On the Geometric Ergodicity of Two-Block Gibbs Samplers

Kjell Nygren

Abstract

We demonstrate that two-block Gibbs samplers for two broad classes of Bayesian models are Geometrically Ergodic. In both cases, our proofs involve demonstrating drift and minorization conditions as in Rosenthal (1995). A key property used in the proof of the Geometric Ergodicity of both models is a variance inequality showing that for models with a multivariate normal prior and a log-concave likelihood function, the difference between the prior and posterior variance-covariance matrices is positive semi-definite.

The first class of models considered have multivariate normal priors and allows one of the two blocks to have data in the form of a log-concave likelihood function. In addition to the above, our proof of the drift condition for this class of models relies on the fact that one of the two conditional densities is a multivariate normal prior as well as the fact that the data (in a general sense) decreases the dependence between the two blocks. As a result, the geometric drift is no slower than the drift to the target density if the two-block Gibbs sampler was applied to the prior density.

The second class of models considered is a general class of Bayesian Hierarchical random effects models with log-concave likelihood functions. Under two key assumptions we show that a general class of two-block Gibbs samplers are Geometrically Ergodic for

such models. The assumptions require that (a) the prior for the variance-covariance matrix is informative, and (b) that the random effects parameters are identified by the likelihood function itself. Our proof of geometric drift involves showing that these assumptions ensures that the conditional expectation for one of the two blocks is a bounded function of the values on the prior iteration. Our proof of the minorization condition, on the other hand, utilizes a Weyl (1912) type eigenvalue inequality for a class of non-hermitian matrices.

Aside from our theoretical results, the paper also includes an application to overdispersed generalized linear models and several examples illustrating the behavior of samplers satisfying and violating our main assumptions. The latter shows that relaxing our assumptions in the second class of models further may be difficult, and that sampling efficiently from models violating our assumptions in general may present serious challenges.

1 Introduction

Sampling from many Bayesian models present serious challenges as posterior densities usually are complex, ruling out direct sampling procedures. A popular approach involves the use of Gibbs or (more generally) Block-Gibbs sampling procedures. The idea behind Block-Gibbs sampling is to replace sampling from the joint density with iterative sampling from conditional densities with known sampling procedures. Under relatively mild assumptions, this yields a Markov chain whose limiting density is the desired joint density. While general convergence of Gibbs samplers are well known, the rate of convergence for the complex models

studies here has not been extensively studied. Work due to Rosenthal (1995) and Meyn and Tweedie (1993) has shown that Markov chains satisfying certain drift and minorization conditions are Geometrically Ergodic. Convergence of such Markov chains is geometrically fast and stable. Geometric Ergodicity has been studied by Roberts and Tweedie (1996) and Mengersen and Tweedie (1996) who study specialized (independent and symmetric) candidates and by Hobert and Geyer (1998) and Jones and Hobert (2004) who study it for a Bayesian Hierarchical Random Effects models with proper conjugate priors.

In this paper, we demonstrate that two-block Gibbs sampling procedures for two important classes of models generally satisfy drift and minorization conditons as in Rosenthal (1995), and hence are Geometrically Ergodic. Our paper starts with some Markov chain background in section 2. This includes a basic discussion of Geometric Ergodicity, as well as Rosenthal's drift and minorization conditions with corresponding theorem. In section 3, we first consider two-block Gibbs sampling from Multivariate Normal models and demonstrate that appropriate drift and minorization conditions are satisfied. We then extend this to a more general class of models where one of the two blocks has data from a log-concave density. In section 4, we consider sampling from a a general class of Bayesian Hierarchical random effects models with log-concave likelihood functions. We consider a two-Block Gibbs sampler in which the variance parameters are updated in the first block and the remaining parameters in the second. We show that such a two-Block Gibbs sampler is Geometrically Ergodic if the model satisfies appropriate assumptions essentially requiring that (a) the prior for the variance-covariance matrix is informative, and (b) that the random effects parameters are identified by the likelihood function itself.

Section 5 contains example models with normal data for which many of the constants in our drift and minorization conditions can be provided in closed form solution. In section 6, we include simulation results from models satisfying our assumptions as well as from a model violating one of the key assumptions for Theorem 4.1. We also include several Appendix sections covering details of our proofs. Section 8 contains details related to the Variance inequality in Theorem 3.1. In section 9, we cover details related to the lemmas in section 3. Section 10 covers results needed to establish Geometric drift for our second class of models, while section 11 relates to results needed to establish the minorization condition.

2 Markov Chain Background

Let $\mathcal{X} \subset \mathbf{R}^m$ for $m \geq 1$ and let $\mathcal{B}$ denote the associated σ-algebra. Suppose that $X = \{X_i, i = 0, 1, \ldots\}$ is a discrete time, time homogeneous Markov chain with state space $\mathcal{X}$ and Markov transition kernel Q; that is for $x \in \mathcal{X}$ and $A \in \mathcal{B}$, $Q(x, A) = Pr(X_{i+1} \in A | X_i = x)$. Also, for $n = 1, 2, 3, \ldots$, let Q^n denote the n-step transition kernel, that is, $Q^n(x, A) = Pr(X_{i+n} \in A | X_i = x)$ so, in particular, $Q = Q^1$. Note that $Q^n(x, .)$ is the probability measure of the random variable X_n conditional on starting the chain at $X_0 = x$.

Let v be a measure on $\mathcal{B}$. We will say that the Markov Chain X satisfies assumption $(\mathcal{A})$ if it is v-irreducible, aperiodic, and positive Harris recurrent with invariant probability measure $\pi(|\mathbf{y})$. It is straightforward to verify that the block Gibbs Samplers considered in this paper satisfies assumption $(\mathcal{A})$ with v equal to the Lebesgue measure. Under assumption $(\mathcal{A})$, for every $x \in \mathcal{X}$ we have

$$||Q^n(x,) - \pi(.|\mathbf{y})|| \downarrow 0 \quad as \quad n \to \infty$$

where $||Q^n(x,.) - \pi(.|\mathbf{y})|| := \sup_{A \in \mathcal{B}} |Q^n(x,A) - \pi(A|y)|$ is the total variation distance between Q^n and π.

Remark 2.1. Let $A_x^n = \{x' \in \mathbf{R}^m | Q^n(x,x') \geq \pi(x'|\mathbf{y})\}$. Under mild assumptions, it is straightforward to verify that

$$\begin{aligned} ||Q^n(x,) - \pi(.|\mathbf{y})|| &= |Q^n(x,A_x^n) - \pi(A_x^n|y)| \\ &= \int_{x' \in A_x^n} Q^n(x,x')dx' - \int_{x' \in A_x^n} \pi(x'|\mathbf{y})dx' \\ &= \int_{x' \in (A_x^n)^c} \pi(x'|\mathbf{y})dx' - \int_{x' \in (A_x^n)^c} Q^n(x,x')dx'. \end{aligned}$$

The chain X is called *geometrically ergodic* if it satisfies assumption $(\mathcal{A})$ and, in addition, there exists a constant $0 < t < 1$ and a function $g : \mathcal{X} \mapsto [0,\infty)$ such that, for any $x \in \mathcal{X}$,

$$||Q^n(x,.) - \pi(.|\mathbf{y})|| \leq g(x)t^n$$

for $n = 1, 2, \ldots$. It has been shown that establishing drift and minoritization conditions for X verifies geometric ergodicity (the existence of g and t) and yields an upper bound on the right hand side above. We will make use of a result due to Rosenthal (1995). A similar condition is provided by Roberts and Tweedie (1999). A slightly simplified result follows.

Theorem 2.1. *Let X be a Markov chain satisfying assumption $(\mathcal{A})$. Suppose X satisfies the following drift condition. For some function $V : \mathcal{X} \mapsto [0,\infty)$, for some $0 < \gamma < 1$ and some $b < \infty$,*

$$E[V(X_{i+1})|X_i = x] \leq \gamma V(x) + b \qquad \forall x \in \mathcal{X} \ \cdot$$

Let $C = \{x \in \mathcal{X} : V(x) \leq d_R\}$, *where* $d_R > 2b/(1-\gamma)$ *and suppose that* X *satisfies the following minorization condition. For some probability measure* S *on* $\mathcal{B}$ *and some* $\epsilon > 0$,

$$Q(x, .) \geq \epsilon S(.) \qquad \forall x \in C.$$

Let $X_0 = x_0$ *and define two constants as follows:*

$$\alpha = (1 + d_R)/(1 + 2b + \gamma d_R) \quad \textit{and} \quad U = 1 + 2(\gamma d_R + b)$$

Then for any $0 < r < 1$,

$$||Q^n(x_0, .) - \pi(.|\mathbf{y})|| \leq (1-\epsilon)^{rn} + (U^r/\alpha^{1-r})^n(1 + b/(1-\gamma) + V(x_0)).$$

3 Multivariate Normal models with extensions

3.1 A Multivariate Normal Model

In this section, we consider a model in which (μ_1, μ_2) are from a Joint Multivariate-Normal density with Variance-Covariance matrix Σ (Precision matrix $P := \Sigma^{-1}$). We consider the Geometric Ergodicity of a two-block Gibbs sampler implemented for this density. The two block Gibbs sampler will consist of an update for μ_1 from its full conditional density, followed by an update of μ_2 from its full conditional density.

3.1.1 Geometric Drift

Our proofs of Geometric drift throughout the paper will make use of the following key Variance inequality.

Theorem 3.1. *Suppose β has a multivariate normal prior with variance-covariance matrix Σ and a log-concave likelihood function. Denote by $Var(\beta|y)$ the posterior variance-covariance matrix. Then the matrix*

$$\Sigma - Var(\beta|y)$$

is positive semidefinite.

Proof. Consider any vector $x \in \mathbf{R}^m$. Then

$$\begin{aligned} x^T(\Sigma - Var(\beta|\mathbf{y}))x &= x^T\Sigma x - x^T Var(\beta|\mathbf{y})x \\ &= Var(x^T\beta) - Var(x^T\beta|\mathbf{y}) \\ &\geq 0 \end{aligned}$$

where the equality follows from fact 8.1, and the inequality from remark 8.3. □

Now, denote by μ_1^* and μ_2^* the unconditional means for the two densities above. Then it is straightforward to verify that the two conditional means are given by

$$E[\mu_1|\mu_2] = \mu_1^* - P_{11}^{-1}P_{12}(\mu_2 - \mu_2^*)$$

and

$$E[\mu_2|\mu_1] = \mu_2^* - P_{22}^{-1}P_{21}(\mu_1 - \mu_1^*)$$

respectively.

Using the above conditional means, it is straightforward to verify that under the two block Gibbs sampler, the expected value of $\mu_2^{(k)}$ (the kth-iteration) for μ_2 conditional upon $\mu_2^{(k-1)}$ is given by

$$E[\mu_2^{(k)}|\mu_2^{(k-1)}] = \mu_2^* + P_{22}^{-1}P_{21}P_{11}^{-1}P_{12}(\mu_2^{(k-1)} - \mu_2^*).$$

For the below, we will make use of the matrix $T := P_{21}P_{11}^{-1}P_{12}P_{22}^{-1}P_{22}^{-1}P_{21}P_{11}^{-1}P_{12}$. The following two Lemmas are critical in order to establish Geometric drift. In the proof of our first Lemma, the ordered eigenvalues of an arbitrary Hermitian matrix A in decreasing order are denoted by $\lambda(A)_1, \lambda(A)_2, \ldots, \lambda(A)_m$.

Lemma 3.1. *The matrix T is a convergent normal matrix for which all eigenvalues are non-negative and less than 1.*

Proof. It is straightforward to verify that T is a normal Hermitian positive semi-definite matrix. Using properties of partitioned matrices (see e.g., Green ()), it is also straightforward to verify that

$$\begin{aligned} T &= \Sigma_{22}^{-1}\Sigma_{21}(\Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21})^{-1}\Sigma_{12}\Sigma_{22}^{-1}[\Sigma_{22}^{-1}\Sigma_{21}(\Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21})^{-1}\Sigma_{12}\Sigma_{22}^{-1} + \Sigma_{22}^{-1}]^{-1} \\ &\quad [\Sigma_{22}^{-1}\Sigma_{21}(\Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21})^{-1}\Sigma_{12}\Sigma_{22}^{-1} + \Sigma_{22}^{-1}]^{-1}\Sigma_{22}^{-1}\Sigma_{21}(\Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21})^{-1}\Sigma_{12}\Sigma_{22}^{-1}. \end{aligned}$$

Let $A = \Sigma_{22}^{-1}\Sigma_{21}(\Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21})^{-1}\Sigma_{12}\Sigma_{22}^{-1}$ and $B = \Sigma_{22}^{-1}$. Then A and B are both positive definite matrices and

$$T = A(A+B)^{-1}(A+B)^{-1}A.$$

To show that T is convergent it suffices to show that all eigenvalues of T are less than 1. We prove this by contradiction. Suppose there exist an eigenvalue $\gamma^* \geq 1$. Then by fact 9.2, we know that $|AA - \gamma^*(A+B)(A+B)| = 0$. We now have that

$$\begin{aligned}
\lambda(AA + (-\gamma^*(A+B)(A+B)))_i &= \lambda(-(\gamma^*-1)A*A - AB - BA - BB)_i \\
&\leq -(\gamma^*-1)\lambda(A*A)_m - \lambda(AB)_m - \lambda(BA)_m \\
&\quad -\lambda(BB)_m \\
&= -(\gamma^*-1)\lambda(A*A)_m - \lambda(A^{1/2}BA^{1/2})_m \\
&\quad -\lambda(A^{1/2}BA^{1/2})_m - \lambda(BB)_m \\
&< 0
\end{aligned}$$

where the second to last inequality follows from Weyl's (1912) inequality, the second equality from a property noted in Merikoski and Kumar (2004), and the last inequality from the fact that each matrix is positive semi-definite and BB positive definite. □

Lemma 3.2. *Let y be an arbitrary vector, then $y^TTy \leq \lambda^* y^T y$, where $0 \leq \lambda^* < 1$ equals the maximal eigenvalue for T.*

Proof. Since T is normal, there exists a unitary matrix U and a diagonal matrix Λ (with the eigenvalues for T on the diagonal) such that $T = U\Lambda U^T$. Now, define $z = U^T y$ and note that

$$\begin{aligned} y^T T y &= z^T \Lambda z \\ &= \textstyle\sum_{i=1}^m \lambda_i z_i^2 \\ &\leq \lambda^* \textstyle\sum_{i=1}^m z_i^2 \\ &= \lambda^* y^T U U^T y \\ &= \lambda^* y^T y. \end{aligned}$$

□

We are no ready to establish a Geometric drift condition for our model.

Theorem 3.2. *Let $V : \mathbf{R}^m \to \mathbf{R}_+$ be the function defined by $V(\mu_2) = (\mu_2 - \mu_2^*)^T(\mu_2 - \mu_2^*)$, and denote by var_i, $i = 1, 2, \ldots, m$ the conditional variances associated with the individual components of the vector $\mu_2^{(k)}$ given $\mu_2^{(k-1)}$. Then the function V has the property that*

$$E[V(\mu_2^{(k)})|\mu_2^{(k-1)}] \leq (\sum_{i=1}^{m} var_i) + \lambda^*(\mu_2^{(k-1)} - \mu_2^*)^T(\mu_2^{(k-1)} - \mu_2^*)$$

where $0 \leq \lambda^* < 1$ is the maximal eigenvalue for the matrix T above.

Proof. We simply note that

$$
\begin{aligned}
E[V(\mu_2^{(k)})|\mu_2^{(k-1)}] &= E[(\mu_2^{(k)} - \mu_2^*)^T(\mu_2^{(k)} - \mu_2^*)|\mu_2^{(k-1)}] \\
&= E[(\mu_2^{(k)} - E[\mu_2^{(k)}|\mu_2^{k-1}] + E[\mu_2^{(k)}|\mu_2^{k-1}] - \mu_2^*)^T \\
&\quad (\mu_2^{(k)} - E[\mu_2^{(k)}|\mu_2^{k-1}] + E[\mu_2^{(k)}|\mu_2^{k-1}] - \mu_2^*)|\mu_2^{(k-1)}] \\
&= E[(\mu_2^{(k)} - E[\mu_2^{(k)}|\mu_2^{k-1}])^T(\mu_2^{(k)} - E[\mu_2^{(k)}|\mu_2^{k-1}]|\mu_2^{(k-1)}] \\
&\quad + E[(E[\mu_2^{(k)}|\mu_2^{k-1}] - \mu_2^*)^T(E[\mu_2^{(k)}|\mu_2^{k-1}] - \mu_2^*)|\mu_2^{(k-1)}] \\
&= E[(\mu_2^{(k)} - E[\mu_2^{(k)}|\mu_2^{k-1}])^T(\mu_2^{(k)} - E[\mu_2^{(k)}|\mu_2^{k-1}]|\mu_2^{(k-1)}] \\
&\quad + (\mu_2^{(k-1)} - \mu_2^*)^T T(\mu_2^{(k-1)} - \mu_2^*) \\
&\leq (\textstyle\sum_{i=1}^m var_i) + \lambda^*(\mu_2^{(k-1)} - \mu_2^*)^T(\mu_2^{(k-1)} - \mu_2^*)
\end{aligned}
$$

where the last inequality follows from Theorem 3.1 in conjunction with Lemmas 3.1 and 3.2. □

Remark 3.1. By rewriting the conditional density for $\mu_2^{(k)}$ given $\mu_2^{(k-1)}$ as a posterior density resulting from the original prior for μ_2 and a log-concave density, it can shown (using Theorem 3.1) that the conditional variances in the above are less than or equal to the prior variances.

