Gibbs sampling for a Bayesian hierarchical general linear model

Alicia A. Johnson
Galin L. Jones, University of Minnesota - Twin Cities
Gibbs sampling for a Bayesian hierarchical general linear model

Alicia A. Johnson
Department of Mathematics, Statistics, and Computer Science
Macalester College
e-mail: ajohns24@macalester.edu

and

Galin L. Jones
School of Statistics
University of Minnesota
e-mail: galin@stat.umn.edu

Abstract: We consider a Bayesian hierarchical version of the normal theory general linear model which is practically relevant in the sense that it is general enough to have many applications and it is not straightforward to sample directly from the corresponding posterior distribution. Thus we study a block Gibbs sampler that has the posterior as its invariant distribution. In particular, we establish that the Gibbs sampler converges at a geometric rate. This allows us to establish conditions for a central limit theorem for the ergodic averages used to estimate features of the posterior. Geometric ergodicity is also a key requirement for using batch means methods to consistently estimate the variance of the asymptotic normal distribution. Together, our results give practitioners the tools to be as confident in inferences based on the observations from the Gibbs sampler as they would be with inferences based on random samples from the posterior. Our theoretical results are illustrated with an application to data on the cost of health plans issued by health maintenance organizations.

Keywords and phrases: Gibbs sampler, convergence rate, drift condition, general linear model, geometric ergodicity, Markov chain, Monte Carlo.

Received October 2009.

1. Introduction

The flexibility of Bayesian hierarchical models makes them widely applicable. One of the most popular (see, e.g., Gelman et al., 2004; Spiegelhalter et al., 2005) is a version of the usual normal theory general linear model. Let $Y$ denote an $N \times 1$ response vector and suppose $\beta$ is a $p \times 1$ vector of regression coefficients, $u$ is a $k \times 1$ vector, $X$ is a known $N \times p$ design matrix having full column rank,
and \( Z \) is a known \( N \times k \) matrix. Then for \( r, s, t \in \{1, 2, \ldots\} \), the hierarchy is

\[
Y | \beta, u, \lambda_R, \lambda_D \sim N_N \left( X \beta + Zu, \lambda_R^{-1} I_N \right)
\]

\[
\beta | u, \lambda_R, \lambda_D \sim \sum_{i=1}^{r} \eta_i N_p \left( b_i, B^{-1} \right)
\]

\[
u | \lambda_R, \lambda_D \sim N_k \left( 0, \lambda_D^{-1} I_k \right)
\]

\[
\lambda_R \sim \sum_{j=1}^{s} \phi_j \text{Gamma} \left( r_j, r_{j2} \right)
\]

\[
\lambda_D \sim \sum_{l=1}^{t} \psi_l \text{Gamma} \left( d_{l1}, d_{l2} \right)
\]

where the mixture parameters \( \eta_i, \phi_j, \) and \( \psi_l \) are known nonnegative constants which satisfy

\[
\sum_{i=1}^{r} \eta_i = \sum_{j=1}^{s} \phi_j = \sum_{l=1}^{t} \psi_l = 1
\]

and we say \( W \sim \text{Gamma}(a, b) \) if it has density proportional to \( w^{a-1} e^{-bw} \) for \( w > 0 \). Further, we require \( \beta \) and \( u \) to be a posteriori conditionally independent given \( \lambda_R, \lambda_D \), and \( y \) which holds if and only if \( X^T Z = 0 \). Finally, \( b_i \in \mathbb{R} \) and positive definite matrix \( B \) are known and the hyperparameters \( r_j, r_{j2}, d_{l1}, \) and \( d_{l2} \) are all assumed to be positive.

Let \( \xi = \left( u^T, \beta^T \right)^T \) and \( \lambda = \left( \lambda_R, \lambda_D \right)^T \). Then the posterior has support \( \mathcal{X} = \mathbb{R}^{k+p} \times \mathbb{R}^2_+ \) and a density characterized by

\[
\pi(\xi, \lambda | y) \propto f(y | \xi, \lambda) f(\xi | \lambda) f(\lambda)
\]

where \( y \) is the observed data and \( f \) denotes a generic density. Posterior inference is often based on the expectation of a function \( g : \mathcal{X} \to \mathbb{R} \) with respect to the posterior. For the model (1.1) we can only rarely calculate the expectation

\[
E_\pi g(\xi, \lambda) := \int_{\mathcal{X}} g(\xi, \lambda) \pi(\xi, \lambda | y) d\xi d\lambda,
\]

since it is a ratio of two potentially high-dimensional intractable integrals. Hence inference regarding the posterior may require Markov chain Monte Carlo (MCMC) methods. We consider two-component Gibbs sampling which produces a Harris ergodic Markov chain \( \Phi = \{ (\xi_0, \lambda_0), (\xi_1, \lambda_1), \ldots \} \) with invariant density \( \pi(\xi, \lambda | y) \).

Suppose \( E_\pi | g | < \infty \) and we obtain \( n \) observations from the Gibbs sampler. Then a natural estimate of \( E_\pi g \) is \( \bar{g}_n = n^{-1} \sum_{i=0}^{n-1} g(\xi_i, \lambda_i) \) since \( \bar{g}_n \to E_\pi g \) with probability 1 as \( n \to \infty \). In other words, the longer we run the Gibbs sampler, the better our estimate is likely to be. However, this gives no indication of how large \( n \) must be to ensure the Monte Carlo error \( \bar{g}_n - E_\pi g \) is sufficiently small. The size of this error is usually judged by appealing to its approximate sampling
distribution via a Markov chain central limit theorem (CLT), which in the cases of current interest takes the form

\[ \sqrt{n}(\bar{g}_n - E_\pi g) \overset{d}{\to} N(0, \sigma^2_g) \quad \text{as} \quad n \to \infty \]  

(1.2)

where \( \sigma^2_g \in (0, \infty) \). Due to the serial correlation in \( \Phi \), the variance \( \sigma^2_g \) will be complicated and require specialized techniques (such as batch means or spectral methods) to estimate consistently with \( \hat{\sigma}^2_g \), say. Suppose \( \hat{\sigma}^2_n \to \sigma^2_g \) with probability 1 as \( n \to \infty \). Then an asymptotically valid Monte Carlo standard error (MCSE) is given by \( \hat{\sigma}_n / \sqrt{n} \). In turn, this can be used to perform statistical analysis of the Monte Carlo error and to implement rigorous sequential stopping rules for determining the length of simulation required (see Flegal et al., 2008; Jones and Hobert, 2001) so that the user will have as much confidence in the simulation results as if the observations were a random sample from the posterior; this is described in more detail in Section 4.

Unfortunately, for Harris ergodic Markov chains simple moment conditions are not sufficient to ensure an asymptotic distribution for the Monte Carlo error or that we can consistently estimate \( \sigma^2_g \). In addition, we need to know that the convergence of \( \Phi \) occurs rapidly. Thus, one of our goals is to establish verifiable conditions under which the Gibbs sampler is geometrically ergodic, that is, it converges to the posterior in total variation norm at a geometric rate.

