Electronic Journal of Statistics
Vol. 11 (2017) 4033–4064
ISSN: 1935-7524
DOI: 10.1214/17-EJS1351

Geometric ergodicity of Gibbs samplers in Bayesian penalized regression models

Dootika Vats

Department of Statistics
University of Warwick
Coventry, CV4 7AL
e-mail: D.Vats@warwick.ac.uk

Abstract: We consider three Bayesian penalized regression models and show that the respective deterministic scan Gibbs samplers are geometrically ergodic regardless of the dimension of the regression problem. We prove geometric ergodicity of the Gibbs samplers for the Bayesian fused lasso, the Bayesian group lasso, and the Bayesian sparse group lasso. Geometric ergodicity along with a moment condition results in the existence of a Markov chain central limit theorem for Monte Carlo averages and ensures reliable output analysis. Our results of geometric ergodicity allow us to also provide default starting values for the Gibbs samplers.

MSC 2010 subject classifications: Primary 60J05; secondary 62F15.
Keywords and phrases: Markov chains, geometric ergodicity, Bayesian lassos, starting values.

Received September 2016.

1. Introduction

Let $y \in \mathbb{R}^n$ be the observed realization of the response Y, X be the $n \times p$ model matrix, and $\beta \in \mathbb{R}^p$ be the regression coefficient vector. The goal, generally, is to identify important predictors amongst the p covariates and estimate the corresponding coefficients in β. However, for many problems in genetics, image processing, chemometrics, economics etc, the number of covariates, p, can be much larger than n, making it difficult to use classical regression techniques. Bayesian and frequentist penalization methods have been useful in such situations. Consider the Bayesian regression model of the form

$$\begin{aligned} Y \mid \beta, \sigma^2 &\sim N_n(X\beta, \sigma^2 I_n) \\ \beta \mid \eta, \sigma^2 &\sim N_p(0, \sigma^2 \Sigma_\eta) \\ \eta &\sim p(\eta) \\ \sigma^2 &\sim \text{Inverse-Gamma}(\alpha, \xi)\,, \end{aligned} \tag{1.1}$$

where $\alpha, \xi \geq 0$ are assumed known, Σ_η is a $p \times p$ covariance matrix determined by $\eta \in \mathbb{R}^s_+$, and $p(\eta)$ is a proper prior on η. Many Bayesian penalized regression and variable selection models can be presented in this framework (see for example [11, 20, 25, 37]). The resulting posteriors are often intractable and Markov chain Monte Carlo (MCMC) is used to estimate model parameters.

Consider the Bayesian fused lasso, the Bayesian group lasso [20], and the Bayesian sparse group lasso [36], all three of which belong to the family of models in (1.1). The Bayesian group lasso and the Bayesian sparse group lasso find use in medical research [5, 10, 23, 27]. The Bayesian fused lasso has been used in breast cancer research [39]. Given the use of these models in medical research, reliable inference is essential.

Reliable estimation from MCMC output rests heavily on the rate of convergence of the Markov chain. In particular, a geometric rate of convergence lets users appeal to the Markov chain central limit theorem (CLT), allowing for the estimation of Monte Carlo error in posterior estimates and consistent estimation of effective sample size. We show that the MCMC samplers used in the three models converge to their respective stationary distribution at a geometric rate. That is, we show that the Gibbs samplers are *geometrically ergodic* (formal definitions are in Section 2).

In the models we study, the full conditionals for β, η, and σ^2 are available in closed form so that it is straightforward to draw samples from $f(\beta \mid \eta, \sigma^2, y)$, $f(\eta \mid \beta, \sigma^2, y)$, and $f(\sigma^2 \mid \beta, \eta, y)$. As a consequence, a three variable deterministic scan Gibbs sampler is implemented to draw approximate samples from the intractable posterior distribution and inference is done using sample statistics. The quality of estimation is affected not only by the size of the Monte Carlo sample, but also by the rate of convergence of the Gibbs sampler. We show that all three Gibbs samplers converge to their respective stationary distribution at a geometric rate under reasonable conditions. Specifically, we only require the number of observations, n, to be larger than three and require no assumptions on the number of covariates, p or the model matrix X. This geometric rate of convergence allows for reliable estimation of posterior quantities in the following way.

Let F denote the posterior distribution of (β, η, σ^2) obtained from (1.1), defined on the space $\mathsf{X} = \mathbb{R}^p \times \mathbb{R}^s_+ \times \mathbb{R}_+$ and let $f(\beta, \eta, \sigma^2 \mid y)$ be the associated density. Let $g : \mathsf{X} \to \mathbb{R}^d$ be an F-integrable function, then interest is in estimating

$$\theta := \int_{\mathsf{X}} g(\beta, \eta, \sigma^2)\ f(\beta, \eta, \sigma^2 \mid y) d\beta\, d\eta\, d\sigma^2 < \infty\,.$$

Typically θ represents means, variance or quantiles of the posterior distribution. For $t = 0, 1, 2, \ldots$, let $(\beta^{(t)}, \eta^{(t)}, \sigma^{2(t)})$ be the samples obtained using a Harris ergodic Gibbs sampler. Then, with probability 1, for every $(\beta^{(0)}, \eta^{(0)}, \sigma^{2(0)}) \in \mathsf{X}$

$$\theta_N := \frac{1}{N} \sum_{t=0}^{N-1} g(\beta^{(t)}, \eta^{(t)}, \sigma^{2(t)}) \to \theta \quad \text{ as } N \to \infty\,.$$

However, in finite samples there is typically a non-zero Monte Carlo error $\theta_N - \theta$ and an approximate sampling distribution of this error may be available via a Markov chain CLT. Let $\|\cdot\|$ denote the Euclidean norm. If the deterministic scan Gibbs sampler is geometrically ergodic and

$$\int_{\mathsf{X}} \left\|g(\beta, \eta, \sigma^2)\right\|^{2+\delta} f(\beta, \eta, \sigma^2 \mid y) d\beta\ d\eta\ d\sigma^2 < \infty\,,$$

then a Markov chain CLT holds as below:

$$\sqrt{n}(\theta_N - \theta) \xrightarrow{d} N_d(0, \Sigma) \quad \text{as } N \to \infty\,, \tag{1.2}$$

where Σ is the $d \times d$ asymptotic covariance matrix that is difficult to calculate due to the serial correlation in the Markov chain. However, if the process is geometrically ergodic, then [34] and [35] provide strongly consistent estimators of Σ. This leads to the construction of asymptotically valid confidence ellipsoids around θ_N and consistent estimation of effective sample size [34]. Under the assumption of geometric ergodicity, the diagonals of Σ were estimated by [6], [7], [8], [12], and [17] leading to reliable univariate analysis of MCMC output. For estimating quantiles, [4] show that geometric ergodicity leads to strongly consistent estimators of the Monte Carlo error.

There has been a considerable amount of work done in establishing geometric ergodicity of Gibbs samplers; many of which are two variable Gibbs samplers. Two variable Gibbs samplers are special because the marginal process for each variable is a Markov chain with the same rate of convergence as the joint chain. Thus, it is sufficient to study the marginal chains to ascertain the properties of the joint chain. Higher variable Gibbs samplers do not benefit from this property and thus studying their rate of convergence is often more challenging. Geometric ergodicity of the three variable Gibbs samplers in the Bayesian lasso and the Bayesian elastic net were shown by [19] and [31], respectively; [24] proved geometric ergodicity of the three variable Gibbs sampler for the normal-gamma model of [9]; [18] demonstrated geometric ergodicity of the three variable Gibbs sampler in Bayesian quantile regression, and [3] and [16] demonstrated geometric ergodicity of the three variable Gibbs sampler in hierarchical random effects models. Recently, [14] established geometric ergodicity of a four variable random scan Gibbs sampler for a hierarchical random effects model.

The rest of the paper is organized as follows. In Section 2 we present important definitions and some relevant Markov chain background. In Section 3, Section 4, and Section 5 we present the models and main results for the Bayesian fused lasso, Bayesian group lasso, and the Bayesian sparse group lasso. We finish with a discussion in Section 6. All proofs are deferred to the appendices.

2. Markov chain background

Recall that F denotes the posterior distribution of (β, η, σ^2) obtained from (1.1) and $f(\beta, \eta, \sigma^2 \mid y)$ is the associated density. Also recall that $\mathsf{X} = \mathbb{R}^p \times \mathbb{R}^s_+ \times \mathbb{R}_+$ is the support of the posterior and let $\mathcal{B}(\mathsf{X})$ denote the Borel σ-algebra. Let $f(\beta \mid \eta, \sigma^2, y)$ be the density of the full conditional distribution of β and similarly denote the densities of the conditional distributions of η and σ^2 with $f(\eta \mid \beta, \sigma^2, y)$ and $f(\sigma^2 \mid \beta, \eta, y)$, respectively. Let $(\beta^{(0)}, \eta^{(0)}, \sigma^{2(0)})$ be the starting value for the Gibbs sampler and define the Markov chain transition density (MTD) for the deterministic scan Gibbs sampler as

$$k\left((\beta^{(1)},\eta^{(1)},\sigma^{2(1)}) \mid (\beta^{(0)},\eta^{(0)},\sigma^{2(0)})\right) = f(\beta^{(1)} \mid \eta^{(1)},\sigma^{2(1)},y) \times f(\eta^{(1)} \mid \beta^{(0)},\sigma^{2(1)},y) \times f(\sigma^{2(1)} \mid \beta^{(0)},\eta^{(0)},y)\,.$$

Then, the one-step transition kernel $P : \mathsf{X} \times \mathcal{B}(\mathsf{X}) \to [0,1]$ is such that for any $A \in \mathcal{B}(\mathsf{X})$,

$$\begin{aligned} &P\left((\beta^{(0)},\eta^{(0)},\sigma^{2(0)}),A\right) \\ &= \Pr\left((\beta^{(1)},\eta^{(1)},\sigma^{2(1)}) \in A \mid (\beta^{(0)},\eta^{(0)},\sigma^{2(0)})\right) \\ &= \int_A k\left((\beta^{(1)},\eta^{(1)},\sigma^{2(1)}) \mid (\beta^{(0)},\eta^{(0)},\sigma^{2(0)})\right) d\beta^{(1)}\, d\eta^{(1)}\, d\sigma^{2(1)}\,. \end{aligned}$$

Similarly, the t-step Markov chain transition kernel for the deterministic scan Gibbs sampler is $P^t : \mathsf{X} \times \mathcal{B}(\mathsf{X}) \to [0,1]$ such that for all $A \in \mathcal{B}(\mathsf{X})$,

$$P^t\left((\beta^{(0)},\eta^{(0)},\sigma^{2(0)}),A\right) = \Pr\left((\beta^{(t)},\eta^{(t)},\sigma^{2(t)}) \in A \mid (\beta^{(0)},\eta^{(0)},\sigma^{2(0)})\right)\,.$$

Let $\|\cdot\|_{TV}$ denote total variation norm. If the Markov chain is aperiodic, irreducible, and Harris recurrent (see [22, 29] for definitions), then for every $(\beta^{(0)},\eta^{(0)},\sigma^{2(0)}) \in \mathsf{X}$

$$\left\|P^t\left((\beta^{(0)},\eta^{(0)},\sigma^{2(0)}),\cdot\right) - F(\cdot)\right\|_{TV} \to 0 \text{ as } t \to \infty\,.$$

However, convergence of the transition kernel to the invariant distribution is not sufficient to ensure reliable inference and a geometric rate of convergence is often required. The Gibbs sampler is *geometrically ergodic* if there exists a function $M : \mathsf{X} \to [0,\infty)$ and $0 \le \rho < 1$ such that for all $(\beta^{(0)},\eta^{(0)},\sigma^{2(0)}) \in \mathsf{X}$,

$$\left\|P^t\left((\beta^{(0)},\eta^{(0)},\sigma^{2(0)}),\cdot\right) - F(\cdot)\right\|_{TV} \le M\left((\beta^{(0)},\eta^{(0)},\sigma^{2(0)})\right)\rho^t\,. \tag{2.1}$$

Since $\rho < 1$, the upper bound in (2.1) decreases at a geometric rate as a function of t. We will show that the three Gibbs samplers are geometrically ergodic by establishing a *drift condition* and an associated *minorization condition*. In effect, we will determine M up to a proportionality constant and minimize this quantity to arrive at default starting values for the Gibbs samplers. Our results can also be used to obtain quantitative upper bounds for (2.1) using the results of [30]; we do not explore that here.

Geometric ergodicity is often demonstrated by establishing a drift condition and an associated minorization condition. A drift condition is said to hold if there exists a function $V : \mathsf{X} \to [0,\infty)$, and constants $0 < \phi < 1$ and $L < \infty$ such that for all $(\beta_0,\eta_0,\sigma_0^2) \in \mathsf{X}$,

$$\mathrm{E}\left[V(\beta,\eta,\sigma^2) \mid \beta_0,\eta_0,\sigma_0^2\right] \le \phi V(\beta_0,\eta_0,\sigma_0^2) + L\,. \tag{2.2}$$

In (2.2), the expectation is with respect to the MTD for the Gibbs sampler.

Consider for $d > 0$, the set $C_d = \{(\beta, \eta, \sigma^2) : V(\beta, \eta, \sigma^2) \leq d\}$. A minorization condition holds if there exists an $\epsilon > 0$ and a distribution Q such that for all $(\beta_0, \eta_0, \sigma_0^2) \in C_d$

$$P\left((\beta_0, \eta_0, \sigma_0^2), \cdot\right) \geq \epsilon\, Q(\cdot)\,. \tag{2.3}$$

It is well know that both (2.2) and (2.3) together imply geometric ergodicity (see [15] and [22]). The *drift rate* ϕ determines how fast the Markov chain drifts back to the *small set* C_d. A drift rate close to one signifies slower convergence and a smaller value indicates faster convergence. See [15] for a heuristic explanation.

When a drift condition holds, [22], [28], and [29, Fact 10] explain that the function M is proportional to the *drift function* V up to an unknown constant. Thus, minimizing V over the state space leads to the tightest bound in (2.1) for our choice of V. This will lead us to default starting values for the three Gibbs samplers.

