\gamma
\]

\[
\propto (\sigma^2)^{-m+1+q_\gamma/2} \exp\left( -\frac{S^2}{2\sigma^2} \right) (\sigma^2)^{q_\gamma/2} \exp\left[ \frac{g}{2\sigma^2(g + 1)}Y'(X_\gamma^T X_\gamma)^{-1}Y \right]
\]

\[
\propto (\sigma^2)^{-m+1} \exp\left\{ -\frac{S^2}{2\sigma^2(g + 1)} \cdot [1 + g(1 - R_\gamma^2)] \right\},
\]

where the next-to-last proportionality relation results from using the formula

\[
\int \exp\left( -\frac{1}{2} \beta_\gamma W^{-1} \beta_\gamma + a' \beta_\gamma \right) \, d\beta_\gamma = (2\pi)^{q_\gamma/2} |W|^{1/2} \exp(a' Wa/2)
\]

which can be shown to hold for any vector \( a \) of length \( q_\gamma \) and any positive definite matrix \( W \) by using a “completing the squares” argument. In practice, we use the distributional relationship

\[
\frac{S^2 [1 + g(1 - R_\gamma^2)]}{\sigma^2(g + 1)} \sim \chi^2_{m-1}
\]

to draw \( \sigma^2 \).
Step 3 Generate $\beta_0^{(i)} | \gamma^{(i)}, \sigma^2, Y$ according to the density

$$p(\beta_0 | \gamma, \sigma^2, Y) \propto \int p(Y | \gamma, \sigma^2, \beta_0, \beta_\gamma) p(\beta_\gamma | \gamma, \sigma^2) d\beta_\gamma \cdot p(\beta_0)$$  \hspace{1cm} (B.2a)

$$\propto \exp\left[-\frac{1}{2\sigma^2}(Y - 1_m\beta_0)'(Y - 1_m\beta_0)\right]$$  \hspace{1cm} (B.2b)

$$\propto N(\bar{Y}, \sigma^2/m).$$

Note that (B.2b) follows from (B.2a) because $1'_mX_\gamma = 0$, since the columns of $X_\gamma$ are centered.

Step 4 Generate $\beta_\gamma^{(i)} | \gamma^{(i)}, \sigma^2, \beta_0^{(i)}, Y$ according to the density

$$p(\beta_\gamma | \gamma, \sigma^2, \beta_0, Y) \propto p(Y | \gamma, \sigma^2, \beta_0, \beta_\gamma) p(\beta_\gamma | \gamma, \sigma^2)$$

$$\propto \exp\left\{-\frac{1}{2\sigma^2}\left[(Y - 1_m\beta_0 - X_\gamma\beta_\gamma)'(Y - 1_m\beta_0 - X_\gamma\beta_\gamma) + \frac{\beta_\gamma'X_\gamma'X_\gamma\beta_\gamma}{g}\right]\right\},$$

which can be shown to be a $q_\gamma$-dimensional normal with mean and covariance matrix given respectively by

$$\mu = g\beta_\gamma / (g + 1) \quad \text{and} \quad \Sigma = g/(g + 1)\sigma^2(X_\gamma'X_\gamma)^{-1}.$$

where $\hat{\beta}_\gamma = (X_\gamma'X_\gamma)^{-1}X_\gamma'Y$ is the usual least squares estimate for model $\gamma$.

We now discuss the computational effort needed to implement our sampler.

Consider generating the first component of $\gamma$. As seen in Step 1, the conditional distribution for this component is Bernoulli with success probability

$$p(\gamma_1 = 1 | \gamma_{\neq 1}, Y) = \frac{p((1, \gamma_2, \ldots, \gamma_q) | Y)}{p((0, \gamma_2, \ldots, \gamma_q) | Y) + p((1, \gamma_2, \ldots, \gamma_q) | Y)},$$