We know of three papers that address geometric ergodicity of Gibbs samplers in the context of the normal theory linear model with proper priors. These are Hobert and Geyer (1998), Jones and Hobert (2004), and Papaspiliopoulos and Roberts (2008). The linear model we consider substantively differs from those in Papaspiliopoulos and Roberts (2008) in that we do not assume the variance components are known. Our model is also much more general than the one-way random effects model in Hobert and Geyer (1998) and Jones and Hobert (2004). Gibbs sampling for the balanced one-way random effects model is also considered in Rosenthal (1995) where coupling techniques were used to establish upper bounds on the total variation distance to stationarity. However, these results fall short of establishing geometric ergodicity of the associated Markov chain.

The rest of this paper is organized as follows. Gibbs sampling for the Bayesian hierarchical general linear model is discussed in Section 2 and geometric ergodicity for these Gibbs samplers is established in Section 3. Conditions for the CLT (1.2) are given in Section 4 along with a description of the method of batch means for estimating the variance of the asymptotic normal distribution. Finally, our results are illustrated with a numerical example in Section 5. Many technical details are deferred to the appendix.

2. The Gibbs samplers

The full conditional densities required for implementation of the two-component block Gibbs sampler are as follows: Conditional on \( \xi \) and \( y \), \( \lambda \) follows the dis-
Gibbs sampling for a Bayesian linear model

The distribution corresponding to density

\[ f(\lambda | \xi, y) = \sum_{j=1}^{s} \sum_{l=1}^{t} \phi_j \psi_l f_{1j}(\lambda_R | \xi, y) f_{2l}(\lambda_D | \xi, y) \]  \hspace{2cm} (2.1)\]

where \( f_{1j}(\cdot | \xi, y) \) denotes a Gamma\((r_{j1} + N/2, r_{j2} + v_1(\xi)/2)\) density and \( f_{2l}(\cdot | \xi, y) \) denotes a Gamma\((d_{l1} + k/2, d_{l2} + v_2(\xi)/2)\) density with

\[ v_1(\xi) := (y - X\beta - Zu)^T (y - X\beta - Zu), \quad v_2(\xi) := u^T u. \hspace{2cm} (2.2)\]

Also,

\[ \xi | \lambda, y \sim \sum_{i=1}^{r} \eta_i N_{k+p}(m_i, \Sigma^{-1}) \]

where

\[ \Sigma^{-1} = \begin{pmatrix} (\lambda_R Z^T Z + \lambda_D I_k)^{-1} & 0 \\ 0 & (\lambda_R X^T X + B)^{-1} \end{pmatrix} \]

\[ m_i = \begin{pmatrix} \lambda_R (\lambda_R Z^T Z + \lambda_D I_k)^{-1} Z^T y \\ \lambda_R X^T X + B \end{pmatrix} (\lambda_R X^T y + B_{bi})^{-1}. \hspace{2cm} (2.3)\]

These follow from our assumption that \( X^T Z = 0 \).

There are two possible update orders for our 2-component Gibbs sampler. First, let \( \Phi_1 \) denote the Markov chain produced by the Gibbs sampler which updates \( \xi \) followed by \( \lambda \) in each iteration so that a one-step transition looks like \((\xi', \lambda') \to (\xi, \lambda') \to (\xi, \lambda)\). Then the one-step Markov transition density (Mtd) for \( \Phi_1 \) is

\[ k_1(\xi, \lambda | \xi', \lambda') = f(\xi | \lambda', y) f(\lambda | \xi, y). \]

Similarly, let \( \Phi_2 \) denote the Markov chain produced by the Gibbs sampler which updates \( \lambda \) followed by \( \xi \) in each iteration so that the one-step transition is \((\xi', \lambda') \to (\xi', \lambda) \rightarrow (\xi, \lambda)\). Then the corresponding Mtd is

\[ k_2(\xi, \lambda | \xi', \lambda') = f(\lambda | \xi', y) f(\xi | \lambda, y). \]

Also, let \( \Phi_\xi = \{\xi_0, \xi_1, \ldots\} \) and \( \Phi_\lambda = \{\lambda_0, \lambda_1, \ldots\} \) denote the associated marginal chains with Mtds

\[ k_\xi(\xi | \xi') = \int_{\mathbb{R}^k_+} f(\lambda | \xi', y) f(\xi, \lambda, y) \, d\lambda \]

and

\[ k_\lambda(\lambda | \lambda') = \int_{\mathbb{R}^{k+p}} f(\xi | \lambda', y) f(\lambda | \xi, y) \, d\xi, \]

respectively.

Because the Mtd’s are strictly positive on the state space it is straightforward to show that \( \Phi_1 \) and \( \Phi_2 \) are Harris ergodic; see also Lemma 1 in Tan and Hobert.
The posterior density \( \pi(\xi, \lambda | y) \) is invariant for \( \Phi_1 \) and \( \Phi_2 \) by construction. Similarly, \( \Phi_\xi \) and \( \Phi_\lambda \) are Harris ergodic with invariant densities the marginal posteriors \( \pi(\xi | y) \) and \( \pi(\lambda | y) \), respectively. Hence all four Markov chains converge in total variation norm to their respective invariant distributions. In the next section we establish conditions under which this convergence occurs at a geometric rate.

### 3. Geometric ergodicity

#### 3.1. Establishing geometric ergodicity

Our main goal in this section is to establish conditions for the geometric ergodicity of \( \Phi_1 \) and \( \Phi_2 \). Before doing so it is useful to acquaint ourselves with a concept introduced by Roberts and Rosenthal (2001). Let \( X = \{X_n, n \geq 0\} \) be a Markov chain on a space \( X \) and \( Y = \{Y_n, n \geq 0\} \) a stochastic process on a possibly different space \( Y \). Then \( Y \) is de-initializing for \( X \) if, for each \( n \geq 1 \), conditionally on \( Y_n \) it follows that \( X_n \) is independent of \( X_0 \). Roughly speaking, Roberts and Rosenthal (2001) use this concept to show that \( Y \) controls the convergence properties of the Markov chain \( X \).

To establish the geometric ergodicity of \( \Phi_1 \) and \( \Phi_2 \) it suffices to work with the marginal chains \( \Phi_\xi \) and \( \Phi_\lambda \). First, \( \Phi_\xi \) is de-initializing for \( \Phi_1 \) and \( \Phi_\lambda \) is de-initializing for \( \Phi_2 \). Results in Roberts and Rosenthal (2001) imply that if \( \Phi_\xi \) (\( \Phi_\lambda \)) is geometrically ergodic, so is \( \Phi_1 \) (\( \Phi_2 \)). Further, \( \Phi_1 \) and \( \Phi_2 \) are co-de-initializing. Hence if one is geometrically ergodic, then they both are and Lemma 3.1 follows directly.

**Lemma 3.1.** If \( \Phi_\xi \) or \( \Phi_\lambda \) is geometrically ergodic, then so are \( \Phi_1 \) and \( \Phi_2 \).