3. Bayesian fused lasso

Recall that $y \in \mathbb{R}^n$ is the observed realization of the response Y, X is the $n \times p$ model matrix, and $\beta \in \mathbb{R}^p$ is the regression coefficient vector. Tibshirani et al. [33] proposed the fused lasso in an effort to account for ordering in the predictors. In addition to penalizing the L_1 norm of the coefficients, the fused lasso also penalizes pairwise differences. That is, for tuning parameters $\lambda_1, \lambda_2 > 0$, the fused lasso estimate is,

$$\hat{\beta}_{\text{fused}} = \arg\max_{\beta} \|y - X\beta\|^2 + \lambda_1 \sum_{j=1}^{p} |\beta_j| + \lambda_2 \sum_{j=1}^{p-1} |\beta_{j+1} - \beta_j|\,. \tag{3.1}$$

A Bayesian formulation of the fused lasso requires a prior on β so that the resulting posterior mode is $\hat{\beta}_{\text{fused}}$. Kyung et al. [20] present the following Bayesian formulation of the fused lasso. Let

$$\begin{aligned}
Y \mid \beta, \sigma^2, \tau^2 &\sim N_n(X\beta, \sigma^2 I_n) \\
\beta \mid \tau^2, w^2, \sigma^2 &\sim N_p(0, \sigma^2\, \Sigma_{\tau,w}) \\
\tau_i^2 &\overset{\text{ind}}{\sim} \frac{\lambda_1^2}{2} e^{-\lambda_1 \tau_i^2/2} d\tau_i^2 \quad \text{for } \tau_i^2 > 0, i = 1, \ldots, p \\
w_i^2 &\overset{\text{ind}}{\sim} \frac{\lambda_2^2}{2} e^{-\lambda_2 w_i^2/2} dw_i^2 \quad \text{for } w_i^2 > 0, i = 1, \ldots, p-1 \\
\sigma^2 &\sim \text{Inverse-Gamma}(\alpha, \xi)\,,
\end{aligned} \tag{3.2}$$

where $\alpha, \xi \geq 0$ are known, $\lambda_1, \lambda_2 > 0$ are fixed, and $\Sigma_{\tau,w}$ is such that $\Sigma_{\tau,w}^{-1}$ is a tridiagonal matrix with main diagonals

$$\left(\frac{1}{\tau_1^2} + \frac{1}{w_1^2}\right),\ \left(\frac{1}{\tau_i^2} + \frac{1}{w_{i-1}^2} + \frac{1}{w_i^2}\right) \text{ for } i = 2, \ldots, p-1, \text{ and } \left(\frac{1}{\tau_p^2} + \frac{1}{w_{p-1}^2}\right),$$

and off diagonals $\{-1/w_i^2 : i = 1, \ldots, p\}$. Specifically,

$$\Sigma_{\tau,w}^{-1} = \begin{bmatrix} \frac{1}{\tau_1^2} + \frac{1}{w_1^2} & -\frac{1}{w_1^2} & \ldots & 0 \\ -\frac{1}{w_1^2} & \frac{1}{\tau_2^2} + \frac{1}{w_1^2} + \frac{1}{w_2^2} & \ldots & 0 \\ 0 & -\frac{1}{w_2^2} & \ldots & 0 \\ \ldots & \ldots & \ddots & \ldots \\ 0 & 0 & \frac{1}{\tau_{p-1}^2} + \frac{1}{w_{p-2}^2} + \frac{1}{w_{p-1}^2} & -\frac{1}{w_{p-1}^2} \\ 0 & 0 & -\frac{1}{w_{p-1}^2} & \frac{1}{\tau_p^2} + \frac{1}{w_{p-1}^2} \end{bmatrix}. \tag{3.3}$$

Let $\tau^2 = (\tau_1^2, \ldots, \tau_p^2)$ and $w^2 = (w_1^2, \ldots, w_{p-1}^2)$. Kyung et al. [20] state that the priors in (3.2) lead to the following marginal prior on β given σ^2.

$$\pi(\beta \mid \sigma^2) \propto \exp\left(-\frac{\lambda_1}{\sigma} \sum_{j=1}^{p} |\beta_j| - \frac{\lambda_2}{\sigma} \sum_{j=1}^{p-1} |\beta_{j+1} - \beta_j| \right). \tag{3.4}$$

However, this is not the case and in particular, the independent exponential priors on τ^2 and w^2 do not lead to the marginal prior in (3.4). Instead, our proposed prior is

$$\pi(\tau^2, w^2) \propto \det\left(\Sigma_{\tau,w}\right)^{1/2} \left(\prod_{i=1}^{p} \left(\tau_i^2\right)^{-1/2} e^{-\lambda_1 \tau_i^2/2} \right) \left(\prod_{i=1}^{p-1} \left(w_i^2\right)^{-1/2} e^{-\lambda_2 w_i^2/2} \right). \tag{3.5}$$

In Appendix B.1, we show that the prior on (τ^2, w^2) in (3.5) is proper and in Appendix B.2 we demonstrate that the marginal prior on β given σ^2 is the appropriate prior in (3.4). Thus, our model formulation is a valid Bayesian fused lasso model.

3.1. Gibbs sampler for the Bayesian fused lasso

The resulting full conditionals from the model in (3.2) with prior (3.5) are,

$$\begin{aligned} \beta \mid \sigma^2, \tau^2, w^2, y &\sim N_p\left((X^TX + \Sigma_{\tau,w}^{-1})^{-1} X^T y, \sigma^2 (X^TX + \Sigma_{\tau,w}^{-1})^{-1}\right) \\ \frac{1}{\tau_i^2} \mid \beta, \sigma^2, y &\overset{\text{ind}}{\sim} \text{Inverse-Gaussian}\left(\sqrt{\frac{\lambda_1^2\sigma^2}{\beta_i^2}}, \lambda_1^2\right), \ i = 1, \ldots, p \\ \frac{1}{w_i^2} \mid \beta, \sigma^2, y &\overset{\text{ind}}{\sim} \text{Inverse-Gaussian}\left(\sqrt{\frac{\lambda_2^2\sigma^2}{(\beta_{i+1} - \beta_i)^2}}, \lambda_2^2\right), \ i = 1, \ldots, p-1 \\ \sigma^2 \mid \beta, \tau^2, w^2, y & \\ &\sim \text{Inverse-Gamma}\left(\frac{n + p + 2\alpha}{2}, \frac{(y - X\beta)^T(y - X\beta) + \beta^T \Sigma_{\tau,w}^{-1} \beta + 2\xi}{2}\right). \end{aligned} \tag{3.6}$$

Here the Inverse-Gaussian(a, b) density is $f(x) \propto x^{-3/2} \exp(-b(x-a)^2/2a^2x)$ and the Inverse-Gamma(a, b) density is $f(x) \propto x^{-a-1} \exp(-b/x)$. Notice that the full conditionals for τ^2 and w^2 are independent and thus can be updated in one block. This reduces the four variable Gibbs sampler to a three variable Gibbs sampler. If $(\beta_{(t)}, \tau^2_{(t)}, w^2_{(t)}, \sigma^2_{(t)})$ is the current state of the Gibbs sampler the $(t+1)$th state is obtained as follows.

---

1. Draw $\sigma^2_{(n+1)}$ from $f(\sigma^2 \mid \beta_{(n)}, \tau^2_{(n)}, w^2_{(n)}, y)$.
2. Draw $\left(1/\tau^2_{(n+1)}, 1/w^2_{(n+1)}\right)$ from $f(1/\tau^2 \mid \beta_{(n)}, \sigma^2_{(n+1)}, y)\ f(1/w^2 \mid \beta_{(n)}, \sigma^2_{(n+1)}, y)$.
3. Draw $\beta_{(n+1)}$ from $f(\beta \mid \tau^2_{(n+1)}, w^2_{(n+1)}, \sigma^2_{(n+1)}, y)$.

This three variable deterministic scan Gibbs sampler has MTD,

$$\begin{aligned} &k_{BFL}(\beta, \tau^2, w^2, \sigma^2 \mid \beta_0, \tau_0^2, w_0^2, \sigma_0^2) \\ &= f(\beta \mid \tau^2, w^2, \sigma^2, y)\ f(\tau^2, w^2 \mid \beta_0, \sigma^2, y) f(\sigma^2 \mid \beta_0, \tau_0^2, w_0^2, y)\,. \end{aligned} \tag{3.7}$$

First we note that the full conditional distribution of $1/\tau_i^2$ is an Inverse-Gaussian with mean parameter $\sqrt{\lambda_1^2\sigma^2/\beta_i^2}$. If the starting value for any β_i is zero, this Inverse-Gaussian is still well defined as it is an Inverse-Gamma distribution with shape parameter $1/2$ and rate parameter $\lambda_1^2/2$. The same is true for the full conditional of $1/w_i^2$. Thus, the MTD is strictly positive and well defined which implies the Markov chain is aperiodic, irreducible almost everywhere, and Harris recurrent.

We define the drift function $V_{BFL} : \mathbb{R}^p \times \mathbb{R}_+^p \times \mathbb{R}_+^{p-1} \times \mathbb{R}_+ \to [0, \infty)$ as

$$V_{BFL}(\beta, \tau^2, w^2, \sigma^2) = (y - X\beta)^T(y - X\beta) + \beta^T \Sigma_{\tau,w}^{-1} \beta + \frac{\lambda_1^2}{4} \sum_{i=1}^{p} \tau_i^2 + \frac{\lambda_2^2}{4} \sum_{i=1}^{p-1} w_i^2\,. \tag{3.8}$$

The following theorem is proved by establishing (2.2) and (2.3) for the drift function V_{BFL}.

Theorem 3.1. *If $n \geq 3$, the three variable Gibbs sampler for the Bayesian fused lasso is geometrically ergodic.*

Proof. See Appendix C. □

Remark 3.1. In Appendix C.1, we arrive at the drift rate

$$\phi_{BFL} = \max\left\{\frac{p}{n+p+2\alpha-2}, \frac{1}{2}\right\}.$$

Thus, ϕ_{BFL} is no better than $1/2$ and as p increases, the drift rate approaches one. Thus, convergence may be slower for problems with large p.

Remark 3.2. Minimizing V_{BFL} yields default starting value of β_0 being the frequentist fused lasso estimate, $\tau^2_{0,i} = 2|\beta_{0,i}|/\lambda_1$ and $w^2_{0,i} = 2|\beta_{0,i+1} - \beta_{0,i}|/\lambda_2$. See Appendix C.3 for details.

4. Bayesian group lasso

Knowledge of correlation among predictors is ignored by the usual lasso. The group lasso of [38] imposes sparsity across grouped predictors. For a fixed K, partition β in K groups of size $m_1, m_2, \ldots, m_K$; the groups being denoted by $\beta_{G_1}, \beta_{G_2}, \ldots, \beta_{G_K}$. Let X_{G_k} denote the matrix of predictors for group k. The group lasso estimate for tuning parameter $\lambda > 0$ is,

$$\hat{\beta}_{\text{group}} = \arg\max_{\beta} \left\| y - \sum_{k=1}^{K} X_{G_k}\beta_{G_k} \right\|^2 + \lambda \sum_{k=1}^{K} \|\beta_{G_k}\| \,. \tag{4.1}$$

Kyung et al. [20] present the following Bayesian analog of the group lasso. Let

$$\begin{aligned} Y \mid \beta, \sigma^2 &\sim N_n(X\beta, \sigma^2 I_n) \\ \beta_{G_k} \mid \sigma^2, \tau_k^2 &\overset{\text{ind}}{\sim} N_{m_k}(0, \sigma^2\tau_k^2 I_{m_k}) \quad k = 1, \ldots, K \\ \tau_k^2 &\overset{\text{ind}}{\sim} \text{Gamma}\left(\frac{m_k+1}{2}, \frac{\lambda^2}{2}\right) \quad k = 1, \ldots, K \\ \sigma^2 &\sim \text{Inverse-Gamma}(\alpha, \xi)\,, \end{aligned} \tag{4.2}$$

where $\lambda > 0$ is fixed, $\alpha, \xi \geq 0$ are known, and the density of a Gamma(a, b) is $f(x) \propto x^{a-1}e^{-bx}$.

4.1. Gibbs sampler for Bayesian group lasso

Let $\tau^2 = (\tau_1^2, \tau_2^2, \ldots, \tau_K^2)$. Define

$$D_\tau = \text{diag}(\underbrace{\tau_1^2, \ldots, \tau_1^2}_{m_1}, \underbrace{\tau_2^2, \ldots, \tau_2^2}_{m_2}, \ldots, \underbrace{\tau_K^2, \ldots, \tau_K^2}_{m_K}) \,.$$

The Bayesian group lasso in (4.2) leads to the following full conditionals for β, τ^2 and σ^2:

$$\begin{aligned} &\beta \mid \sigma^2, \tau^2, y \sim N_p\left((X^TX + D_\tau^{-1})^{-1}X^Ty, \sigma^2(X^TX + D_\tau^{-1})^{-1}\right) \\ &\frac{1}{\tau_k^2} \mid \beta, \sigma^2, y \overset{\text{ind}}{\sim} \text{Inverse-Gaussian}\left(\sqrt{\frac{\lambda^2\sigma^2}{\beta_{G_k}^T\beta_{G_k}}}, \lambda^2\right), k = 1, \ldots, K \\ &\sigma^2 \mid \beta, \tau^2, y \\ &\quad \sim \text{Inverse-Gamma}\left(\frac{n+p+2\alpha}{2}, \frac{(y - X\beta)^T(y - X\beta) + \beta^T D_\tau^{-1}\beta + 2\xi}{2}\right). \end{aligned} \tag{4.3}$$

These full conditionals lead to a three variable Gibbs sampler where the variables are β, τ^2, and σ^2.

Remark 4.1. Kyung et al. [20] propose a $K+2$ variable Gibbs sampler where the variables are $\beta_{G_1}, \beta_{G_2}, \ldots, \beta_{G_K}, \tau^2$, and σ^2. For this sampler, the full conditionals for σ^2 and τ^2 are the same as above, but the full conditional for each β_{G_k} is

$$\beta_{G_k} \mid \beta_{-G_k}, \sigma^2, \tau^2, y \sim N_{m_k}\Bigg(\left(X_{G_k}^T X_{G_k} + \tau_k^{-2} I_{m_k}\right)^{-1} X_{G_k}^T \left(y - \sum_{k' \neq k} X_{G'_k} \beta_{G_{k'}} \right) , \sigma^2 \left(X_{G_k}^T X_{G_k} + \tau_k^{-2} I_{m_k}\right)^{-1} \Bigg).$$

Kyung et al. [20] had an error in their full conditional where they had

$$\left(y - \frac{1}{2} \sum_{k' \neq k} X_{G'_k} \beta_{G_{k'}} \right) \text{ instead of } \left(y - \sum_{k' \neq k} X_{G'_k} \beta_{G_{k'}} \right).$$

The motivation for using the $K+2$ sampler is to avoid the $p \times p$ matrix inversion of $(X^T X + D_\tau^{-1})$, and instead do K matrix inversions each of size $m_k \times m_k$. This reduces the computational cost from $O(p^3)$ to $O(\sum_{k=1}^K m_k^3)$. Such a technique was also discussed in [13]. However, it is known that a blocked Gibbs sampler mixes as well as or better than a full Gibbs sampler (see [21]). In addition, [2] recently proposed a linear time sampling algorithm to sample from high-dimensional normal distributions of the form in (4.3). Using their method, the computational cost of drawing from the full conditional of β is $O(n^2 p)$, and thus the $K+2$ variable Gibbs sampler is not required.

We will study the rate of convergence of the three variable Gibbs sampler. If $(\beta_{(t)}, \tau^2_{(t)}, \sigma^2_{(t)})$ is the current state of the Gibbs sampler, the $(t+1)$th state is obtained as follows.

---

1. Draw $\sigma^2_{(n+1)}$ from $f(\sigma^2 \mid \beta_{(n)}, \tau^2_{(n)}, y)$.
2. Draw $1/\tau^2_{(n+1)}$ from $f(1/\tau^2 \mid \beta_{(n)}, \sigma^2_{(n+1)}, y)$.
3. Draw $\beta_{(n+1)}$ from $f(\beta \mid \tau^2_{(n+1)}, \sigma^2_{(n+1)}, y)$.

The MTD for the above three variable deterministic scan Gibbs sampler is

$$k_{BGL}(\beta, \tau^2, \sigma^2 \mid \beta_0, \tau_0^2, \sigma_0^2) = f(\beta \mid \tau^2, \sigma^2, y)\, f(\tau^2 \mid \beta_0, \sigma^2, y)\, f(\sigma^2 \mid \beta_0, \tau_0^2, y)\,. \tag{4.4}$$

As in the Bayesian fused lasso, the MTD is well defined and strictly positive leading to an aperiodic, irreducible almost everywhere, and Harris recurrent Markov chain.