Remark 3.2. In the Theorem above, the drift function V can be replaced by any function $\tilde{V}$ of the form $\tilde{V}(\mu_2^{(k)}) = (\mu_2^{(k)} - \mu_2^*)^T\tilde{P}(\mu_2^{(k)} - \mu_2^*)$, where $\tilde{P}$ is a positive definite matrix. Moreover, there exists a constant λ^* such that $\forall \mu_2^{(k-1)}$, $(E[\mu_2^{(k)}|\mu_2^{(k-1)}] - \mu_2^*)^T\tilde{P}(E[\mu_2^{(k)}|\mu_2^{(k-1)}] - \mu_2^*) \leq \lambda^*(\mu_2^{(k-1)} - \mu_2^*)^T\tilde{P}(\mu_2^{(k-1)} - \mu_2^*)$.

Proof. We note that there exists a real valued square root matrix A such that $A^TA = \tilde{P}$. Define $z = A\mu_2$ and $z^* = A\mu_2^*$. Then for all μ_2, $V(\mu_2) = (\mu_2 - \mu_2^*)^T\tilde{P}(\mu_2 - \mu_2^*) = (z - z^*)^T(z - z^*)$. Applying the method of proof used in Theorem 3.2 to z and converting the final results back to μ_2 now yields the desired result. □

3.1.2 Minorization

To construct the density $S(.)$ and to find the constant ϵ needed for the minorization condition, we first note that for every $\mu_2^{(k-1)}$, $Q(\mu_2^{(k-1)}, .)$ is a multivariate normal density. Moreover, the variance-covariance matrix (and hence the precision P^*) is the same for every $\mu_2^{(k-1)}$.

We now establish the following:

Theorem 3.3. *Let $\tilde{P}$ be a positive definite matrix and let $S(.)$ be the multivariate normal density with precision $P^* + \tilde{P}$ and mean vector μ_2^*. Consider any $\gamma > 0$ and define $C_\gamma := \{\mu_2 | V(\mu_2) \leq \gamma\}$. Then there exists a positive number ϵ_γ such that $\forall \mu_2^{(k-1)} \in C_\gamma$:*

$$Q(\mu_2^{(k-1)}, .) \geq \epsilon_\gamma S(.)$$

Proof. We first note that for every $\mu_2^{(k-1)}$, the function $log[Q(\mu_2^{(k-1)}, .)] - log[S(.)]$ is a strictly convex and quadratic function. Hence it obtains a unique minimum for each $\mu_2^{(k-1)}$. We can hence define a function $\epsilon : C_\gamma \to \mathbf{R}$, where $\epsilon(\mu_2^{(k-1)})$ denotes the exponentiated value of the minimum of the function $log[Q(\mu_2^{(k-1)}, .)] - log[S(.)]$. We note that this is continuous function on a convex compact set which in turn obtains its minimum on C_γ. Denote this minimum by ϵ_γ. Then for every $\mu_2^{(k-1)} \in C_\gamma$ and $\mu_2^{(k)}$, we have

$$(Q(\mu_2^{(k-1)}, \mu_2^{(k)})/S(\mu_2^{(k)})) \geq \epsilon(\mu_2^{(k-1)}) \geq \epsilon_\gamma.$$

Rearrangement of terms completes the proof. □

Lemma 3.3. *Let P^* be the precision matrix for μ_2 under $Q(\mu_2^{(k-1)}, .)$, a constant function of $\mu_2^{(k-1)}$, let $\tilde{P}$ be a positive definite matrix, and define the drift function in the previous section by $V(\mu) = (\mu_2 - \mu_2^*)^T (P^* \tilde{P}^{-1} P^* + P^*)(\mu_2 - \mu_2^*)$. Let $S(.)$ be the multivariate normal*

density with mean μ_2^* and precision matrix $P^* + \tilde{P}$. Then for every γ, every $\mu^{(k-1)} \in C_\gamma$, and every $\mu^{(k)}$

$$Q(\mu^{(k-1)}, \mu^{(k)})/S(\mu^{(k)}) \geq (|P^*|/|P^* + \tilde{P}|)^{0.5} exp(-0.5\lambda^*\gamma)$$

Proof.

$$\begin{aligned}
Q(\mu^{(k-1)}, \mu^{(k)})/S(\mu^{(k)}) &= (|P^*|/|P^* + \tilde{P}|)^{0.5} \\
&\quad exp(-0.5((\mu^{(k)} - \mu^*) - E[\mu^{(k)} - \mu^*|\mu^{(k-1)}])^T P^* \\
&\quad ((\mu^{(k)} - \mu^*) - E[\mu^{(k)} - \mu^*|\mu^{(k-1)}])) \\
&\quad exp(0.5(\mu^{(k)} - \mu^*)^T (P^* + \tilde{P})(\mu^{(k)} - \mu^*)) \\
&= (|P^*|/|P^* + \tilde{P}|)^{0.5} \\
&\quad exp(0.5[((\mu^{(k)} - \mu^*) - (-(\tilde{P})^{-1}(P^* E[\mu^{(k)} - \mu^*|\mu^{(k-1)}])))^T \tilde{P} \\
&\quad ((\mu^{(k)} - \mu^*) - (-(\tilde{P})^{-1}(P^* E[\mu^{(k)} - \mu^*|\mu^{(k-1)}])))]) \\
&\quad exp(-0.5[((\tilde{P})^{-1}(P^* E[\mu^{(k)} - \mu^*|\mu^{(k-1)}]))^T \tilde{P} \\
&\quad ((\tilde{P})^{-1}(P^* E[\mu^{(k)} - \mu^*|\mu^{(k-1)}]))]) \\
&\quad exp(-0.5E[\mu^{(k)} - \mu^*|\mu^{(k-1)}]^T P^* E[\mu^{(k)} - \mu^*|\mu^{(k-1)}]) \\
&\geq (|P^*|/|P^* + \tilde{P}|)^{0.5} \\
&\quad exp(-0.5(E[\mu^{(k)} - \mu^*|\mu^{(k-1)}])^T \\
&\quad (P^* \tilde{P}^{-1} P^* + P^*)(E[\mu^{(k)} - \mu^*|\mu^{(k-1)}])) \\
&\geq (|P^*|/|P^* + \tilde{P}|)^{0.5} \\
&\quad exp(-0.5\lambda^* V(\mu^{(k-1)})) \\
&\geq (|P^*|/|P^* + \tilde{P}|)^{0.5} \\
&\quad exp(-0.5\lambda^*\gamma).
\end{aligned}$$

□

Remark 3.3. In the above theorem, $P^* = P_{22} - P_{21}[P_{11} + P_{12}P_{22}^{-1}P_{21}]^{-1}P_{12}$.

3.2 Models with Log-Concave Data

We now consider the following Bayesian model.

$$\begin{aligned} \mu &\sim MNorm(\nu, P_0^{-1}) \\ \beta &\sim MNorm(X\mu, P^{-1}), \end{aligned}$$

(where P is a diagonal matrix) and a likelihood function

$$f(y|\beta).$$

We will consider the two-block Gibbs-sampler in which β is updated in first block and μ in the second. The assumption of P diagonal is in a sense without loss of generality as a simple transformation always can be applied to ensure this property.

3.2.1 Geometric Drift

Remark 3.4. The conditional density for μ given β and y is a multivariate normal density with mean vector and Variance-Covariance matrix given by

$$E[\mu|\beta, y] = (P_0 + X^T P X)^{-1}(P_0\nu + X^T P\beta)$$

and

$$E[(\mu - E[\mu|\beta, y])^T(\mu - E[\mu|\beta, y])|\beta, y] = (P_0 + X^T P X)^{-1}$$

respectively.

Remark 3.5. The conditional expectation for β given μ and y depends on μ only through $X\mu$. Hence we will write the conditional mean vector and variance-covariance matrices for β by $E[\beta|X\mu, y]$ and $Var[\beta|X\mu, y]$.

In the below, we will make use of the prior precision matrix P^* defined by

$$\begin{aligned} P^*_{11} &= P \\ P^*_{12} &= PX \\ P^*_{21} &= X^T P \\ P^*_{22} &= P_0 + X^T P X. \end{aligned}$$

Theorem 3.4. *Denote by var_i, $i = 1, 2, \ldots, m$ the prior variances associated with the individual components of the vector μ. If the likelihood function is log-concave, then there exists a fixed point μ^* such that*

$$\mu^* = (P_0 + X^T P X)^{-1}(P_0\nu + X^T P E[\beta|X\mu^*, y])$$

and a function $V : \mathbf{R}^m \to \mathbf{R}_+$, defined by $V(\mu) = (\mu - \mu^*)^T(\mu - \mu^*)$, such that the function V has the property that

$$E[V(\mu^{(k)})|\mu^{(k-1)}] \leq (\sum_{i=1}^{m} var_i) + \lambda^* V(\mu^{(k-1)})$$

where $0 \leq \lambda^* < 1$ is the maximal eigenvalue of the matrix

$$T^* := P^*_{21}(P^*_{11})^{-1}P^*_{12}(P^*_{22})^{-1}(P^*_{22})^{-1}P^*_{21}(P^*_{11})^{-1}P^*_{12}.$$

Proof. Pick any μ^{**} as in Lemma 9.2. Define a function $g : \mathbf{R}^m \to \mathbf{R}^m$ by

$$g(\mu) = (P_0 + X^T P(P^{-1} - V_{\mu^{**}}(\mu))PX)^{-1}(P_0\nu + X^T PE[\beta|X_i\mu^{**}, y] - X^T PV_{\mu^{**}}(\mu)PX\mu^{**}).$$

Theorem 3.1 states that $P^{-1} - V_{\mu^{**}}(\mu)$ is a positive semi-definite matrix. Hence it is straightforward to verify that $P_0 + X^T P(P^{-1} - V_{\mu^{**}}(\mu))PX$ is a positive definite matrix and that $g(.)$ is a continuous function. Since Theorem 3.1 also implies that $V_{\mu^{**}}(.)$ is a bounded function, it also follows that $g(.)$ is a bounded function. Denote by D the closure of the convex hull of the image $g(\mathbf{R}^m)$. It is straightforward to verify that D is a convex compact set. Define a continuous function $\tilde{g} : D \to D$ by $\tilde{g}(\mu) = g(\mu)$. By a simple application of Brouwer's [] fixed point theorem, this function $\tilde{g}$ has a fixed point μ^* such that $\mu^* = \tilde{g}(\mu^*)$. This fixed point is easily verified to be the fixed point required by our theorem.

Consider any $\mu^{(k-1)} \in \mathbf{R}^m$. It then follows that

$$\begin{aligned}
E[V(\mu^{(k)})|\mu^{(k-1)}] &= \textstyle\sum_{i=1}^m E[((\mu_i^{(k)} - E[\mu_i^{(k)}|\mu^{(k-1)}])^2 + (E[\mu_i^{(k)}|\mu^{(k-1)}] - \mu_i^*))^2|\mu^{(k-1)}] \\
&= \textstyle\sum_{i=1}^m [E[(\mu_i^{(k)} - E[\mu_i^{(k)}|\mu^{(k-1)}])^2|\mu^{(k-1)}] \\
&\quad + E[(E[\mu_i^{(k)}|\mu^{(k-1)}] - \mu_i^*)^2|\mu^{(k-1)}]] \\
&\leq [\textstyle\sum_{i=1}^m var_i] + \textstyle\sum_{i=1}^m [E[(E[\mu_i^{(k)}|\mu^{(k-1)}] - \mu_i^*)^2|\mu^{(k-1)}]] \\
&= [\textstyle\sum_{i=1}^m var_i] + [(\mu^{(k-1)} - \mu^*)^T T(V_{\mu^{**}}(\mu^{(k-1)}))(\mu^{(k-1)} - \mu^*)] \\
&\leq [\textstyle\sum_{i=1}^m var_i] + \lambda^* V(\mu^{(k-1)})
\end{aligned}$$

where the last inequality follows from Claims 9.1 and 9.2. □

Remark 3.6. As in the previous subsection, the drift function in Theorem 3.4 can be replaced by a function $\tilde{V}$ defined by $V(\mu) = (\mu - \mu^*)^T \tilde{P}(\mu - \mu^*)$, where $\tilde{P}$ is any positive definite

matrix. Moreover, there exists a constant λ^* $(0 \leq \lambda^* < 1)$ such that

$$(E[\mu^{(k)}|\mu^{(k-1)}] - \mu^*)^T \tilde{P}(E[\mu^{(k)}|\mu^{(k-1)}] - \mu^*) \leq \lambda^*(\mu^{(k-1)} - \mu^*)^T \tilde{P}(\mu^{(k-1)} - \mu^*).$$

3.2.2 Minorization

To construct the density $S(.)$ and to find the constant ϵ needed for the minorization condition, we first note a couple of properties of the densities.

Fact 3.1. Let $\tilde{P}$ be a positive definite matrix, and let $S(.)$ be a multivariate normal density with mean vector μ^* and precision matrix $P_{22}^* + \tilde{P}$. Then for every $\mu^{(k-1)}$ and $\beta^{(k)}$, the function $h(.|\mu^{(k-1)}, \beta^{(k)}) : \mathbf{R}^m \to \mathbf{R}$ defined by $h(\mu^{(k)})|\mu^{(k-1)}, \beta^{(k)}) = \pi(\beta^{(k)}|\mu^{(k-1)})\pi(\mu^{(k)}|\beta^{(k)})/S(\mu^{(k)})$ is a strictly log-convex function for which the difference between the Hessian for the log of the function and $\tilde{P}$ is positive semi-definite.

Fact 3.2. Let $\tilde{P}$ be a positive definite matrix, and let $S(.)$ be a multivariate normal density with any mean vector μ^{**} and precision matrix $P_{22}^* + \tilde{P}$. Then for every $\mu^{(k-1)}$ the function $g(.|\mu^{(k-1)}) : \mathbf{R}^m \to \mathbf{R}$ defined by $g(\mu^{(k)})|\mu^{(k-1)}) = \int \pi(\beta^{(k)}|\mu^{(k-1)})\pi(\mu^{(k)}|\beta^{(k)})/S(\mu^{(k)})d\beta^{(k)}$ is a strictly log-convex function for which the difference between the Hessian for the log of the function and $\tilde{P}$ is positive semi-definite.

Proof. From properties of log-convex functions. See e.g., An (1996). □

Denote by $V(.)$ the drift function used in the previous section. We now establish the following:

Theorem 3.5. *Let $\tilde{P}$ be a positive definite matrix and let $S(.)$ be the multivariate normal density with precision $P_{22}^* + \tilde{P}$ and mean vector μ^{**}. Consider any $\gamma > 0$ and define $C_\gamma :=$*

$\{\mu_2 | V(\mu_2) \leq \gamma\}$. *Then there exists a positive number* ϵ_γ *such that* $\forall \mu_2^{(k-1)} \in C_\gamma$*:*

$$Q(\mu^{(k-1)}, .) \geq \epsilon_\gamma S(.)$$

Proof. We first note that for every $\mu^{(k-1)}$, it follows from fact 3.2 that the function $log[Q(\mu^{(k-1)}, .)] - log[S(.)]$ is a strictly convex function for which the difference between the Hessian for the function and $\tilde{P}$ is positive semi-definite. It can be shown that this implies that the function obtains a unique minimum for each $\mu^{(k-1)}$. We can hence define a function $\epsilon : C_\gamma \rightarrow \mathbf{R}$, where $\epsilon(\mu^{(k-1)})$ denotes the exponentiated value of the minimum of the function $log[Q(\mu^{(k-1)}, .)] - log[S(.)]$. We note that this is a continuous function on a convex compact set which in turn obtains its minimum on C_γ. Denote this minimum by ϵ_γ. Then for every $\mu^{(k-1)} \in C_\gamma$ and $\mu^{(k)}$, we have

$$(Q(\mu_2^{(k-1)}, \mu^{(k)})/S(\mu^{(k)})) \geq \epsilon(\mu^{(k-1)}) \geq \epsilon_\gamma.$$

Rearrangement of terms completes the proof. □

Remark 3.7. Unlike the previous section, we can not derive an explicit bound for ϵ_γ. If the posterior density is approximately multivariate normal, however, we can utilize a drift function defined by $V(\mu) = (\mu - \mu^*)^T (H^*((P_{22}^* - H^*) + \tilde{P})^{-1} H^* + H^*)(\mu - \mu^*)$, where H^* is an approximation for the precision of μ under the density $Q(\mu^{(k-1)}, .)$. Using Theorem 3.3 in the previous subsection, it follows that ϵ_γ should be close to the below

$$\epsilon_\gamma \approx (|H^*|/|H^* + (P_{22}^* - H^*) + \tilde{P}|)^{0.5} exp(-0.5\lambda^* \gamma)$$

The approximate precision H^* could be based on a quadratic approximation to the log-posterior density at a maximum a posteriori estimate. We note that it is straightforward to verify that $(P_{22}^* - H^*)$ is a positive semi-definite matrix.

Remark 3.8. Let $P_D(\beta^*)$ denote the hessian for the negative of the log-likelihood function at β^*. We then have

$$\begin{aligned} H^* &= P_{22}^* - P_{21}^*(P + P_D(\beta^*))^{-1}P_{12}^* \\ &= P_0 + X^T PX - X^T P(P + P_D(\beta^*))^{-1}PX \end{aligned}$$

and

$$\epsilon_\gamma \approx (|P_0 + X^T PX - X^T P(P + P_D(\beta^*))^{-1}PX|/|P_0 + X^T PX + \tilde{P}|)^{0.5} exp(-0.5\lambda^*\gamma).$$

It is interesting here to note the dependence on $P_D(\beta^*)$. As $P_D(\beta^*)$ gets small in magnitude, we have

$$\epsilon_\gamma \approx (|P_0|/|P_0 + X^T PX + \tilde{P}|)^{0.5} exp(-0.5\lambda^*\gamma)$$

while as the magnitude of $P_D(\beta^*)$ gets large, we approach

$$\epsilon_\gamma \approx (|P_0 + X^T PX|/|P_0 + X^T PX + \tilde{P}|)^{0.5} exp(-0.5\lambda^*\gamma).$$

4 Bayesian Hierarchical Random Effects Models with Log-concave Likelihood Functions

We will consider the following general class of models (where β_i is p-dimensional).