Accordingly, we can proceed by studying the convergence behavior of the marginal chains. We establish geometric ergodicity for \( \Phi_\xi \) by establishing a drift condition. That is we need to specify a function \( V : \mathbb{R}^{k+p} \rightarrow \mathbb{R}_+ \) and constants \( 0 < \gamma < 1 \) and \( L < \infty \) such that

\[
E[V(\xi) \mid \xi'] \leq \gamma V(\xi') + L \quad \text{for all } \xi' \in \mathbb{R}^{k+p}
\]  

(3.1)

where the expectation is taken with respect to the Mtd \( k_\xi \). Let \( W(\xi) = 1 + V(\xi) \), \( b = L + 1 - \gamma \) and \( C = \{\xi : W(\xi) \leq 4b/(1 - \gamma)\} \). Jones and Hobert (2004, Lemma 3.1) show that equation (3.1) implies

\[
\Delta W(\xi') := E[W(\xi) \mid \xi'] - W(\xi') \leq -\frac{1 - \gamma}{2} W(\xi') + 2bI(\xi' \in C).
\]

Here \( \Delta W(\xi') \) is the drift, \( V \) (or \( W \)) is a drift function and \( \gamma \) a drift rate. If \( \xi' \notin C \) the expected change in \( W \) is negative so \( \Phi_\xi \) will tend to "drift" to \( C \), that is, where the value of \( W \) is small. Moreover, it also does it in such a way that the drift towards \( C \) is faster when \( \gamma \) is small. On the other hand, if \( \gamma \approx 1 \) the drift will be slow. Thus the value of \( \gamma \) is intimately connected to the convergence rate of \( \Phi_\xi \); for a thorough accessible discussion of the connection
see Jones and Hobert (2001, Section 3.3). Hence examination of $\gamma$ can give us some intuition for the convergence behavior of $\Phi_x$. However, drift functions are not unique so this examination generally will not lead to definitive conclusions.

One method for using the drift condition (3.1) to establish geometric ergodicity generally requires consideration of petite sets; see Lemma 15.2.8 of Meyn and Tweedie (1993). However, this may be avoided in the current setting. A function $V : \mathbb{R}^{k+p} \to \mathbb{R}$ is unbounded off compact sets if the set \{\(\xi \in \mathbb{R}^{k+p} : V(\xi) \leq d\)\} is compact for any $d > 0$. Note that the maximal irreducibility measure for $\Phi_x$ is equivalent to Lebesgue on $\mathbb{R}^{k+p}$ so that its support certainly has a non-empty interior. A straightforward application of Fatou’s lemma shows that $\Phi_x$ is Feller and hence if $V$ is unbounded off compact sets it is also unbounded off petite sets by Theorem 6.0.1 in Meyn and Tweedie (1993). The following proposition now follows easily from Lemma 15.2.8 of Meyn and Tweedie (1993) and our Lemma 3.1.

**Proposition 3.1.** Suppose (3.1) holds for a drift function that is unbounded off compact sets. Then $\Phi_x$ is geometrically ergodic and so are $\Phi_1$ and $\Phi_2$.

In Section 3.2 we develop conditions on our Bayesian model (1.1) which are sufficient for the conditions of Proposition 3.1.

### 3.2. Drift for $\Phi_x$

For all $j \in \{1, \ldots, s\}$ and $l \in \{1, \ldots, t\}$, define constants

$$
\delta_{j1} = \frac{\sum_{i=1}^{N} z_i \left(Z^T Z\right)^{-1} z_i^T}{2r_{j1} + N - 2};
\quad \delta_{l2} = \frac{k}{2d_{l1} + k - 2},
\quad \delta_{j3} = \frac{\sum_{i=1}^{N} x_i \left(X^T X\right)^{-1} x_i^T}{4(2r_{j1} + N - 2)};
\quad \delta_{l4} = \frac{k + \sum_{i=1}^{N} z_i z_i^T}{2d_{l1} + k - 2}.
$$

(3.2)

Also, let $x_i$ and $z_i$ denote the $i$th rows of matrices $X$ and $Z$, respectively, and let $y_i$ and $u_i$ denote the $i$th elements of vectors $y$ and $u$, respectively. Next, for $i \in \{1, \ldots, r\}$ define

$$
G_i(\lambda) := \sum_{m=1}^{N} |E_i(y_m - x_m \beta - z_m u(\lambda, y))|^2 + \sum_{m=1}^{k} |E_i(u_m|\lambda, y)|^2
$$

where $E_i$ denotes expectation with respect to the $N_{k+p}(m_i, \Sigma^{-1})$ distribution.

**Proposition 3.2.** Assume there exists some $K < \infty$ such that $G_i(\lambda) \leq K$ for all $\lambda \in \mathbb{R}^r_+$ and $i \in \{1, \ldots, r\}$. Let $V(\xi) = v_1(\xi) + v_2(\xi)$ where $v_1(\cdot)$ and $v_2(\cdot)$ are defined at (2.2).

1. If $Z^T Z$ is nonsingular, $d_{l1} > 1$ for all $l \in \{1, \ldots, t\}$, and

$$
r_{j1} > 0 \vee 0.5 \left( \sum_{i=1}^{N} z_i (Z^T Z)^{-1} z_i^T - N + 2 \right)
$$

for all $j \in \{1, \ldots, s\}$,
then (3.1) holds for drift function $V(\xi)$ with $\max_{j,l}\{\delta_{j3}, \delta_{l4}\} \leq \gamma < 1$ and

$$L = \sum_{i=1}^{N} x_i B^{-1} x_i^T + \max_{j,l} \{2r_{j2}\delta_{j1} + 2d_{l2}\delta_{l2}\} + K.$$  

2. If for all $j \in \{1, \ldots, s\}$ and $l \in \{1, \ldots, t\}$

$$r_{j1} > 0 \vee 0.5 \left[ 0.25 \sum_{i=1}^{N} x_i (X^T X)^{-1} x_i^T - N + 2 \right]$$

and

$$d_{l1} > 0.5 \left[ 2 + \sum_{i=1}^{N} z_i z_i^T \right]$$

then (3.1) holds for drift function $V(\xi)$ with $\max_{j,l}\{\delta_{j3}, \delta_{l4}\} \leq \gamma < 1$ and

$$L = \frac{1}{4} \sum_{i=1}^{N} x_i B^{-1} x_i^T + \max_{j,l} \{2r_{j2}\delta_{j3} + 2d_{l2}\delta_{l3}\} + K.$$  

Proof. See Appendix A.2.

Notice that the formulations of $\gamma$ given by Proposition 3.2 depend on the Bayesian model setting through $\delta_{j1}$, $\delta_{l2}$, $\delta_{j3}$, and $\delta_{l4}$. Therefore, the drift and convergence rates of the $\Phi_{\xi}$ marginal chain (hence the Gibbs samplers) may be sensitive to changes in the dimension $k$ of $u$, the total number of observations $N$, or the hyperparameter setting. However, it is interesting that the dimension of $\beta$, which is $p$, has only an indirect impact on this result. Specifically, when $Z^T Z$ is nonsingular the value of $p$ has no impact, that is, the drift rate is unaffected by changes in $p$. Of course, changing $p$ does mean that $X$ changes which may impact $\delta_{j3}$ which in turn can change the permissible hyperparameters $r_{j1}$ and the drift rate when $Z^T Z$ is singular.