Define the drift function $V_{BGL} : \mathbb{R}^p \times \mathbb{R}_+^K \times \mathbb{R}_+ \to [0, \infty)$ as

$$V_{BGL}(\beta, \tau^2, \sigma^2) = (y - X\beta)^T (y - X\beta) + \beta^T D_\tau^{-1} \beta + \frac{\lambda^2}{4} \sum_{k=1}^K \tau_k^2\,. \tag{4.5}$$

Theorem 4.1. *If $n \geq 3$, the three variable Gibbs sampler for the Bayesian group lasso is geometrically ergodic.*

Proof. See Appendix D. □

Remark 4.2. As in the Bayesian fused lasso Gibbs sampler, the drift rate,

$$\phi_{BGL} = \max\left\{\frac{p}{n+p+2\alpha-2}, \frac{1}{2}\right\},$$

is no better than $1/2$ and approaches 1 as p increases.

Remark 4.3. Minimizing V_{BGL} yields default starting values for the Markov chain as β_0 being the frequentist group lasso estimate and $\tau^2_{0,k} = 2\sqrt{\beta^T_{0,G_k}\beta_{0,G_k}}/\lambda$. See Appendix D.3 for details.

Remark 4.4. Since for $K = p$, the Bayesian group lasso is the Bayesian lasso, our result of geometric ergodicity holds for the Bayesian lasso as well. Geometric ergodicity of the Bayesian lasso was demonstrated by [19] under exactly the same conditions. Our result on the starting values in Remark 4.3 also holds for the Bayesian lasso Gibbs sampler.

5. Bayesian sparse group lasso

The group lasso induces sparsity across groups but does not induce sparsity within a group. Simon et al. [32] added an L_1 penalty on the individual coefficients to the group lasso to arrive at the sparse group lasso. As before, for a fixed K, partition β in K groups each of size $m_1, m_2, \ldots, m_K$, the groups being denoted by $\beta_{G_1}, \beta_{G_2}, \ldots, \beta_{G_K}$. For tuning parameters $\lambda_1 > 0$ and $\lambda_2 > 0$, the sparse group lasso estimate is

$$\hat{\beta}_{\text{sgroup}} = \arg\max_{\beta} \left\| y - \sum_{k=1}^{K} X_{G_k}\beta_{G_k} \right\|^2 + \lambda_1\|\beta\|_1 + \lambda_2 \sum_{k=1}^{K} \|\beta_{G_k}\|, \tag{5.1}$$

where $\|\cdot\|_1$ is the L_1 norm. The Bayesian sparse group lasso was introduced by [36]. Let $\gamma^2_{1,1}, \gamma^2_{1,2}, \ldots, \gamma^2_{1,m_1}, \ldots, \gamma^2_{K,m_K}$ and $\tau^2_1, \ldots, \tau^2_p$ be variables defined on the positive reals. For each group k define,

$$V_k = \text{diag}\left\{ \left(\frac{1}{\tau^2_k} + \frac{1}{\gamma^2_{k,j}} \right)^{-1} : j = 1, \ldots, m_k \right\}.$$

The notation $\gamma^2_{k,j}$ is used for convenience and may be replaced with γ^2_i for all i. Let $\tau^2 = (\tau^2_1, \tau^2_2, \ldots, \tau^2_K)$ and let $\gamma^2 = (\gamma^2_{1,1}, \ldots, \gamma^2_{1,m_1}, \ldots, \gamma^2_{K,1}, \ldots, \gamma^2_{K,m_K})$. The Bayesian sparse group lasso model formulated by [36] is

$$\begin{aligned} Y \mid \beta, \sigma^2 &\sim N_n(X\beta, \sigma^2 I_n) \\ \beta_{G_k} \mid \sigma^2, \tau^2, \gamma^2 &\overset{\text{iid}}{\sim} N_{m_k}(0, \sigma^2 V_k) \quad \text{for } k = 1, \ldots, K \end{aligned} \tag{5.2}$$

$$\begin{aligned}\pi(\gamma_{k,1},\ldots,\gamma_{k,m_k},\tau_k^2) &= \pi_k \quad \text{independently for } k=1,\ldots,K\\ \sigma^2 &\sim \text{Inverse-Gamma}(\alpha,\xi)\,,\end{aligned}$$

where $\alpha,\xi\geq 0$ are fixed and the independent prior on each $(\gamma_{k,1},\ldots,\gamma_{k,m_k},\tau_k^2)$ is

$$\pi_k \propto \prod_{j=1}^{m_k}\left[\left(\gamma_{k,j}^2\right)^{-\frac{1}{2}}\left(\frac{1}{\gamma_{k,j}^2}+\frac{1}{\tau_k^2}\right)^{-\frac{1}{2}}\right]\left(\tau_k^2\right)^{-\frac{1}{2}}\exp\left\{-\frac{\lambda_2^2}{2}\sum_{j=1}^{m_k}\gamma_{k,j}^2-\frac{\lambda_1^2}{2}\tau_k^2\right\}. \tag{5.3}$$

Here $\lambda_1,\lambda_2>0$ are fixed. Xu and Ghosh [36] show that the prior in (5.3) is proper with the normalizing constant being a function of λ_1 and λ_2.

5.1. Gibbs sampler for Bayesian sparse group lasso

Define $V_{\tau,\gamma}$ to be the diagonal matrix with diagonals being that of $V_1,\ldots,V_K$ in that sequence. In addition, let $\beta_{k,j}$, refer to the jth coefficient in the kth group. The Bayesian sparse group lasso model in (5.2) leads to the following full conditionals for β,τ^2,γ^2 and σ^2:

$$\begin{aligned}\beta\mid\sigma^2,\tau^2,\gamma^2,y &\sim N_p\left((X^TX+V_{\tau,\gamma}^{-1})^{-1}X^Ty,\sigma^2(X^TX+V_{\tau,\gamma}^{-1})^{-1}\right)\\ \frac{1}{\tau_k^2}\mid\beta,\sigma^2,y &\overset{\text{ind}}{\sim} \text{Inverse-Gaussian}\left(\sqrt{\frac{\lambda_1^2\sigma^2}{\beta_{G_k}^T\beta_{G_k}}},\lambda_1^2\right),\ \text{for all } k\\ \frac{1}{\gamma_{k,j}^2}\mid\beta,\sigma^2,y &\overset{\text{ind}}{\sim} \text{Inverse-Gaussian}\left(\sqrt{\frac{\lambda_2^2\sigma^2}{\beta_{k,j}^2}},\lambda_2^2\right),\ \text{for all } k,j\\ \sigma^2\mid\beta,\tau^2,\gamma^2,y &\sim \text{Inverse-Gamma}\left(\frac{n+p+2\alpha}{2},\frac{(y-X\beta)^T(y-X\beta)+\beta^TV_{\tau,\gamma}^{-1}\beta+2\xi}{2}\right).\end{aligned} \tag{5.4}$$

Note that the full conditionals for τ^2 and γ^2 are independent and thus can be updated in one block yielding a three variable Gibbs sampler. If $(\beta_{(t)},\tau_{(t)}^2,\gamma_{(t)}^2,\sigma_{(t)}^2)$ is the current state of the Gibbs sampler, the $(t+1)$th state is obtained as follows.

---

1. Draw $\sigma_{(n+1)}^2$ from $f(\sigma^2\mid\beta_{(n)},\tau_{(n)}^2,\gamma_{(n)}^2,y)$.
2. Draw $\left(1/\tau_{(n+1)}^2,1/\gamma_{(n+1)}^2\right)$ from $f(1/\tau^2\mid\beta_{(n)},\sigma_{(n+1)}^2,y)\ f(1/\gamma^2\mid\beta_{(n)},\sigma_{(n+1)}^2,y)$.
3. Draw $\beta_{(n+1)}$ from $f(\beta\mid\tau_{(n+1)}^2,\gamma_{(n+1)}^2,\sigma_{(n+1)}^2,y)$.

The MTD for the three variable Gibbs sampler is

$$\begin{aligned}&k_{BSGL}(\beta,\tau^2,\gamma^2,\sigma^2\mid\beta_0,\tau_0^2,\gamma_0^2,\sigma_0^2)\\ &= f(\beta\mid\tau^2,\gamma^2,\sigma^2,y)\ f(\tau^2,\gamma^2\mid\beta_0,\sigma^2,y)\ f(\sigma^2\mid\beta_0,\tau_0^2,\gamma_0^2,y)\,.\end{aligned} \tag{5.5}$$

As in the Bayesian group lasso Gibbs sampler, the MTD is strictly positive and thus aperiodic, irreducible almost everywhere, and the chain is Harris recurrent. We will prove geometric ergodicity by establishing a drift and an associated minorization condition.

Define the drift function $V_{BSGL} : \mathbb{R}^p \times \mathbb{R}_+^K \times \mathbb{R}_+^p \times \mathbb{R}_+ \to [0, \infty)$ as,

$$V_{BSGL}(\beta, \tau^2, \gamma^2, \sigma^2) = (y - X\beta)^T(y - X\beta) + \beta^T V_{\tau,\gamma}^{-1}\beta + \frac{\lambda_1^2}{4}\sum_{k=1}^{K}\tau_k^2 + \frac{\lambda_2^2}{4}\sum_{k=1}^{K}\sum_{j=1}^{m_k}\gamma_{k,j}^2 . \tag{5.6}$$

Theorem 5.1. *If $n \geq 3$, the three variable Gibbs sampler for the Bayesian sparse group lasso is geometrically ergodic.*

Proof. See Appendix E. □

Remark 5.1. Define $M = \max_k m_k$. In Appendix E.1 the drift rate is determined to be

$$\phi_{BSGL} = \max\left\{ \frac{p}{n + p + 2\alpha - 2}, \frac{\left(1 + \frac{\lambda_2^2}{\lambda_1^2}\right)}{2\left(1 + \frac{\lambda_1^2}{\lambda_2^2} + \frac{\lambda_2^2}{\lambda_1^2}\right)}, \frac{\left(1 + \frac{\lambda_1^2}{\lambda_2^2}\right)}{2M\left(1 + \frac{\lambda_1^2}{\lambda_2^2} + \frac{\lambda_2^2}{\lambda_1^2}\right)} \right\} .$$

Unlike the drift rate in the previous two models, the drift rate here can be lower than 1/2. However, it is likely that p is large enough so that ϕ_{BSGL} is determined by the first term $p/(n + p + 2\alpha - 2)$. In this case again, the drift rate will tend to 1 as p increases and thus convergence may be slower for large p problems.

Remark 5.2. A reasonable starting value for this Markov chain is β_0 being the sparse group lasso estimate, $\tau_{0,k}^2 = 2\sqrt{\beta_{0,G_k}^T \beta_{0,G_k}}/\lambda_1$ and $\gamma_{0,k}^2 = 2|\beta_{0,k,j}|/\lambda_2$. See Appendix E.3.

6. Discussion

As discussed in Section 1, reliable estimation from MCMC output rests heavily on the rate of convergence of the Markov chain. Our geometric ergodicity results immediately imply the existence of a Markov chain CLT and strong consistency of some estimators of the asymptotic covariance matrix in this CLT. As a consequence, practitioners can use tools such as effective sample size to understand the quality of the Monte Carlo estimates.

Our results of geometric ergodicity hold under reasonable conditions. We require no conditions on p, and only need n to be larger than 3. However, our results suggest that it may be possible for the Gibbs samplers to converge at a slower rate if $p \gg n$. This agrees with the results in [26]. Users might then be inclined to first use Bayesian variable selection alternatives to these models. For example, [36] introduced the Bayesian variable selection alternatives to the

group and the sparse group lasso by using spike-and-slab type priors. A natural direction for future research would be to investigate the convergence rate for the Gibbs samplers in these Bayesian variable selection models.

Appendix A: Preliminaries

In general, $\mathrm{E}_{(k)}$ represents expectation with respect to the MTD being studied in the section. Expectations with respect to a full conditional is denoted by E.. The index 0 on variables denotes starting values for the Markov chain.

Below are some properties of known distributions that will be used often.

- If $1/X \sim$ Inverse-Gaussian(a, b), then $\mathrm{E}[X] = 1/a + 1/b$.
- If $X \sim N_p(\mu, \Sigma)$, then $\mathrm{E}[XX^T] = \Sigma + \mu\mu^T$.
- If $X \sim$ Inverse-Gamma(a, b), then $\mathrm{E}[X] = b/(a-1)$.
- If $X \sim$ Inverse-Gamma(a, b), then $\mathrm{E}[1/X] = a/b$.

A.1. Useful lemmas

We present some results that will used in the proofs of geometric ergodicity for all three samplers. Most of the results are generalizations of the results in [19] and the proofs are presented here for completeness.

Lemma A.1. *Let y, X, and β be the observed $n \times 1$ response, the $n \times p$ matrix of covariates and the $p \times 1$ vector of regression coefficients. Let Σ be the $p \times p$ positive definite matrix such that*

$$\beta \sim N_p\left((X^TX + \Sigma^{-1})^{-1}X^Ty, \sigma^2(X^TX + \Sigma^{-1})^{-1}\right),$$

for $\sigma^2 > 0$. Then,

$$E\left[(y - X\beta)^T(y - X\beta) + \beta^T\Sigma^{-1}\beta\right] \leq y^Ty + p\sigma^2.$$

Proof. Consider,

$$\begin{aligned}
&\mathrm{E}\left[(y - X\beta)^T(y - X\beta) + \beta^T\Sigma^{-1}\beta\right] \\
&\quad = y^Ty - 2y^TX\,\mathrm{E}\left[\beta\right] + \mathrm{E}\left[\beta^T(X^TX + \Sigma^{-1})\beta\right] \\
&\quad = y^Ty - 2y^TX(X^TX + \Sigma^{-1})^{-1}X^Ty + \mathrm{E}\left[tr(\beta^T(X^TX + \Sigma^{-1})\beta)\right] \\
&\quad = y^Ty - 2y^TX(X^TX + \Sigma^{-1})^{-1}X^Ty \\
&\qquad + tr\left(\sigma^2(X^TX + \Sigma^{-1})(X^TX + \Sigma^{-1})^{-1}\right) \\
&\qquad + tr\left((X^TX + \Sigma^{-1})(X^TX + \Sigma^{-1})^{-1}X^Tyy^TX(X^TX + \Sigma^{-1})^{-1}\right) \\
&\quad = y^Ty - 2y^TX(X^TX + \Sigma^{-1})^{-1}X^Ty + p\sigma^2 \\
&\qquad + tr\left(y^TX(X^TX + \Sigma^{-1})^{-1}X^Ty\right) \\
&\quad \leq y^Ty + p\sigma^2\,.
\end{aligned}$$

□

Lemma A.2. *For $\alpha = (\alpha_1, \ldots, \alpha_p) \in \mathbb{R}^p$ and $\delta = (\delta_1, \ldots, \delta_p)$ such that $\delta_i \neq 0$,*

$$\frac{\sum_{i=1}^p \alpha_i^2}{\sum_{i=1}^p \alpha_i^2/\delta_i^2} \leq \sum_{i=1}^p \delta_i^2.$$

Proof. Using the fact that the square of a number is non-negative,

$$\frac{\sum_{i=1}^p \alpha_i^2}{\sum_{i=1}^p \alpha_i^2/\delta_i^2} = \frac{\sum_{i=1}^p \frac{\alpha_i^2}{\delta_i^2}\delta_i^2}{\sum_{i=1}^p \alpha_i^2/\delta_i^2} \leq \frac{\sum_{i=1}^p \frac{\alpha_i^2}{\delta_i^2}\left(\sum_{i=1}^p \delta_i^2\right)}{\sum_{i=1}^p \alpha_i^2/\delta_i^2} = \sum_{i=1}^p \delta_i^2 \,. \qquad \square$$

Lemma A.3. *For $\lambda^2, a^2, \sigma^2 > 0$, if X has a probability density function $f(x)$ such that*

$$f(x) \propto x^{-1/2} \exp\left\{-\frac{\lambda^2 x}{2} - \frac{a^2}{2\sigma^2 x}\right\},$$

then $1/X \sim$ Inverse-Gaussian distribution with mean parameter $\sqrt{\lambda^2\sigma^2/a^2}$ and scale parameter λ^2.