(i) A likelihood function

$$\pi(\mathbf{y}|\beta, \mu)$$

(ii) A prior specification, where

$$\mu \sim Multivariate - Normal(\mathbf{0}, \Sigma_\mu)$$

$$\beta_i \sim Multivariate - Normal(\mathbf{0}, \Sigma_\beta), i = 1, \ldots, m_1.$$

$$\mathbf{\Sigma}_\beta \sim Inverse - Wishart(m_0 \mathbf{Var_0}, m_0).$$

We will study properties of a two-block Gibbs sampler where $\mathbf{\Sigma}_\beta$ is updated in the first block, while β and μ are updated in the second. We note that the conditional density for β and μ can be updated using the procedure in Nygren and Nygren (2006), while the conditional density for $\mathbf{\Sigma}_\beta$ is an Inverse-Wishart density.

In establishing Geometric Ergodicity, we will utilize some of the properties noted in Roberts (1995) as well as in Cowles and Rosenthal (1998) for more general sequentially

updated Gibbs Samplers.

4.1 Geometric Drift

In order to establish geometric drift, we will make use of the following Lemma.

Lemma 4.1. *Using the model above, define a function* $g(.)$ *by* $g(\beta,\mu) = Log(\pi(y|\beta,\mu)) - 0.5\mu^T\Sigma_\mu^{-1}\mu$ *and assume that it satisfies the following assumptions*

(i) The function $g(.)$ *is a strictly concave function.*

(ii) $\forall t \in \mathbf{R}$*, the set* $\{(\beta,\mu) \in \mathbf{R}^{L+m_1*p} | g(\beta,\mu) \geq t\}$ *is compact.*

(iii) Log $g(.)$ *is twice continuously differentiable.*

Then $\exists b \in \mathbf{R}_+ : \forall(\beta^{(k-1)}, \mu^{(k-1)})$:

$$\frac{1}{m}\sum_{j=1}^{p}\sum_{i=1}^{m} E[E[\beta_{i,j}^{(k)}|\Sigma_\beta^{(k)}]^2|(\beta^{(k-1)},\mu^{(k-1)})] \leq b.$$

Proof. See section 10. □

Conditions (i)-(ii) together implies that the function $g(.)$ obtains its maximum. Condition (iii) is a technical regularity condition which is likely to hold in most applications. We suspect that it may be possible to relax this condition, but do not pursue the matter here as it would substantially complicate our proof.

Remark 4.1. Conditions (i)-(ii) above imply the existence of a pair (β^*,μ^*) and a vector $c(\beta^*,\mu^*)$ such that $\pi(\mathbf{y}|\beta,\mu)) \leq \pi(\mathbf{y}|\beta^*,\mu^*))exp(-c(\beta^*,\mu^*)\mu)$ for each pair (β,μ). This in turn implies that $\pi(\mathbf{y}|P) \leq \pi(\mathbf{y}|\beta^*,\mu^*)MGF(-c(\beta^*,\mu^*))$, where $MGF(-c(\beta^*,\mu^*)) = \int exp(-c(\beta^*,\mu^*)\mu)\pi(\mu)d\mu$.

Cowles and Rosenthal (1998) notes that the function V only needs to depend on those parameters that are "remembered" at the next iteration. Given our update order, the only parameters remembered can be shown to be the β's. This is so since we are working with a two-block Gibbs-sampler and the update for P depends only on the β's and not on μ.

Define a function $V : \mathbf{R}^{m*p} \to [0, \infty)$ by

$$V(\beta^{(k)}) = \frac{1}{m_1} \sum_{j=1}^{p} \sum_{i=1}^{m_1} (\beta_{i,j}^{(k)})^2.$$

Remark 4.2. $E[V(\beta^{(k)})|\beta^{(k-1)}, \mu^{(k-1)}] = E[V(\beta^{(k)})|\beta^{(k-1)}]$.

Proof. Follows since the update for P does not depend on μ. □

Theorem 4.1 (Geometric Drift)**.** *Assume the following*

(i) $m_0 > p + 1$ *(informative prior for variance parameters).*

(ii) The function $g(.)$ *is a strictly concave function.*

(iii) $\forall t \in \mathbf{R}$*, the set* $\{(\beta, \mu) \in \mathbf{R}^{L+m_1*p} | g(\beta, \mu) \geq t\}$ *is compact.*

(iv) Log $g(.)$ *is twice continuously differentiable.*

Then

$$E[V(\beta^{(k)})|\beta^{(k-1)}, \mu^{(k-1)}] \leq b + (\frac{m_0}{m + m_0 - p - 1}) \sum_{j=1}^{p} Var_{0,j,j} + (\frac{m}{m + m_0 - p - 1}) V(\beta^{(k-1)}).$$

where b *is as in Claim 4.1.*

Proof. We note that

$$\begin{aligned}
E[V(\beta^{(k)})|\beta^{(k-1)},\mu^{(k-1)}] &= E[V(\beta^{(k)})|\beta^{(k-1)}] \\
&= E\{E[V(\beta^{(k)})|\Sigma^{(k)}]|\beta^{(k-1)}\} \\
&= \tfrac{1}{m}\textstyle\sum_{j=1}^{p}\sum_{i=1}^{m} E\{E[(\beta_{i,j}^{(k)} \\
&\quad -E[\beta_{i,j}^{(k)}|\Sigma^{(k)}] + E[\beta_{i,j}^{(k)}|\Sigma^{(k)}])^2|\Sigma^{(k)}]|\beta^{(k-1)}\} \\
&= \tfrac{1}{m}\textstyle\sum_{j=1}^{p}\sum_{i=1}^{m} E\{E[(\beta_{i,j}^{(k)} - E[\beta_{i,j}^{(k)}|\Sigma^{(k)}])^2|\Sigma^{(k)}] \\
&\quad +E[(\beta_{i,j}^{(k)})|\Sigma^{(k)}]^2|\beta^{(k-1)}\} \\
&\leq \tfrac{1}{m}\textstyle\sum_{j=1}^{p}\sum_{i=1}^{m}\{E[\Sigma_{j,j}^{(k)}|\beta^{(k-1)}] + E[E[\beta_{i,j}^{(k)}|\Sigma^{(k)}]^2|\beta^{(k-1)}]\} \\
&= \textstyle\sum_{j=1}^{p} E[\Sigma_{j,j}^{(k)}|\beta^{(k-1)}] + \tfrac{1}{m}\sum_{j=1}^{p}\sum_{i=1}^{m} E[E[\beta_{i,j}^{(k)}|\Sigma^{(k)}]^2|\beta^{(k-1)}] \\
&= \textstyle\sum_{j=1}^{p}(\frac{1}{m+m_0-p-1}(m_0 Var_{0,j,j} + \sum_{i=1}^{m}(\beta_{i,j}^{(k-1)})^2)) \\
&\quad +\tfrac{1}{m}\textstyle\sum_{j=1}^{p}\sum_{i=1}^{m} E[E[\beta_{i,j}^{(k)}|\Sigma^{(k)}]^2|\beta^{(k-1)}] \\
&\leq b + \textstyle\sum_{j=1}^{p}(\frac{1}{m+m_0-p-1}(m_0 Var_{0,j,j} + \sum_{i=1}^{m}(\beta_{i,j}^{(k-1)})^2)) \\
&= b + (\tfrac{m_0}{m+m_0-p-1})\textstyle\sum_{j=1}^{p} Var_{0,j,j} + (\frac{m}{m+m_0-p-1})V(\beta^{(k-1)}).
\end{aligned}$$

In the above, the first inequality follows from Claim 3.1 , the second from Claim 4.1, and the substitution of $(\frac{1}{m+m_0-p-1}(m_0 Var_{0,j,j} + \sum_{i=1}^{m}(\beta_{i,j}^{(k-1)})^2)$ for $E[\Sigma_{j,j}^{(k)}|\beta^{(k-1)}]$ is valid due to properties of the Inverse-Wishart density. □

4.2 Minorization

We will make use of Lemma 7 in Rosenthal (1995).

Remark 4.3. The transition kernel Q for the Markov chain associated with β and μ is independent of μ. Hence for a given $\beta^{(k-1)}$, we can define $\tilde{Q}$ by $\tilde{Q}(\beta^{(k-1)},) = Q(\beta^{(k-1)},\mu^{k-1})$, where the equality holds for all μ^{k-1}.

The following inequality is a consequence of Claim 11.2 in our appendix and plays a critical role in establishing our minorization condition.

Claim 4.1. Let A be a positive definite matrix, and let B and C be positive semi-definite matrices. Then $|A+C|/|A+B+C| \geq |A|/|A+B|$.

Proof. Since A is a positive definite matrix, there exists uniquely a square root matrix $\tilde{A}$ such that $A = \tilde{A}\tilde{A}$. It thus follows that

$$\begin{aligned}
|A+C|/|A+B+C| &\geq |A|/|A+B| \\
&\Updownarrow \\
(|\tilde{A}||I+\tilde{A}^{-1}C\tilde{A}^{-1}||\tilde{A}|)/(|\tilde{A}||I+\tilde{A}^{-1}(B+C)\tilde{A}^{-1}||\tilde{A}|) &\geq (|\tilde{A}||\tilde{A}|)/(|\tilde{A}||I+\tilde{A}^{-1}B\tilde{A}^{-1}||\tilde{A}|) \\
&\Updownarrow \\
(|I+\tilde{A}^{-1}C\tilde{A}^{-1}|)/(|I+\tilde{A}^{-1}B\tilde{A}^{-1}+\tilde{A}^{-1}C\tilde{A}^{-1}|) &\geq 1/(|I+\tilde{A}^{-1}B\tilde{A}^{-1}|) \\
&\Updownarrow \\
(|I+\tilde{A}^{-1}C\tilde{A}^{-1}|)(|I+\tilde{A}^{-1}B\tilde{A}^{-1}|) &\geq (|I+\tilde{A}^{-1}B\tilde{A}^{-1}+\tilde{A}^{-1}C\tilde{A}^{-1}|).
\end{aligned}$$

where the last inequality is of the form in Claim 11.2. □

Denote by $\pi(\mathbf{y}, \beta, \mu, \mathbf{P})$ the joint density function for $\mathbf{y}$, β, μ, and $\mathbf{P}$. Before proceeding to the statement of our minorization claim, we note the following properties holds for some of the conditional densities.

$$
\begin{aligned}
\pi(\beta^{(k)}|P^{(k)}) &= \textstyle\prod_{i=1}^{m}[[|P^{(k)}|^{1/2}exp(-0.5(\beta_i^{(k)})^T P^{(k)}\beta_i^{(k)})]/(2\pi)^{p/2}] \\
\pi(P^{(k)}|\beta^{(k-1)},\mu^{(k-1)}) &= [|P^{(k)}|^{(m+m_0-p-1)/2}|m_0V_0+ \\
&\quad \textstyle\sum_{i=1}^{m}\beta_i^{(k-1)}(\beta_i^{(k-1)})^T|^{(m+m_0)/2} \\
&\quad exp(-0.5tr((m_0V_0+\textstyle\sum_{i=1}^{m}\beta_i^{(k-1)}(\beta_i^{(k-1)})^T)P^{(k)}))] \\
&\quad /(\textstyle\prod_{j=1}^{p}\Gamma((m+m_0+1-j)/2))2^{(m+m_0)p/2}\pi^{p(p-1)/4} \\
\pi(\beta^{(k)}|P^{(k)})\pi(P^{(k)}|\beta^{(k-1)}) &= [|P^{(k)}|^{(2m+m_0-p-1)/2}|m_0V_{0,Pop} \\
&\quad +\textstyle\sum_{i=1}^{m}\beta_i^{(k-1)}(\beta_i^{(k-1)})^T|^{(m+m_0)/2} \\
&\quad exp(-0.5tr((m_0V_{0,Pop} \\
&\quad +\textstyle\sum_{i=1}^{m}(\beta_i^{(k-1)}(\beta_i^{(k-1)})^T+\beta_i^{(k)}(\beta_i^{(k)})^T))P^{(k)}))] \\
&\quad /((\textstyle\prod_{j=1}^{p}\Gamma((m+m_0+1-j)/2)) \\
&\quad (2^{(2m+m_0)p/2})\pi^{(mp/2)+(p(p-1)/4)}) \\
\textstyle\int \pi(\beta^{(k)}|P^{(k)})\pi(P^{(k)}|\beta^{(k-1)})dP^{(k)} &= [(\textstyle\prod_{j=1}^{p}\Gamma((2m+m_0+1-j)/2)) \\
&\quad |m_0V_{0,Pop}+\textstyle\sum_{i=1}^{m}\beta_i^{(k-1)}(\beta_i^{(k-1)})^T|^{(m+m_0)/2}] \\
&\quad /[\pi^{(mp/2)}(\textstyle\prod_{j=1}^{p}\Gamma((m+m_0+1-j)/2)) \\
&\quad |m_0V_{0,Pop}+\textstyle\sum_{i=1}^{m}(\beta_i^{(k-1)}(\beta_i^{(k-1)})^T \\
&\quad +\beta_i^{(k)}(\beta_i^{(k)})^T)|^{(2m+m_0)/2}]
\end{aligned}
$$

Using these properties, we now establish the following.

Theorem 4.2 (Minorization)**.** *Assume that conditions (i)-(ii) in Claim 4.1 are satisfied and define*

$\tilde{f} := \int\int\int \pi(y|\beta,\mu)\pi(\mu)\pi(\beta|P)\pi(P|\mathbf{0})dPd\beta d\mu$. *Let* $\gamma \geq 0$ *and define* $C_\gamma := \{\beta|V(\beta) \leq \gamma\}$. *Then* $\beta^{(k-1)} \in C_\gamma$ *implies*

$$
\begin{aligned}
Q((\beta^{k-1}, \mu^{(k-1)}), (\beta^k, \mu^{(k)})) \quad &\geq \quad [\tilde{f}/(\pi(\mathbf{y}|\beta^*, \mu^*) MGF(-c(\beta^*, \mu^*)))] \\
&\quad [|m_0 Var_{0,Pop}|/(\sup_{\beta \in C_\gamma} |m_0 Var_{0,Pop} \\
&\quad + \textstyle\sum_{i=1}^{m} \beta_i^{(k-1)} (\beta_i^{(k-1)})^T|)]^{m/2} \\
&\quad S(\beta^{(k)}, \mu^{(k)}).
\end{aligned}
$$

Where $S(.)$ is the density defined by

$$
\begin{aligned}
S(\beta^{(k)}, \mu^{(k)}) \quad &= \quad \int \pi(y|\beta^{(k)}, \mu^{(k)}) \pi(\mu^{(k)}) \pi(\beta^{(k)}|P) \pi(P|\mathbf{0}) dP \\
&\quad / \int \int \int \pi(y|\beta, \mu) \pi(\mu) \pi(\beta|P) \pi(P|\mathbf{0}) dP d\beta d\mu.
\end{aligned}
$$

Proof. We note that

$$
\begin{aligned}
Q((\beta^{k-1},\mu^{(k-1)}),(\beta^{k},\mu^{(k)}))/S(\beta^{(k)},\mu^{(k)}) &= \tilde{f} * \int \pi(\beta^{(k)},\mu^{(k)}|y,P)\pi(P|\beta^{(k-1)})dP \\
&\quad / \int \pi(y|\beta^{(k)},\mu^{(k)})\pi(\mu^{(k)})\pi(\beta^{(k)}|P)\pi(P|\mathbf{0})dP \\
&= \tilde{f} * \int [\pi(y|\beta^{(k)},\mu^{(k)})\pi(\mu^{(k)})\pi(\beta^{(k)}|P)/\pi(y|P)] \\
&\quad \pi(P|\beta^{(k-1)})dP \\
&\quad / \int \pi(y|\beta^{(k)},\mu^{(k)}))\pi(\mu^{(k)})\pi(\beta^{(k)}|P)\pi(P|\mathbf{0})dP \\
&\geq [\tilde{f}/(\pi(\mathbf{y}|\beta^{*},\mu^{*})MGF(-c(\beta^{*},\mu^{*})))] \\
&\quad [\int \pi(y|\beta^{(k)},\mu^{(k)})\pi(\mu^{(k)})\pi(\beta^{(k)}|P)\pi(P|\beta^{(k-1)})dP] \\
&\quad / \int \pi(y|\beta^{(k)},\mu^{(k)})\pi(\mu^{(k)})\pi(\beta^{(k)}|P)\pi(P|\mathbf{0})dP \\
&= [\tilde{f}/(\pi(\mathbf{y}|\beta^{*},\mu^{*})MGF(-c(\beta^{*},\mu^{*})))] \\
&\quad [\int \pi(\beta^{(k)}|P)\pi(P|\beta^{(k-1)})dP]/[\int \pi(\beta^{(k)}|P)\pi(P|\mathbf{0})dP] \\
&= [\tilde{f}/(\pi(\mathbf{y}|\beta^{*},\mu^{*})MGF(-c(\beta^{*},\mu^{*})))] \\
&\quad (|m_o V_{0,pop} + \textstyle\sum_{i=1}^{m} \beta_{i,k-1}\beta_{i,k-1}^{T}|/|m_o V_{0,pop}|)^{(m+m_0)/2} \\
&\quad *(|m_o V_{0,pop} + \textstyle\sum_{i=1}^{m} \beta_{i,k}\beta_{i,k}^{T}| \\
&\quad /|m_o V_{0,pop} + \textstyle\sum_{i=1}^{m} \beta_{i,k-1}\beta_{i,k-1}^{T} + \sum_{i=1}^{m} \beta_{i,k}\beta_{i,k}^{T}|)^{(2m+m_0)/2} \\
&\geq [\tilde{f}/(\pi(\mathbf{y}|\beta^{*},\mu^{*})MGF(-c(\beta^{*},\mu^{*})))] \\
&\quad (|m_o V_{0,pop} + \textstyle\sum_{i=1}^{m} \beta_{i,k-1}\beta_{i,k-1}^{T}|/|m_o V_{0,pop}|)^{(m+m_0)/2} \\
&\quad *(|m_o V_{0,pop}|/|m_o V_{0,pop} + \textstyle\sum_{i=1}^{m} \beta_{i,k-1}\beta_{i,k-1}^{T}|)^{(2m+m_0)/2} \\
&= [\tilde{f}]/(\pi(\mathbf{y}|\beta^{*},\mu^{*})MGF(-c(\beta^{*},\mu^{*})))] \\
&\quad (|m_o V_{0,pop}|/|m_o V_{0,pop} + \textstyle\sum_{i=1}^{m} \beta_{i,k-1}\beta_{i,k-1}^{T}|)^{m/2} \\
&\geq [\tilde{f}/(\pi(\mathbf{y}|\beta^{*},\mu^{*})MGF(-c(\beta^{*},\mu^{*})))] \\
&\quad (|m_o V_{0,pop}|/(\sup_{\beta_{k-1}\in C_\gamma} |m_0 Var_{0,Pop} + \textstyle\sum_{i=1}^{m} \beta_{i,k-1}\beta_{i,k-1}^{T}|)
\end{aligned}
$$

where the first inequality follows from remark 4.1, the second from Claim 4.1, and the third from the compactness of C_γ. Rearrangement of terms completes the proof. □

A couple of remarks regarding the terms on the right hand side are in order.