Example 3.1. Consider the balanced random intercept model derived from (1.1) for $k$ subjects with $m$ observations each. In this case, $Z = I_k \otimes 1_m$ where $\otimes$ denotes the Kronecker product and $1_m$ represents a vector of ones of length $m$. Hence $Z^T Z = mI_k$ is nonsingular. Define

$$M_{N,k} := \max_{l} \left\{ \frac{k}{2r_{j1} + N - 2}, \frac{k}{2d_{l1} + k - 2} \right\}.$$  

If $d_{l1} > 1$ for all $l$, Condition 1 of Proposition 3.2 establishes drift rate $M_{N,k} \leq \gamma < 1$. Notice that $M_{N,k} \to 1$ as $k \to \infty$ and hence $\gamma \to 1$ as well. This supports our intuition that the Gibbs sampler should converge more slowly as its dimension increases. On the other hand, if $k$ is held constant but $m$ increases so that $N = km \to \infty$, then

$$M_{N,k} = \frac{k}{2d_{l1} + k - 2}.$$
Thus increasing the number of observations per subject does not have the same negative, qualitative impact as increasing the number of subjects. Finally, \( M_{N,k} \rightarrow 1 \) (hence \( \gamma \rightarrow 1 \)) when \( k \) is held constant and \( d_{11} \rightarrow 1 \) for any \( l \).

Consider the condition that \( G_i(\lambda) \leq K \) for all \( \lambda \in \mathbb{R}^2_+ \) and \( i \in \{1, \ldots, r\} \). In our experience it is often straightforward to show that \( G_i \) is bounded and, if desired, numerical optimization methods yield appropriate \( K \). The following result establishes this condition for important special cases of (1.1).

**Proposition 3.3.**

1. If \( Z = 0 \), then \( G_i(\lambda_R) \) is bounded for all \( \lambda_R \in \mathbb{R}_+ \) and \( i \in \{1, \ldots, r\} \).
2. Assume \( b_i = 0 \) for all \( i \in \{1, \ldots, r\} \) and \( Z^T Z \) is nonsingular. Then

\[
G_i(\lambda) \leq y^T y + y^T Z (Z^T Z)^{-2} Z^T y
\]

for all \( \lambda \in \mathbb{R}^2_+ \) and \( i \in \{1, \ldots, r\} \).

**Proof.** See Appendix A.2.

We are now in position to state conditions on (1.1) guaranteeing geometric ergodicity of the Gibbs samplers \( \Phi_1 \) and \( \Phi_2 \). This follows easily from Propositions 3.1 and 3.2 if the drift function \( V(\xi) = v_1(\xi) + v_2(\xi) \) is unbounded off compact sets on \( \mathbb{R}^{k+p} \). Define \( S = \{ \xi \in \mathbb{R}^{k+p} : V(\xi) = v_1(\xi) + v_2(\xi) \leq d \} \) where \( d > 0 \). Notice that \( V \) is continuous so it is sufficient to show that, on \( S \), \( |\beta_i| \) is bounded for \( i \in \{1, 2, \ldots, p\} \) and \( |u_j| \) is bounded for \( j \in \{1, 2, \ldots, k\} \).

Clearly, \( S_1 \subset S_2 = \{ \xi : u^T u \leq d \} \) and it is obvious that each \( |u_j| \) is bounded on \( S_2 \) hence also on \( S \). Moreover, note that \( v_2 \rightarrow \infty \) as \( |u_j| \rightarrow \infty \). Given that the \( |u_j| \) are bounded it is easy to see that \( v_1 \rightarrow \infty \) and \( |\beta_i| \rightarrow \infty \). Putting this together we see that \( V \) is unbounded off compact sets. The main result of this section follows.

**Theorem 3.1.** Assume the conditions of Proposition 3.1. Then the Markov chain \( \Phi_\xi \) and the Gibbs samplers \( \Phi_1 \) and \( \Phi_2 \) are geometrically ergodic.

4. Interval estimation

Suppose we want to estimate an expectation \( E_\pi g := \int_X g(\xi, \lambda) \pi(\xi, \lambda | y) \Pi \Pi d\xi d\lambda \) where \( g \) is real-valued and \( \pi \)-integrable. It is straightforward to estimate \( E_\pi g \) with \( \hat{g}_n := n^{-1} \sum_{i=0}^{n-1} g(\xi_i, \lambda_i) \). A key step in the statistical analysis of \( \hat{g}_n \) is the assessment of the Monte Carlo error \( \hat{g}_n - E_\pi g \) through its approximate sampling distribution.

**Theorem 4.1.** Assume the conditions of Theorem 3.1. If \( E_\pi |g|^{2+\epsilon} < \infty \) for some \( \epsilon > 0 \), then there is a constant \( \sigma_\pi^2 \in (0, \infty) \) such that for any initial distribution \( \sqrt{n}(\hat{g}_n - E_\pi g) \overset{d}{\rightarrow} N(0, \sigma_\pi^2) \) as \( n \rightarrow \infty \).
The proof of this theorem follows easily from Theorem 3.1, Theorem 2 of Chan and Geyer (1994) and Section 1 of Flegal and Jones (2010). Roughly speaking, results in Hobert et al. (2002), Jones et al. (2006) and Bednorz and Latuszynski (2007) show that, under conditions comparable to those required for Theorem 4.1, techniques such as regenerative simulation and batch means can be used to construct an estimator of \( \sigma^2_g \), say \( \hat{\sigma}^2_n \), such that \( \hat{\sigma}^2_n \to \sigma^2_g \) as \( n \to \infty \) almost surely. See Flegal and Jones (2010) for the conditions required to ensure consistency of overlapping batch means and spectral estimators of \( \sigma^2_g \).

Before giving a precise discussion of the conditions for consistency we need a preliminary definition and result. Let \( X \subseteq \mathbb{R}^d \) for \( d \geq 1 \) and \( k : X \times X \to [0, \infty) \) be an Mtd with respect to Lebesgue measure. Suppose there exist a function \( s : X \to [0, 1) \) and a density \( q \) such that for all \( x, x' \in X \)

\[
k(x|x') \geq s(x')q(x).
\]

Then we say there is a minorization condition for \( k \).

**Lemma 4.1.** Let \( C \subseteq X \) be compact and assume \( c > 0 \) where

\[
c = \int_C k(x|x^*) \, dx
\]

for some \( x^* \in X \). If for each \( x' \), \( k(\cdot|x') \) is positive and continuous on \( C \), then there exists a minorization condition for \( k \).

**Proof.** The proof follows a technique first introduced by Mykland et al. (1995). Fix \( x^* \in X \). Then for all \( x \in C \)

\[
k(x|x') = \frac{k(x|x^*)}{k(x|x^*)}k(x|x') \geq \left[ \inf_{x \in C} \frac{k(x|x')}{k(x|x^*)} \right] k(x|x^*).
\]

Let \( x_m \) be the point where the infimum is achieved. Then the minorization follows by setting \( q(x) = c^{-1}k(x|x^*)I(x \in C) \) and

\[
s(x') = c\frac{k(x_m|x')}{k(x_m|x^*)}.
\]

The conditions of Lemma 4.1 are not the weakest that ensure the existence of a minorization condition but they will suffice for our purposes. In particular, it is straightforward to use Lemma 4.1 to see that there exists a minorization condition for both \( k_1 \) and \( k_2 \) the Mtd’s for \( \Phi_1 \) and \( \Phi_2 \), respectively. Also, Hobert et al. (2006) derived an explicit closed form expression for a minorization for a Markov chain for which \( \Phi_2 \) is a special case.