Proof. For the change of variable $z = 1/x$,

$$\begin{aligned} f(z) &\propto z^{-2} z^{\frac{1}{2}} \exp\left\{-\frac{\lambda^2}{2z} - \frac{a^2 z}{2\sigma^2}\right\} = z^{-\frac{3}{2}} \exp\left\{-\frac{a^2\left(\frac{\lambda^2\sigma^2}{a^2} + z^2\right)}{2\sigma^2 z}\right\} \\ &= \exp\left\{-\sqrt{\frac{\lambda^2 a^2}{\sigma^2}}\right\} z^{-\frac{3}{2}} \exp\left\{-\frac{a^2\left(\frac{\lambda^2\sigma^2}{a^2} - 2\sqrt{\frac{\lambda^2\sigma^2}{a^2}}z + z^2\right)}{2\sigma^2 z}\right\} \\ &\propto z^{-\frac{3}{2}} \exp\left\{-\frac{\lambda^2\left(\frac{\lambda^2\sigma^2}{a^2} - 2\sqrt{\frac{\lambda^2\sigma^2}{a^2}}z + z^2\right)}{2\frac{\lambda^2\sigma^2}{a^2} z}\right\}. \end{aligned}$$

Thus, $Z \sim$ Inverse-Gaussian with mean parameter $\sqrt{\lambda^2\sigma^2/a^2}$ and scale parameter λ^2. $\square$

Lemma A.4. *If $1/X \sim$ Inverse-Gaussian with mean parameter $\sqrt{\lambda^2\sigma^2/a^2}$ and scale parameter λ^2 and $a^2 \leq d^2$ for some $d^2 > 0$, then*

$$f(x) \geq \exp\left\{-\sqrt{\frac{\lambda^2 d^2}{\sigma^2}}\right\} q(x)\,,$$

where $f(x)$ is the pdf of X and $q(x)$ is the pdf of the reciprocal of the Inverse-Gaussian distribution with mean parameter $\sqrt{\lambda^2\sigma^2/d^2}$ and scale parameter λ^2.

Proof. By Lemma A.3, we have

$$f(x) = \sqrt{\frac{\lambda^2}{2\pi}}\,(x)^{-\frac{1}{2}} \exp\left\{ -\frac{\lambda^2\left(\frac{\lambda^2\sigma^2}{a^2} - 2\sqrt{\frac{\lambda^2\sigma^2}{a^2}}\frac{1}{x} + \frac{1}{x^2}\right)}{2\frac{\lambda^2\sigma^2}{a^2}\frac{1}{x}} \right\}$$

$$= \exp\left\{\sqrt{\frac{\lambda^2 a^2}{\sigma^2}}\right\} \sqrt{\frac{\lambda^2}{2\pi}}\,(x)^{-\frac{1}{2}} \exp\left\{ -\frac{\lambda^2\left(\frac{\lambda^2\sigma^2}{a^2} + \frac{1}{x^2}\right)}{2\frac{\lambda^2\sigma^2}{a^2}\frac{1}{x}} \right\}$$

$$\geq \sqrt{\frac{\lambda^2}{2\pi}}\,(x)^{-\frac{1}{2}} \exp\left\{ -\frac{\lambda^2\left(\frac{\lambda^2\sigma^2}{a^2} + \frac{1}{x^2}\right)}{2\frac{\lambda^2\sigma^2}{a^2}\frac{1}{x}} \right\} = \exp\left\{-\sqrt{\frac{\lambda^2 d^2}{\sigma^2}}\right\} q(x)\,. \qquad \square$$

Lemma A.5. *Let y, X, and β be the observed $n \times 1$ response, the $n \times p$ matrix of covariates and the $p \times 1$ vector of regression coefficients respectively. Let Σ be a $p \times p$ positive definite matrix. Then,*

$$(y - X\beta)^T(y - X\beta) + \beta^T\Sigma^{-1}\beta \geq y^T y - y^T X\left(X^T X + \Sigma^{-1}\right)^{-1} X^T y\,.$$

Proof. The proof mainly requires completing the square in the following way,

$$\begin{aligned}
&(y - X\beta)^T(y - X\beta) + \beta^T\Sigma^{-1}\beta \\
&\quad = y^T y - 2y^T X(X^T X + \Sigma^{-1})(X^T X + \Sigma^{-1})^{-1}\beta + \beta^T(X^T X + \Sigma^{-1})\beta \\
&\qquad + y^T X(X^T X + \Sigma^{-1})^{-1}(X^T X + \Sigma^{-1})(X^T X + \Sigma^{-1})^{-1}X^T y \\
&\qquad - y^T X(X^T X + \Sigma^{-1})^{-1}(X^T X + \Sigma^{-1})(X^T X + \Sigma^{-1})^{-1}X^T y \\
&\quad = y^T y - y^T X(X^T X + \Sigma^{-1})^{-1}X^T y \\
&\qquad + (\beta - (X^T X + \Sigma^{-1})^{-1}X^T y)^T(X^T X + \Sigma^{-1})(\beta - (X^T X + \Sigma^{-1})^{-1}X^T y) \\
&\quad \geq y^T y - y^T X(X^T X + \Sigma^{-1})^{-1}X^T y\,.
\end{aligned}$$

$\square$

Appendix B: Bayesian fused lasso prior

B.1. Propriety of the prior

First note that $\det(\Sigma_{\tau,w}) = \left(\det\left(\Sigma_{\tau,w}^{-1}\right)\right)^{-1}$. We decompose $\Sigma_{\tau,w}^{-1}$ into

$$\Sigma_{\tau,w}^{-1} = L_1 + L_2\,, \tag{B.1}$$

where

$$L_1 = \text{diag}\left(\frac{1}{2\tau_1^2}, \frac{1}{2\tau_2^2}, \ldots, \frac{1}{2\tau_p^2}\right) \text{ and},$$

$$L_2 = \begin{bmatrix} \frac{1}{2\tau_1^2} + \frac{1}{w_1^2} & -\frac{1}{w_1^2} & \ldots & 0 \\ -\frac{1}{w_1^2} & \frac{1}{2\tau_2^2} + \frac{1}{w_1^2} + \frac{1}{w_2^2} & \ldots & 0 \\ 0 & -\frac{1}{w_2^2} & \ldots & 0 \\ & & \ddots & \\ \ldots & \ldots & & \ldots \\ 0 & 0 & \frac{1}{2\tau_{p-1}^2} + \frac{1}{w_{p-2}^2} + \frac{1}{w_{p-1}^2} & -\frac{1}{w_{p-1}^2} \\ 0 & 0 & -\frac{1}{w_{p-1}^2} & \frac{1}{2\tau_p^2} + \frac{1}{w_{p-1}^2} \end{bmatrix}.$$

The diagonal matrix L_1 is clearly positive definite. The tridiagonal matrix L_2 is also positive definite since L_2 is real symmetric, has positive diagonals, and is strictly diagonally dominant [1, Theorem 1.2]. Here the condition of strict diagonal dominance is satisfied since

$$\frac{1}{2\tau_i^2} + \frac{1}{w_{i-1}^2} + \frac{1}{w_i^2} > \frac{1}{w_{i-1}^2} + \frac{1}{w_i^2}.$$

Thus,

$$\det(\Sigma_{\tau,w}^{-1}) = \det(L_1 + L_2) \geq \det(L_1) + \det(L_2) \geq \det(L_1) = \prod_{i=1}^{p}\left(\frac{1}{2\tau_i^2}\right)$$

$$\Rightarrow \ (\det(\Sigma_{\tau,w}))^{\frac{1}{2}} \leq \prod_{i=1}^{p}\left(2\tau_i^2\right)^{1/2}.$$

Thus, the joint prior on (τ^2, w^2) satisfies,

$$\begin{aligned} \pi(\tau^2, w^2) &\propto \det(\Sigma_{\tau,w})^{1/2} \left(\prod_{i=1}^{p}\left(\tau_i^2\right)^{-1/2} e^{-\lambda_1\tau_i^2/2}\right)\left(\prod_{i=1}^{p-1}\left(w_i^2\right)^{-1/2} e^{-\lambda_2 w_i^2/2}\right) \\ &\leq \prod_{i=1}^{p}\left(2\tau_i^2\right)^{1/2} \left(\prod_{i=1}^{p}\left(\tau_i^2\right)^{-1/2} e^{-\lambda_1\tau_i^2/2}\right)\left(\prod_{i=1}^{p-1}\left(w_i^2\right)^{-1/2} e^{-\lambda_2 w_i^2/2}\right) \\ &= 2^{p/2}\left(\prod_{i=1}^{p} e^{-\lambda_1\tau_i^2/2}\right)\left(\prod_{i=1}^{p-1}\left(w_i^2\right)^{-1/2} e^{-\lambda_2 w_i^2/2}\right). \end{aligned}$$

This is the product of p exponentials densities and $p-1$ Gamma densities. Thus, the prior is proper.

B.2. Validity of the prior

In this section we demonstrate that our choice of prior in the Bayesian fused lasso leads to the Laplace prior in (3.4). First we expand $\beta^T\Sigma_{\tau,w}^{-1}\beta$ in the following way:

$$
\begin{aligned}
&\beta^T \Sigma_{\tau,w}^{-1} \beta \\
&= \begin{bmatrix}
\beta_1 \left(\frac{1}{\tau_1^2} + \frac{1}{w_1^2}\right) - \frac{\beta_2}{w_1^2} \\
-\frac{\beta_1}{w_1^2} + \beta_2 \left(\frac{1}{\tau_2^2} + \frac{1}{w_1^2} + \frac{1}{w_2^2}\right) - \frac{\beta_3}{w_2^2} \\
-\frac{\beta_2}{w_2^2} + \beta_3 \left(\frac{1}{\tau_3^2} + \frac{1}{w_2^2} + \frac{1}{w_3^2}\right) - \frac{\beta_4}{w_3^2} \\
\vdots \\
-\frac{\beta_{p-1}}{w_{p-1}^2} + \beta_p \left(\frac{1}{\tau_p^2} + \frac{1}{w_{p-1}^2}\right)
\end{bmatrix}^T
\begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \\ \vdots \\ \beta_p \end{bmatrix} \\
&= \beta_1^2 \left(\frac{1}{\tau_1^2} + \frac{1}{w_1^2}\right) - \frac{\beta_1\beta_2}{w_1^2} - \frac{\beta_1\beta_2}{w_1^2} + \beta_2^2 \left(\frac{1}{\tau_2^2} + \frac{1}{w_1^2} + \frac{1}{w_2^2}\right) - \frac{\beta_2\beta_3}{w_2^2} \\
&\quad - \frac{\beta_2\beta_3}{w_2^2} + \beta_3^2 \left(\frac{1}{\tau_3^2} + \frac{1}{w_2^2} + \frac{1}{w_3^2}\right) - \frac{\beta_3\beta_4}{w_3^2} + \dots \\
&\quad - \frac{\beta_{p-1}\beta_p}{w_{p-1}^2} + \beta_p^2 \left(\frac{1}{\tau_p^2} + \frac{1}{w_{p-1}^2}\right) \\
&= \sum_{i=1}^{p} \frac{\beta_i^2}{\tau_i^2} + \frac{\beta_1^2 + \beta_2^2 - 2\beta_1\beta_2}{w_1^2} + \frac{\beta_2^2 + \beta_3^2 - 2\beta_2\beta_3}{w_2^2} + \dots + \frac{\beta_p^2 + \beta_{p-1}^2 - 2\beta_p\beta_{p-1}}{w_{p-1}^2} \\
&= \sum_{i=1}^{p} \frac{\beta_i^2}{\tau_i^2} + \sum_{i=1}^{p-1} \frac{(\beta_{i+1} - \beta_i)^2}{w_i^2}. \qquad \text{(B.2)}
\end{aligned}
$$

Using (B.2),

$$
\begin{aligned}
&\pi(\beta \mid \sigma^2) \\
&\propto \int_{\mathbb{R}_+^p} \int_{\mathbb{R}_+^{p-1}} (2\pi\sigma^2)^{-\frac{p}{2}} \det\left(\Sigma_{\tau,w}^{-1}\right)^{1/2} \exp\left\{-\frac{\beta^T \Sigma_{\tau,w}^{-1} \beta}{2\sigma^2}\right\} \\
&\quad \times \det\left(\Sigma_{\tau,w}\right)^{\frac{1}{2}} \left(\prod_{i=1}^{p} \left(\tau_i^2\right)^{-1/2} e^{-\lambda_1 \tau_i^2/2}\right) \left(\prod_{i=1}^{p-1} \left(w_i^2\right)^{-1/2} e^{-\lambda_2 w_i^2/2}\right) dw^2 d\tau^2 \\
&\propto \int \prod_{i=1}^{p} \left(\tau_i^2\right)^{-1/2} \exp\left\{-\frac{\lambda_1 \tau_i^2}{2} - \frac{\beta_i^2}{2\sigma^2\tau_i^2}\right\} d\tau^2 \\
&\quad \times \int \prod_{i=1}^{p-1} \left(w_i^2\right)^{-1/2} \exp\left\{-\frac{\lambda_2 w_i^2}{2} - \frac{\left(\beta_{i+1} - \beta_i\right)^2}{2\sigma^2 w_i^2}\right\} dw^2 \\
&= \exp\left\{-\frac{\lambda_1}{\sigma} \sum_{i=1}^{p} |\beta_i| - \frac{\lambda_2}{\sigma} \sum_{i=1}^{p-1} |\beta_{i+1} - \beta_i|\right\} \\
&\quad \times \int \prod_{i=1}^{p} \left(\tau_i^2\right)^{-1/2} \exp\left\{-\frac{\lambda_1 \tau_i^2}{2} - \frac{\beta_i^2}{2\sigma^2\tau_i^2} + \frac{\lambda_1}{\sigma}|\beta_i|\right\} d\tau^2
\end{aligned}
$$

$$\times \int \prod_{i=1}^{p-1} \left(w_i^2\right)^{-1/2} \exp\left\{-\frac{\lambda_2 w_i^2}{2} - \frac{(\beta_{i+1}-\beta_i)^2}{2\sigma^2 w_i^2} + \frac{\lambda_2}{\sigma}|\beta_{i+1}-\beta_i|\right\} dw^2$$

$$\propto \exp\left\{-\frac{\lambda_1}{\sigma}\sum_{i=1}^{p}|\beta_i| - \frac{\lambda_2}{\sigma}\sum_{i=1}^{p-1}|\beta_{i+1}-\beta_i|\right\},$$

where the last equality is due to the integrands being the densities of the reciprocal of Inverse-Gaussian distributions; see Lemma A.3.

Appendix C: Proof of geometric ergodicity in Bayesian fused lasso

We will establish geometric ergodicity of the three variable Gibbs sampler for the Bayesian fused lasso by establishing a drift condition and an associated minorization condition.