with the expression for $p(\gamma | Y)$ given by (B.1). The other components of $\gamma$ can be in turn similarly generated, and then the other components of $\theta$ can be generated according to the conditional distributions from Steps 2–4. The main computational burden is in (i) forming $R_\gamma^2$, (ii) forming $\hat{\beta}_\gamma$, and (iii) generating from $N(c_1\hat{\beta}_\gamma, c_2(X_\gamma'X_\gamma)^{-1})$, where $c_1$ and $c_2$ are constants. All of these ostensibly require calculation of $(X_\gamma'X_\gamma)^{-1}$, for which $O(q_\gamma^3)$
operations are required. In fact, (i) and (ii) require only $\beta$, which can be calculated by solving $(X'_\gamma X_\gamma)\beta = X'_\gamma Y$, requiring only $O(q^2)$ operations. Now the essence of (iii) is generating from a $\mathcal{N}(0, (X'_\gamma X_\gamma)^{-1})$ distribution, and to do this we do not need to form $(X'_\gamma X_\gamma)^{-1}$. We need only express $X'_\gamma X_\gamma = U'U$, where $U$ is upper triangular. For if $Z \sim \mathcal{N}(0, I_q)$, then $U^{-1}Z \sim \mathcal{N}(0, (X'_\gamma X_\gamma)^{-1})$, and $U^{-1}Z$ is obtained without calculating $U^{-1}$, and simply by solving for $\beta$ in the equation $U\beta = Z$ (which requires only $O(q^2)$ operations, since $U$ is upper triangular). We note that, if we start with $X_\gamma$, finding the factorization $X'_\gamma X_\gamma = U'U$ requires $O(q^2)$ operations.

Now if $\gamma^+\gamma$ differ in a single component (this is the case for example when cycling through Step 1 of the algorithm), then a factorization of $X'_\gamma X_\gamma^*$ can be obtained from the factorization $X'_\gamma X_\gamma$ very efficiently: there are well known methods for updating the fit (and related quantities) of a linear regression model when a predictor is added or dropped from the model. These rely on fast updates of QR, Cholesky, or singular value decompositions when the design matrix is changed by the addition or deletion of a column. Smith and Kohn (1996) rely on fast updating of the Cholesky decomposition of $X'_\gamma X_\gamma$ in order to update $R^2_\gamma$. Our Markov chain is more involved than that of Smith and Kohn (1996) since our chain runs on $\theta = (\gamma, \sigma, \beta_0, \beta)$. Our implementation uses the “sweep operator,” a well-known method for updating a linear fit, because this provides all the quantities needed for our chain in one shot. We now describe this in more detail.

We first define the sweep operator. Let $T$ be a symmetric matrix. The sweep of $T$ on its $k^{th}$ diagonal entry $t_{kk} \neq 0$ is the symmetric matrix $S$ with

$$s_{kk} = -\frac{1}{t_{kk}}, \quad s_{ik} = \frac{t_{ik}}{t_{kk}}, \quad s_{kj} = \frac{t_{kj}}{t_{kk}}, \quad s_{ij} = t_{ij} - \frac{t_{ik} t_{kj}}{t_{kk}},$$

for $i \neq k$ and $j \neq k$. The sweep operator has an obvious inverse operator.

If we apply the sweep operator to the matrix

$$T = \begin{pmatrix}
X'X & X'Y \\
Y'X & Y'Y
\end{pmatrix}$$

(B.3)
on all 1 through $q$ diagonal entries, then we obtain the matrix

$$S = \begin{pmatrix}
- (X'X)^{-1} & (X'X)^{-1}X'Y \\
Y'X(X'X)^{-1} & Y'Y - Y'X(X'X)^{-1}X'Y
\end{pmatrix}.$$

If we sweep the augmented matrix $T$ defined in (B.3) on the diagonal entries corresponding to the covariates in $\gamma$, then from the resulting matrix $S$ we can obtain all the important quantities needed by Steps 1–4: $(X'_\gamma X_\gamma)^{-1}$ is the negative of the submatrix of $S$ corresponding to rows and columns in $\gamma$, $(X'_\gamma X_\gamma)^{-1}X'_\gamma Y$ is the submatrix of $S$ corresponding to rows in $\gamma$ and column $q + 1$, and $Y'X_\gamma(X'_\gamma X_\gamma)^{-1}X'_\gamma Y$ can be obtained by subtracting the $(q + 1, q + 1)$ element of $S$ from $Y'Y$ (with other methods, we may need to compute separately the last three quantities).

To illustrate the use of this operator, suppose that we have already swept $T$ over the covariates in $\gamma = (0, \gamma_2, \ldots, \gamma_q)$. Then we only need to perform one sweep on the first diagonal entry to get the swept matrix corresponding to the first predictor being added to the previous model ($\gamma = (1, \gamma_2, \ldots, \gamma_q)$). Conversely, since the sweep operator has an inverse, the latter matrix could be “unswept” over the first diagonal entry to get the swept matrix corresponding to dropping the first predictor.
APPENDIX C
PROOF OF THE UNIFORM ERGODICITY AND DEVELOPMENT OF THE MINORIZATION CONDITION FROM CHAPTER 5