The consistency results for \( \hat{\sigma}^2_n \) in Flegal and Jones (2010), Hobert et al. (2002), Jones et al. (2006) and Bednorz and Latuszynski (2007) all require that a minorization condition hold. The efficacy of regenerative simulation is utterly dependent upon the minorization while minorization is irrelevant to the implementation of batch means and spectral methods. That is, the minorization is
purely a technical device used in the proofs of consistency for batch means and spectral estimators.

We use the method of batch means in Section 5 to estimate $\sigma_g^2$. Let $n$ be the simulation length, $b_n = \lfloor n^a \rfloor$ and $a_n = \lfloor n/b_n \rfloor$. Now define

$$\bar{Y}_j := \frac{1}{b_n} \sum_{i=(j-1)b_n}^{jb_n-1} g(\xi_i, \lambda_i) \quad \text{for } j = 1, \ldots, a_n.$$ 

The batch means estimate of $\sigma_g^2$ is

$$\hat{\sigma}_n^2 = \frac{b_n}{a_n-1} \sum_{j=1}^{a_n} (\bar{Y}_j - \bar{g}_n)^2. \quad (4.1)$$

Putting together our Theorem 3.1 and Lemma 4.1 with results in Jones et al. (2006) and Bednorz and Latuszynski (2007) we have the following consistency result.

**Theorem 4.2.** Assume the conditions of Theorem 3.1. If $E[|g|^{2+\epsilon}] < \infty$ for some $\epsilon > 0$ set $\epsilon = \epsilon_1 + \epsilon_2$ and let $\left(1 + \epsilon_1/2\right)^{-1} < a < 1$, then for any initial distribution for either $\Phi_1$ or $\Phi_2$ we have that $\hat{\sigma}_n^2 \to \sigma_g^2$ with probability 1 as $n \to \infty$. Using Theorems 4.1 and 4.2 we can use (4.1) to form an asymptotically valid confidence interval for $E_g$ in the usual way

$$\bar{g}_n \pm t_{a_n-1} \frac{\hat{\sigma}_n}{\sqrt{n}} \quad (4.2)$$

where $t_{a_n-1}$ is a quantile from a Student’s $t$ distribution with $a_n - 1$ degrees of freedom. Moreover, we can use batch means to implement the fixed-width methods of Jones et al. (2006) to determine how long to run the simulation. Following Flegal et al. (2008) let $\epsilon$ be the desired half-width of the interval in (4.2) and $n^*$ be a minimum simulation size specified by the user. Then we can terminate the simulation the first time

$$t_{a_n-1} \frac{\hat{\sigma}_n}{\sqrt{n}} + \epsilon I(n \geq n^*) + \frac{1}{n} \leq \epsilon.$$

The final interval estimate will be asymptotically valid in the sense that the interval will have the desired coverage probability for sufficiently small $\epsilon$; see also Flegal et al. (2008), Flegal and Jones (2010), Glynn and Whitt (1992) and Jones et al. (2006).

5. A numerical example

In this section we illustrate our theoretical results in the analysis of US government health maintenance organization (HMO) data. To study the cost-effectiveness of transferring military retirees from a Defense Department health
Fig 1. Individual monthly HMO premiums are plotted against the average expenses per admission in the state in which the HMO operates. Solid circles represent states in New England.

plan to health plans for government employees, information was gathered from 341 state-based health maintenance organizations (HMOs). These plans represent 42 states, the District of Columbia, Puerto Rico, and Guam. An HMO plan’s cost is measured by its monthly premium for individual subscribers. Two possible factors in this cost are (1) the typical hospital expenses in the state in which the HMO operates; and (2) the region in which the HMO operates. In Figure 1, the individual monthly premiums for the 341 HMOs are plotted against the average expenses per admission in the state of operation (both in US dollars).

Let $y_i$ denote the individual monthly premium of the $i$th HMO plan. To analyze these data, Hodges (1998) considered a Bayesian version of the following frequentist model:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \varepsilon_i$$

(5.1)
Table 1

Least squares regression results for (5.1)

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Estimate</th>
<th>Standard Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\beta_0$</td>
<td>164.989</td>
<td>1.322</td>
</tr>
<tr>
<td>$\beta_1$</td>
<td>3.910</td>
<td>1.508</td>
</tr>
<tr>
<td>$\beta_2$</td>
<td>32.799</td>
<td>5.961</td>
</tr>
</tbody>
</table>

$N = 341$

degrees of freedom = 338

MSE = SSE/338 = $\sum_{i=1}^{N}(y_i - \hat{y}_i)^2/338 = 23.79^2$

where the $\varepsilon_i$ are iid $N\left(0, \lambda_R^{-1}\right)$, $x_{i1}$ denotes the centered and scaled average expenses per admission in the state in which the $i$th HMO operates, and $x_{i2}$ is an indicator for New England. The $x_{i1}$ values were centered and scaled to avoid collinearity. Specifically, if $\tilde{x}_{i1}$ is the raw average expense per admission and $\bar{x}_1$ is the overall average expense per admission, $x_{i1} = (\tilde{x}_{i1} - \bar{x}_1)/1000$. The results of fitting (5.1) using least squares regression are summarized in Table 1.

We perform a Bayesian regression analysis based on the following hierarchical version of (5.1):

$$
\begin{align*}
    y | \beta, \lambda_R &\sim N_N \left(X \beta, \lambda_R^{-1} I_N \right) \\
    \beta | \lambda_R &\sim N_3 \left(b, B^{-1} \right) \\
    \lambda_R &\sim \text{Gamma}(r_1, r_2)
\end{align*}
$$

(5.2)

where $N = 341$, $y$ is the $N \times 1$ vector of individual premiums, $\beta = (\beta_0, \beta_1, \beta_2)$ is the vector of regression parameters, and $X$ is the $N \times 3$ data matrix. Complete specification of this model requires values for hyperparameters $(b, B, r_1, r_2)$. We chose the following prior mean and covariance matrix for $\beta$:

$$
    b = \begin{pmatrix} 0 \\
    0 \\
    0 \end{pmatrix} \quad \text{and} \quad B^{-1} = \begin{pmatrix} 100 & 0 & 0 \\
    0 & 100 & 0 \\
    0 & 0 & 100 \end{pmatrix}.
$$

Next, using an approach which is empirical Bayesian in spirit, we set the prior mean and variance for $\lambda_R$ to

$$
    E(\lambda_R) = \frac{r_1}{r_2} = \frac{1}{\text{MSE}} = 0.00177; \quad \text{and} \quad \text{Var}(\lambda_R) = \frac{r_1}{r_2^2} = 1
$$

where MSE is the least squares estimate of $\lambda_R^{-1}$ given in Table 1. Solving for $r_1$ and $r_2$ gives $r_1 = 3.122 \times 10^{-6}$ and $r_2 = 0.00177$.

