

Fast Bayesian Lasso for High-Dimensional Regression

Bala Rajaratnam and Doug Sparks
Stanford University

October 12, 2018

Abstract

The lasso (Tibshirani, 1996) is an essential tool in modern high-dimensional regression and variable selection. The Bayesian lasso of Park and Casella (2008) interprets the lasso objective function as a posterior under a Laplace prior and proposes a three-step Gibbs sampler to sample from this posterior. The Bayesian lasso yields a natural approach to quantifying the uncertainty of lasso estimates. Furthermore, the Gibbs sampler for the Bayesian lasso has been shown to be geometrically ergodic (Khare and Hobert, 2013). The geometric rate constant of this Markov chain, however, tends to 1 if the number of regression coefficients grows faster than the sample size (Rajaratnam and Sparks, 2015). Thus, convergence of the Bayesian lasso Gibbs sampler can still be quite slow in modern high-dimensional settings despite the apparent theoretical safeguard of geometric ergodicity. In order to address this challenge, we propose a new method to draw from the same posterior via a tractable two-step blocked Gibbs sampler, which we call the *fast Bayesian lasso*. We provide a theoretical underpinning to the new method by proving rigorously that the fast Bayesian lasso is geometrically ergodic. We then demonstrate numerically that this blocked sampler exhibits vastly superior convergence behavior in high-dimensional regimes.

1 Introduction

The lasso provides a method for simultaneously inducing shrinkage and sparsity in the estimation of regression coefficients (Tibshirani, 1996). These goals are especially desirable when the number of coefficients to be estimated is greater than the sample size, and thus the lasso has become a widely used tool in such high-dimensional settings. One drawback of the lasso method is that it is not immediately obvious how to provide meaningful uncertainty quantification for the coefficient estimates. An alternative solution is to interpret the lasso objective function (or some monotone transformation thereof) as the posterior under a certain Bayesian model with a particular prior, as was noted immediately by Tibshirani. Then the lasso estimate corresponds to the posterior mode, and the uncertainty of this estimate can be quantified in a natural way through the usual Bayesian framework (e.g., credible intervals).

However, the Bayesian lasso posterior is not adequately tractable to permit the closed-form evaluation of integrals. To address this problem, [Park and Casella \(2008\)](#) proposed a Gibbs sampler for the Bayesian lasso, which is based on a hierarchical formulation of the prior structure. This structure, which is essentially a type of data augmentation, leads to a tractable three-step Gibbs sampler that can be used to draw (approximately) from the desired posterior. These posterior samples can then be used to construct credible intervals or other quantities of interest.

[Khare and Hobert \(2013\)](#) proved that the Bayesian lasso Gibbs sampler is geometrically ergodic for arbitrary values of the sample size n and the number of coefficients p , and they provide a quantitative upper bound for the geometric rate constant. However, this upper bound tends to 1 in the asymptotic limit as either $p \rightarrow \infty$ or $n \rightarrow \infty$ ([Rajaratnam and Sparks, 2015](#)). Still, since the bound is one-sided, it does not necessarily prove that convergence is slow in regimes of large n or p . However, it has been demonstrated empirically that the geometric rate constant does indeed tend to 1 if $p/n \rightarrow \infty$ ([Rajaratnam and Sparks, 2015](#)). Thus, despite the apparent theoretical safeguard of geometric ergodicity, the Bayesian lasso Gibbs sampler can take arbitrarily long to converge (to within a given total variation distance of the true posterior) if p/n is large enough. This fact is problematic since the so-called “small n , large p ” setting is precisely where the use of the lasso and other regularized regression methods is most beneficial and hence most commonly espoused.

Since the convergence properties of the original three-step Bayesian lasso Gibbs sampler deteriorate in high-dimensional regimes, it may be asked whether there exist alternative schemes for sampling from the same posterior that maintain a reasonably fast (i.e., small) geometric convergence rate when p is large compared to n . Two commonly employed approaches to constructing such alternative MCMC schemes within the Gibbs sampling context are known as *collapsing* and *blocking*. In a collapsed Gibbs sampler, one or more parameters are integrated out of the joint posterior, and a Gibbs sampler is constructed on the posterior of the remaining parameters. Although a collapsed Gibbs sampler converges at least as fast as its uncollapsed counterpart ([Liu et al., 1994](#)), the resulting distributions may not be adequately tractable to permit the implementation of such a scheme in practice. In a blocked Gibbs sampler (also called a grouped Gibbs sampler), multiple parameters are combined and sampled simultaneously in a single step of the cycle that forms each iteration. While no universal comparison between blocked and unblocked Gibbs samplers is possible, it can nevertheless be said that blocking usually does improve the convergence rate with a careful choice of which parameters to group into the same step ([Liu et al., 1994](#)).