Proof of Proposition 1

Let $\nu_\gamma$ and $p(\gamma \mid Y)$ denote the posterior distribution of $\theta$ and $\gamma$, respectively, under the prior $\nu$ on $\theta$ (we are suppressing the subscript $h$, since the hyperparameter is fixed throughout). We use $K^n$ and $V^n$ to denote the $n$-step Markov transition functions for the $\theta$ and $\gamma$ chains, respectively. Also, letting $\lambda$ denote the product of counting measure on $\{0, 1\}^q$ and Lebesgue measure on $(0, \infty) \times \mathbb{R}^{q+1}$, we use $k^n$ to denote the density of $K^n$ with respect to $\lambda$ and $v^n$ to denote the probability mass function of $V^n$. We now show that the $\theta$-chain and the $\gamma$-chain converge to their corresponding posterior distributions at exactly the same rate. For any starting state $\theta_0$ and $n \in \mathbb{N}$, we have

$$
\|K^n(\cdot) - \nu_\gamma(\cdot)\| = \sup_A |K^n(\theta_0, A) - \nu_\gamma(A)|
= \frac{1}{2} \int |k^n(\theta_0, \theta) - \nu_\gamma(\theta)| \, d\lambda
= \frac{1}{2} \int \left| v^n(\gamma_0, \gamma) p(\sigma^2, \beta_0, \beta_\gamma \mid \gamma, Y) - p(\gamma \mid Y) p(\sigma^2, \beta_0, \beta_\gamma \mid \gamma, Y) \right| \, d\lambda
= \frac{1}{2} \sum_{\gamma \in \Gamma} \left| v^n(\gamma_0, \gamma) - p(\gamma \mid Y) \right| \left[ \int \int \int p(\sigma^2, \beta_0, \beta_\gamma \mid \gamma, Y) \, d\sigma^2 \, d\beta_0 \, d\beta_\gamma \right]
= \frac{1}{2} \sum_{\gamma \in \Gamma} \left| v^n(\gamma_0, \gamma) - p(\gamma \mid Y) \right|
= \sup_B \left| V^n(\gamma_0, B) - p(\gamma \in B \mid Y) \right|.
$$

Hence, the $\theta$ chain inherits the convergence rate of its $\gamma$-subchain, a uniformly ergodic Gibbs sampler on a finite state space. \qed

Description of the Regeneration Scheme

For regeneration purposes, it is enough to restrict our attention to the Markov chain that runs on $\gamma$. This is because, as we will see later, whenever this subchain regenerates, the augmented chain that produces draws from the posterior distribution of $\theta = (\gamma, \sigma, \beta_0, \beta_\gamma)$ also regenerates. We will find a function $s : \Gamma \to [0, 1)$ and a probability
mass function \( d \) on \( \Gamma \) such that \( \nu(\gamma, \gamma') \) satisfies the minorization condition

\[
\nu(\gamma, \gamma') \geq s(\gamma) d(\gamma') \quad \text{for all } \gamma, \gamma' \in \Gamma.
\]

(C.1)

We proceed via the “distinguished point” technique introduced in Mykland et al. (1995). Let \( \gamma^* \) denote a fixed model, which we will refer to as a distinguished model, and let \( D \subset \Gamma \) be a set of models. The model \( \gamma^* \) and the set \( D \) are arbitrary, but below we give guidelines for making a practical choice of \( \gamma^* \) and \( D \). For all \( \gamma, \gamma' \in \Gamma \) we have

\[
\nu(\gamma, \gamma') = \frac{\nu(\gamma, \gamma')}{\nu(\gamma^*, \gamma')} \cdot \nu(\gamma^*, \gamma') \geq \min_{\gamma'' \in D} \frac{\nu(\gamma, \gamma'')}{\nu(\gamma^*, \gamma'')} \cdot \nu(\gamma^*, \gamma') I(\gamma' \in D).
\]

If we let \( c(\gamma^*) \) denote the normalizing constant for \( \nu(\gamma^*, \gamma') I(\gamma' \in D) \), that is, \( c(\gamma^*) = \sum_{\gamma' \in D} \nu(\gamma^*, \gamma') \), then we get

\[
\nu(\gamma, \gamma') \geq c(\gamma^*) \min_{\gamma'' \in D} \frac{\nu(\gamma, \gamma'')}{\nu(\gamma^*, \gamma'')} \cdot \frac{\nu(\gamma^*, \gamma') I(\gamma' \in D)}{c(\gamma^*)} = s(\gamma) \cdot d(\gamma'),
\]

where

\[
s(\gamma) = c(\gamma^*) \min_{\gamma'' \in D} \frac{\nu(\gamma, \gamma'')}{\nu(\gamma^*, \gamma'')} \quad \text{and} \quad d(\gamma') = \frac{\nu(\gamma^*, \gamma') I(\gamma' \in D)}{c(\gamma^*)}.
\]