Remark 4.4. A closed form expression for $\tilde{f}$ is typically not available. In principle, $\tilde{f}$, can be estimated through Monte Carlo simulation from $\pi(\mu)\pi(\beta|P)\pi(P|\mathbf{0})$ and calculation of the expectation for $\pi(y|\beta,\mu)$.

Remark 4.5. A closed form expression for $\pi(\mathbf{y}|\beta^*,\mu^*)MGF(-c(\beta^*,\mu^*)))$ can typically be found through optimization of the function $g(.)$ in Claim 4.1.

Remark 4.6. In the univariate case, $\sup_{\beta\in C_\gamma}|m_0Var_{0,Pop}+\sum_{i=1}^m \beta_i^{(k-1)}(\beta_i^{(k-1)})^T| = m_0Var_{0,Pop}+ m_1\gamma$.

Remark 4.7. In the minorization condition, $\sup_{\beta\in C_\gamma}|m_0Var_{0,Pop}+\sum_{i=1}^m \beta_i^{(k-1)}(\beta_i^{(k-1)})^T|$ can be replaced by an upper bound for $|m_0Var_{0,Pop}+\sum_{i=1}^m \beta_i^{(k-1)}(\beta_i^{(k-1)})^T|$ on C_γ. The Haddamard [Insert reference] inequality, which states that the determinant of any positive definite matrix is bounded above by the product of its diagonal elements is particularly useful in this regard.

Claim 4.2. If m_0V_0 is equal to the identity matrix, then $\sup_{\beta\in C_\gamma}|m_0Var_{0,Pop}+\sum_{i=1}^m \beta_i^{(k-1)}(\beta_i^{(k-1)})^T| \leq (1+(m_1\gamma/p))^p$.

Proof. We first note that for an arbitrary β, we have $|I+\sum_{i=1}^m \beta_i(\beta_i)^T| \leq \prod_{j=1}^p(1+\sum_{i=1}^m \beta_{i,j}^2)$ by the Hadamard inequality. We also note that for all β's in C_γ, we have $\frac{1}{m_1}\sum_{j=1}^p\sum_{i=1}^{m_1}\beta_{i,j}^2 \leq \gamma$. We note that $\prod_{j=1}^p(1+\sum_{i=1}^m \beta_{i,j}^2)$ obtains its maximum on C_γ if and only $\sum_{j=1}^p log(1+\sum_{i=1}^m \beta_{i,j}^2)]$ obtains its maximum on C_γ. It is straightforward to verify

that any such maximum must occur where (i) $\sum_{i=1}^{m} \beta_{i,j1}^2 = \sum_{i=1}^{m} \beta_{i,j2}^2$ for all $j1$ and $j2$, and (ii) $\frac{1}{m_1} \sum_{j=1}^{p} \sum_{i=1}^{m_1} \beta_{i,j}^2 = \gamma$. These two properties together implies that $\sum_{i=1}^{m} \beta_{i,j}^2 = m_1\gamma/p$ for all j. Hence we have $|I + \sum_{i=1}^{m} \beta_i(\beta_i)^T| \leq \prod_{j=1}^{p}(1 + \sum_{i=1}^{m} \beta_{i,j}^2) \leq (1 + (m_1\gamma/p))^p$ for all $\beta \in C_\gamma$. □

Remark 4.8. Claim 4.2 suggests that it may be useful to standardize variables so that m_0V_0 equals the identity matrix. This can always be done for the models presented here.

5 Examples

5.1 Example 1: Drift For the Models in section 3

Consider the following Bayesian model (A special case of the model in subsection 3.2). A prior specification,

$$\begin{aligned} \mu &\sim Normal(\nu, 1/P_0) \\ \beta_i &\sim Normal(\mu, 1/P), \end{aligned}$$

for $i = 1, 2, \ldots, n$.

In addition, normal data for each β_i

$$y_i \sim Normal(\beta_i, 1/P_D).$$

We will consider the two-block Gibbs-sampler in which β is updated in first block and μ in the second.

The conditional density for μ is

$$\mu|\beta \sim Normal(\nu(P_0/(P_0+nP)) + \sum_{i=1}^{n} \beta_i(P/(P_0+nP), 1/(P_0+nP)$$

and the conditional densities for β_i, $i = 1, 2, \ldots, n$ are given by

$$\beta_i|\mu \sim Normal(\mu(P/(P+P_D)) + y_i(P_D/(P+P_D)), (1/(P+P_D))).$$

This implies that the conditional mean for $\mu^{(k)}$ given $\mu^{(k-1)}$ is given by

$$E[\mu^{(k)}|\mu^{(k-1)}] = \nu(P_0/(P_0+nP)) + \overline{y}(nP/(P_0+nP)) + (nP/(P_0+nP))(P/(P+P_D))\mu^{(k-1)}.$$

It is straightforward to verify that μ^* is given by

$$\mu^* = (1/(1-(nP/(P_0+nP))(P/(P+P_D))))(\nu(P_0/(P_0+nP)) + \overline{y}(nP/(P_0+nP)))$$

and that

$$E[\mu^{(k)}|\mu^{(k-1)}] - \mu^* = (nP/(P_0+nP))(P/(P+P_D))(\mu^{(k-1)} - \mu^*).$$

We note that $Var_{\mu^*}(\mu^{(k-1)})$ is an $n \times n$ diagonal matrix with diagonal elements given by $1/(P+P_D)$. P^* in subsection 3.2 satisfies the properties that

(i) P^*_{11} is a $n \times n$ diagonal matrix with diagonal elements given by P;

(ii) P^*_{12} is an $n \times 1$ matrix with individual elements given by P;

(iii) P^*_{21} is an $1 \times n$ matrix with individual elements given by P;

(iv) P^*_{22} is a 1×1 matrix with single element given by $P_0 + nP$.

It is straightforward to verify that

$$P_{21}^* Var_{\mu^*}(\mu^{(k-1)}) P_{12}^* = (P/(P+P_D))nP$$

and

$$P_{21}^* (P_{11}^*)^{-1} P_{12}^* = (P/P)nP.$$

Hence we have

$$\begin{aligned}
T(Var_{\mu^*}(\mu^{(k-1)})) &= (P/(P+P_D))nP(1/(P_0+nP))(1/(P_0+nP))nP(P/(P+P_D)) \\
&= ((P/(P+P_D))(nP/(P_0+NP)))^2 \\
&\leq ((P/(P))(nP/(P_0+NP)))^2 \\
&= (P/P)nP(1/(P_0+nP))(1/(P_0+nP))nP(P/P) \\
&= T^*
\end{aligned}$$

Using properties of partitioned matrices, it is also possible to verify that the conditional variance of $\mu^{(k)}$ given $\mu^{(k-1)}$ is given by

$$Var(\mu^{(k-1)}|\mu^{(k-1)}) = (P_0 + nP - P_{21}^*[(P+P_D)I + P_{12}^*(P_0+nP)^{-1}P_{21}^*]^{-1}P_{12}^*)^{-1}.$$

The matrix $P_{12}^*(P_0+nP)^{-1}P_{21}^*$ is easily shown to be an $n \times n$ matrix where all elements equal to $(P/(P_0+nP))P$. In most cases, $P_{12}^*(P_0+nP)^{-1}P_{21}^*$ is therefore small relative to $(P+P_D)I$, and hence

$$\begin{aligned}
Var(\mu^{(k-1)}|\mu^{(k-1)}) &\approx (P_0 + nP - P_{21}^*[(P+P_D)I]^{-1}P_{12}^*)^{-1} \\
&= (P_0 + nP - (P/(P+P_D))nP)^{-1}.
\end{aligned}$$

We note the strong dependence on P_D. As P_D gets small, $Var(\mu^{(k-1)}|\mu^{(k-1)})$ approaches P_0^{-1}. On other hand, $Var(\mu^{(k-1)}|\mu^{(k-1)})$ approaches $(P_0 + nP)^{-1}$ as P_D gets large.

Applying Theorem 3.2 to the above results, we get the drift condition

$$\begin{aligned} E[V(\mu^k)|\mu^{(k-1)}] &\leq (P_0 + nP - P_{21}^*[(P + P_D)I + P_{12}^*(P_0 + nP)^{-1}P_{21}^*]^{-1}P_{12}^*)^{-1} \\ &\quad +((P/(P + P_D))(nP/(P_0 + nP)))^2 V(\mu^{(k-1)}) \\ &\approx (P_0 + nP - (P/(P + P_D))nP)^{-1} + ((P/(P + P_D))(nP/(P_0 + nP)))^2 V(\mu^{(k-1)}) \end{aligned}$$

Alternatively, applying Theorem 3.4 we get the drift condition

$$E[V(\mu^k)|\mu^{(k-1)}] = (P_0)^{-1} + (nP/(P_0 + NP))^2 V(\mu^{(k-1)}.$$

Theorem 3.2 hence gives a tighter drift condition for this model. The advantage of Theorem 3.4 is of course that it is applicable even when the data is non-normal. In a sense, it represents the limiting drift in Theorem 3.2 as P_D approaches 0. Log-concave likelihood functions can be quite deceptive and tend to have fat tails. While the precision of the data in such models can be large near the center of the density, it tends to die of in the tails. Hence model with log-concave likelihood functions are likely to drift quickly near the center of the density, while it drifts slowly(close to that in Theorem 3.4) in the tails of the density.

Turning to the minorization conditions, we get expressions for ϵ_γ given by

$$\begin{aligned} \epsilon_\gamma &= (\tfrac{|P_0+nP-P_{21}^*[(P+P_D)I+P_{12}^*(P_0+nP)^{-1}P_{21}^*]^{-1}P_{12}^*|}{|P_0+nP-P_{21}^*[(P+P_D)I+P_{12}^*(P_0+nP)^{-1}P_{21}^*]^{-1}P_{12}^*+\tilde{P}|})^{0.5} exp(-((P/(P + P_D))(nP/(P_0 + nP)))^2\gamma) \\ &\approx (\tfrac{|P_0+nP-(P/(P+P_D))nP|}{|P_0+nP-(P/(P+P_D))nP+\tilde{P}|})^{0.5} exp(-((P/(P + P_D))(nP/(P_0 + nP)))^2\gamma) \end{aligned}$$

if we use the results in subsection [] and

$$\begin{aligned}
\epsilon_\gamma &= \left(\frac{|P_0+nP-P_{21}^*[(P+P_D)I+P_{12}^*(P_0+nP)^{-1}P_{21}^*]^{-1}P_{12}^*|}{|P_0+nP+\tilde{P}|}\right)^{0.5} exp(-(nP/(P_0+nP))^2\gamma) \\
&\approx \left(\frac{|P_0+nP-(P/(P+P_D))nP|}{|P_0+nP+\tilde{P}|}\right)^{0.5} exp(-(nP/(P_0+nP))^2\gamma) \\
&\geq \left(\frac{|P_0|}{|P_0+nP+\tilde{P}|}\right)^{0.5} exp(-(nP/(P_0+nP))^2\gamma)
\end{aligned}$$

if we use the results in subsection []. Again, we note the strong dependence on P_D, with ϵ larger (which generally will be better) when P_D is large as opposed to when it is small. Our last expression represents a very conservative lower bound for ϵ_γ. Due to the fat tail of many log-concave likelihood functions, however, it may be the tightest possible bound for many such models.

5.2 Example 2: Normal Data

Consider a block Gibbs Sampler for the posterior density resulting from the following model:

$$\sigma_{Pop}^2 \sim Inverse - Gamma(m_0/2, m_0 Var_{0,Pop}/2)$$

$$\beta_i \sim Normal(0, \sigma_{Pop}^2), i = 1, \ldots, m_1.$$

and for all $i = 1, \ldots, m_1$,

$$y_{i,j} \sim Normal(\beta_i, \sigma_{Data}^2), j = 1, \ldots, n_i.$$

We let $\overline{\beta} := \sum_{i=1}^{m_1} \beta_i/m_1$, and $\overline{y}_i := \sum_{j=1}^{n_i} y_{i,j}/n_i$ for every $i = 1, \ldots, m_1$.

Define a function $V : \mathbf{R}^m \to [0, \infty)$ by

$$V(\beta^{(k)}) = (1/m)\sum_{i=1}^{m}((\beta_i^{(k)}))^2.$$

Remark 5.1 (Geometric Drift). If $m_0 > 2$, then

$$E[V(\beta^{(k)})|\beta^{(k-1)}] \leq [(1/m)\sum_{i=1}^{m}(\overline{y}_i)^2] + (m_0/(m+m_0-2))Var_{0,Pop} + (m/(m+m_0-2))V(\beta^{(k-1)}).$$

Proof. We note that the conditional expectation for β_i is given by

$$E[\beta_i|\sigma^2_{Pop}] = \overline{y}_i[(n_i/\sigma^2_{Data})/((1/\sigma^2_{Pop}) + (n_i/\sigma^2_{Data}))].$$

It follows that $[E[\beta_i|\sigma^2_{Pop}]]^2 \leq \overline{y}_i^2$. Substituting this into the formula in claim 4.1 yields the desired inequality. □

Remark 5.2 (Minorization). For the model considered here,

$$\pi(\mathbf{y}|\beta^*, \mu^*)MGF(-c(\beta^*, \mu^*)) = \prod_{i=1}^{m_1}\prod_{j=1}^{n_i}[(1/(\sqrt{2\pi}\sigma_{Data}))exp(-0.5(\bar{y}_i - y_{i,j})^2/\sigma^2_{Data})].$$

Hence $\beta^{(k-1)} \in C_\gamma$ implies

$$\begin{aligned} Q((\beta^{k-1}, \mu^{(k-1)}), (\beta^k, \mu^{(k)})) \quad \geq \quad & [\tilde{f}/[\textstyle\prod_{i=1}^{m_1}\prod_{j=1}^{n_i}[(1/(\sqrt{2\pi}\sigma_{Data}))exp(-0.5(\bar{y}_i - y_{i,j})^2/\sigma^2_{Data})]] \\ & [(m_0 Var_0)/((m_0 Var_0 + m_1\gamma)]^{m/2} \\ & S(\beta^{(k)}, \mu^{(k)}). \end{aligned}$$

6 Simulation

In this section, we first consider a variation on a hierarchical random effects model using data first considered in Crowder (1978). Our data consists of the proportion of seeds germinating in $n = 20$ plates. Relevant covariates are seed (2 types), root extract (2 types), and an

interaction term. We consider a Hierarchical logit model where the success probabilities p_i are related to β_i via $logit(p_i) = \beta_i$ for $i = 1, \ldots, 20$. The hierarchcial specification is given by $\beta_i \sim N(X\mu, |\sigma^2_{Pop})$, where there are independent priors for μ and σ^2_{Pop} respectively given by

$$\mu \sim N(\nu, \sigma_0^2)$$

$$1/\sigma^2_{Pop} := P_{Pop} \sim Gamma((n0 * V0)/2, n0/2).$$

We simulate from this model using an inner two-block Gibbs sampler like that in section () and an outer two-block Gibbs sampler like that in section (). Please note that one of the observations in Crowder's original data is excluded from the simulation as it would cause our outerloop sampler below to violate one of the assumptions in Theorem []. The observation excluded is the only observation in which the count for one of the categories actually is zero. In subsection [], we illustrate the convergence problems such models may present for two-block Gibbs samplers using dummied up data, and discuss alternate approaches for such models.

6.1 Inner loop

For fixed P_{Pop}, we simulate using a two-block Gibbs sampler in which μ and β are updated in separate blocks. This two-block Gibbs sampler is in the same form as the model in subsection (). In order to determine the required number of iterations for this inner loop, we use a simulation approach detailed in appendix () in order to find the minimal number of iteration required to achieve the desired accuracy (a total variation distance of no more than 0.01) if we instead used a two-block Gibbs sampler to sample from the prior density.

In general, the correlation between the two blocks should decrease as data is added to the model, and hence one would expect the number of iterations needed for the model with data to be no greater than the number of iterations needed for the model without data. Our proof of the drift condition in subsection () makes use of this relationship in order to show that the drift in the model with data is no slower than the drift in the model without data. Our example in subsection () also demonstrates the relationship between the two models. Graph () below shows the estimated minimal number of iterations required at each of the values for P_{Pop} simulated in the outerloop. Interestingly, the minimal number of iterations appear to increase linearly with the value of P_{Pop}. Our graph includes a line fitted through the estimated number of iterations showing the apparent linear relationship.