Since (5.2) does not contain any random effects, it follows from Theorem 3.1 that the Gibbs sampler for $\pi(\beta, \lambda_R|y)$ is geometrically ergodic since

$$
    r_1 > 0 \lor 0.5 \left[2 - N \right] = 0
$$
and for any $\lambda_R \in \mathbb{R}_+$ the function $G(\lambda_R)$ is bounded (recall Proposition 3.3) where

$$G(\lambda_R) = \sum_{i=1}^{N} [E(y_i - x_i\beta|\lambda_R, y)]^2 = (y - XE(\beta|\lambda_R, y))^T (y - XE(\beta|\lambda_R, y)).$$

Consider estimating the posterior means of $\beta_0$, $\beta_1$, and $\beta_2$. By Lemma A.6, the fourth posterior moments of these parameters are finite. Thus Theorems 4.1 and 4.2 in conjunction with geometric ergodicity guarantee the existence of CLTs and consistent estimators of the asymptotic variance via batch means with $b_n = \lfloor n^{0.501} \rfloor$ which was chosen based on recommendations in Jones et al. (2006).

To begin our analysis of the posterior means, we simulated independent realizations of $\Phi_2$ (i.e. we updated $\lambda_R$ followed by $\beta$ in each iteration) from a variety of starting values. In each case, we required a minimum simulation length of 1000. At each successive iteration, we calculated the approximate half-widths of the Bonferroni-corrected 95% intervals for the posterior means of $\beta_0$, $\beta_1$, and $\beta_2$,

$$t_{n-1, 0.025/3} \sqrt{n} / \sqrt{n} + \frac{1}{n}.$$

Simulation continued until the half-widths for $\beta_0$, $\beta_1$, and $\beta_2$, were below 0.10, 0.02, and 0.10, respectively. The results were consistent across starting values. That is, Gibbs samplers with different starting values produced similar estimates and required similar simulation effort to meet the above specifications. Here, we present the results for the chain started from $b$, the prior mean of $\beta$, as well as for the chain started from the vector of least squares estimates of $\beta$. Under these settings, the interval half-width thresholds were met after 32089 iterations and 29584 iterations, respectively. Further, the estimates of the posterior means and corresponding standard errors based on the two chains are the same up to the specified number of significant digits. These are reported in Table 2.

### Appendix A: Appendix

#### A.1. Proof of Proposition 3.2

We will require the following general results in our proof. A proof of Lemma A.1 is given in Henderson and Searle (1981) and Lemma A.2 follows from the convexity of the inverse function.
Lemma A.1. Let $A$ be a nonsingular $n \times n$ matrix, $B$ be a nonsingular $s \times s$ matrix, $U$ be an $n \times s$ matrix, and $V$ be an $s \times n$ matrix. Then

$$(A + UBV)^{-1} = A^{-1} - A^{-1}U(B^{-1} + VA^{-1}U)^{-1}VA^{-1}.$$ 

When $U = V$ this implies

$$x^T (A + UBV)^{-1} x \leq x^T A^{-1} x$$

for any $n \times 1$ vector $x$.

Lemma A.2. Let $x$ be an $m \times 1$ vector. Also, let $A$ and $B$ be nonsingular, $m \times m$ matrices. Then

$$x^T (A + B)^{-1} x \leq \frac{1}{4} x^T (A^{-1} + B^{-1}) x .$$

We begin the proof of Proposition 3.2. Recall that

$$v_1(\xi) := (y - X\beta - Zu)^T (y - X\beta - Zu), \quad \text{and} \quad v_2(\xi) := u^T u .$$

We must show that for all $\xi' \in \mathbb{R}^{k+p}$

$$E_{k\xi}[V(\xi)|\xi'] = E_{k\xi}[v_1(\xi) + v_2(\xi)|\xi'] \leq \gamma V(\xi') + L$$

where the constants $\gamma$ and $L$ are given in the statement of Proposition 3.2. Let $E_i$ and $\text{Var}_i$ denote expectation and variance with respect to the $N_{k+p}(m_i, \Sigma^{-1})$ distribution. Similarly, let $E_{jl}$ and $\text{Var}_{jl}$ denote expectation and variance with respect to density

$$f_{ij}(\lambda_{R}|\xi, y)f_{2l}(\lambda_{D}|\xi, y)$$

defined by (2.1). Notice that

$$E_{k\xi}[V(\xi)|\xi'] = E[\text{E}(V(\xi)|\lambda)|\xi'] = \sum_{j=1}^{s} \sum_{i=1}^{t} \sum_{l=1}^{r} \phi_j \psi_i \eta_l E_{jl}(V(\xi)|\lambda)|\xi'] \quad (A.1)$$

where the first equality holds by the construction of $\Phi_{\xi}$. Thus we focus on $E_{jl}[E_i(V(\xi)|\lambda)|\xi']$ in the next 3 lemmas.

Lemma A.3. Suppose $Z^T Z$ is nonsingular. Then for all $i, j, l$

$$E_{jl}[E_i(v_1(\xi)|\lambda)|\xi'] \leq \delta_{jl} v_1(\xi') + L_1$$

where

$$L_1 = E_{jl} \left[ \sum_{m=1}^{N} [E_i(y_m - x_m\beta - z_m u|\lambda)]^2 \right] \left[ \xi' \right] + \sum_{m=1}^{N} x_m B^{-1} x_m + 2r_j \delta_{jl} . \quad (A.2)$$
Proof. Consider the inner expectation \( E_i(v_1(\xi)|\lambda) \). For any \( i \) we have

\[
E_i(v_1(\xi)|\lambda) = \sum_{m=1}^{N} E_i \left[ (y_m - x_m \beta - z_m u)^2 | \lambda \right]
\]

\[
= \sum_{m=1}^{N} [E_i(y_m - x_m \beta - z_m u|\lambda)]^2 + \sum_{m=1}^{N} \text{Var}_i(y_m - x_m \beta - z_m u|\lambda)
\]

and

\[
\text{Var}_i(y_m - x_m \beta - z_m u|\lambda) = x_m \left( \lambda_R X^T X + B \right)^{-1} x_m^T + z_m \left( \lambda_R Z^T Z + \lambda_D I_k \right)^{-1} z_m^T
\]

\[
\leq x_m B^{-1} x_m^T + \lambda_R^{-1} z_m \left( Z^T Z \right)^{-1} z_m^T
\]

by Lemma A.1. It follows that for any \( i, j, l \) we have

\[
E_{jl} \left[ E_i(v_1(\xi)|\lambda) \mid \xi' \right] \leq E_{jl} \left[ \sum_{m=1}^{N} [E_i(y_m - x_m \beta - z_m u|\lambda)]^2 \mid \xi' \right] + E_{jl} \left( \lambda_R^{-1} | \xi' \right) \sum_{m=1}^{N} z_m \left( Z^T Z \right)^{-1} z_m + \sum_{m=1}^{N} x_m B^{-1} x_m^T.
\]