In this paper, we propose a two-step blocked Gibbs sampler for the Bayesian lasso in which the regression coefficients β and the residual variance σ^2 are drawn in the same step of the Gibbs sampling cycle. This method, which we call the *fast Bayesian lasso*, turns out to be just as tractable as the original sampler of [Park and Casella \(2008\)](#). Indeed, the distributional forms of our proposed sampler coincide with those of [Park and Casella](#), differing only in the shape and scale of the inverse gamma distribution from which σ^2 is drawn. We demonstrate empirically that in regimes where p is much larger than n , the convergence rate of the fast Bayesian lasso is vastly superior to that of the original scheme of [Park and Casella](#).

The remainder of the paper is organized as follows. In [Section 2](#), we revisit the original Bayesian lasso. We propose the fast Bayesian lasso in [Section 3](#) and establish its geometric

ergodicity in Section 4. Section 5 provides a numerical comparison of the original Bayesian lasso and the fast Bayesian lasso. Some concluding remarks are given in Section 6.

2 The original Bayesian lasso

Consider the model

$$\mathbf{Y} \mid \boldsymbol{\beta}, \sigma^2 \sim N_n(\mu \mathbf{1}_n + \mathbf{X}\boldsymbol{\beta}, \sigma^2 \mathbf{I}_n), \quad (1)$$

where $\mathbf{Y} \in \mathbb{R}^n$ is a response vector, \mathbf{X} is a known $n \times p$ design matrix of standardized covariates, $\boldsymbol{\beta} \in \mathbb{R}^p$ is an unknown vector of regression coefficients, $\sigma^2 > 0$ is an unknown residual variance, and $\mu \in \mathbb{R}$ is an unknown intercept. The lasso (Tibshirani, 1996) proposes to estimate $\boldsymbol{\beta}$ in (1) by minimizing $\|\tilde{\mathbf{Y}} - \mathbf{X}\boldsymbol{\beta}\|_2^2$ subject to the constraint that $\|\boldsymbol{\beta}\|_1 \leq t$ for some specified $t > 0$, where $\tilde{\mathbf{Y}} = \mathbf{Y} - n^{-1}\mathbf{1}_n\mathbf{1}_n^T\mathbf{Y}$, and where $\|\mathbf{v}\|_2^2 = \sum_{i=1}^m v_i^2$ and $\|\mathbf{v}\|_1 = \sum_{i=1}^m |v_i|$ for any $\mathbf{v} = (v_1, \dots, v_m) \in \mathbb{R}^m$. An alternative Lagrangian formulation of the same estimator is to define it as

$$\hat{\boldsymbol{\beta}}_L = \arg \min_{\boldsymbol{\beta} \in \mathbb{R}^p} (\|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda \|\boldsymbol{\beta}\|_1) \quad (2)$$

for some specified $\lambda > 0$. It has been noted (Tibshirani, 1996; Park and Casella, 2008) that $\hat{\boldsymbol{\beta}}_L$ as defined in (2) may be viewed as the posterior mean of $\boldsymbol{\beta}$ under the likelihood in (1) with the prior

$$\beta_j \mid \sigma^2 \sim \text{iid Laplace}(\lambda/\sigma), \quad j \in \{1, \dots, p\}, \quad (3)$$

where the $\text{Laplace}(\alpha)$ distribution has density $f(u) = (\alpha/2) \exp(-\alpha|u|)$ with respect to Lebesgue measure and where we write σ as shorthand for the square root of σ^2 .