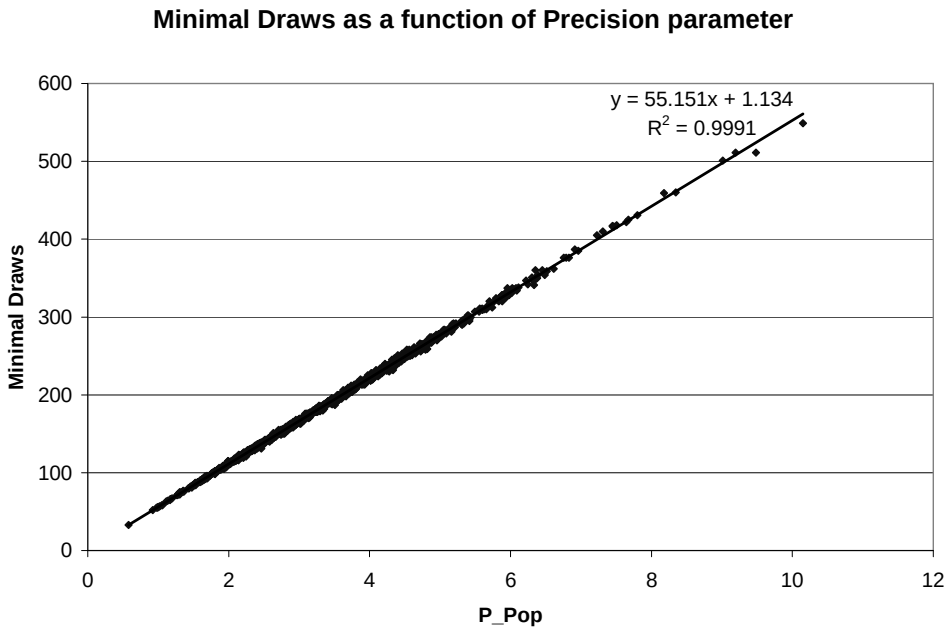

Figure 1: Minimal iterations as a function of population precision

We also include a set of graphs showing the difference in mixing behavior between a low value for the precision ($P_{Pop} = 2$) and a large value ($P_{Pop} = 10$) for P_{Pop}. In all cases, the

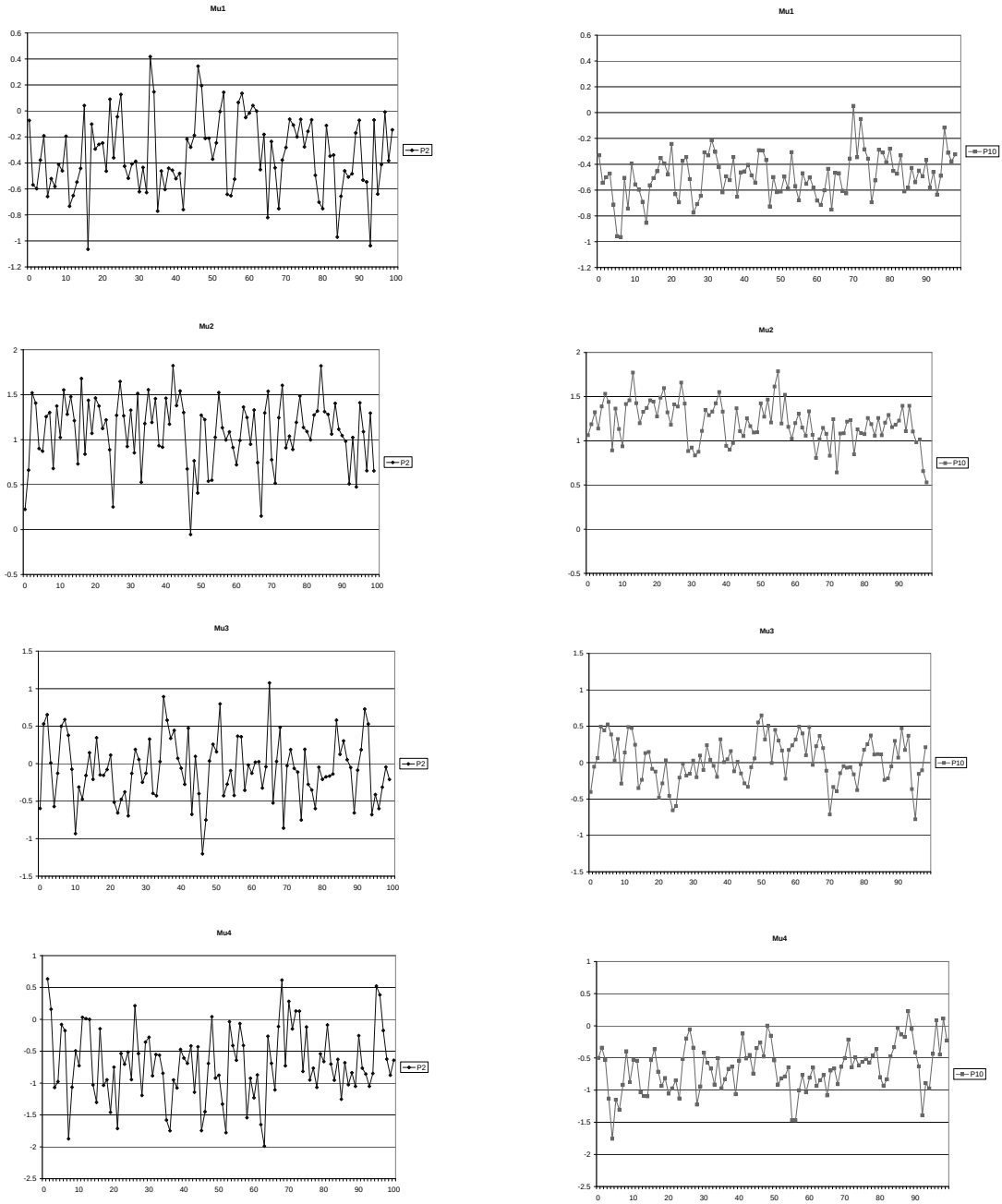

Figure 2: Mixing for P_{Pop}=2 and $P_{Pop} = 10$, first 100 Iterations

graphs include 100 iterations. From the graphs it can be seen that the mixing is substantially slower when $P_{Pop} = 10$. The estimated one step autocorrelation functions also reveal this, as can be seen from table 1.

Table 1: Autocorrelation Functions

Parameter	$P_{Pop} = 2$	$P_{Pop} = 10$
Intercept	0.194840454	0.39168583
Seed Coef.	0.120731396	0.446093684
Extract Coef.	0.197714902	0.531434867
Interaction Coef.	0.247909305	0.564863541

6.2 Outer Loop

The outerloop consists of an approximate two-block Gibbs sampler in which P_{Pop} is updated in one block and (μ, β) is updated in the second, the latter through the use of the inner loop. It can be seen, through a re-parameterization of the (μ, β) block, that this two block Gibbs sampler satisfies all the assumptions in Theorem (), and hence it is a Geometrically Ergodic Markov Chain. Our graph below shows the mixing behavior for some of the parameters in the model and table 2 shows the parameters estimates based on iterations $101 - 1100$. Generally, mixing of the Chain appears to be quite fast.

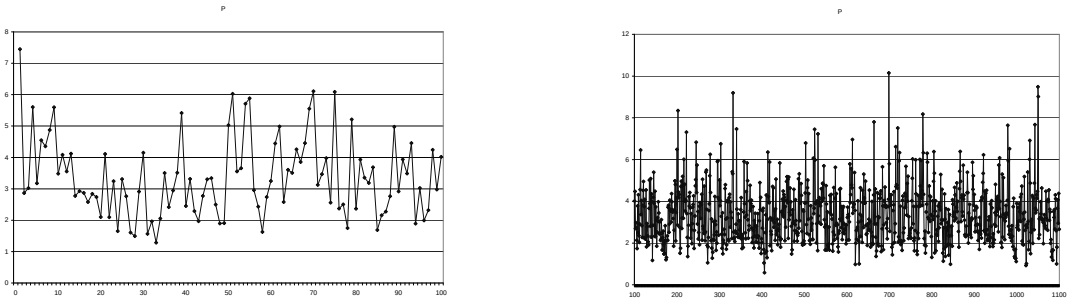

Figure 3: Outerloop mixing for first 100 iterations and iterations 101-1000

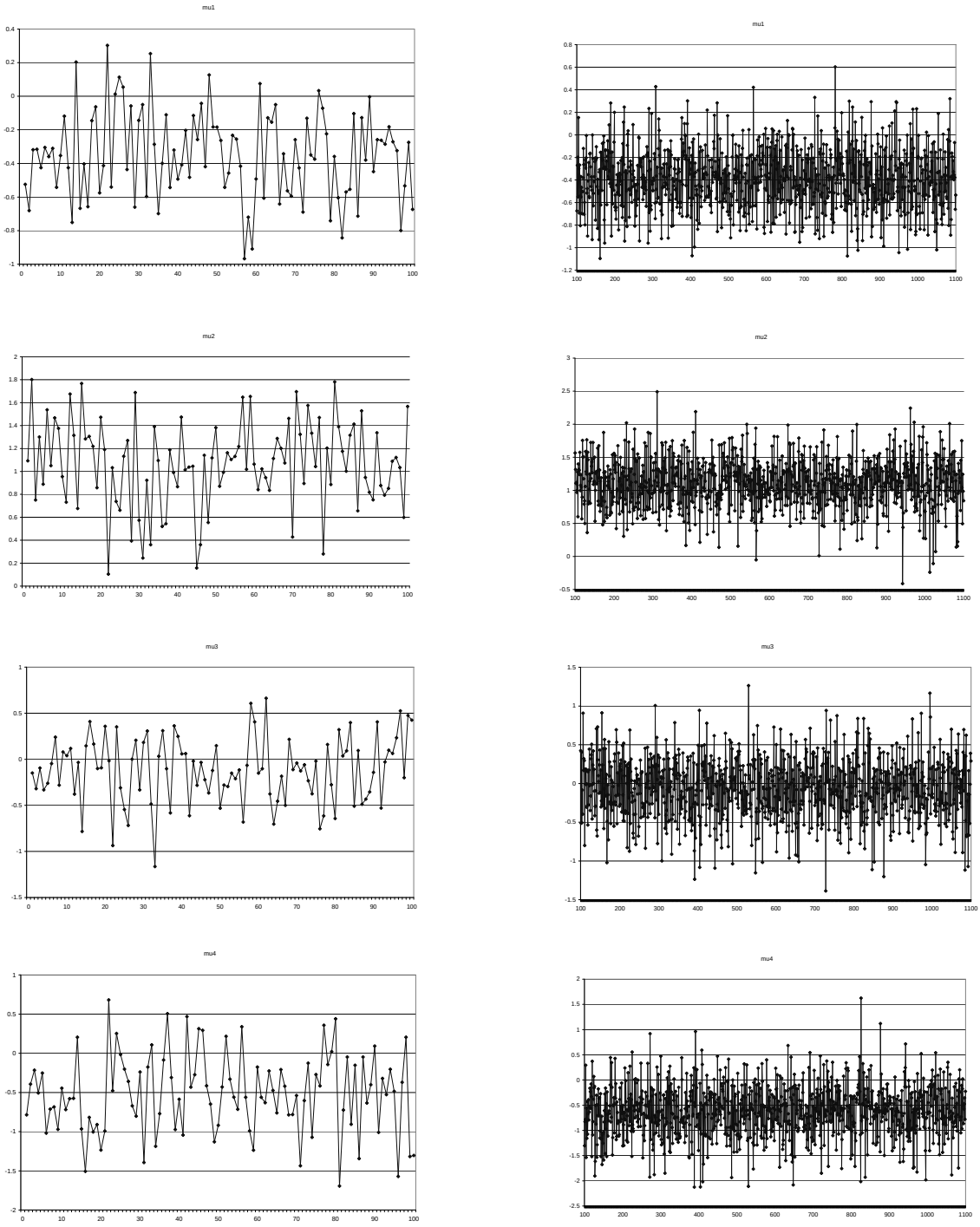

Figure 4: Outerloop mixing for first 100 iterations and iterations 101-1000

6.3 A Non-Geometrically Ergodic Model

We now illustrate the problematic behavior of the two-block Gibbs sampler in the previous section when assumption (iii) in claims () and () is violated. In order to do so, we work

Table 2: Coefficient Estimates

Parameter	Estimate ± SE
Intercept	-0.38738532 ± 0.262458467
Seed Coef.	1.104614007 ± 0.360594236
Extract Coef.	-0.067273229 ± 0.381275926
Interaction Coef.	-0.616930373 ± 0.500028425
P_{Pop}	3.32898998 ± 1.30896072

with a modified version of the data in the previous section, in which the entire count is assigned to the largest category. Figure 5 shows the behavior of the variance parameter and the first 5 β parameters over the first 220 iterations. The chain exhibits a behavior inconsistent with geometric convergence. The challenge with this type of data appears to be that the random effects parameters are not identified by the likelihood function itself. The maximum likelihood estimates for this model are not well defined. As a result, the values for the β parameters grow more extreme as the variance grows large. At the same time, the variance tends to grow larger as the β parameters grow more extreme. All this suggests that extreme caution is needed when sampling from models where the random effects parameters are not identified by the likelihood function. On possibility for getting around these sampling difficulties may be by utilizing alternative functional forms as priors for the variance or variance-covariance parameters. In particular, utilizing priors which bound the variance parameters would likely allow the two-block Gibbs sampler to retain the property of geometric ergodicity. In the univariate case, there are a number of alternative to the

inverse-gamma prior utilized in the current model. In the multivariate case, alternatives to the inverse-Wishart prior may be more difficult to utilize.

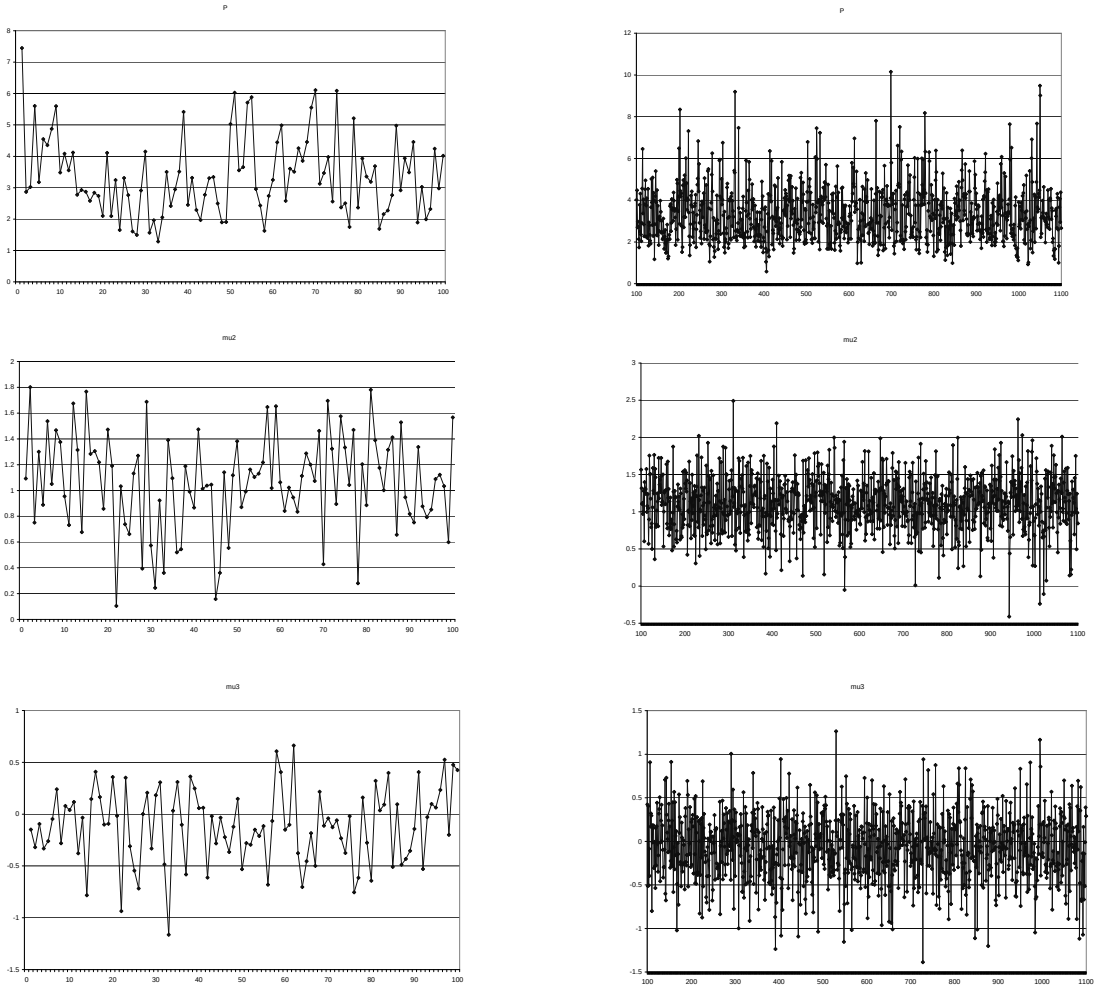

Figure 5: Outerloop mixing for first 220 iterations

7 Concluding Discussion

8 Appendix A: Variance Inequalities

In this section, we derive several intermediate results before establishing Claim 3.1. The main arguments are contained in subsection 8.1, while some of the lengthier details are contained in subsection 8.2.

8.1 Main Results

Throughout this section, we will denote the mean of densities $f(.)$ and $q(.)$ by $\bar{x}_f$ and $\bar{x}_q$ respectively. We start by noting the following basic property of the variance of any density $f(.)$.

Remark 8.1. $\int(\bar{x}_f - x)^2 f(x)dx \leq \int(\bar{x}_q - x)^2 f(x)dx$

Proof. Well known property of the mean. □

We also note the following properties for two densities $f(.)$ and $q(.)$.