C.1. Drift condition

Consider the drift function

$$V_{BFL}(\beta,\tau^2,w^2,\sigma^2) = (y-X\beta)^T(y-X\beta) + \beta^T\Sigma_{\tau,w}^{-1}\beta + \frac{\lambda_1^2}{4}\sum_{i=1}^{p}\tau_i^2 + \frac{\lambda_2^2}{4}\sum_{i=1}^{p-1}w_i^2\,. \tag{C.1}$$

Then $V_{BFL}:\mathbb{R}^p\times\mathbb{R}_+^p\times\mathbb{R}_+^{p-1}\times\mathbb{R}_+\to[0,\infty)$. To establish the drift condition we need to show that there exists a $0<\phi_{BFL}<1$ and $L_{BFL}>0$ such that,

$$\mathrm{E}_{(k)}\left[V_{BFL}(\beta,\tau^2,w^2,\sigma^2)\,|\,\beta_0,\tau_0^2,w_0^2,\sigma_0^2\right] \le \phi_{BFL}V_{BFL}(\beta_0,\tau_0^2,w_0^2,\sigma_0^2) + L_{BFL}\,,$$

for every $(\beta_0,\tau_0^2,w_0^2,\sigma_0^2)\in\mathbb{R}^p\times\mathbb{R}_+^p\times\mathbb{R}_+^{p-1}\times\mathbb{R}_+$. The left hand side is the expectation with respect to the MTD, that is,

$$\begin{aligned}
&\mathrm{E}_{(k)}\left[V_{BFL}(\beta,\tau^2,w^2,\sigma^2)\;|\;\beta_0,\tau_0^2,w_0^2,\sigma_0^2\right]\\
&=\int V_{BFL}(\beta,\tau^2,w^2,\sigma^2)f(\sigma^2\mid\beta_0,\tau_0^2,w_0^2,y)\\
&\qquad\times f(\tau^2,w^2\mid\beta_0,\sigma^2,y)f(\beta\mid\tau^2,w^2,\sigma^2,y)d\beta\,d\tau^2\,dw^2\,d\sigma^2\\
&=\int f(\sigma^2|\beta_0,\tau_0^2,w_0^2,y)\int f(\tau^2,w^2|\beta_0,\sigma^2,y)\\
&\quad\times\int V_{BFL}(\beta,\tau^2,w^2,\sigma^2)f(\beta|\tau^2,w^2,\sigma^2,y)d\beta\,d\tau^2\,dw^2\,d\sigma^2\\
&=\mathrm{E}_{\sigma^2}\left[\mathrm{E}_{\tau^2,w^2}\left[\mathrm{E}_{\beta}\left[V_{BFL}(\beta,\tau^2,w^2,\sigma^2)\,|\,\tau^2,w^2,\sigma^2,y\right]\,|\,\beta_0,\sigma^2,y\right]\,|\,\beta_0,\tau_0^2,w_0^2,y\right].
\end{aligned}$$

We will evaluate these sequentially, starting with the innermost expectation. By Lemma A.1,

$$\mathrm{E}_{(k)}\left[V_{BFL}(\beta,\tau^2,w^2,\sigma^2)\mid\tau^2,w^2,\sigma^2,y\right] \le y^Ty + \frac{\lambda_1^2}{4}\sum_{i=1}^{p}\tau_i^2 + \frac{\lambda_2^2}{4}\sum_{i=1}^{p-1}w_i^2 + p\sigma^2\,.$$

Next we move on to the expectation with respect to the full conditional of τ^2, w^2. Note that

$$\begin{aligned}
&\mathrm{E}_{\tau^2,w^2}\left[\mathrm{E}_{\beta}\left[V_{BFL}(\beta,\tau^2,w^2,\sigma^2) \mid \tau^2,w^2,\sigma^2,y\right] \mid \beta_0,\sigma^2,y\right] \\
&\leq y^T y + p\sigma^2 + \frac{\lambda_1^2}{4}\sum_{i=1}^{p}\left[\sqrt{\frac{\beta_{0,i}^2}{\lambda_1^2\sigma^2}} + \frac{1}{\lambda_1^2}\right] + \frac{\lambda_2^2}{4}\sum_{i=1}^{p-1}\left[\sqrt{\frac{(\beta_{0,i+1}-\beta_{0,i})^2}{\lambda_2^2\sigma^2}} + \frac{1}{\lambda_2^2}\right],
\end{aligned} \tag{C.2}$$

using the properties of the Inverse-Gaussian distribution mentioned in Appendix A. Since for $a, b > 0$, $2ab \leq a^2 + b^2$,

$$\begin{aligned}
&\mathrm{E}_{\tau^2,w^2}\left[\mathrm{E}_{\beta}\left[V_{BFL}(\beta,\tau^2,w^2,\sigma^2) \mid \tau^2,w^2,\sigma^2,y\right] \mid \beta_0,\sigma^2,y\right] \\
&\leq y^T y + p\sigma^2 + \frac{\lambda_1^2}{4}\sum_{i=1}^{p}\left[\frac{\beta_{0,i}^2}{2\sigma^2(n+p+2\alpha)} + \frac{(n+p+2\alpha)}{2\lambda_1^2} + \frac{1}{\lambda_1^2}\right] \\
&\quad + \frac{\lambda_2^2}{4}\sum_{i=1}^{p-1}\left[\frac{(\beta_{0,i+1}-\beta_{0,i})^2}{2\sigma^2(n+p+2\alpha)} + \frac{(n+p+2\alpha)}{2\lambda_2^2} + \frac{1}{\lambda_2^2}\right] \\
&\leq y^T y + \frac{p}{4}\left(2 + (n+p+2\alpha)\right) + p\sigma^2 + \frac{\lambda_1^2}{8(n+p+2\alpha)}\sum_{i=1}^{p}\frac{\beta_{0,i}^2}{\sigma^2} \\
&\quad + \frac{\lambda_2^2}{8(n+p+2\alpha)}\sum_{i=1}^{p-1}\frac{(\beta_{0,i+1}-\beta_{0,i})^2}{\sigma^2}.
\end{aligned}$$

Finally, the last expectation,

$$\begin{aligned}
&\mathrm{E}_{\sigma^2}\left[\mathrm{E}_{\tau^2,w^2}\left[\mathrm{E}_{\beta}\left[V_{BFL}(\beta,\tau^2,w^2,\sigma^2) \mid \tau^2,w^2,\sigma^2,y\right] \mid \beta_0,\sigma^2,y\right] \mid \beta_0,\tau_0^2,w_0^2,y\right] \\
&\leq y^T y + \frac{p}{4}\left(n+p+2\alpha+2\right) + p\mathrm{E}_{\sigma^2}[\sigma^2 \mid \beta_0,\tau_0^2,w_0^2,y] \\
&\quad + \frac{\lambda_1^2}{8(n+p+2\alpha)}\sum_{i=1}^{p}\mathrm{E}_{\sigma^2}\left[\frac{\beta_{0,i}^2}{\sigma^2} \mid \beta_0,\tau_0^2,w_0^2,y\right] \\
&\quad + \frac{\lambda_2^2}{8(n+p+2\alpha)}\sum_{i=1}^{p-1}\mathrm{E}_{\sigma^2}\left[\frac{(\beta_{0,i+1}-\beta_{0,i})^2}{\sigma^2} \mid \beta_0,\tau_0^2,w_0^2,y\right] \\
&\leq y^T y + \frac{p}{4}\left(n+p+2\alpha+2\right) \\
&\quad + p\frac{(y-X\beta_0)^T(y-X\beta_0) + \beta_0^T\Sigma_{\tau_0,w_0}^{-1}\beta_0 + 2\xi}{n+p+2\alpha-2} \\
&\quad + \frac{\lambda_1^2}{8(n+p+2\alpha)}\sum_{i=1}^{p}\frac{(n+p+2\alpha)\beta_{0,i}^2}{(y-X\beta_0)^T(y-X\beta_0) + \beta_0^T\Sigma_{\tau_0,w_0}^{-1}\beta_0 + 2\xi} \\
&\quad + \frac{\lambda_2^2}{8(n+p+2\alpha)}\sum_{i=1}^{p-1}\frac{(n+p+2\alpha)(\beta_{0,i+1}-\beta_{0,i})^2}{(y-X\beta_0)^T(y-X\beta_0) + \beta_0^T\Sigma_{\tau_0,w_0}^{-1}\beta_0 + 2\xi}
\end{aligned}$$

$$\leq y^T y + p\frac{(y - X\beta_0)^T(y - X\beta_0) + \beta_0^T \Sigma_{\tau_0, w_0}^{-1}\beta_0 + 2\xi}{n + p + 2\alpha - 2} + \frac{p}{4}(n + p + 2\alpha + 2) + \frac{\lambda_1^2}{8}\frac{\sum_{i=1}^p \beta_{0,i}^2}{\beta_0^T \Sigma_{\tau_0, w_0}^{-1}\beta_0} + \frac{\lambda_2^2}{8}\frac{\sum_{i=1}^{p-1}(\beta_{0,i+1} - \beta_{0,i})^2}{\beta_0^T \Sigma_{\tau_0, w_0}^{-1}\beta_0}\,. \tag{C.3}$$

Using (B.2),

$$\beta_0^T \Sigma_{\tau_0, w_0}^{-1}\beta_0 \geq \sum_{i=1}^p \frac{\beta_{0,i}^2}{\tau_{0,i}^2} \quad \text{and} \quad \beta_0^T \Sigma_{\tau_0, w_0}^{-1}\beta_0 \geq \sum_{i=1}^{p-1} \frac{(\beta_{0,i+1} - \beta_{0,i})^2}{w_{0,i}^2}\,. \tag{C.4}$$

Using (C.4) in (C.3),

$$\begin{aligned}
&\mathrm{E}_{\sigma^2}\left[\mathrm{E}_{\tau^2, w^2}\left[\mathrm{E}_{\beta}\left[V_{BFL}(\beta, \tau^2, w^2, \sigma^2) \mid \tau^2, w^2, \sigma^2, y\right] \mid \beta_0, \sigma^2, y\right] \mid \beta_0, \tau_0^2, w_0^2, y\right] \\
&\leq y^T y + p\frac{(y - X\beta_0)^T(y - X\beta_0) + \beta_0^T \Sigma_{\tau_0, w_0}^{-1}\beta_0 + 2\xi}{n + p + 2\alpha - 2} \\
&\quad + \frac{p}{4}(n + p + 2\alpha + 2) + \frac{\lambda_1^2}{8}\frac{\sum_{i=1}^p \beta_{0,i}^2}{\sum_{i=1}^p \beta_{0,i}^2/\tau_{0,i}^2} + \frac{\lambda_2^2}{8}\frac{\sum_{i=1}^{p-1}(\beta_{0,i+1} - \beta_{0,i})^2}{\sum_{i=1}^{p-1}(\beta_{0,i+1} - \beta_{0,i})^2/w_{0,i}^2}\,.
\end{aligned}$$

By Lemma A.2,

$$\begin{aligned}
&\mathrm{E}_{\sigma^2}\left[\mathrm{E}_{\tau^2, w^2}\left[\mathrm{E}_{\beta}\left[V_{BFL}(\beta, \tau^2, w^2, \sigma^2) \mid \tau^2, w^2, \sigma^2, y\right] \mid \beta_0, \sigma^2, y\right] \mid \beta_0, \tau_0^2, w_0^2, y\right] \\
&\leq y^T y + \frac{p}{4}(n + p + 2\alpha + 2) + \frac{2p\xi}{n + p + 2\alpha - 2} \\
&\quad + \frac{p}{n + p + 2\alpha - 2}\left((y - X\beta_0)^T(y - X\beta_0) + \beta_0^T \Sigma_{\tau_0, w}^{-1}\beta_0\right) \\
&\quad + \frac{\lambda_1^2}{8}\sum_{i=1}^p \tau_{0,i}^2 + \frac{\lambda_2^2}{8}\sum_{i=1}^{p-1} w_{0,i}^2 \\
&\leq \phi_{BFL} V(\beta_0, \tau_0^2, w_0^2, \sigma_0^2) + L_{BFL}\,,
\end{aligned}$$

where

$$\phi_{BFL} = \max\left\{\frac{p}{n + p + 2\alpha - 2}, \frac{1}{2}\right\} < 1 \text{ for } n \geq 3 \quad \text{and} \tag{C.5}$$

$$L_{BFL} = y^T y + \frac{p}{2}(n + p + 2\alpha + 2) + \frac{2p\xi}{n + p + 2\alpha - 2}\,. \tag{C.6}$$

C.2. Minorization

To establish a one-step minorization, we need to show that for all sets C_d defined as

$$C_d = \{(\beta, \tau^2, w^2, \sigma^2) : V_{BFL}(\beta, \tau^2, w^2, \sigma^2) \leq d\}\,,$$

there exists an $\epsilon > 0$ and a density q such that for all $(\beta_0, \tau_0^2, w_0^2, \sigma_0^2) \in C_d$

$$k_{BFL}(\beta, \tau^2, w^2, \sigma^2 \mid \beta_0, \tau_0^2, w_0^2, \sigma_0^2) \geq \epsilon\, q(\beta, \tau^2, w^2, \sigma^2)\,.$$

To establish this condition, recall that,

$$k_{BFL}(\beta, \tau^2, w^2, \sigma^2 \mid \beta_0, \tau_0^2, w_0^2, \sigma_0^2) = f(\beta \mid \tau^2, w^2, \sigma^2, y)\, f(\tau^2, w^2 \mid \beta_0, \sigma^2, y) \times f(\sigma^2 \mid \beta_0, \tau_0^2, w_0^2, y)\,.$$

For our drift function, for all $(\beta_0, \tau_0^2, w_0^2, \sigma_0^2) \in C_d$ the following relation holds due to (B.2):

$$(y - X\beta_0)^T(y - X\beta_0) + \beta_0^T \Sigma_{\tau_0, w_0}^{-1} \beta_0 + \frac{\lambda_1^2}{4} \sum_{i=1}^{p} \tau_{0,i}^2 + \frac{\lambda_2^2}{4} \sum_{i=1}^{p-1} w_{0,i}^2 \leq d$$

$$\|(y - X\beta_0)\|^2 + \sum_{i=1}^{p} \frac{\beta_{0,i}^2}{\tau_{0,i}^2} + \sum_{i=1}^{p-1} \frac{(\beta_{0,i+1} - \beta_{0,i})^2}{w_{0,i}^2} + \frac{\lambda_1^2}{4} \sum_{i=1}^{p} \tau_{0,i}^2 + \frac{\lambda_2^2}{4} \sum_{i=1}^{p-1} w_{0,i}^2 \leq d\,.$$

Using the above and Lemma A.2, for each $\beta_{0,j}$,

$$\beta_{0,j}^2 \leq \sum_{i=1}^{p} \beta_{0,i}^2 \leq \left(\sum_{i=1}^{p} \tau_{0,i}^2 \right) \left(\sum_{i=1}^{p} \frac{\beta_{0,i}^2}{\tau_{0,i}^2} \right) \leq \frac{4d^2}{\lambda_1^2} := d_1^2\,, \tag{C.7}$$

and similarly for each $i = 1, \ldots, p-1$

$$(\beta_{0,j+1} - \beta_{0,j})^2 \leq \left(\sum_{i=1}^{p-1} w_{0,i}^2 \right) \left(\sum_{i=1}^{p-1} \frac{(\beta_{0,i} - \beta_{0,i})^2}{w_{0,i}^2} \right) \leq \frac{4d^2}{\lambda_2^2} := d_2^2\,. \tag{C.8}$$

With these bounds involving β_0 and using Lemma A.4,

$$\begin{aligned}
&f(\tau^2, w^2 \mid \beta_0, \sigma^2, y) \\
&= f(\tau^2 \mid \beta_0, \sigma^2, y)\, f(w^2 \mid \beta_0, \sigma^2, y) \\
&\geq \prod_{i=1}^{p} \exp\left\{ -\sqrt{\frac{\lambda_1^2 d_1^2}{\sigma^2}} \right\} q_i(\tau_i^2 \mid \sigma^2) \prod_{i=1}^{p-1} \exp\left\{ -\sqrt{\frac{\lambda_2^2 d_2^2}{\sigma^2}} \right\} h_i(w_i^2 \mid \sigma^2) \\
&= \exp\left\{ -p\sqrt{\frac{\lambda_1^2 d_1^2}{\sigma^2}} - p\sqrt{\frac{\lambda_2^2 d_2^2}{\sigma^2}} \right\} \left[\prod_{i=1}^{p} q_i(\tau_i^2 \mid \sigma^2) \right] \left[\prod_{i=1}^{p-1} h_i(w_i^2 \mid \sigma^2) \right]\,.
\end{aligned}$$

Since for $a, b \geq 0, 2ab \leq a^2 + b^2$,

$$\begin{aligned}
f(\tau^2, w^2 \mid \beta_0, \sigma^2, y) \geq \exp&\left\{ -1 - \frac{p^2 \lambda_2^2 d_2^2}{2\sigma^2} - \frac{p^2 \lambda_1^2 d_1^2}{2\sigma^2} \right\} \\
&\times \left[\prod_{i=1}^{p} q_i(\tau_i^2 \mid \sigma^2) \right] \left[\prod_{i=1}^{p-1} h_i(w_i^2 \mid \sigma^2) \right]\,,
\end{aligned} \tag{C.9}$$

where q_i and h_i are densities of the reciprocal of an Inverse-Gaussian distribution with parameters $\sqrt{\lambda_1^2 \sigma^2 / d_1^2}$ and λ_1^2, and $\sqrt{\lambda_2^2 \sigma^2 / d_2^2}$ and λ_2^2, respectively.