Park and Casella (2008) observed that the Laplace prior in (3) can be represented as a scale mixture of normal distributions by introducing positive parameters $\boldsymbol{\tau} = (\tau_1, \dots, \tau_p)$ and taking

$$\boldsymbol{\beta} \mid \sigma^2, \boldsymbol{\tau} \sim N_p(\mathbf{0}_p, \sigma^2 \mathbf{D}_{\boldsymbol{\tau}}), \quad \tau_j \sim \text{iid Exp}(\lambda^2/2), \quad (4)$$

where $\mathbf{D}_{\boldsymbol{\tau}} = \text{Diag}(\tau_1, \dots, \tau_p)$. Suppose further that the prior on σ^2 and μ is improper, $\pi(\sigma^2, \mu) = 1/\sigma^2$, and independent of the prior on $\boldsymbol{\tau}$. Then after integrating out μ from the joint posterior, the remaining full conditional distributions are

$$\begin{aligned} (1/\tau_j) \mid \boldsymbol{\beta}, \sigma^2, \mathbf{Y} &\sim \text{ind InverseGaussian}(\lambda\sigma/|\beta_j|, \lambda^2), \\ \sigma^2 \mid \boldsymbol{\beta}, \boldsymbol{\tau}, \mathbf{Y} &\sim \text{InverseGamma}\left[(n+p-1)/2, \|\tilde{\mathbf{Y}} - \mathbf{X}\boldsymbol{\beta}\|_2^2/2 + \boldsymbol{\beta}^T \mathbf{D}_{\boldsymbol{\tau}}^{-1} \boldsymbol{\beta}/2\right], \\ \boldsymbol{\beta} \mid \sigma^2, \boldsymbol{\tau}, \mathbf{Y} &\sim N_p\left(\mathbf{A}_{\boldsymbol{\tau}}^{-1} \mathbf{X}^T \tilde{\mathbf{Y}}, \sigma^2 \mathbf{A}_{\boldsymbol{\tau}}^{-1}\right), \end{aligned} \quad (5)$$

where $\mathbf{A}_{\boldsymbol{\tau}} = \mathbf{X}^T \mathbf{X} + \mathbf{D}_{\boldsymbol{\tau}}^{-1}$ (Park and Casella, 2008). These conditionals can then be used to construct a useful three-step Gibbs sampler to draw from the posterior of $\boldsymbol{\beta}$, σ^2 , and $\boldsymbol{\tau}$. This framework of model, priors, and Gibbs sampler is popularly known as the *Bayesian lasso*.

Importantly, the Bayesian lasso framework allows for the construction of credible intervals for the lasso.

Khare and Hobert (2013) showed that the Bayesian lasso Gibbs sampler is geometrically ergodic for arbitrary n and p , and they established a quantitative upper bound for the geometric rate constant. However, this upper bound tends to 1 exponentially fast as $n \rightarrow \infty$ or $p \rightarrow \infty$ (Rajaratnam and Sparks, 2015). This bound is one-sided, but Rajaratnam and Sparks (2015) also empirically examined the rate constant itself (as opposed to merely a bound for it) and demonstrated that it tends to 1 in the limit as $p/n \rightarrow \infty$. Thus, while the Markov chain is indeed geometrically ergodic for all n and p , it may nevertheless take arbitrarily long to converge if p is sufficiently large relative to n .

3 The fast Bayesian lasso

We now propose a novel way to sample from the Bayesian lasso posterior. Consider a blocked Gibbs sampler for the Bayesian lasso that alternates between drawing $(\boldsymbol{\beta}, \sigma^2) \mid \boldsymbol{\tau}$ and $\boldsymbol{\tau} \mid (\boldsymbol{\beta}, \sigma^2)$. In particular, $(\boldsymbol{\beta}, \sigma^2) \mid \boldsymbol{\tau}$ may be drawn by first drawing $\sigma^2 \mid \boldsymbol{\tau}$ and then drawing $\boldsymbol{\beta} \mid \sigma^2, \boldsymbol{\tau}$. The posterior distribution of $\sigma^2 \mid \boldsymbol{\tau}$ is provided by the following lemma. Its proof may be found in the Supplementary Material.

lem 1. *For the Bayesian lasso, $\sigma^2 \mid \boldsymbol{\tau}, \mathbf{Y}$ has the inverse gamma distribution with shape parameter $(n - 1)/2$ and scale parameter $\tilde{\mathbf{Y}}^T(\mathbf{I}_n - \mathbf{X}\mathbf{A}_{\boldsymbol{\tau}}^{-1}\mathbf{X}^T)\tilde{\mathbf{Y}}/2$.*