Claim 8.1. Suppose densities $f(.)$ and $q(.)$ satisfies the properties that there exists r such that

(i)

$$[q(x) < f(x)] \Rightarrow [|x - \bar{x}_q| < r]$$

and (ii)

$$[q(x) > f(x)] \Rightarrow [|x - \bar{x}_q| > r]$$

then

$$\int (x - \bar{x}_q)^2 f(x)dx \leq \int (x - \bar{x}_q)^2 q(x)dx.$$

Proof. Let $A := \{x \in \mathbf{R} | f(x) > q(x)\}$ and $B := \{x \in \mathbf{R} | f(x) < q(x)\}$. Note that $\int_{x \in A}(f(x) - q(x))dx = \int_{x \in B}(q(x) - f(x))dx$. Under the present assumptions, it then follows that

$$\begin{aligned} \int_{x \in A}(x - \bar{x}_q)^2(f(x) - q(x))dx &\leq \int_{x \in A} r^2(f(x) - q(x))dx \\ &= \int_{x \in B} r^2(q(x) - f(x))dx \\ &\leq \int_{x \in B}(x - \bar{x}_q)^2(q(x) - f(x))dx \end{aligned}$$

Rearranging terms, we have

$$\begin{aligned} \int_{x \in A \bigcup B}(x - \bar{x}_q)^2 f(x)dx &\leq \int_{x \in A \bigcup B}(x - \bar{x}_q)^2 q(x)dx \\ &\Updownarrow \\ \int_{x \in \mathbf{R}}(x - \bar{x}_q)^2 f(x)dx &\leq \int_{x \in \mathbf{R}}(x - \bar{x}_q)^2 q(x)dx \end{aligned}$$

which completes the proof. □

Combining Remark 8.1 with Claim 8.1 allows us to state the following result relating the variances of two densities $f(.)$ and $q(.)$.

Claim 8.2. Suppose densities $f(.)$ and $q(.)$ satisfies the properties that there exists r such that

(i)

$$[q(x) < f(x)] \Rightarrow [|x - \bar{x}_q| < r]$$

and (ii)

$$[q(x) > f(x)] \Rightarrow [|x - \bar{x}_q| > r]$$

then

$$\int (x - \bar{x}_f)^2 f(x)dx \leq \int (x - \bar{x}_q)^2 q(x)dx.$$

Proof. Follows from Remark 8.1 and Claim 8.1. □

Our next result relates the variance of posterior densities with normal priors to normal densities with the same variance as the prior.

Claim 8.3. Let $f(.)$ be a posterior density with a normal prior, and let q be a normal density with the same variance as the prior. Assume that the likelihood function is log-concave and that there exists $r > 0$ such that

$$q(\bar{x}_q - r) = f(\bar{x}_q - r)$$

and

$$q(\bar{x}_q + r) = f(\bar{x}_q + r).$$

then

$$\int (x - \bar{x}_f)^2 f(x)dx \leq \int (x - \bar{x}_q)^2 q(x)dx.$$

Proof. We will show that the two assumptions of claim 8.2 are satisfied. First note that $log(f(x)) - log(q(x))$ is a concave function (in fact it is a linear transformation of the log-likelihood function). It follows that for any a, the set $\{x \in \mathbf{R} | log(f(x)) - log(q(x)) \geq a\}$ is a

convex set. In particular, this holds when $a = 0$, which implies that the set $\{x \in \mathbf{R} | f(x) \geq q(x)\}$ is convex. Hence $f(x)$ must be greater than or equal to $q(x)$ for every $x \in [\bar{x}_q - r, \bar{x}_q - r]$. Hence property (ii) in claim 8.2 must hold.

To see that $f(x) <= q(x)$ outside of the interval, first note that $log(f(x)) - log(q(x))$ is a concave function and that $log(f(x)) - log(q(x)) >= 0$ for any x' in the interior of the interval $[\bar{x}_q - r, \bar{x}_q - r]$. Now, suppose there existed a point x'' outside of the interval where $log(f(x'')) - log(q(x'')) > 0$. Since $\bar{x}_q - r$ or $\bar{x}_q + r$ is a strict convex combination of x' and x'', it follows from the concavity that either $log(f(\bar{x}_q - r)) - log(q(\bar{x}_q - r)) > 0$ or $log(f(\bar{x}_q + r)) - log(q(\bar{x}_q + r)) > 0$, a contradiction. Hence property (i) of claim 8.2 must also hold. Our result hence follows from Claim 8.2. □

Our next claim shows the the required $r > 0$ indeed do exist for strictly log-concave likelihood functions. It relies on rather extensive arguments in subsection 8.2.

Claim 8.4. Assume that the likelihood function is strictly log-concave and the prior is normal. Then there exists a normal density $q^*()$ with the same variance as the prior, and a constant $r > 0$ such that

$$q^*(\bar{x}_q - r) = f(\bar{x}_q - r)$$

and

$$q^*(\bar{x}_q + r) = f(\bar{x}_q + r).$$

Proof. We make use of claim 8.15 in subsection 8.2. We first note that for $\mu \in M$ sufficient close to $\mu_1 := \inf\{\mu \in M\}$, we have $h(\mu) < 0$. Likewise, for μ sufficiently close to $\mu_2 :=$

$\sup\{\mu \in M\}$, we have $h(\mu) > 0$. Since $h(.)$ is a continuous function, and M is a connected set, it follows from the intermediate value theorem that there exist μ^* such that $h(\mu^*) = 0$. It straightforward to verify that the density $q^{\mu^*}(.)$ satisfies all of the requirement for claim 8.3. □

Our next set of results now establish the required result. The first for univariate strictly log-concave likelihood functions, while the second and third extend this to univariate log-concave and multivariate log-concave likelihood functions.

Claim 8.5. Let $f(\mathbf{y}|.)$ be a strictly log-concave likelihood function with a normal prior $\pi(.)$. Then

$$\int (\beta - E^{\pi(.|y)}[\beta])^2 \pi(\beta|y) d\beta \leq \int (\beta - E^{\pi(.)}[\beta])^2 \pi(\beta) d\beta.$$

Proof. Follows from claims 8.3 and 8.4. □

Theorem 8.1. *Let $f(\mathbf{y}|.)$ be a log-concave likelihood function with a normal prior $\pi(.)$. Then*

$$\int (\beta - E^{\pi(.|y)}[\beta])^2 \pi(\beta|y) d\beta \leq \int (\beta - E^{\pi(.)}[\beta])^2 \pi(\beta) d\beta.$$

Proof. We note that the set of concave functions is the closure of the set of strictly concave functions. Hence for any log-concave likelihood function $f(\mathbf{y}|.)$, there exists a sequence $\{f^k(.)\}_{k=1}^{\infty}$ of strictly log-concave likelihood functions such that $\lim_{k\to\infty} f^k(\beta) = f(\beta)$ for every β. From claim 8.5 we know that the desired inequality holds for every $f^k(.)$. Hence it follows that the inequality also must hold for the limiting function $f(\mathbf{y}|.)$. □

Theorem 8.2. *Let $f(\mathbf{y}|.)$ be a log-concave likelihood function with a multivariate normal prior $\pi(.)$. Then for every dimension i,*

$$\int (\beta_i - E^{\pi(.|y)}[\beta_i])^2 \pi(\beta|y) d\beta \leq \int (\beta_i - E^{\pi(.)}[\beta_i])^2 \pi(\beta) d\beta.$$

Proof. For multivariate normal priors, it is straightforward to verify, using the result from Prekopa (1973), that the marginal likelihood for each i is well defined and itself log-concave. Our result then follows from Theorem 8.1. □

We now further extend these results. Let us first start with a couple of remarks.

Remark 8.2. Suppose β has a multivariate normal prior and a log-concave likelihood function. Consider a one-to-one linear transformation of β, $\tilde{\beta} = A\beta$, for which $\beta = A^{-1}\tilde{\beta}$. Then the prior density for $\tilde{\beta}$ is also multivariate normal with a log-concave likelihood function. Moreover, each of the components of $\tilde{\beta}$ satisfies Theorem 8.2.

Remark 8.3. Let x be a vector (not all components zero), and let $z = x^T\beta$. Then the prior density for z is a normal distribution, the likelihood function log-concave, and the posterior variance less than or equal to the prior variance.

Proof. Construct a matrix A as in our above remark, with x^T as the first row. The result then follows immediately from our earlier remark. □

Fact 8.1. Let $x \in \mathbf{R}^m$ be any vector. Then $x^T Var(\beta) x = Var(x^T\beta)$.

Proof. We note that

$$
\begin{aligned}
x^T Var(\beta)x &= \sum_{i=1}^m \sum_{j=1}^m x_i Cov(\beta_i, \beta_j) x_j \\
&= \sum_{i=1}^m \sum_{j=1}^m x_i (E[\beta_i \beta_j] - E[\beta_i]E[\beta_j]) x_j \\
&= \sum_{i=1}^m \sum_{j=1}^m (E[x_i \beta_i x_j \beta_j] - E[x_i \beta_i]E[x_j \beta_j]) \\
&= \sum_{i=1}^m \sum_{j=1}^m (E[x_i \beta_i x_j \beta_j]) - \sum_{i=1}^m \sum_{j=1}^m (E[x_i \beta_i]E[x_j \beta_j]) \\
&= E[(\sum_{i=1}^m x_i \beta_i) * (\sum_{j=1}^m x_j \beta_j)] - (\sum_{i=1}^m E[x_i \beta_i]) * (\sum_{j=1}^m E[x_j \beta_j]) \\
&= E[(x^T \beta) * (x^T \beta)] - (E[x^T \beta]) * (E[x^T \beta]) \\
&= Var(x^T \beta).
\end{aligned}
$$

□

8.2 Technical Details

Remark 8.4. A posterior density with a normal prior and a log-concave likelihood function has a unique posterior mode.

Proof. By log-concavity we have that for any a, the set $\{x \in \mathbf{R} | log(\pi(x|\mathbf{y})) \geq a\}$ is convex, and hence an interval. Moreover, this interval must have finite measure in order for the posterior density to be well defined. From the continuity of a concave function, it is also straightforward to verify that the interval must be closed. It then follows from the Minkovski theorem that $log\pi(x|\mathbf{y})$ obtains its maximum on any such non-empty interval (which clearly exists). Uniqueness of the maximum then follows from the strict log-concavity of the posterior density. □

Let us now introduce some notation. Define

$$M := \left\{ \mu \in \mathbf{R} \;\middle|\; \exists r1, r2 \in \mathbf{R}_{++} : \begin{array}{l} (i)\pi(\mu|y) > q^{\mu}(\mu) \\ (ii)\pi(\mu - r1|y) = q^{\mu}(\mu - r1) \\ (iii)\pi(\mu + r2|y) = q^{\mu}(\mu + r2) \end{array} \right\}$$

where $q^{\mu}(.)$ is a normal density with mean μ and the same variance as the prior. We also define $g^{\mu}(\beta) := log(\pi(\beta|y)) - log(q^{\mu}(\beta))$.

Claim 8.6. If $\pi(\mathbf{y}|.)$ is strictly log-concave, then M is non-empty and contains an open interval around β^*.

Proof. : Let β^* be the posterior mode. Under our assumptions, it exists and is unique. Let $\mu = \beta^*$. Then $q^{\mu}(\mu) < \pi(\mu|y)$. The function $g^{\beta^*}(.)$ is strictly concave and obtains its unique maximum at β^* (since the gradients for both $log(\pi(\beta|y))$ and $log(q^{\mu}(\beta))$ vanish at β^*). For every $\beta < \beta^*$, every super-gradient is positive, and for every $\beta > \beta^*$ every super-gradient is negative. Hence there exists $\beta 1 < \beta^*$ such that $g(\beta 1) > 0$ and a super-gradient at $\beta 1$ is positive. There also exists $\beta 2 > \beta^*$ such that $g(\beta 2) > 0$ and $dg(\beta 2)/d\beta < 0$. It is straightforward to verify that this implies the existence of the required $r1$ and $r2$ (use upper bounds on $g(.)$ derived using the super-gradients at the two points combined with the intermediate value theorem). Using the same method, it is also straightforward to verify that the super-gradients continue to have the same properties even if μ is chosen from some neighborhood of β^*. Q.E.D.

□

Claim 8.7. Show that $\mu \in M$ if and only if the following two properties hold:

(i) $\exists \beta 1 < \mu$: (a) $g^{\mu}(\beta 1) > 0$, and (b) a super-gradient for $g^{\mu}(.)$ at $\beta 1$ is positive.

(i) $\exists \beta 2 > \mu$: (a) $g^{\mu}(\beta 2) > 0$, and (b) a super-gradient for $g^{\mu}(.)$ at $\beta 2$ is negative.

Proof. Suppose $\mu \in M$. We note that $g^{\mu}(.)$ is a strictly concave function which obtains its maximum at some $\tilde{\beta}$. Pick $\beta 1 < min(\tilde{\beta}, \mu)$ and $\beta 2 > max(\tilde{\beta}, \mu)$ such that $g^{\mu}(\beta 1) > 0$ and $g^{\mu}(\beta 2) > 0$. This is possible since $g^{\mu}(\tilde{\beta}) \geq g^{\mu}(\mu) > 0$. It is straightforward to verify that every super-gradient at $\beta 1$ is positive and every super-gradient at $\beta 2$ is negative (otherwise $\tilde{\beta}$ could not be a maximum). This establishes the only if part of our claim. To see the if part, note that the existence of a $\beta 1$ as above with a positive super-gradient implies that $g^{\mu}(\beta) < 0$ for β small enough. Since $g^{\mu}(.)$ is a continuous function, it follows from the intermediate value theorem that there exist β' between β and $\beta 1$ such that $g^{\mu}(\beta') = 0$. Hence we can set $r1 = \beta 1 - \beta'$. By a symmetric argument, we can also find a suitable $r2$. □

Claim 8.8. The set M is open.

Proof. Pick any $\mu \in M$. Then there exists $\beta 1$ and $\beta 2$ as above. By continuity, $\beta 1$ and $\beta 2$ satisfy the same properties for all μ' in some open neighborhood of μ. Hence some neighborhood of μ is also contained in M. □

Definition 8.1. Assume that $\pi(\mathbf{y}|\mu)$ is strictly log-concave. Then for a given $\mu \in M$, the numbers r_1 and r_2 in the definition of M are unique. We will denote by $r_1(\mu)$ the unique number $r > 0$ such that $\pi(\mu - r|\mathbf{y}) = q^{\mu}(\mu - r)$. Likewise, we will denote by $r_2(\mu)$ the unique number $r > 0$ such that $\pi(\mu + r|\mathbf{y}) = q^{\mu}(\mu + r)$.

Claim 8.9. Let $\mu_1, \mu_2 \in M$ satisfy $\mu_1 < \mu_2$ then $\mu_2 < \mu_1 + r_2(\mu_1)$ and $\mu_1 > \mu_2 - r_1(\mu_2)$.

Proof. We will show that $\mu_2 < \mu_1 + r_2(\mu_1)$, a symmetric argument establishes the second inequality. Suppose $\mu_2 \geq \mu_1 + r_2(\mu_1)$. Then we have

$$\begin{aligned}
\pi(\mu_2|y) &\leq \pi(\mu_1 + r_2(\mu_1)|\mathbf{y}) \\
&= q^{\mu_1}(\mu_1 + r_2(\mu_1)) \\
&< q^{\mu_1}(\mu_1) \\
&= q^{\mu_2}(\mu_2)
\end{aligned}$$

a contradiction. □

Claim 8.10. The set M is convex (and hence connected).

Proof. Let $\mu_1, \mu_2 \in M$ satisfy $\mu_1 < \mu_2$ and denote by $r_2(\mu_1)$ and $r_2(\mu_2)$ the two positive numbers such that $g^{\mu_1}(\mu_1 + r_2(\mu_1)) = 0$ and $g^{\mu_2}(\mu_2 + r_2(\mu_2)) = 0$. Now, pick any $\mu \in (\mu_1, \mu_2)$. We will show that there exists $r_2 > 0$ such that $g^{\mu}(\mu + r_2) = 0$. A symmetric argument can be used to show the existence of $r_1 > 0$ such that $g^{\mu}(\mu - r_1) = 0$.

Define

$$\begin{aligned}
\tilde{r}_{2,1} &= \mu_1 + r_2(\mu_1) - \mu \quad \Rightarrow \quad \mu + \tilde{r}_{2,1} = \mu_1 + r_2(\mu_1) \\
\tilde{r}_{2,2} &= \mu_2 + r_2(\mu_2) - \mu \quad \Rightarrow \quad \mu + \tilde{r}_{2,2} = \mu_2 + r_2(\mu_2)
\end{aligned}$$

Using our above claim, we now have

$$\begin{aligned}
g^{\mu}(\mu + \tilde{r}_{2,1}) &= g^{\mu}(\mu_1 + r_2(\mu_1)) \\
&= log(\pi(\mu_1 + r_2(\mu_1)|y)) + k + 0.5(\mu_1 + r_2(\mu_1) - \mu)^2/\sigma^2 \\
&> log(\pi(\mu_1 + r_2(\mu_1)|y)) + k + 0.5(r_2(\mu_1))^2/\sigma^2 \\
&= g^{\mu_1}(\mu_1 + r_2(\mu_1)) \\
&= 0.
\end{aligned}$$

and

$$\begin{aligned} g^{\mu}(\mu + \tilde{r}_{2,2}) &= g^{\mu}(\mu_2 + r_2(\mu_2)) \\ &= log(\pi(\mu_2 + r_2(\mu_2)|y)) + k + 0.5(\mu_2 + r_2(\mu_2) - \mu)^2/\sigma^2 \\ &< log(\pi(\mu_2 + r_2(\mu_2)|y)) + k + 0.5(r_2(\mu_2))^2/\sigma^2 \\ &= g^{\mu_2}(\mu_2 + r_2(\mu_2)) \\ &= 0. \end{aligned}$$

If $\tilde{r}_{2,1} < \tilde{r}_{2,2}$, then it follows from the continuity of $g^{\mu}()$ and the intermediate value theorem that there exists $r_2 \in (\tilde{r}_{2,1}, \tilde{r}_{2,2})$ such that $g^{\mu}(\mu + r_2) = 0$. If $\tilde{r}_{2,1} > \tilde{r}_{2,2}$ then it follows likewise that there exists $r_2 \in (\tilde{r}_{2,2}, \tilde{r}_{2,1})$ such that $g^{\mu}(\mu + r_2) = 0$. Hence the required r_2 exists. Since a symmetric argument yields the required r_1, we can conclude that $\mu \in M$ and hence M is a convex set. □

Claim 8.11. If $\mu \in M$ satisfies $\mu < \beta^*$ then $r_2(\mu) > \beta^* - \mu$. Likewise, if $\mu \in M$ satisfies $\mu > \beta^*$ then $r_1(\mu) > \mu - \beta^*$.