Recall the decomposition $\Sigma^{-1}_{\tau_0,w_0} = L_{0,1} + L_{0,2}$ in (B.1); here the 0 in the index indicates τ_0^2 and w_0^2 entries. Here $L_{0,1}$ is the diagonal matrix with entries $1/(2\tau^2_{0,i})$. Then since

$$\begin{aligned} & y^T X(X^TX + L_{0,1} + L_{0,2})X^T y \geq y^T X(X^TX + L_{0,1})X^T y \\ \Rightarrow\ & y^T X(X^TX + L_{0,1} + L_{0,2})^{-1}X^T y \leq y^T X(X^TX + L_{0,1})^{-1}X^T y\,. \end{aligned}$$

Using the above, the fact that for each $i = 1, \ldots, p$, $2\tau^2_{0,i} \leq 8d/\lambda_1^2$, and Lemma A.5,

$$\begin{aligned} & (y - X\beta_0)^T(y - X\beta_0) + \beta_0^T \Sigma^{-1}_{\tau_0,w_0}\beta_0 \\ & \geq y^T y - y^T X \left(X^TX + \Sigma^{-1}_{\tau,w}\right)^{-1} X^T y \\ & \geq y^T y - y^T X(X^TX + L_{0,1})^{-1}X^T y \\ & \geq y^T y - y^T X \left(X^TX + \frac{\lambda_1^2}{8d} I_p\right)^{-1} X^T y\,. \end{aligned} \tag{C.10}$$

Using (C.10) and the fact that for $(\beta_0, \tau_0^2, w_0^2, \sigma_0^2) \in C_d$, $(y - X\beta_0)^T(y - X\beta_0) + \beta_0^T \Sigma^{-1}_{\tau_0,w_0}\beta_0 \leq d$,

$$\begin{aligned} & \exp\left\{-\frac{1}{2} - \frac{p^2\lambda_2^2 d_2^2}{2\sigma^2} - \frac{1}{2} - \frac{p^2\lambda_1^2 d_1^2}{2\sigma^2}\right\} f(\sigma^2 \mid \beta_0, \tau_0^2, w_0^2, y) \\ & = \exp\left\{-1 - \frac{p^2\lambda_2^2 d_2^2}{2\sigma^2} - \frac{p^2\lambda_1^2 d_1^2}{2\sigma^2}\right\} \frac{\left(\frac{(y-X\beta_0)^T(y-X\beta_0) + \beta_0^T\Sigma^{-1}_{\tau_0,w_0}\beta_0 + 2\xi}{2}\right)^{\frac{n+p}{2}+\alpha}}{\Gamma\left(\frac{n+p}{2} + \alpha\right)} \\ & \quad \times \left(\sigma^2\right)^{-\frac{n+p}{2}-\alpha-1} \exp\left\{-\frac{(y - X\beta_0)^T(y - X\beta_0) + \beta_0^T\Sigma^{-1}_{\tau_0,w_0}\beta_0 + 2\xi}{2\sigma^2}\right\} \\ & \geq e^{-1}\left(\frac{y^Ty - y^TX(X^TX + \lambda_1^2(8d)^{-1}I_p)^{-1}X^Ty + 2\xi}{2}\right)^{\frac{n+p}{2}+\alpha} \frac{1}{\Gamma\left(\frac{n+p}{2} + \alpha\right)} \\ & \quad \times \left(\sigma^2\right)^{-\frac{n+p}{2}-\alpha-1} \exp\left\{-\frac{d + 2\xi + p^2\lambda_2^2 d_2^2 + p^2\lambda_1^2 d_1^2}{2\sigma^2}\right\} \\ & = e^{-1}\left(\frac{y^Ty - y^TX(X^TX + \lambda_1^2(8d)^{-1}I_p)^{-1}X^Ty + 2\xi}{d + 2\xi + p^2\lambda_2^2 d_2^2 + p^2\lambda_1^2 d_1^2}\right)^{\frac{n+p}{2}+\alpha} q(\sigma^2)\,, \end{aligned} \tag{C.11}$$

where $q(\sigma^2)$ is the Inverse-Gamma density with parameters, $(n+p)/2 + \alpha$ and $d + 2\xi + p^2\lambda_2^2 d_2^2 + p^2\lambda_1^2 d_1^2$. Finally, using (C.9) and (C.11),

$$\begin{aligned} k_{BFL}(\beta, \tau^2, w^2, \sigma^2 \mid \beta_0, \tau_0^2, w_0^2, \sigma_0^2) \geq \epsilon\ & f(\beta \mid \tau^2, w^2, \sigma^2, y)\ q(\sigma^2) \\ & \times \left[\prod_{i=1}^{p} q_i(\tau_i^2 \mid \sigma^2)\right] \left[\prod_{i=1}^{p-1} h_i(w_i^2 \mid \sigma^2)\right], \end{aligned}$$

where

$$\epsilon = e^{-1}\left(\frac{y^T y - y^T X(X^T X + \lambda_1(8d)^{-1}I_p)^{-1}X^T y + 2\xi}{d + 2\xi + p^2\lambda_2^2 d_2^2 + p^2\lambda_1^2 d_1^2}\right)^{\frac{n+p}{2}+\alpha}.$$

C.3. Starting values

Starting value $(\beta_0, \tau_0^2, w_0^2, \sigma_0^2)$ can be chosen so that

$$(\beta_0, \tau_0^2, w_0^2, \sigma_0^2) = \arg\min V_{BFL}(\beta, \tau^2, w^2, \sigma^2).$$

We will find the minimum by profiling out τ^2 and w^2. By (B.2) in Appendix B.2,

$$\frac{\partial V_{BFL}}{\partial \tau_{0,i}^2} = 0 \Rightarrow = -\frac{\beta_{0,i}^2}{\tau_{0,i}^4} + \frac{\lambda_1^2}{4} = 0 \Rightarrow \tau_{0,i}^2 = \sqrt{\frac{4\beta_{0,i}^2}{\lambda_1^2}}$$

$$\frac{\partial V_{BFL}}{\partial w_{0,i}^2} = 0 \Rightarrow = -\frac{(\beta_{0,i+1} - \beta_{0,i})^2}{w_{0,i}^4} + \frac{\lambda_2^2}{4} = 0 \Rightarrow w_{0,i}^2 = \sqrt{\frac{4(\beta_{0,i+1} - \beta_{0,i})^2}{\lambda_2^2}}.$$

The β_0 that minimizes V_{BFL} is,

$$\begin{aligned}\beta_0 &= \arg\min_{\beta\in\mathbb{R}^p}\left\{(y - X\beta)^T(y - X\beta) + \sum_{i=1}^{p}\frac{\lambda_1\beta_i^2}{2\sqrt{\beta_i^2}} + \sum_{i=1}^{p-1}\frac{\lambda_2(\beta_{i+1} - \beta_i)^2}{2\sqrt{(\beta_{i+1} - \beta_i)^2}}\right.\\ &\quad\left. + \sum_{i=1}^{p}\frac{\lambda_1^2}{4}\sqrt{\frac{4\beta_i^2}{\lambda_1^2}} + \frac{\lambda_2^2}{4}\sum_{i=1}^{p-1}\sqrt{\frac{4(\beta_{i+1} - \beta_i)^2}{\lambda_2^2}}\right\}\\ &= \arg\min_{\beta\in\mathbb{R}^p}\left\{(y - X\beta)^T(y - X\beta) + \lambda_1\sum_{i=1}^{p}|\beta_i| + \lambda_2\sum_{i=1}^{p-1}|\beta_{i+1} - \beta_i|\right\},\end{aligned}$$

which is equivalent to the fused lasso solution. Thus, a reasonable starting value is β_0 being the fused lasso estimate, $\tau_{0,i}^2 = 2|\beta_{0,i}|/\lambda_1$ and $w_{0,i}^2 = 2|\beta_{0,i+1} - \beta_{0,i}|/\lambda_2$.

Appendix D: Proof of geometric ergodicity in the Bayesian group lasso

D.1. Drift condition

Consider the drift function

$$V_{BGL}(\beta, \tau^2, \sigma^2) = (y - X\beta)^T(y - X\beta) + \beta^T D_\tau^{-1}\beta + \frac{\lambda^2}{4}\sum_{k=1}^{K}\tau_k^2. \tag{D.1}$$

For the drift condition we need to show that there exists a $0 < \phi_{BGL} < 1$ and $L_{BGL} > 0$ such that,

$$\mathrm{E}_{(k)}\left[V_{BGL}(\beta, \tau^2, \sigma^2)\,|\,\beta_0, \tau_0^2, \sigma_0^2\right] \le \phi_{BGL}V_{BGL}(\beta_0, \tau_0^2, \sigma_0^2) + L_{BGL},$$

for every $(\beta_0, \tau_0^2, \sigma_0^2) \in \mathbb{R}^p \times \mathbb{R}_+^K \times \mathbb{R}_+$. Just as in the proof for BFL,

$$\begin{aligned}&\mathrm{E}_{(k)}\left[V_{BGL}(\beta, \tau^2, \sigma^2) \mid \beta_0, \tau_0^2, \sigma_0^2\right] \\ &= \mathrm{E}_{\sigma^2}\left[\mathrm{E}_{\tau^2}\left[\mathrm{E}_{\beta}\left[V_{BGL}(\beta, \tau^2, \sigma^2) \mid \tau^2, \sigma^2, y\right] \mid \beta_0, \sigma^2, y\right] \mid \beta_0, \tau_0^2, y\right].\end{aligned}$$

We will evaluate the expectations sequentially, starting with the innermost expectation. By Lemma A.1 and following the steps as before (C.2),

$$\begin{aligned}&\mathrm{E}_{\tau^2}\left[\mathrm{E}_{\beta}\left[V_{BGL}(\beta, \tau^2, \sigma^2) \mid \tau^2, \sigma^2, y\right] \mid \beta_0, \sigma^2, y\right] \\ &\leq y^T y + p\sigma^2 + \frac{\lambda^2}{4}\sum_{k=1}^{K}\left[\sqrt{\frac{\beta_{0,G_k}^T \beta_{0,G_k}}{\lambda^2\sigma^2}} + \frac{1}{\lambda^2}\right].\end{aligned}$$

Let $M = \max\{m_1, \ldots, m_K\}$. Then,

$$\begin{aligned}&\mathrm{E}_{\tau^2}\left[\mathrm{E}_{\beta}\left[V_{BGL}(\beta, \tau^2, \sigma^2) \mid \tau^2, \sigma^2, y\right] \mid \beta_0, \sigma^2, y\right] \\ &\leq y^T y + p\sigma^2 + \frac{\lambda^2}{4}\sum_{k=1}^{K}\left[\frac{\beta_{0,G_k}^T \beta_{0,G_k}}{2\sigma^2 M(n+p+2\alpha)} + \frac{M(n+p+2\alpha)}{2\lambda^2} + \frac{1}{\lambda^2}\right] \\ &\leq y^T y + p\sigma^2 + \frac{p}{4}\left(1 + \frac{M(n+p+2\alpha)}{2}\right) + \frac{\lambda^2 \sum_{k=1}^{K} \beta_{0,G_k}^T \beta_{0,G_k}}{8\sigma^2 M(n+p+2\alpha)}.\end{aligned}$$

For the last expectation, using steps as before (C.3), we get

$$\begin{aligned}&\mathrm{E}_{\sigma^2}\left[\mathrm{E}_{\tau^2}\left[\mathrm{E}_{\beta}\left[V_{BGL}(\beta, \tau^2, \sigma^2) \mid \tau^2, \sigma^2, y\right] \mid \beta_0, \sigma^2, y\right] \mid \beta_0, \tau_0^2, y\right] \\ &\leq y^T y + \frac{p}{4}\left(1 + \frac{M(n+p+2\alpha)}{2}\right) + \frac{\lambda^2}{8M}\left(\frac{\sum_{k=1}^{K} \beta_{0,G_k}^T \beta_{0,G_k}}{\beta_0^T D_{\tau_0}^{-1} \beta_0}\right) \\ &\quad + p\frac{\|y - X\beta_0\|^2 + \beta_0^T D_{\tau_0}^{-1}\beta_0 + 2\xi}{n+p+2\alpha-2}.\end{aligned}$$

Recall that,

$$D_{\tau_0} = \mathrm{diag}(\underbrace{\tau_{0,1}^2, \ldots, \tau_{0,1}^2}_{m_1}, \underbrace{\tau_{0,2}^2, \ldots, \tau_{0,2}^2}_{m_2}, \ldots, \underbrace{\tau_{0,K}^2, \ldots, \tau_{0,K}^2}_{m_K}).$$