Then a blocked Gibbs sampler for the Bayesian lasso may be constructed by replacing the draw of $\sigma^2 \mid \boldsymbol{\beta}, \boldsymbol{\tau}, \mathbf{Y}$ in (5) with a draw of $\sigma^2 \mid \boldsymbol{\tau}$ as given by Lemma 1. The resulting blocked Gibbs sampler, which we call the *fast Bayesian lasso*, cycles through drawing from the distributions

$$\begin{aligned} (1/\tau_j) \mid (\sigma^2, \boldsymbol{\beta}), \mathbf{Y} &\sim \text{ind InverseGaussian}(\lambda\sigma/|\beta_j|, \lambda^2), \\ (\sigma^2, \boldsymbol{\beta}) \mid \boldsymbol{\tau}, \mathbf{Y} &\sim \begin{cases} \sigma^2 \mid \boldsymbol{\tau}, \mathbf{Y} \sim \text{InverseGamma}\left[(n - 1)/2, \tilde{\mathbf{Y}}^T(\mathbf{I}_n - \mathbf{X}\mathbf{A}_{\boldsymbol{\tau}}^{-1}\mathbf{X}^T)\tilde{\mathbf{Y}}/2\right], \\ \boldsymbol{\beta} \mid \sigma^2, \boldsymbol{\tau}, \mathbf{Y} \sim N_p\left(\mathbf{A}_{\boldsymbol{\tau}}^{-1}\mathbf{X}^T\tilde{\mathbf{Y}}, \sigma^2\mathbf{A}_{\boldsymbol{\tau}}^{-1}\right). \end{cases} \end{aligned} \quad (6)$$

Note that the fast Bayesian lasso is just as tractable as the original sampler since the only the parameters of the inverse gamma distribution are modified. In particular, although the fast Bayesian lasso requires inversion of the matrix $\mathbf{A}_{\boldsymbol{\tau}}$ to draw σ^2 , this inversion must be carried out anyway to draw $\boldsymbol{\beta}$, so no real increase in computation is required.

rem. We have chosen to use the term *blocked* rather than *grouped* in the exposition above in order to avoid confusion with an existing method called the *group Bayesian lasso* (Kyung et al., 2010), in which the notion of grouping is entirely unrelated to the concept considered here. Note also that in some sense the original Bayesian lasso could already be considered a blocked Gibbs sampler since the β_j and τ_j are not drawn individually. Thus, our use of the term *blocked* in describing the fast Bayesian lasso should be understood as meaning that the Gibbs sampling cycle is divided into fewer blocks than in the original Gibbs sampler to which it may be compared.

4 Geometric ergodicity of the fast Bayesian lasso

We now proceed to establish geometric ergodicity of the fast Bayesian lasso Gibbs sampler using the method of [Rosenthal \(1995\)](#). This approach provides a quantitative upper bound on the geometric convergence rate. To express this result rigorously, we first define some notation. For every $k \geq 1$, let $F_k(\sigma_0^2, \boldsymbol{\beta}_0)$ denote the distribution of the k th iterate of $(\boldsymbol{\tau}, \sigma^2, \boldsymbol{\beta})$ for the fast Bayesian lasso Gibbs sampler initialized at $\sigma^2 = \sigma_0^2$ and $\boldsymbol{\beta} = \boldsymbol{\beta}_0$. Let F denote the stationary distribution of this Markov chain, i.e., the true joint posterior distribution of $(\boldsymbol{\tau}, \sigma^2, \boldsymbol{\beta})$. Finally, let d_{TV} denote total variation distance. Then we have the following result, which we express using notation similar to that of Theorem 2 of [Jones and Hobert \(2001\)](#).

thm 2. *Let $0 < \gamma < 1$,*

$$b = \frac{(n+3)p}{4} \left(1 + \frac{n+3}{16\gamma} \right), \quad d > \frac{2b}{1-\gamma}.$$