Proof. Consider any $\mu \in M$ satisfying $\mu < \beta^*$. It is straightforward to verify that $\pi(.|y)$ is an nondecreasing function on the interval on $[\mu, \beta^*]$ and that $q^{\mu}(.)$ is a decreasing function. Since $q^{\mu}(\mu) < \pi(\mu|y)$, it follows that $q^{\mu}(\mu + r_2(\mu)) = \pi(\mu + r_2(\mu)|y)$ requires $\mu + r_2(\mu) > \beta^*$, which is equivalent to the required inequality for $r_2(\mu)$. A symmetric argument yields the corresponding inequality for $r_1(\mu)$. □

Claim 8.12. If $\mu < \beta^*$ and $g^{\mu}(\mu) > 0$ then there exists $r_1 > 0$ such that $g^{\mu}(\mu - r_1) = 0$. Likewise, if $\mu > \beta^*$ and $g^{\mu}(\mu) > 0$ then there exists $r_2 > 0$ such that $g^{\mu}(\mu + r_2) = 0$.

Proof. Suppose $\mu < \beta^*$. It is straightforward to verify that every supergradient for $g^{\mu}(.)$ at μ is positive. It follows that for β small enough $g^{\mu}(\beta) < 0$. Pick any such $\underline{\beta}$. Then it follows

from the intermediate value theorem that there exist $\tilde{\beta} \in (\underline{\beta}, \mu)$ such that $g^{\mu}(\tilde{\beta}) = 0$. Hence $r_1 = \mu - \tilde{\beta}$ satisfies the required property. A symmetric argument yields the corresponding result for r_2. □

Claim 8.13. The set M is bounded.

Proof. We note that for any μ, $g^{\mu}(.)$ is a concave function. Hence the set $\{\beta | g^{\mu}(\beta) \geq 0\}$ is a convex set. If $\mu \in M$, then the existence of $r_1(\mu)$ and $r_2(\mu)$ implies that $g^{\mu}(\beta) \geq 0$ for all $\beta \in [\mu - r_1(\mu), \mu + r_2(\mu)]$. In particular, $g^{\mu}(\mu) \geq 0$. Hence we have that $log(\pi(\mu|y)) \geq log(q^{\mu}(\mu))$, which is equivalent to $log(\pi(\mu|y)) \geq log(q^0(0))$. Hence $M \subset \{\mu | log(\pi(\mu|y)) \geq log(q^0(0))\}$. The latter set is clearly bounded since the $\pi(.|\mathbf{y})$ is a proper density, hence M is also bounded. □

Claim 8.14. M is a non-empty, bounded, and open interval which contains β^*.

Proof. Follows from our earlier claims. □

Claim 8.15. Let $\mu_1 := inf\{\mu \in M\}$ and $\mu_2 := sup\{\mu \in M\}$ and define $h : M \to \mathbf{R}$ by $h(\mu) = r_1(\mu) - r_2(\mu)$. Then $lim_{\mu^k \to \mu_1} h(\mu^k) < 0$ and $lim_{\mu^k \to \mu_2} h(\mu^k) > 0$.

Proof. Suppose $lim_{k \to \infty} \mu^k = \mu_1$. Note that μ_1 is not in M. There are two possible cases, which we deal with separately.

Case 1: $g^{\mu_1}(\mu_1) > 0$. By our above claim, there exists $r_1 > 0$ such that $g^{\mu_1}(\mu_1 - r_1) = 0$. But we know that μ_1 is not in M, hence there can not exist $r_2 > 0$ such that $g^{\mu_1}(\mu_1 + r_2) = 0$. We note that for every k, the corresponding $r_2^k > \beta^* - \mu^k$. Hence $\lim_{k \to \infty} r_2^k \geq \beta^* - \mu_1$. If the sequence $\{r_2^k\}_{k=1}^{\infty}$ had a subsequence converging to a finite value, then the limiting value

would be an $r_2 > 0$ such that $g^{\mu_1}(\mu_1 + r_2) = 0$, an impossibility. Hence we conclude that the sequence $\{r_2^k\}_{k=1}^{\infty}$ goes to infinity. Hence $lim_{\mu^k \to \mu_1} h(\mu^k) < 0$.

Case 2: $g^{\mu_1}(\mu_1) \leq 0$. Simply note that every supergradient for $g^{\mu_1}(.)$ at μ_1 is positive. Hence for every $\beta < \mu_1$, it follows that $g^{\mu_1}(\beta) < 0$. Hence $lim_{k \to \infty} r_1^k = 0$ and $lim_{\mu^k \to \mu_1} h(\mu^k) < 0$.

This establishes the desired inequality for μ_1. The inequality for μ_2 follows using a symmetric argument. □

9 Appendix B: Technical Details For Section 3

9.1 Details For Subsection 3.1

We first note a few facts used in the proof of the below. These are all straightforward applications of Weyl's (1912) eigenvalue inequality. For details see e.g., Fisk(1996).

Fact 9.1. If B is a positive definite matrix, then $\lambda(A)_i < \lambda(A+B)_i$.

Fact 9.2. If A is a positive semi-definite matrix and B is a positive definite matrix then a constant γ^* is an eigenvalue for $A(A+B)^{-1}(A+B)^{-1}A$ and hence satisfies $|A(A+B)^{-1}(A+B)^{-1}A - \gamma^* I| = 0$ if and only if $|AA - \gamma^*(A+B)(A+B)| = 0$.

9.2 Details For Subsection 3.2

Lemma 9.1. *Consider any μ^* and μ and consider the expectation $E[\beta|X\mu^* + tX(\mu - \mu^*), y]$. Then the derivative vector for the mean vector of β with respect to t is given by*

$$dE[\beta|X\mu^* + tX(\mu - \mu^*), y]/dt \;=\; Var[\beta|X\mu^* + tX(\mu - \mu^*), y]PX(\mu - \mu^*).$$

Proof. We note that the *ith* component of the derivative vector satisfies the following:

$$
\begin{aligned}
\frac{dE[\beta_i|X\mu^*+tX(\mu-\mu^*),y]}{dt} &= \frac{d[\int \beta_i(\pi(\beta|X\mu^*+tX(\mu-\mu^*))\pi(y|\beta)/\pi(y|X\mu^*+tX(\mu-\mu^*)))d\beta]}{dt} \\
&= \textstyle\sum_{j=1}^m [\int [\beta_i(\beta_j - (X_j\mu + tX_j(\mu-\mu^*)))P_jX_j(\mu-\mu^*)\frac{\pi(\beta|X\mu^*+tX(\mu-\mu^*))\pi(y|\beta)}{\pi(y|X\mu^*+tX(\mu-\mu^*))}]d\beta \\
&\quad - \int [\beta_i \frac{\pi(\beta|X\mu^*+tX(\mu-\mu^*))\pi(y|\beta)}{\pi(y|X\mu^*+tX(\mu-\mu^*))}] \\
&\quad *[\int (\tilde{\beta}_j - (X_j\mu + tX_j(\mu-\mu^*)))P_jX_j(\mu-\mu^*)\frac{\pi(\tilde{\beta}|\mu,\sigma^2)\pi(y|\tilde{\beta})}{\pi(y|X\mu^*+tX(\mu-\mu^*))}d\tilde{\beta}]d\beta]] \\
&= \textstyle\sum_{j=1}^m [[E[\beta_i\beta_j|X\mu^* + tX(\mu-\mu^*), y] \\
&\quad - E[\beta_i|X\mu^* + tX(\mu-\mu^*), y] * (X_j\mu + tX_j(\mu-\mu^*))] \\
&\quad - \int [\beta_i \frac{\pi(\beta|X\mu^*+tX(\mu-\mu^*))\pi(y|\beta)}{\pi(y|X\mu^*+tX(\mu-\mu^*))}] \\
&\quad *[(E[\beta_j|X\mu^* + tX(\mu-\mu^*), y] - (X_j\mu + tX_j(\mu-\mu^*)))]d\beta]P_jX_j(\mu-\mu^*)] \\
&= \textstyle\sum_{j=1}^m [[E[\beta_i\beta_j|X\mu^* + tX(\mu-\mu^*), y] \\
&\quad - E[\beta_i|X\mu^* + tX(\mu-\mu^*), y] * (X_j\mu + tX_j(\mu-\mu^*))] \\
&\quad - E[\beta_i|X\mu^* + tX(\mu-\mu^*), y] * E[\beta_j|X\mu^* + tX(\mu-\mu^*), y] \\
&\quad + E[\beta_i|X\mu^* + tX(\mu-\mu^*), y] * (X_j\mu + tX_j(\mu-\mu^*))]P_jX_j(\mu-\mu^*)] \\
&= \textstyle\sum_{j=1}^m [[E[\beta_i\beta_j|X\mu^* + tX(\mu-\mu^*), y] \\
&\quad - E[\beta_i|X\mu^* + tX(\mu-\mu^*), y] * E[\beta_j|X\mu^* + tX(\mu-\mu^*), y]]P_jX_j(\mu-\mu^*)] \\
&= \textstyle\sum_{j=1}^m [Cov[\beta_i, \beta_j|X\mu^* + tX(\mu-\mu^*), y]P_jX_j(\mu-\mu^*)] \\
&= Var_{i,\cdot}[\beta|X\mu^* + tX(\mu-\mu^*), y]PX(\mu-\mu^*).
\end{aligned}
$$

□

Lemma 9.2. *Let μ^* be any point, and let $V_{\mu^*}(\mu^{(k-1)})$ be the matrix with element ij given by $\int_0^1 Cov[\beta_i, \beta_j|X_i\mu^* + tX_i(\mu^{(k-1)} - \mu^*), y]dt$. Then the expectation of $\mu^{(k)}$ given $\mu^{(k-1)}$ can be*

expressed as:

$$E[\mu^{(k)}|\mu^{(k-1)}] = (P_0 + X^T PX)^{-1}(P_0\nu + X^T PE[\beta|X\mu^*, y] + X^T PV_{\mu^*}(\mu^{(k-1)})PX(\mu^{(k-1)} - \mu^*))$$

Proof. By our two earlier claims, we have

$$\begin{aligned}
E[\mu^{(k)}|\mu^{(k-1)}] &= (P_0 + X^T PX)^{-1}(P_0\nu + X^T PE[\beta|X\mu^{(k-1)}, y]) \\
&= (P_0 + X^T PX)^{-1}(P_0\nu + X^T P(E[\beta|X\mu^*, y] + X^T P[\textstyle\int_0^1 dE[\beta|X\mu^* + tX(\mu^{(k-1)} - \mu^*), \\
&= (P_0 + X^T PX)^{-1}(P_0\nu + X^T PE[\beta|X\mu^*, y] \\
&\quad + X^T P[\textstyle\int_0^1 Var[\beta|X\mu^* + tX(\mu^{(k-1)} - \mu^*), y]PX(\mu^{(k-1)} - \mu^*)dt]) \\
&= (P_0 + X^T PX)^{-1}(P_0\nu + X^T PE[\beta|X\mu^*, y] + X^T PV_{\mu^*}(\mu^{(k-1)})PX(\mu^{(k-1)} - \mu^*))
\end{aligned}$$

□

We define matrices

$$\begin{aligned}
A^* &= P_{21}^*(P_{11}^*)^{-1}P_{12}^* \\
A(V) &= P_{21}^* V P_{12}^* \\
Q^* &= A^*(P_{22}^*)^{-1}(P_{22}^*)^{-1}A^* \\
Q(V) &= A(V)(P_{22}^*)^{-1}(P_{22}^*)^{-1}A(V).
\end{aligned}$$

We note a couple of facts which are straightforward implications of Weyl's inequality and general properties of positive semi-definite and positive definite matrices.

Fact 9.3. $A^* - A(V)$ is a positive semi-definite matrix and hence we have $\lambda(A^*)_i \geq \lambda(A(V))_i$ for all $i = 1, 2 \ldots, m$.

Fact 9.4. For every $i = 1, 2 \ldots, m$, we have $\lambda(A^*A^*)_i \geq \lambda(A(V)A(V))_i$ which implies that $A^*A^* - A(V)A(V)$ is a positive semi-definite matrix.

Fact 9.5. For any constant γ, $|A(V)A(V) - \gamma(P_{22}^* P_{22}^*)| \leq |A^* A^* - \gamma(P_{22}^* P_{22}^*)|$.

Claim 9.1. The maximal eigenvalue for $Q(V)$ is less than or equal to the maximal eigenvalue for Q^* which in turn is less than 1.

Proof. We note that as γ goes to infinity, the determinant $|A^* A^* - \gamma(P_{22}^* P_{22}^*)|$ goes to negative infinity. Continuity of the determinant as a function of γ implies that $|A^* A^* - \gamma(P_{22}^* P_{22}^*)|$ is negative for all γ greater than the maximal eigenvalue for Q^*. Fact [] above then implies that the maximal eigenvalue for $Q(V)$ must be less than or equal to the maximal eigenvalue of Q^*. That the maximal eigenvalue for Q^* is less than 1 now follows from Claim []. □

Claim 9.2. Let y be an arbitrary vector, then $y^T Q(V) y \leq \lambda^* y^T y$, where $0 \leq \lambda_* < 1$ equals the maximal eigenvalue for Q^*.

10 Appendix C: Bounded Expectations

10.1 Main Results

Let us start by introducing some notation:

N1. $\tilde{P} := \Sigma_\beta^{-1}$.

N2. P - Block diagonal matrix, where the first m blocks are equal to $\tilde{P}$ and the last block is equal to Σ_μ^{-1}.

N3. $\gamma := (\beta, \mu)$.

N4. $\gamma^*(P)$ - posterior mode if the prior precision is P. Unique by strict concavity.

N5. $P_0(\gamma)$ - Hessian for negative of log-likelihood function at γ.

N6. $A^*(P) := [P_0(\gamma^*(P)) + P]^{-0.5} P_0(\gamma^*(P))[P_0(\gamma^*(P)) + P]^{-0.5}$.

N7. $B^*(P) := [P_0(\gamma^*(P)) + P]^{-0.5} P [P_0(\gamma^*(P)) + P]^{-0.5}$.

Remark 10.1. Note that $A^*(P)$ and $B^*(P)$ are both positive semi-definite ($B^*(P)$ positive definite) and that $A^*(P) + B^*(P) = I$. This in turn implies that the diagonal elements of both matrices are bounded between 0 and 1 and that the off diagonal elements are bounded between -1 and 1. Hence $A^*(.)$ and $B^*(.)$ are both bounded matrix valued functions.

N8. $D^*(P) := P_0(\gamma^*(P))^{0.5} [P_0(\gamma^*(P)) + P]^{-0.5}$.

Remark 10.2. Note that $A^*(P) = D^*(P)^T D^*(P)$. It is straightforward to verify that the diagonal elements of $A^*(.)$ are bounded only if $D^*(.)$ is bounded. Hence it follows from our above remark that $D^*(.)$ is bounded.

N10. PD - Set of positive definite matrices $p \times p$ matrices.

N11. $C := \{\gamma \in \mathbf{R}^{m1*p+L} | g(\gamma) \geq g(\mathbf{0})\}$. Closed by log-concavity, compactness by assumption.

Remark 10.3. For every P, $\gamma^*(P)$ is in the set C. Hence the function $\gamma^*(.)$ is a bounded function.

N12. $\underline{\gamma}^*(P) := (P_0(\gamma^*(P)) + P)^{0.5} \gamma^*(P)$.

Claim 10.1. $\underline{\gamma}^*(.)$ is a bounded function.

Proof. See subsection 10.2.1. □

N11. $\mathcal{A} := \{(D, \gamma, \underline{\gamma}) | \exists \tilde{P} \in PD : D = D^*(P(\tilde{P})), \gamma = \gamma^*(P(\tilde{P})), \underline{\gamma} = \underline{\gamma}^*(P(\tilde{P}))\}$.

N12. $\underline{\mathcal{A}}$ closure of $\mathcal{A}$.

Remark 10.4. $\underline{\mathcal{A}}$ is a compact set since the functions $D^*(.)$, $\gamma^*(.)$, and $\underline{\gamma}^*(.)$ are all bounded. Moreover, for any $(D, \gamma, \underline{\gamma}) \in \underline{\mathcal{A}}$, there exists a sequence $\{(D^k, \gamma^k, \underline{\gamma}^k)\}_{k=1}^{\infty} \in \mathcal{A}$ which converges to $(D, \gamma, \underline{\gamma})$. Likewise, there is a sequence $\{P^k\}_{k=1}^{\infty}$ such that $D^k = D^*(P^k)$, $\gamma^k = \gamma^*(P^k)$, and $\underline{\gamma}^k = \underline{\gamma}^*(P^k)$ for all k.

We are now ready to outline the proof. Details for some of the steps are in subsubsections 10.2.2, 10.2.3, and 10.2.4.

Step 1: Show that for any sequences $\{(D^k, \beta^k, \underline{\beta}^k)\}_{k=1}^{\infty} \in \mathcal{A}$ and $\{P^k\}_{k=1}^{\infty}$ as above, the limiting function h below is well defined with well defined continuous and gradient functions for the log of h. Furthermore, log h(.) is a concave function and has a unique mode. For details of this step, see subsubsection 10.2.2.

$$h(\underline{\gamma}) = Lim_{k\to\infty}\pi(y|(P_0(\gamma^*(P_k))+P_k)^{-0.5}\underline{\gamma})exp[-0.5((P_0(\gamma^*(P_k))+P_k)^{-0.5}\underline{\gamma})^T P_k((P_0(\gamma^*(P_k))+P_k)^{-0.5}\underline{\gamma})]$$

Step 2: Show that the function h above is Lebesgue integrable for each element $(D, \gamma, \underline{\gamma}) \in \underline{\mathcal{A}}$. Note that this implies that (viewed as a function of the elements of $\underline{\mathcal{A}}$), the value of the integral is bounded (follows from being a continuous function on a compact set). For details of this step, see subsubsection 10.2.3.