Let the diagonals of D_{τ_0} be $\tau_{0,*i}^2$ for $i = 1, \ldots, p$. Then $\beta_0^T D_{\tau_0}^{-1}\beta_0 = \sum_{i=1}^{p} \beta_{0,i}^2 / \tau_{0,*i}^2$ and $\sum_{i=1}^{p} \tau_{0,*i}^2 \leq M \sum_{k=1}^{K} \tau_{0,k}^2$. Using this and Lemma A.2,

$$\begin{aligned}&\mathrm{E}_{\sigma^2}\left[\mathrm{E}_{\tau^2}\left[\mathrm{E}_{\beta}\left[V_{BGL}(\beta, \tau^2, \sigma^2) \mid \tau^2, \sigma^2, y\right] \mid \beta_0, \sigma^2, y\right] \mid \beta_0, \tau_0^2, y\right] \\ &\leq y^T y + \frac{p}{4}\left(1 + \frac{M(n+p+2\alpha)}{2}\right) + \frac{\lambda^2}{8}\sum_{k=1}^{K}\tau_{0,k}^2 \\ &\quad + p\frac{(y - X\beta_0)^T(y - X\beta_0) + \beta_0^T D_{\tau_0}^{-1}\beta_0 + 2\xi}{n+p+2\alpha-2} \\ &\leq \phi_{BGL} V_{BGL}(\beta_0, \tau_0^2, \sigma_0^2) + L_{BGL},\end{aligned}$$

where

$$\phi_{BGL} = \max\left\{\frac{p}{n+p+2\alpha-2}, \frac{1}{2}\right\} < 1 \text{ for } n \geq 3 \quad \text{and} \tag{D.2}$$

$$L_{BGL} = y^T y + \frac{p}{4}\left(1 + \frac{M(n+p+2\alpha)}{2}\right) + \frac{2p\xi}{n+p+2\alpha-2}\,. \tag{D.3}$$

D.2. Minorization condition

For $d > 0$, define $C_d = \{(\beta, \tau^2, \sigma^2) : V_{BGL}(\beta, \tau^2, \sigma^2) \leq d\}$. To establish the minorization condition, we recall that,

$$k_{BGL}(\beta, \tau^2, \sigma^2 \mid \beta_0, \tau_0^2, \sigma_0^2) = f(\beta \mid \tau^2, \sigma^2, y)\; f(\tau^2 \mid \beta_0, \sigma^2, y)\; f(\sigma^2 \mid \beta_0, \tau_0^2, y)\,. \tag{D.4}$$

By our choice of drift function, for all $(\beta_0, \tau_0^2, \sigma_0^2) \in C_d$ the following relation holds,

$$(y - X\beta_0)^T(y - X\beta_0) + \beta_0^T D_{\tau_0}^{-1}\beta_0 + \frac{\lambda^2}{4}\sum_{k=1}^{K}\tau_{0,k}^2 \leq d\,. \tag{D.5}$$

By (D.5), each of $\beta_0^T D_{\tau_0}^{-1}\beta_0$ and $(\lambda^2/4)\sum_{k=1}^{K}\tau_{0,k}^2$ is less than or equal to d, so $\beta_{0,G_k}^T\beta_{0,G_k} \leq 4d^2/\lambda^2 := d_1^2$ for all $k = 1, \ldots, K$. By Lemma A.4,

$$f(\tau^2 \mid \beta_0, \sigma^2, y) \geq \exp\left\{-\frac{1}{2} - \frac{K^2\lambda^2 d_1^2}{2\sigma^2}\right\}\prod_{k=1}^{K} q_k(\tau_k^2 \mid \sigma^2)\,, \tag{D.6}$$

where q_k is the density of the reciprocal of an Inverse-Gaussian distribution with parameters $\sqrt{\lambda^2\sigma^2/d_1^2}$ and λ^2. Now, since for each $i = 1, \ldots, p$, $\tau_{0,i}^2 \leq 4d/\lambda^2$, by Lemma A.5

$$(y - X\beta_0)^T(y - X\beta_0) + \beta_0^T D_{\tau_0}^{-1}\beta_0 \geq y^T y - y^T X\left(X^T X + \frac{\lambda^2}{4d}I_p\right)^{-1} X^T y\,. \tag{D.7}$$

Using (D.7) and following steps as before (C.11), we arrive at the following,

$$\exp\left\{-\frac{1}{2} - \frac{K^2\lambda^2 d_1^2}{2\sigma^2}\right\} f(\sigma^2 \mid \beta_0, \tau_0^2, y)$$

$$\geq e^{-\frac{1}{2}}\left(\frac{y^T y - y^T X(X^T X + \lambda^2(4d)^{-1}I_p)^{-1}X^T y + 2\xi}{d + 2\xi + K^2\lambda^2 d_1^2}\right)^{\frac{n+p}{2}+\alpha} q(\sigma^2)\,, \tag{D.8}$$

where $q(\sigma^2)$ is the Inverse-Gamma density with parameters, $(n+p)/2 + \alpha$ and $d + 2\xi + K^2\lambda^2 d_1^2$.

Finally, using (D.6) and (D.8) in (D.4)

$$k_{BGL}(\beta, \tau^2, \sigma^2 \mid \beta_0, \tau_0^2, \sigma_0^2) \geq\; \epsilon\, f(\beta \mid \tau^2, \sigma^2, y)\, q(\sigma^2)\prod_{k=1}^{K} q_k(\tau^2 \mid \sigma^2)\,, \tag{D.9}$$

where

$$\epsilon = e^{-\frac{1}{2}}\left(\frac{y^T y - y^T X(X^T X + \lambda^2(4d)^{-1}I_p)^{-1}X^T y + 2\xi}{d + 2\xi + 4K^2 d^2}\right)^{\frac{n+p}{2}+\alpha}\,. \tag{D.10}$$

D.3. Starting values

As before, we first differentiate with respect to τ^2 and then with respect to β. Note that

$$\frac{\partial V_{BGL}}{\partial \tau^2_{0,k}} = 0 \Rightarrow = -\frac{\beta^T_{0,G_k}\beta_{0,G_k}}{\tau^4_{0,k}} + \frac{\lambda^2}{4} = 0 \Rightarrow \tau^2_{0,k} = \sqrt{\frac{4\beta^T_{0,G_k}\beta_{0,G_k}}{\lambda^2}}\,.$$

Thus, the β_0 that minimizes V_{BGL} is then,

$$\begin{aligned}\beta_0 &= \arg\min_{\beta\in\mathbb{R}^p}\ (y-X\beta)^T(y-X\beta) + \sum_{k=1}^K \frac{\lambda\,\beta^T_{G_k}\beta_{G_k}}{2\sqrt{\beta^T_{G_k}\beta_{G_k}}} + \frac{\lambda^2}{4}\sum_{k=1}^K \sqrt{\frac{4\beta^T_{G_k}\beta_{G_k}}{\lambda^2}} \\ &= \arg\min_{\beta\in\mathbb{R}^p}\ (y-X\beta)^T(y-X\beta) + \lambda\sum_{k=1}^K \sqrt{\beta^T_{G_k}\beta_{G_k}}\,,\end{aligned}$$

which is equivalent to the group lasso solution. Thus a reasonable starting value for the Markov chain is β_0 being the group lasso estimate and $\tau^2_{0,k} = 2\sqrt{\beta^T_{0,G_k}\beta_{0,G_k}}/\lambda$.

Appendix E: Proof of geometric ergodicity in the Bayesian sparse group lasso

E.1. Drift condition

Consider the drift function

$$V_{BSGL}(\beta,\tau^2,\gamma^2,\sigma^2) = (y-X\beta)^T(y-X\beta) + \beta^T V^{-1}_{\tau,\gamma}\beta + \frac{\lambda_1^2}{4}\sum_{k=1}^K \tau_k^2 + \frac{\lambda_2^2}{4}\sum_{k=1}^K\sum_{j=1}^{m_k}\gamma^2_{k,j}\,. \tag{E.1}$$

By Lemma A.1 and following the steps as before (C.2)

$$\begin{aligned}&\mathrm{E}_{\tau^2,\gamma^2}\left[\mathrm{E}_\beta\left[V_{BSGL}(\beta,\tau^2,\gamma^2,\sigma^2)\ \middle|\ \tau^2,\gamma^2,\sigma^2,y\right]\ \middle|\ \beta_0,\sigma^2,y\right] \\ &\leq y^Ty + p\sigma^2 + \frac{\lambda_1^2}{4}\sum_{k=1}^K\left[\sqrt{\frac{\beta^T_{0,G_k}\beta_{0,G_k}}{\lambda_1^2\sigma^2}} + \frac{1}{\lambda_1^2}\right] + \frac{\lambda_2^2}{4}\sum_{k=1}^K\sum_{j=1}^{m_k}\left[\sqrt{\frac{\beta^2_{0,k,j}}{\lambda_2^2\sigma^2}} + \frac{1}{\lambda_2^2}\right]\,.\end{aligned}$$

Define $M = \max\{m_1,\ldots,m_K\}$. In addition, define

$$A = \left(1 + \frac{\lambda_1^2}{\lambda_2^2} + \frac{\lambda_2^2}{\lambda_1^2}\right)(n+p+2\alpha)\,.$$

Then,

$$\begin{aligned}&\mathrm{E}_{\tau^2,\gamma^2}\left[\mathrm{E}_\beta\left[V_{BSGL}(\beta,\tau^2,\sigma^2)\ \middle|\ \tau^2,\sigma^2,y\right]\ \middle|\ \beta_0,\sigma^2,y\right] \\ &\leq y^Ty + p\sigma^2 + \frac{\lambda_1^2}{4}\sum_{k=1}^K\left[\frac{\beta^T_{0,G_k}\beta_{0,G_k}}{2\sigma^2 AM} + \frac{AM}{2\lambda_1^2} + \frac{1}{\lambda_1^2}\right]\end{aligned}$$

$$+\frac{\lambda_2^2}{4}\sum_{k=1}^{K}\sum_{j=1}^{m_k}\left[\frac{\beta_{0,k,j}^2}{2\sigma^2 AM}+\frac{AM}{2\lambda_2^2}+\frac{1}{\lambda_2^2}\right]$$
$$=y^Ty+p\sigma^2+\frac{p}{4}(2+AM)+\left[\frac{\lambda_1^2+\lambda_2^2}{8AM}\right]\frac{\beta_0^T\beta_0}{\sigma^2}\,.$$

For the last expectation, using steps as before (C.3), we get

$$\begin{aligned}&\mathrm{E}_{\sigma^2}\left[\mathrm{E}_{\tau^2,\gamma^2}\left[\mathrm{E}_{\beta}\left[V_{BSGL}(\beta,\tau^2,\gamma^2,\sigma^2)\mid\tau^2,\gamma^2,\sigma^2,y\right]\mid\beta_0,\sigma^2,y\right]\mid\beta_0,\tau_0^2,\gamma_0^2,y\right]\\&\le y^Ty+\frac{p}{4}(2+AM)+(\lambda_1^2+\lambda_2^2)\left[8M\left(1+\frac{\lambda_1^2}{\lambda_2^2}+\frac{\lambda_2^2}{\lambda_1^2}\right)\right]^{-1}\left(\frac{\beta_0^T\beta_0}{\beta_0^TV_{\tau_0,\gamma_0}^{-1}\beta_0}\right)\\&\quad+p\frac{(y-X\beta_0)^T(y-X\beta_0)+\beta_0^TV_{\tau_0,\gamma_0}^{-1}\beta_0+2\xi}{n+p+2\alpha-2}\,.\end{aligned}\tag{E.2}$$

Let $v_{0,i}$ denote the diagonals of V_{τ_0,γ_0}. Then by Lemma A.2, and the fact that the harmonic mean of positive numbers is less than their arithmetic mean,

$$\begin{aligned}\frac{\beta_0^T\beta_0}{\beta_0^TV_{\tau_0,\gamma_0}^{-1}\beta_0}\le\sum_{i=1}^{p}v_{0,i}&=\frac{1}{2}\sum_{k=1}^{K}\sum_{j=1}^{m_k}2\left(\frac{1}{\tau_{0,k}^2}+\frac{1}{\gamma_{0,k,j}^2}\right)^{-1}\\&\le\frac{1}{2}\sum_{k=1}^{K}\sum_{j=1}^{m_k}\frac{\tau_{0,k}^2+\gamma_{0,k,j}^2}{2}\le\frac{M}{4}\sum_{k=1}^{K}\tau_{0,k}^2+\frac{1}{4}\sum_{k=1}^{K}\sum_{j=1}^{m_k}\gamma_{0,k,j}^2\,.\end{aligned}\tag{E.3}$$

Using (E.3) in (E.2),

$$\begin{aligned}&\mathrm{E}_{\sigma^2}\left[\mathrm{E}_{\tau^2}\left[\mathrm{E}_{\beta}\left[V_{BSGL}(\beta,\tau^2,\sigma^2)\mid\tau^2,\sigma^2,y\right]\mid\beta_0,\sigma^2,y\right]\mid\beta_0,\tau_0^2,y\right]\\&\le y^Ty+\frac{p}{4}(2+AM)\\&\quad+(\lambda_1^2+\lambda_2^2)\left[8M\left(1+\frac{\lambda_1^2}{\lambda_2^2}+\frac{\lambda_2^2}{\lambda_1^2}\right)\right]^{-1}\left(\frac{M}{4}\sum_{k=1}^{K}\tau_{0,k}^2+\frac{1}{4}\sum_{k=1}^{K}\sum_{j=1}^{m_k}\gamma_{0,k,j}^2\right)\\&\quad+p\frac{(y-X\beta_0)^T(y-X\beta_0)+\beta_0^TV_{\tau_0,\gamma_0}^{-1}\beta_0+2\xi}{n+p+2\alpha-2}\\&\le y^Ty+\frac{p}{4}(2+AM)+\frac{2p\xi}{n+p+2\alpha-2}\\&\quad+\frac{p}{n+p+2\alpha-2}\left[(y-X\beta_0)^T(y-X\beta_0)+\beta_0^TV_{\tau_0,\gamma_0}^{-1}\beta_0\right]\\&\quad+\left(1+\frac{\lambda_2^2}{\lambda_1^2}\right)\left[8\left(1+\frac{\lambda_1^2}{\lambda_2^2}+\frac{\lambda_2^2}{\lambda_1^2}\right)\right]^{-1}\left(\frac{\lambda_1^2}{4}\sum_{k=1}^{K}\tau_{0,k}^2\right)\\&\quad+\left(1+\frac{\lambda_1^2}{\lambda_2^2}\right)\left[8M\left(1+\frac{\lambda_1^2}{\lambda_2^2}+\frac{\lambda_2^2}{\lambda_1^2}\right)\right]^{-1}\left(\frac{\lambda_2^2}{4}\sum_{k=1}^{K}\sum_{j=1}^{m_k}\gamma_{0,k,j}^2\right)\\&\le\phi_{BSGL}\,V_{BSGL}(\beta_0,\tau_0^2,\gamma_0^2,\sigma_0^2)+L_{BSGL}\,,\end{aligned}$$

where for $n \geq 3$

$$\phi_{BSGL} = \max\left\{ \frac{p}{n+p+2\alpha-2}, \frac{\left(1+\frac{\lambda_2^2}{\lambda_1^2}\right)}{8\left(1+\frac{\lambda_1^2}{\lambda_2^2}+\frac{\lambda_2^2}{\lambda_1^2}\right)}, \frac{\left(1+\frac{\lambda_1^2}{\lambda_2^2}\right)}{8M\left(1+\frac{\lambda_1^2}{\lambda_2^2}+\frac{\lambda_2^2}{\lambda_1^2}\right)} \right\} < 1, \tag{E.4}$$

and

$$L_{BSGL} = y^T y + \frac{p}{4}(2+AM) + \frac{2p\xi}{n+p+2\alpha-2}\,. \tag{E.5}$$

E.2. Minorization

For $d > 0$, define $C_d = \{(\beta, \tau^2, \gamma^2, \sigma^2) : V(\beta, \tau^2, \gamma^2, \sigma^2) \leq d\}$. Recall that,

$$\begin{aligned} k_{BSGL}(\beta, \tau^2, \gamma^2, \sigma^2 | \beta_0, \tau_0^2, \gamma_0^2, \sigma_0^2) &= f(\beta|\tau^2, \gamma^2, \sigma^2, y)\; f(\tau^2, \gamma^2|\beta_0, \sigma^2, y) \\ &\quad \times\; f(\sigma^2|\beta_0, \tau_0^2, \gamma_0^2, y)\,. \end{aligned} \tag{E.6}$$