Evaluating both \( s \) and \( d \) requires computing transition probabilities of the form \( \nu(\gamma, \gamma') \) which, due to the fact that the Markov chain on \( \gamma \) is a Gibbs sampler, can be expressed as

\[
\nu(\gamma, \gamma') = p(\gamma_1 | \gamma_2, \gamma_3, \ldots, \gamma_q, Y) \times p(\gamma_2 | \gamma_1', \gamma_3, \ldots, \gamma_q, Y) \times \cdots \times p(\gamma_q | \gamma_1', \gamma_2', \ldots, \gamma_{q-1}, Y),
\]

where the formula for the right side terms is given by (B.1). Since the \( \gamma \)'s in the terms on the right side differ in at most one component, the fast updating techniques discussed in Appendix A can be applied here too to speed up the computations.
By (2–11), \( P(\delta_i = 1 \mid \gamma^{(i)}, \gamma^{(i+1)}) = s(\gamma^{(i)})d(\gamma^{(i+1)})/v(\gamma^{(i)}, \gamma^{(i+1)}) \), and the normalizing constant \( c(\gamma^*) \) cancels in the numerator, so in practice there is no need to compute it, and the success probability of the regeneration indicator is simply
\[
P(\delta_i = 1 \mid \gamma^{(i)}, \gamma^{(i+1)}) = \left[ \min_{\gamma'' \in D} \frac{v(\gamma'^{\prime}, \gamma^{(i)})v(\gamma^*, \gamma^{(i+1)})} {v(\gamma^{(i)}, \gamma^{(i+1)})} \right] \frac{v(\gamma^{(i)}, \gamma^{(i+1)})I(\gamma^{(i+1)} \in D)} {v(\gamma_{\gamma'}^{\prime}, \gamma^{(i+1)})}.
\] (C.2)

The choice of \( \gamma^* \) and \( D \) affects the regeneration rate. Ideally we would like the regeneration probability to be as big as possible. Notice that regeneration can occur only if \( \gamma \) is in \( D \). This suggests making \( D \) large. However, increasing the size of \( D \) makes the first term in brackets in (C.2) smaller. We have found that a reasonable tradeoff consists of taking \( D \) to be the smallest set of models that encompasses 25\% of the posterior probability. Also, the obvious choice for \( \gamma^* \) is the HPM model. The distinguished model and the set \( D \) are selected from the output of an initial chain.

For the all-inclusive chain that runs not only on the model space but also on the space of error variance and model coefficients, we can obtain a minorization condition if we multiply the condition in (C.1) on both sides by
\[
p(\sigma^2 | \gamma', Y)p(\beta_0' | \gamma', \sigma^2, Y)p(\beta'_\gamma | \gamma', \sigma^2, \beta_0', Y).
\]
This yields
\[
k(\theta, \theta') \geq s_1(\theta)d_1(\theta') \quad \text{for all } \theta = (\gamma, \sigma, \beta_0, \beta_\gamma), \theta' = (\gamma', \sigma', \beta_0', \beta'_\gamma), \quad \text{(C.3)}
\]
where
\[
s_1(\theta) = s(\gamma) \quad \text{and} \quad d_1(\theta') = d(\gamma')p(\sigma^2 | \gamma', Y)p(\beta_0' | \gamma', \sigma^2, Y)p(\beta'_\gamma | \gamma', \sigma^2, \beta_0', Y).
\]
Hence for this bigger chain the regeneration indicator has, according to (2–11), success probability
\[
P(\delta_i = 1 \mid \theta^{(i)}, \theta^{(i+1)}) = \frac{s_1(\theta^{(i)})d_1(\theta^{(i+1)})}{k(\theta^{(i)}, \theta^{(i+1)})} = \frac{s(\gamma^{(i)})d(\gamma^{(i+1)})}{v(\gamma^{(i)}, \gamma^{(i+1)})},
\]
97
which is exactly the same as the regeneration success probability for the chain on \( \Gamma \). Hence, the augmented \( \theta \)-chain and the \( \gamma \)-chain regenerate simultaneously. Note that sampling from \( d_1 \) (which is needed to start the regeneration) is trivial. We first sample \( \gamma \) from \( d \), which is done by sampling from \( v(\gamma^*, \cdot) \) and retaining \( \gamma \) only if it is in \( D \) (and to do this we do not need to know the normalizing constant \( c(\gamma^*) \)); then we sequentially sample \( \sigma^2 \), \( \beta_0 \), and \( \beta_\gamma \) from \( p(\sigma^2 | \gamma, Y) \), \( p(\beta_0 | \gamma, \sigma^2, Y) \), and \( p(\beta_\gamma | \gamma, \sigma^2, \beta_0, Y) \), respectively.
APPENDIX D
MAP FOR THE OZONE PREDICTORS IN FIGURE 5-5

Table D-1. The 44 predictors used in the ozone illustration. The symbol "." represents an interaction.