Combining this with the fact that

\[
E_{jl} \left( \lambda_R^{-1} | \xi' \right) = \frac{2r_{j2} + v_1(\xi')}{2r_{j1} + N - 2} = \frac{\delta_{j1}(2r_{j2} + v_1(\xi'))}{\sum_{m=1}^{N} z_m \left( Z^T Z \right)^{-1} z_m}
\]

gives

\[
E_{jl} \left[ E_i(v_1(\xi)|\lambda) \mid \xi' \right] \leq E_{jl} \left[ \sum_{m=1}^{N} [E_i(y_m - x_m \beta - z_m u|\lambda)]^2 \mid \xi' \right] + \delta_{j1}(2r_{j2} + v_1(\xi')) + \sum_{m=1}^{N} x_m B^{-1} x_m^T
\]

\[
= \delta_{j1} v_1(\xi') + L_1.
\]

Lemma A.4. For any \( i, j, l \)

\[
E_{jl} \left[ E_i(v_1(\xi)|\lambda) \mid \xi' \right] \leq \delta_{j3} v_1(\xi') + (\delta_{l4} - \delta_{l2}) v_2(\xi') + L_2
\]

where

\[
L_2 = E_{jl} \left[ \sum_{m=1}^{N} [E_i(y_m - x_m \beta - z_m u|\lambda)]^2 \mid \xi' \right] + \frac{1}{4} \sum_{m=1}^{N} x_m B^{-1} x_m^T + 2r_{j2} \delta_{j3} + 2d_{l2}(\delta_{l4} - \delta_{l2}).
\]
Proof. Notice that for any $i, j, l$

$$E_{jl} \left[ E_{i} (y_1 | \lambda) \middle| \xi' \right] = E_{jl} \left[ \sum_{m=1}^{N} E_{i} \left[ (y_m - x_m \beta - z_m u)^2 \middle| \lambda \right] \middle| \xi' \right]$$

$$= E_{jl} \left[ \sum_{m=1}^{N} \left[ E_{i} (y_m - x_m \beta - z_m u | \lambda) \right]^2 \middle| \xi' \right]$$

$$+ \sum_{m=1}^{N} E_{jl} \left[ \text{Var}_{i} (y_m - x_m \beta - z_m u | \lambda) \middle| \xi' \right]$$

where from Lemmas A.1 and A.2

$$\text{Var}_{i} (y_m - x_m \beta - z_m u | \lambda) = x_m \left( \lambda_R X^T X + B \right)^{-1} x_m^T$$

$$+ z_m \left( \lambda_R Z^T Z + \lambda_D I_k \right)^{-1} z_m^T$$

$$\leq \frac{1}{4} x_m \left( \lambda_R^{-1} (X^T X)^{-1} + B^{-1} \right) x_m^T + \lambda_D^{-1} z_m z_m^T.$$  

Also, by (2.1) we have

$$E_{jl} (\lambda_R^{-1} | \xi') = \frac{2r_{j2} + v_1 (\xi')}{2r_{j1} + N - 2} \quad \text{and} \quad E_{jl} (\lambda_D^{-1} | \xi') = \frac{2d_{l2} + v_2 (\xi')}{2d_{l1} + k - 2}.$$ 

Therefore

$$\sum_{m=1}^{N} E_{jl} \left[ \text{Var}_{i} (y_m - x_m \beta - z_m u | \lambda) \middle| \xi' \right]$$

$$\leq \sum_{m=1}^{N} \left[ \frac{1}{4} x_m \left( \frac{2r_{j2} + v_1 (\xi')}{2r_{j1} + N - 2} (X^T X)^{-1} + B^{-1} \right) x_m^T + \frac{2d_{l2} + v_2 (\xi')}{2d_{l1} + k - 2} z_m z_m^T \right]$$

$$= \delta_{j3} (2r_{j2} + v_1 (\xi')) + \frac{1}{4} \sum_{m=1}^{N} x_m B^{-1} x_m + (2d_{l2} + v_2 (\xi')) \sum_{m=1}^{N} z_m z_m^T$$

$$= \delta_{j3} (2r_{j2} + v_1 (\xi')) + \frac{1}{4} \sum_{m=1}^{N} x_m B^{-1} x_m + (2d_{l2} + v_2 (\xi')) (\delta_{l4} - \delta_{l2}).$$  

(A.5)

The result holds by combining (A.4) and (A.5). □

Lemma A.5. For any $i, j, l$

$$E_{jl} \left[ E_{i} (y_2 | \lambda) \middle| \xi' \right] \leq \delta_{l2} v_2 (\xi') + L_3$$

where

$$L_3 = E_{jl} \left[ \sum_{m=1}^{k} \left[ E_{i} (u_m | \lambda) \right]^2 \middle| \xi' \right] + 2d_{l2} \delta_{l2}.$$  

(A.6)
Proof. First, for any \(i, j, l\)

\[
E_{jl} \left[ E_i(v_2(\xi)|\lambda) \right] = E_{jl} \left[ k \sum_{m=1}^{k} E_i \left( u_m^2 | \lambda \right) \right] + \sum_{m=1}^{k} E_{jl} \left[ \text{Var}_i(u_m|\lambda) | \xi' \right].
\]

(A.7)

Let \(e_m\) denote the \(k \times 1\) vector with the \(m\)th element being 1 and the rest of the elements being 0. Thus by Lemma A.1,

\[
\text{Var}_i(u_m|\lambda) = e^T_m (\lambda R Z^T Z + \lambda D I_k)^{-1} e_m \leq \lambda^{-1}_D e^T_m e_m = \lambda^{-1}_D.
\]

(A.8)

Also,

\[
E_{jl} \left( \lambda^{-1}_D | \xi' \right) = \frac{2d_{i2} + v_2(\xi')}{2d_{i1} + k - 2} = \delta_{i2} (2d_{i2} + v_2(\xi')).
\]

(A.9)

Putting (A.7)–(A.9) together gives

\[
E_{jl} \left[ E_i(v_2(\xi)|\lambda) \right] \leq E_{jl} \left[ k \sum_{m=1}^{k} E_i \left( u_m^2 | \lambda \right) \right] + \sum_{m=1}^{k} E_{jl} \left[ \lambda^{-1}_D \right] + \delta_{i2} v_2(\xi') + L_3.
\]

(A.10)

We are now ready to finish the proof of Proposition 3.2. We consider the case with nonsingular \(Z^T Z\) and the case in which no restrictions are placed on \(Z\) separately.

1. Case 1: \(Z^T Z\) nonsingular

   Notice that \(L_1 + L_3 \leq L\) for \(L_1\) and \(L_3\) given by (A.2) and (A.6), respectively. Then by Lemmas A.3 and A.5 we have that for any \(i, j, l\)

   \[
   E_{jl} \left[ E_i(V(\xi)|\lambda) | \xi' \right] = E_{jl} \left[ E_i(v_1(\xi) + v_2(\xi)|\lambda) \right] \leq \delta_{i1} v_1(\xi') + \delta_{i2} v_2(\xi') + L_1 + L_3 \leq \gamma V(\xi') + L.
   \]

   Combining (A.1) and (A.10) establishes the drift condition.