By our definition of the drift function, for all $(\beta_0, \tau_0^2, \gamma_0^2, \sigma_0^2) \in C_d$ the following relation holds:

$$\begin{aligned} d \geq (y - X\beta_0)^T(y - X\beta_0) &+ \sum_{k=1}^{K} \frac{\beta_{0,G_k}^T \beta_{0,G_k}}{\tau_{0,k}^2} + \sum_{k=1}^{K}\sum_{j=1}^{m_k} \frac{\beta_{0,k,j}^2}{\gamma_{0,k,j}^2} \\ &+ \frac{\lambda_1^2}{4}\sum_{k=1}^{K}\tau_{0,k}^2 + \frac{\lambda_2^2}{4}\sum_{k=1}^{K}\sum_{j=1}^{m_k}\gamma_{0,k,j}^2\,. \end{aligned}$$

Using the above and following on the lines of (D.5) we get for all $k = 1, \ldots, K$ and $j = 1, \ldots, m_k$

$$\beta_{0,G_k}^T \beta_{0,G_k} \leq \frac{4d^2}{\lambda_1^2} := d_1^2 \quad \text{and} \quad \beta_{0,k,j}^2 \leq \frac{4d^2}{\lambda_2^2} := d_2^2\,. \tag{E.7}$$

Using Lemma A.4 and (E.7) and following steps as before (C.9),

$$\begin{aligned} f(\tau^2, \gamma^2 \mid \beta_0, \sigma^2, y) \geq \exp\left\{-1 - \frac{p^2\lambda_2^2 d_2^2}{2\sigma^2} - \frac{K^2\lambda_1^2 d_1^2}{2\sigma^2}\right\} \\ \times \prod_{k=1}^{K}\left[q_k(\tau_k^2 \mid \sigma^2) \prod_{j=1}^{m_k} q_{k,j}(\gamma_{k,j}^2 \mid \sigma^2) \right], \end{aligned} \tag{E.8}$$

where $q_k(\tau_k^2 \mid \sigma^2)$ and $q_{k,j}(\gamma_{k,j}^2 \mid \sigma^2)$ are the densities of the reciprocal of an Inverse-Gaussian distribution with parameters $\sqrt{\lambda_1^2\sigma^2/d_1^2}$ and λ_1^2, and $\sqrt{\lambda_2^2\sigma^2/d_2^2}$ and λ_2^2, respectively. Since each $\tau_{0,k}^2 \leq 4d/\lambda_1^2$ and each $\gamma_{0,k,j}^2 \leq 4d/\lambda_2^2$, so

$$\left(\frac{1}{\tau_{0,k}^2} + \frac{1}{\gamma_{0,k,j}^2}\right)^{-1} \leq \left(\frac{\lambda_1^2}{4d} + \frac{\lambda_2^2}{4d}\right)^{-1} := d_3\,.$$

By Lemma A.5

$$(y - X\beta_0)^T(y - X\beta_0) + \beta_0^T V_{\tau_0,\gamma_0}^{-1}\beta_0 \geq y^T y - y^T X\left(X^T X + \frac{1}{d_3}I_p\right)^{-1} X^T y\,. \tag{E.9}$$

Using (E.9) and following steps as before (C.11)

$$\begin{aligned}&\exp\left\{-1 - \frac{p^2\lambda_2^2 d_2^2}{2\sigma^2} - \frac{K^2\lambda_1^2 d_1^2}{2\sigma^2}\right\} f(\sigma^2 \mid \beta_0, \tau_0^2, \gamma_0^2, y)\\ &= e^{-1}\left(\frac{y^T y - y^T X(X^T X + d_3^{-1}I_p)^{-1}X^T y + 2\xi}{d + 2\xi + p^2\lambda_2^2 d_2^2 + K^2\lambda_1^2 d_1^2}\right)^{\frac{n+p}{2}+\alpha} q(\sigma^2)\,,\end{aligned} \tag{E.10}$$

where $q(\sigma^2)$ is the density of the Inverse-Gamma distribution with parameters, $(n+p)/2 + \alpha$ and $d + 2\xi + p^2\lambda_2^2 d_2^2 + K^2\lambda_1^2 d_1^2$. Using (E.8) and (E.10) in (E.6),

$$\begin{aligned}k_{BSGL}(\beta, \tau^2, \gamma^2, \sigma^2 \mid \beta_0, \tau_0^2, \gamma_0^2, \sigma_0^2) \geq\ & \epsilon\, f(\beta \mid \tau^2, \gamma^2, \sigma^2, y)\, q(\sigma^2)\\ & \times \prod_{k=1}^{K}\left[q_k(\tau_k^2 \mid \sigma^2)\prod_{j=1}^{m_k} q_{k,j}(\gamma_{k,j}^2 \mid \sigma^2)\right],\end{aligned}$$

where

$$\epsilon = e^{-1}\left(\frac{y^T y - y^T X(X^T X + d_3^{-1}I_p)^{-1}X^T y + 2\xi}{d + 2\xi + p^2\lambda_2^2 d_2^2 + K^2\lambda_1^2 d_1^2}\right)^{\frac{n+p}{2}+\alpha}. \tag{E.11}$$

E.3. Starting values

To minimize V_{BSGL},

$$\begin{aligned}\frac{\partial V_{BSGL}}{\partial \tau_{0,k}^2} = 0 \Rightarrow = -\frac{\beta_{0,G_k}^T\beta_{0,G_k}}{\tau_{0,k}^4} + \frac{\lambda_1^2}{4} = 0 \Rightarrow \tau_{0,k}^2 = \sqrt{\frac{4\beta_{0,G_k}^T\beta_{0,G_k}}{\lambda_1^2}}\\ \frac{\partial V_{BSGL}}{\partial \gamma_{0,k,j}^2} = 0 \Rightarrow = -\frac{\beta_{0,k,j}^2}{\gamma_{0,k,j}^4} + \frac{\lambda_2^2}{4} = 0 \Rightarrow \gamma_{0,k,j}^2 = \sqrt{\frac{4\beta_{0,k,j}^2}{\lambda_2^2}}\,.\end{aligned}$$

For the starting value for β,

$$\begin{aligned}\beta_0 = \arg\min_{\beta\in\mathbb{R}^p}\Bigg\{ & (y - X\beta)^T(y - X\beta) + \sum_{k=1}^{K}\frac{\lambda_1\beta_{G_k}^T\beta_{G_k}}{2\sqrt{\beta_{G_k}^T\beta_{G_k}}} + \sum_{k=1}^{K}\sum_{j=1}^{m_k}\frac{\lambda_2\beta_{k,j}^2}{2\sqrt{\beta_{k,j}^2}}\\ & + \sum_{k=1}^{K}\frac{\lambda_1^2}{4}\sqrt{\frac{4\beta_{G_k}^T\beta_{G_k}}{\lambda_1^2}} + \frac{\lambda_2^2}{4}\sum_{k=1}^{K}\sum_{j=1}^{m_k}\sqrt{\frac{4\beta_{k,j}^2}{\lambda_2^2}}\Bigg\}\\ = \arg\min_{\beta\in\mathbb{R}^p}\ & (y - X\beta)^T(y - X\beta) + \lambda_1\sum_{k=1}^{K}\sqrt{\beta_{G_k}^T\beta_{G_k}} + \lambda_2\sum_{k=1}^{K}\sum_{j=1}^{m_k}|\beta_{k,j}|\,,\end{aligned}$$

which corresponds to the sparse group lasso solutions. Thus a reasonable starting value for is β_0 being the sparse group lasso estimate, $\tau^2_{0,k} = 2\sqrt{\beta^T_{0,G_k}\beta_{0,G_k}}/\lambda_1$ and $\gamma^2_{0,k} = 2|\beta_{0,k,j}|/\lambda_2$.

Acknowledgements

The author is grateful to Galin Jones and Sakshi Arya for helpful conversations and suggestions. The author was supported by the Alumni Fellowship, School of Statistics, University of Minnesota and the Louise T. Dosdall Fellowship, University of Minnesota.

References

[1] Andelić, M. and Da Fonseca, C. (2011). Sufficient conditions for positive definiteness of tridiagonal matrices revisited. *Positivity* **15** 155–159. MR2782752

[2] Bhattacharya, A., Chakraborty, A. and Mallick, B. K. (2016). Fast sampling with Gaussian scale mixture priors in high-dimensional regression. *Biometrika* **103** 985–991. MR3620452

[3] Doss, H. and Hobert, J. P. (2010). Estimation of Bayes factors in a class of hierarchical random effects models using a geometrically ergodic MCMC algorithm. *Journal of Computational and Graphical Statistics* **19** 295–312. MR2675092

[4] Doss, C. R., Flegal, J. M., Jones, G. L. and Neath, R. C. (2014). Markov Chain Monte Carlo estimation of quantiles. *Electronic Journal of Statistics* **8** 2448–2478. MR3285872

[5] Fan, Y., Wang, X. and Peng, Q. (2017). Inference of gene regulatory networks using Bayesian nonparametric regression and topology information. *Computational and Mathematical Methods in Medicine* **2017**. MR3596739

[6] Flegal, J. M. and Gong, L. (2015). Relative fixed-width stopping rules for Markov chain Monte Carlo simulations. *Statistica Sinica* **25** 655–676. MR3379093

[7] Flegal, J. M. and Jones, G. L. (2010). Batch means and spectral variance estimators in Markov Chain Monte Carlo. *The Annals of Statistics* **38** 1034–1070. MR2604704

[8] Gong, L. and Flegal, J. M. (2016). A practical sequential stopping rule for high-dimensional Markov chain Monte Carlo. *Journal of Computational and Graphical Statistics* 684–700. MR3533633

[9] Griffin, J. E. and Brown, P. J. (2010). Inference with normal-gamma prior distributions in regression problems. *Bayesian Analysis* **5** 171–188. MR2596440

[10] Gu, X., Yin, G. and Lee, J. J. (2013). Bayesian two-step lasso strategy for biomarker selection in personalized medicine development for time-to-event endpoints. *Contemporary clinical trials* **36** 642–650.

[11] Guan, Y. and Stephens, M. (2011). Bayesian variable selection regression for genome-wide association studies and other large-scale problems. *The Annals of Applied Statistics* 1780–1815. MR2884922

[12] Hobert, J. P., Jones, G. L., Presnell, B. and Rosenthal, J. S. (2002). On the applicability of regenerative simulation in Markov chain Monte Carlo. *Biometrika* **89** 731–743. MR1946508

[13] Ishwaran, H. and Rao, J. S. (2005). Spike and slab variable selection: frequentist and Bayesian strategies. *Annals of Statistics* 730–773. MR2163158

[14] Johnson, A. A. and Jones, G. L. (2015). Geometric ergodicity of random scan Gibbs samplers for hierarchical one-way random effects models. *Journal of Multivariate Analysis* **140** 325–342. MR3372572

[15] Jones, G. L. and Hobert, J. P. (2001). Honest exploration of intractable probability distributions via Markov chain Monte Carlo. *Statistical Science* **16** 312–334. MR1888447

[16] 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** 784–817. MR2060178

[17] Jones, G. L., Haran, M., Caffo, B. S. and Neath, R. (2006). Fixed-width output analysis for Markov chain Monte Carlo. *Journal of the American Statistical Association* **101** 1537–1547. MR2279478

[18] Khare, K. and Hobert, J. P. (2012). Geometric ergodicity of the Gibbs sampler for Bayesian quantile regression. *Journal of Multivariate Analysis* **112** 108–116. MR2957289

[19] Khare, K. and Hobert, J. P. (2013). Geometric ergodicity of the Bayesian lasso. *Electronic Journal of Statistics* **7** 2150–2163. MR3104915

[20] Kyung, M., Gill, J., Ghosh, M. and Casella, G. (2010). Penalized regression, standard errors, and Bayesian lassos. *Bayesian Analysis* **5** 369–411. MR2719657

[21] Liu, J. S., Wong, W. H. and Kong, A. (1994). Covariance structure of the Gibbs sampler with applications to the comparisons of estimators and augmentation schemes. *Biometrika* **81** 27–40. MR1279653

[22] Meyn, S. P. and Tweedie, R. L. (2009). *Markov Chains and Stochastic Stability*. Cambridge University Press. MR2509253

[23] Nathoo, F. S., Greenlaw, K. and Lesperance, M. (2016). Regularization parameter selection for a Bayesian multi-level group lasso regression model with application to imaging genomics. *arXiv preprint arXiv:1603.08163*.

[24] Pal, S. and Khare, K. (2014). Geometric ergodicity for Bayesian shrinkage models. *Electronic Journal of Statistics* **8** 604–645. MR3211026

[25] Park, T. and Casella, G. (2008). The Bayesian lasso. *Journal of the American Statistical Association* **103** 681–686. MR2524001

[26] Rajaratnam, B. and Sparks, D. (2015). MCMC-based inference in the era of big data: A fundamental analysis of the convergence complexity of high-dimensional chains. *arXiv preprint arXiv:1508.00947*.

[27] Raman, S., Fuchs, T. J., Wild, P. J., Dahl, E., Buhmann, J. M. and Roth, V. (2010). Infinite mixture-of-experts model for sparse sur-

vival regression with application to breast cancer. *BMC bioinformatics* **11** 1.

[28] Roberts, G. O. and Rosenthal, J. S. (1997). Geometric ergodicity and hybrid Markov chains. *Electronic Communications in Probability* **2** 13–25. MR1448322

[29] Roberts, G. O. and Rosenthal, J. S. (2004). General state space Markov chains and MCMC algorithms. *Probability Surveys* **1** 20–71. MR2095565

[30] Rosenthal, J. S. (1995). Minorization conditions and convergence rates for Markov chain Monte Carlo. *Journal of the American Statistical Association* **90** 558–566. MR1340509

[31] Roy, V. and Chakraborty, S. (2017). Selection of tuning parameters, solution paths and standard errors for Bayesian lassos. *Bayesian Analysis* **12** 753–778. MR3655875

[32] Simon, N., Friedman, J., Hastie, T. and Tibshirani, R. (2013). A sparse-group lasso. *Journal of Computational and Graphical Statistics* **22** 231–245. MR3173712

[33] Tibshirani, R., Saunders, M., Rosset, S., Zhu, J. and Knight, K. (2005). Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **67** 91–108. MR2136641

[34] Vats, D., Flegal, J. M. and Jones, G. L. (2015a). Multivariate output analysis for Markov chain Monte Carlo. *arXiv preprint arXiv:1512.07713*.

[35] Vats, D., Flegal, J. M. and Jones, G. L. (2015b). Strong consistency of multivariate spectral variance estimators in Markov chain Monte Carlo. *Bernoulli* (to appear).

[36] Xu, X. and Ghosh, M. (2015). Bayesian variable selection and estimation for group lasso. *Bayesian Analysis* **10** 909–936. MR3432244

[37] Yang, Y., Wainwright, M. J. and Jordan, M. I. (2016). On the computational complexity of high-dimensional Bayesian variable selection. *The Annals of Statistics* **44** 2497–2532. MR3576552

[38] Yuan, M. and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **68** 49–67. MR2212574

[39] Zhang, L., Baladandayuthapani, V., Mallick, B. K., Manyam, G. C., Thompson, P. A., Bondy, M. L. and Do, K.-A. (2014). Bayesian hierarchical structured variable selection methods with application to molecular inversion probe studies in breast cancer. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* **63** 595–620. MR3258055