<table>
<thead>
<tr>
<th>Number</th>
<th>Predictor</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>vh (Vandenburg 500 millibar pressure height (m))</td>
</tr>
<tr>
<td>2</td>
<td>wind (Wind speed (mph) at Los Angeles International Airport (LAX))</td>
</tr>
<tr>
<td>3</td>
<td>humid (Humidity (percent) at LAX)</td>
</tr>
<tr>
<td>4</td>
<td>temp (Sandburg Air Force Base temperature (F))</td>
</tr>
<tr>
<td>5</td>
<td>ibh (Inversion base height at LAX)</td>
</tr>
<tr>
<td>6</td>
<td>dpg (Daggett pressure gradient (mm Hg) from LAX to Daggett, CA)</td>
</tr>
<tr>
<td>7</td>
<td>ibt (Inversion base temperature at LAX)</td>
</tr>
<tr>
<td>8</td>
<td>vis (Visibility (miles) at LAX)</td>
</tr>
<tr>
<td>9</td>
<td>vh(^2)</td>
</tr>
<tr>
<td>10</td>
<td>wind(^2)</td>
</tr>
<tr>
<td>11</td>
<td>humid(^2)</td>
</tr>
<tr>
<td>12</td>
<td>temp(^2)</td>
</tr>
<tr>
<td>13</td>
<td>ibh(^2)</td>
</tr>
<tr>
<td>14</td>
<td>dpg(^2)</td>
</tr>
<tr>
<td>15</td>
<td>ibt(^2)</td>
</tr>
<tr>
<td>16</td>
<td>vis(^2)</td>
</tr>
<tr>
<td>17</td>
<td>vh.wind</td>
</tr>
<tr>
<td>18</td>
<td>vh.humid</td>
</tr>
<tr>
<td>19</td>
<td>wind.humid</td>
</tr>
<tr>
<td>20</td>
<td>vh.temp</td>
</tr>
<tr>
<td>21</td>
<td>wind.temp</td>
</tr>
<tr>
<td>22</td>
<td>humid.temp</td>
</tr>
<tr>
<td>23</td>
<td>vh.ibh</td>
</tr>
<tr>
<td>24</td>
<td>wind.ibh</td>
</tr>
<tr>
<td>25</td>
<td>humid.ibh</td>
</tr>
<tr>
<td>26</td>
<td>temp.ibh</td>
</tr>
<tr>
<td>27</td>
<td>vh.dpg</td>
</tr>
<tr>
<td>28</td>
<td>wind.dp</td>
</tr>
<tr>
<td>29</td>
<td>humid.d</td>
</tr>
<tr>
<td>30</td>
<td>temp.dp</td>
</tr>
<tr>
<td>31</td>
<td>ibh.dpg</td>
</tr>
<tr>
<td>32</td>
<td>vh.ibt</td>
</tr>
<tr>
<td>33</td>
<td>wind.ib</td>
</tr>
<tr>
<td>34</td>
<td>humid.i</td>
</tr>
<tr>
<td>35</td>
<td>temp.ib</td>
</tr>
<tr>
<td>36</td>
<td>ibh.ibt</td>
</tr>
<tr>
<td>37</td>
<td>dpg.ibt</td>
</tr>
<tr>
<td>38</td>
<td>vh.vis</td>
</tr>
<tr>
<td>39</td>
<td>wind.vi</td>
</tr>
<tr>
<td>40</td>
<td>humid.v</td>
</tr>
<tr>
<td>41</td>
<td>temp.vi</td>
</tr>
<tr>
<td>42</td>
<td>ibh.vis</td>
</tr>
<tr>
<td>43</td>
<td>dpg.vis</td>
</tr>
<tr>
<td>44</td>
<td>ibt.vis</td>
</tr>
</tbody>
</table>
REFERENCES


BIOGRAPHICAL SKETCH

Eugenia Buta was born in 1982 in Romania. In 2000, she was admitted to the University of Oradea, Romania, from where she earned her Bachelor’s degree in Mathematics-Informatics in 2004. She then joined the Department of Statistics at the University of Florida to pursue a Ph.D. degree. During her graduate student years, she served as a Teaching Assistant for several undergraduate and graduate level courses in the Department of Statistics. She expects to receive her doctorate degree in Statistics in August 2010.