2. Case 2: \(Z^T Z\) is possibly singular

   Observe that \(L_2 + L_3 \leq L\) for \(L_2\) and \(L_3\) given by (A.3) and (A.6), respectively. Further, it follows from Lemmas A.4 and A.5 that for any \(i, j, l\)
Define \( Z \). Then for all \( \lambda \),

\[ G_i(\lambda_R) = \sum_{i=1}^{N} [E_i(y_i - x_i\beta|\lambda_R, y)]^2 = (y - XE_i(\beta|\lambda_R, y))^T (y - XE_i(\beta|\lambda_R, y)) \]

and \( E_i(\beta|\lambda_R, y) = (\lambda_R X^T X + B)^{-1}(\lambda_R X^T y + Bb_i) \). Then for all \( i \in \{1, \ldots, r\} \), \( G_i(\lambda_R) \) is continuous and finite both as \( \lambda_R \to 0 \) and \( \lambda_R \to \infty \). Hence \( G_i(\lambda_R) \) is bounded for all \( \lambda_R \).

Next, consider the case with \( b_i = 0 \) for all \( i \) and \( Z^T Z \) is nonsingular. Then \( \xi|\lambda, y \sim N_{k+p}(m_0, \Sigma^{-1}) \) where

\[
\Sigma^{-1} = \begin{pmatrix}
(\lambda_R Z^T Z + \lambda_D I_k)^{-1} & 0 \\
0 & (\lambda_R X^T X + B)^{-1}
\end{pmatrix}
\]

\[ m_0 = \begin{pmatrix}
\lambda_R (\lambda_R Z^T Z + \lambda_D I_k)^{-1} Z^T y \\
\lambda_R (\lambda_R X^T X + B)^{-1} X^T y
\end{pmatrix}.
\]

Define \( A_g := \lambda_R X^T X + B \) and \( A_h := \lambda_R Z^T Z + \lambda_D I_k \). Then \( \text{E}(\beta|\lambda) = \lambda_R A_g^{-1} X^T y \) and \( \text{E}(u|\lambda) = \lambda_R A_h^{-1} Z^T y \).

We must establish that there exists \( K \) for which

\[
\sum_{m=1}^{N} [E (y_m - x_m\beta - z_m u|\lambda, y)]^2 + \sum_{m=1}^{k} [E (u_m|\lambda, y)]^2 \leq K.
\]

Let

\[ f(\lambda) = (y - X\text{E}(\beta|\lambda) - Z\text{E}(u|\lambda))^T (y - X\text{E}(\beta|\lambda) - Z\text{E}(u|\lambda)) + \text{E}(u|\lambda)^T \text{E}(u|\lambda) \]

and note that the claim will be proven if we can show that \( f(\lambda) \leq K \) for all \( \lambda \). To this end, define functions \( g \) and \( h \) as

\[ g(\lambda) = (y - X\text{E}(\beta|\lambda))^T (y - X\text{E}(\beta|\lambda)) \]

\[ h(\lambda) = \text{E}(u|\lambda)^T Z^T Z\text{E}(u|\lambda) + \text{E}(u|\lambda)^T \text{E}(u|\lambda) - 2y^T Z\text{E}(u|\lambda). \]

Since the conditional independence of \( \beta \) and \( u \) given \( \lambda \) implies \( X^T Z = 0 \), a little algebra shows that \( f(\lambda) = g(\lambda) + h(\lambda) \). Thus, it suffices to find \( K_g \) and \( K_h \) such that for all \( g(\lambda) \leq K_g \) and \( h(\lambda) \leq K_h \).
First,\\n\\ng(\lambda) = y^T y + E(\beta|\lambda)^T X^T X E(\beta|\lambda) - 2y^T X E(\beta|\lambda)\\n= y^T y + \lambda^2 y^T X A_g^{-1} X^T X A_g^{-1} X^T y - 2\lambda^2 y^T X A_g^{-1} X^T y\\n= y^T y - \lambda^2 y^T X A_g^{-1} B A_g^{-1} X^T y + \lambda^2 y^T X A_g^{-1} A_g^{-1} X^T y\\n- 2\lambda^2 y^T X A_g^{-1} X^T y\\n= y^T y - \lambda^2 y^T X A_g^{-1} B A_g^{-1} X^T y - \lambda^2 y^T X A_g^{-1} X^T y\\n\leq y^T y\\n:= K_g\\nby the positive definiteness of B and A_g^{-1}.\\n
Next, we have\\n\\nh(\lambda) = \lambda^2 y^T Z A_h^{-1} Z^T Z A_h^{-1} Z^T y + \lambda^2 y^T Z A_h^{-2} Z^T y - 2\lambda^2 y^T Z A_h^{-1} Z^T y\\n= \lambda^2 y^T Z A_h^{-1} A_h A_h^{-1} Z^T y - \lambda^2 y^T Z A_h^{-2} Z^T y + \lambda^2 y^T Z A_h^{-1} Z^T y\\n- 2\lambda^2 y^T Z A_h^{-1} Z^T y\\n= (\lambda^2 - \lambda^2 A_D) y^T Z A_h^{-2} Z^T y - \lambda^2 y^T Z A_h^{-1} Z^T y.\\n
Since A_h^{-1} and A_h^{-2} are positive semidefinite we have\\n\\nh(\lambda) \leq \lambda^2 y^T Z A_h^{-2} Z^T y\\n= \lambda^2 y^T Z \left( (\lambda^2 Z^T Z)^2 + \lambda^2 (2\lambda Z^T Z + \lambda Z) \right)^{-1} Z^T y\\n\leq \lambda^2 y^T Z (\lambda^2 Z^T Z)^{-2} Z^T y\\n= y^T Z (Z^T Z)^{-2} Z^T y\\n:= K_h\\nwhere the last inequality holds by Lemma A.1. The result now follows by setting\\nK = K_g + K_h.\\n
A.3. Lemma A.6\\n
Lemma A.6. The fourth posterior moments of \( \beta_0, \beta_1, \) and \( \beta_2 \) are each finite.\\n
Proof. We present the proof for \( \beta_2 \). The proofs for \( \beta_0 \) and \( \beta_1 \) are similar. The finiteness of \( E [ \beta_2^2 | y ] \) will follow from establishing that \( E [ \beta_2^2 | \lambda_R, y ] \) is finite since\\n\\nE [ \beta_2^2 | y ] = E [ E ( \beta_2^2 | \lambda_R, y ) | y ] .\\n
To this end, recall that\\n\\n\( \beta | \lambda_R, y \sim N \left( (\lambda_R X^T X + B)^{-1} (\lambda_R X^T y + Bb), (\lambda_R X^T X + B)^{-1} \right) .\)
Also, let \( \mu_2 = \mathbb{E}(\beta_2 | \lambda_R, y) \) and \( e_3 \) denote a vector of zeroes with a one in the third position. Then

\[
\mathbb{E}[(\beta_2 - \mu_2)^4 | \lambda_R, y] = 3 \left[(\lambda_R X^T X + B)^{-1}\right]_{33}^2 \\
= 3 \left[e_3^T (\lambda_R X^T X + B)^{-1} e_3\right]^2 \\
\leq 3 \left[e_3^T B^{-1} e_3\right]^2 \\
= 3B_{33}^2
\]

where the inequality follows from Lemma A.1. It follows that the fourth (non-central) moment \( \mathbb{E}[\beta_2^4 | \lambda_R, y] \) is finite. \( \square \)

References