Step 3: Show that the value of the integral (again viewed as a function of the elements of $\underline{\mathcal{A}}$) is bounded below by a strictly positive constant. For details of this step, see subsubsection 10.2.4.

Step 4: Note that steps 2 and 3 implies that for any element in $\underline{\mathcal{A}}$, the corresponding function h can be normalized into a proper log-concave density, where the normalizing constant is a continuous function on the set $\underline{\mathcal{A}}$ which obtains both its minimum and its maximum.

Step 5: Note (using the results due to Prekopa (1973) and An (1998)) that all marginal densities of log-concave densities are log-concave and that the moments of one-dimensional log-concave densities are finite. Conclude that the moments of the normalized densities corresponding to $h(.)$ are continuous functions of the elements of $\underline{\mathcal{A}}$ that obtain both their maximal and minimal values on the set $\underline{\mathcal{A}}$.

Step 6: Let $\{(D^k, \gamma^k, \underline{\gamma}^k)\}_{k=1}^{\infty} \in \mathcal{A}$ and $\{P^k\}_{k=1}^{\infty}$ be sequences as above. Note that for each k,

$$\begin{aligned} E[\gamma] &= (P_0(\gamma^*(P_k)) + P_k)^{-0.5} E[\underline{\gamma}] \\ &= [(P_0(\gamma^*(P_k)) + P_k)^{-1}(P_0(\gamma^*(P_k)))(P_0(\gamma^*(P_k)))^{-1}]^{0.5} E[\underline{\gamma}] \\ &= [A_k(P_0(\gamma^*(P_k)))^{-1} E[\underline{\gamma}] \end{aligned}$$

Since A_k, $(P_0(\gamma^*(P_k)))^{-1}$, and $E[\underline{\gamma}]$ are bounded functions on the set $\underline{\mathcal{A}}$, it follows that $E[\gamma]$ also is bounded. The existence of the bound L then follows immediately.

10.2 Technical Details

10.2.1 Boundedness of $\underline{\gamma}^*(P)$.

Proof. Note that $\underline{\gamma}^*(P)$ maximizes the Taylor series approximation

$$[Log(\pi(y|\gamma^*(P))) - 0.5\gamma^*(P)^T P\gamma^*(P)] + [\nabla Log(\pi(y|\gamma^*(P)))]^T[(P_0(\gamma^*(P)) + P)^{-0.5}\underline{\gamma} - \gamma^*]$$

$$-[P\gamma^*(P)]^T[(P_0(\gamma^*(P)) + P)^{-0.5}\underline{\gamma} - \gamma^*(P)]$$

$$-0.5[(P_0(\gamma^*(P)) + P)^{-0.5}\underline{\gamma} - \gamma^*(P)]^T[(P_0(\gamma^*(P)) + P)][(P_0(\gamma^*(P)) + P)^{-0.5}\underline{\gamma} - \gamma^*(P)].$$

This in turns implies that $\underline{\gamma}^*(P)$ must satisfy the first order condition

$$\begin{aligned}
(P_0(\gamma^*(P)) + P)^{-0.5}[[\nabla Log(\pi(\mathbf{y}|\gamma^*(P)))] - [P\gamma^*(P)] & \\
-[(P_0(\gamma^*(P)) + P)][(P_0(\gamma^*(P)) + P)^{-0.5}\underline{\gamma}^*(P) - \gamma^*(P)]] &= 0 \\
\Updownarrow & \\
\underline{\gamma}^*(P) = (P_0(\gamma^*(P)) + P)^{-0.5}[\nabla Log(\pi(\mathbf{y}|\gamma^*(P))) - P\gamma^*(P)] & \\
+(P_0(\gamma^*(P)) + P)^{-0.5}(P_0(\gamma^*(P)) + P)\gamma^*(P) & \\
\Updownarrow & \\
\underline{\gamma}^*(P) = (P_0(\gamma^*(P)) + P)^{-0.5}P_0(\gamma^*(P))^{0.5}P_0(\gamma^*(P))^{-0.5}\nabla Log(\pi(\mathbf{y}|\gamma^*(P))) & \\
+(P_0(\gamma^*(P)) + P)^{-0.5}(P_0(\gamma^*(P)))^{0.5}(P_0(\gamma^*(P)))^{0.5}\gamma^*(P) &
\end{aligned}$$

Now simply note that the right hand side is bounded since each of the two terms is a matrix product of bounded matrix functions (including $\nabla Log(\pi(\mathbf{y}|\gamma^*(P)))$). □

10.2.2 Details Step 1

Here we provide details around Step 1

1A: The limiting function h is given explicitly by

$$\begin{aligned}
& Lim_{k\to\infty}h_k(\underline{\beta}) \\
= \; & Lim_{k\to\infty}\pi(\mathbf{y}|(P_0(\gamma^*(P_k)) + P_k)^{-0.5}\underline{\gamma})exp[-0.5((P_0(\gamma^*(P_k)) + P_k)^{-0.5}\underline{\gamma})^T P_k((P_0(\gamma^*(P_k)) + P_k)^{-0.5} \\
= \; & \pi(\mathbf{y}|D^T P_0(\gamma)^{-0.5}\underline{\gamma})exp(-0.5\underline{\gamma}^T(I - D^T D)\underline{\gamma}).
\end{aligned}$$

1B: The gradient of the log of the function $h()$ is explicitly given by

$$Lim_{k\to\infty}\nabla Logh_k(\underline{\gamma}) = D^T P_0(\gamma)^{-0.5}\nabla Log(\pi(\mathbf{y}|(P_0(\gamma^*(P))) + P)^{-0.5}\underline{\gamma}) - [I - D^T D]\underline{\gamma}$$

1C: The Hessian of the log of the function $h()$ is explicitly given by

$$Lim_{k\to\infty} H^{log(h_k())}(\underline{\gamma}) = D^T P_0(\gamma)^{-0.5}[H^{log(\pi(\mathbf{y}|.))}(D^T P_0(\gamma)^{-0.5}\underline{\gamma})]P_0(\gamma)^{-0.5}D - [I - D^T D]$$

1D: Note that concavity for h follows since h is the limiting function of a series of strictly concave functions.

1E: Note that the gradient of the log of $h()$ must vanish when $D^T P_0(\gamma)^{-0.5}\underline{\gamma} = \gamma$ since it does so for every element of the sequence.

1F: Show that the Hessian of the log of $h()$ is equal $-I$ when $D^T P_0(\gamma)^{-0.5}\underline{\gamma} = \gamma$ [Simply substitute $-P_0(\gamma)$ for $H^{log(\pi(\mathbf{y}|.))}(D^T P_0(\gamma)^{-0.5}\underline{\gamma})$ in expression for Hessian and simplify]. Conclude that h is strictly concave at that point.

1G: Note that it follows from the differentiability assumption that the function log $h()$ is strictly concave in some neighborhood of the point for which $D^T P_0(\gamma)^{-0.5}\underline{\gamma} = \gamma$.

1H: Show that strict concavity in some neighborhood of the maximum together with overall concavity implies that the maximum is unique.

10.2.3 Details Step 2

Here we provide details around step 2.

2A: Note that the set $B := \{\underline{\gamma} | \|\underline{\gamma}^* - \underline{\gamma}\| = 1\}$ (where $\underline{\gamma}^*$ is the unique mode) is a compact set.

2B: Note that $c := min\{r | \exists \underline{\gamma} \in B : log[h(\underline{\gamma}^*)] - log[h(\underline{\gamma})] = r\}$ exists and is strictly positive.

2C: Show that concavity implies that if $\|\underline{\gamma}^* - \underline{\gamma}\| \geq 1$ then $log[h(\underline{\gamma}^*)] - log[h(\underline{\gamma})] \geq c\|\underline{\gamma}^* - \underline{\gamma}\|$. The latter inequality can also be reformulated as $h(\underline{\gamma}) \leq h(\underline{\gamma}^*)exp(-c\|\underline{\gamma}^* - \underline{\gamma}\|)$.

2D: Show that there exists a constant A such that

$$\begin{aligned}\int exp(-c\|\underline{\gamma}^* - \underline{\gamma}\|)d\underline{\gamma} &\leq A + \sum_{k=1}^{\infty} exp(-ck)\{[2\pi^{n/2}/\Gamma((1/2)n)][1/n][(k+1)^n] - [2\pi^{n/2}/\Gamma((1/2)n)][1 \\ &= A + [2\pi^{n/2}/\Gamma((1/2)n)][1/n]\sum_{k=1}^{\infty} exp(-ck)\{(k+1)^n - k^n\}\end{aligned}$$

Note: Terms in the sum above represents differences in volume between hyper-spheres multiplied by an upper bound for the value of $exp(-c\|\underline{\gamma}^* - \underline{\gamma}\|)$ over the hypervolume represented by that difference [Show 2-dimensional picture to illustrate]

2E: Define $h(k) := exp(-ck)\{(k+1)^n - k^n\}$ and show that the derivative of $h()$ is given by

$$\begin{aligned}dh(k)/dk &= -cexp(-ck)\{(k+1)^n - k^n\} + exp(-cl)\{n(k+1)^{n-1} - nk^{n-1}\} \\ &= exp(-ck)\{(k+1)^n[(n/(k+1)) - c] - k^n[(n/(k)) - c]\} \\ &< exp(-ck)\{(k+1)^n[(n/(k+1)) - c] - k^n[(n/(k+1)) - c].\}\end{aligned}$$

2F: Note that $dh(k)/dk$ is negative whenever $k > (n-c)/c$.

2G: Let $k^* > (n-c)/c$ be an integer and define a new function

$$h_2(k_2) := exp(-c(k_2 + k^*))\{((k_2 + k^*) + 1)^n - (k_2 + k^*)^n\}$$

and note that $dh_2(k_2)/dk_2$ is negative whenever $k_2 > 0$.

2H: Show that $\int_{k_2>0} h_2(k_2)dk_2$ is finite.

2I: Conclude (using the integral test) that the sequence $\sum_{k2=1}^{\infty} h_2(k_2)$ is finite.

2J: Conclude that the sequence $\sum_{k=1}^{\infty} h(k)$ as well as the integrals $\int exp(-c\|\underline{\gamma}^* - \underline{\gamma}\|)d\underline{\gamma}$ and $\int h(\underline{\gamma})d\underline{\gamma}$ are also bounded.

10.2.4 Details of Step 3

Here we provide details for Step 3.

3A: Note that

$$\begin{aligned} h(\underline{\beta}) &= \pi(\mathbf{y}|D^T P_0(\gamma^*(P))^{-0.5}\underline{\gamma})exp(-0.5\underline{\gamma}^T(I - D^T D)\underline{\gamma}) \\ &\geq \pi(\mathbf{y}|D^T P_0(\gamma^*(P))^{-0.5}\underline{\gamma})exp(-0.5\underline{\gamma}^T I\underline{\gamma}) \end{aligned}$$

3B: Note that $\pi(\mathbf{y}|D^T P_0(\gamma)^{-0.5})exp(-0.5\underline{\gamma}^T(I-D^T D)\underline{\gamma})$ is a strictly positive function. Consider a compact set $\mathcal{L}$ around 0. Then $\pi(\mathbf{y}|D^T P_0(\gamma)^{-0.5}\underline{\gamma})$ is a bounded from below on the set L (viewed as a function of both $\underline{\gamma}$ and the elements in $\underline{\mathcal{A}}$. Hence on the set $\mathcal{L}$, we have a strictly positive lower bound $f_{\mathcal{L}}$ for $\pi(\mathbf{y}|D^T P_0(\gamma)^{-0.5}\underline{\gamma})$. Hence $\int h(\underline{\gamma})d\underline{\gamma} > \int_{\mathcal{L}} h(\underline{\gamma})d\underline{\gamma} >= \int_{\mathcal{L}} f_{\mathcal{L}} exp(-0.5\underline{\gamma}^T I\underline{\gamma})d\underline{\gamma} > 0$. Q.E.D.

11 Appendix D: Determinant Inequality

We first state a couple of remarks for positive semi-definite matrices A and B and a theorem due to Fisk (1996) for subspaces of n-dimensional vector spaces.

Remark 11.1. The matrices $I + A + B$, AB, and $(I + A)(I + B)$ have real non-negative eigenvalues (follows from properties of positive semidefinite matrices and properties of products of positive semi-definite matrices).

Remark 11.2. The matrices $I + A + B$, AB, and $(I + A)(I + B)$ have corresponding real right and left eigenvectors.

Theorem 11.1 (Fisk 1996)**.** *If $S_1, S_2, \ldots S_k$ are subspaces of an n-dimensional vector space*

V and if $dim(S_1) + dim(S_2) + \ldots + dim(S_k) \leq n(k-1)$ then the intersection of all the S_i's has dimension greater than zero.

For an arbitrary $n \times n$ matrix C with real eigenvalues, we denote its ordered eigenvalues by $\lambda_1(C) \leq \lambda_2(C) \geq \ldots \leq \lambda_n(C)$.

Claim 11.1 (Weyl Type Inequality). If A and B are positive semidefinite matrices, then $\lambda_i(I + A + B) + \lambda_j(AB) \leq \lambda_{i+j-1}((I + A)(I + B))$ for any $i, j \in \{1, 2, \ldots, n\}$ such that $i+j \leq n+1$. In particular, it follows that for any $i \in \{1, 2, \ldots, n\}$, $\lambda_i(I+A+B)+\lambda_1(AB) \leq \lambda_i((I + A)(I + B))$.

Proof. Our proof closely follows the proof of Theorem 1 in Fisk (1996). Set $C_1 = I + A + B$, $C_2 = AB$, $C_3 = -(I + A)(I + B)$ and set $i_1 = i$, $i_2 = j$, $i_3 = n + 1 - (i + j - 1)$. Finally, let S_j denote the subspace spanned by the right eigenvectors corresponding to the eigenvalues $\lambda_{i_j}(C_j), \lambda_{i_{j+1}}(C_j), \ldots, \lambda_n(C_j)$. Then

$$\begin{aligned} \textstyle\sum_{j=1}^{3} dim(S_j) &= \textstyle\sum_{j=1}^{3}(n + 1 - i_j) \\ &= n * 3 + 3 - i - j - (n + 2 - i - j) \\ &= n * 3 + 3 - (n + 2) \\ &> n * (3 - 1). \end{aligned}$$

Fisk's Lemma 1 now ensures that there exists a unit vector x in the intersection of all the S_j's. Now, $\lambda_{i_j}(C_j)$ is the smallest eigenvalue of C_j restricted to S_j and therefore,

$$\lambda_{i_j}(C_j) \leq x^T C_j x$$

since each S_j is invariant under C_j for $j = 1, 2, 3$. Since $C_1 + C_2 + C_3 = \mathbf{0}$, it follows that

$$\lambda_{i_1}(C_1) + \lambda_{i_2}(C_2) + \lambda_{i_3}(C_3) \leq 0$$

.

From the definitions of C_1, C_2, and C_3 combined with the fact that $\lambda_{i+j-1}((I+A)(I+B)) = -\lambda_{i_3}(C_3)$, it follows that

$$\lambda_i(I+A+B) + \lambda_j(AB) \leq \lambda_{i+j-1}((I+A)(I+B)).$$

□

The following claim is key to our proof of the minorization condition.

Claim 11.2. Let A and B be positive semidefinite matrices. Then $|I+A|*|I+B| \geq |I+A+B|$.

Proof. Using the properties in Remark 11.1 and Claim 11.1, we have

$$\begin{aligned} |I+A||I+B| &= \textstyle\prod_{i=1}^n \lambda_i((I+A)(I+B)) \\ &\geq \textstyle\prod_{i=1}^n (\lambda_i(I+A+B) + \lambda_1(AB)) \\ &\geq \textstyle\prod_{i=1}^n \lambda_i(I+A+B) \\ &= |I+A+B| \end{aligned}$$

□

References

[1] Rosenthal, J.S. (1995), "Minorization Conditions and Convergence rates for Markov Chain Monte Carlo," *Journal of the American Statistical Association*, 90, 558-566. Correction, p. 1136.

[2] Cowles, M.K. and Rosenthal, J.S. (1998), "A simulation approach to convergence rates for Markov Chain Monte Carlo Algorithms," *Statistics and Computing*, 115-124.

[3] Jones, G.L. and Hobert, J.P. (2004), "Sufficient Burn-In for Gibbs Samplers for a Hierarchical Random Effects Model," *The Annals of Statistics*, 32, No. 2, 784-817.

[4] Hobert, J.P. and Geyer, C.J. (1998), "Geometric ergodicity of Gibbs and block Gibbs samplers for a hierarchical random effects model," *The Annals of Statistics*, 32, No. 2, 784-817.

[5] Roberts, G.O. and Tweedie, R.L. (1999), "Bounds on regeneration times and convergence rates for Markov Chains," *Stochastic Process. Appl*, 80, 211-229.

[6] Roberts, G.O. and Tweedie, R.L. (1999), Corrigendum to "Bounds on regeneration times and convergence rates for Markov Chains," *Stochastic Process. Appl*, 91, 337-338.

[7] Prekopa, (1973), "Bounds on regeneration times and convergence rates for Markov Chains," *Stochastic Process. Appl*, 91, 337-338.

[8] Fisk, (1996), "Bounds on regeneration times and convergence rates for Markov Chains," *Stochastic Process. Appl*, 91, 337-338.

[9] An, (1996), "Bounds on regeneration times and convergence rates for Markov Chains," *Stochastic Process. Appl*, 91, 337-338.

[10] Roberts, G.O, and Tweedie, R.L. (1996), "Geometric convergence and central limit theorems for multidimensional Hastings and Metropolis Algorithms," *Biometrica*, 83, 95-110.

[11] Roberts, G.O, "Rates of Convergence of for Gibbs Sampling for variance component models," *The Annals of Statistics*, 23, 740-761.