Then for any $0 < r < 1$,

$$d_{\text{TV}}[F_k(\sigma_0^2, \boldsymbol{\beta}_0), F] \leq (1-\varepsilon)^{rk} + \left(\frac{U^r}{\alpha^{1-r}} \right)^k \left(1 + \frac{b}{1-\gamma} + \frac{\lambda \|\boldsymbol{\beta}_0\|_2^2}{\sigma_0^2} \right),$$

for every $k \geq 1$, where

$$\varepsilon = \exp(-p d^{1/2}), \quad U = 1 + 2(\gamma d + b), \quad \alpha = (1+d)/(1+2b+\gamma d).$$

To prove Theorem 2, we first introduce the following lemma. Its proof may be found in the Supplementary Material.

lem 3. *Let $C_\tau = \tilde{\mathbf{Y}}^\text{T}(\mathbf{I}_n - \mathbf{X}\mathbf{A}_\tau^{-1}\mathbf{X}^\text{T})\tilde{\mathbf{Y}}$. Then $\|\mathbf{A}_\tau^{-1}\mathbf{X}^\text{T}\tilde{\mathbf{Y}}\|_2^2/C_\tau \leq \|\boldsymbol{\tau}\|_1/4$.*

With Lemma 3 in place, we can now prove Theorem 2.

of Theorem 2. We appeal to Theorem 2 of [Jones and Hobert \(2001\)](#) (see also Theorem 12 of [Rosenthal, 1995](#)). To apply this result, it is necessary to establish a drift condition and an associated minorization condition. Let $V(\boldsymbol{\tau}, \sigma^2, \boldsymbol{\beta}) = \lambda^2 \boldsymbol{\beta}^\text{T} \boldsymbol{\beta} / \sigma^2$. Let $\sigma_0^2 > 0$ and $\boldsymbol{\beta}_0 \in \mathbb{R}^p$, and let $(\boldsymbol{\tau}_1, \sigma_1^2, \boldsymbol{\beta}_1)$ be the first iteration of the chain.

To establish the drift condition, observe that

$$\begin{aligned} E[V(\boldsymbol{\tau}_1, \sigma_1^2, \boldsymbol{\beta}_1)] &= \lambda^2 E[\sigma_1^{-2} E(\boldsymbol{\beta}_1^\text{T} \boldsymbol{\beta}_1 \mid \boldsymbol{\tau}_1, \sigma_1^2)] \\ &= \lambda^2 E\left[\text{tr}(\mathbf{A}_{\boldsymbol{\tau}_1}^{-1}) + \|\mathbf{A}_{\boldsymbol{\tau}_1}^{-1} \mathbf{X}^\text{T} \tilde{\mathbf{Y}}\|_2^2 E(\sigma_1^{-2} \mid \boldsymbol{\tau}_1)\right] \\ &= \lambda^2 E\left[\text{tr}(\mathbf{A}_{\boldsymbol{\tau}_1}^{-1}) + \|\mathbf{A}_{\boldsymbol{\tau}_1}^{-1} \mathbf{X}^\text{T} \tilde{\mathbf{Y}}\|_2^2 (n-1)/C_{\boldsymbol{\tau}_1}\right] \\ &\leq \lambda^2 E[\text{tr}(\mathbf{D}_{\boldsymbol{\tau}_1}) + \|\boldsymbol{\tau}_1\|_1 (n-1)/4] = E(\|\boldsymbol{\tau}_1\|_1) (n+3)\lambda^2/4, \end{aligned} \quad (7)$$

where the inequality is by Lemma 3 and the fact that $\text{tr}(\mathbf{A}_{\boldsymbol{\tau}_1}^{-1}) = \text{tr}[(\mathbf{X}^\text{T} \mathbf{X} + \mathbf{D}_{\boldsymbol{\tau}_1}^{-1})^{-1}] \leq \text{tr}[(\mathbf{D}_{\boldsymbol{\tau}_1}^{-1})^{-1}] = \text{tr}(\mathbf{D}_{\boldsymbol{\tau}_1})$. Then continuing from (7), we have

$$E[V(\boldsymbol{\tau}_1, \sigma_1^2, \boldsymbol{\beta}_1)] \leq \frac{(n+3)\lambda^2}{4} \sum_{j=1}^p \left(\frac{|\beta_{0,j}|}{\lambda \sigma_0} + \frac{1}{\lambda^2} \right) = \frac{n+3}{4} \sum_{j=1}^p \frac{\lambda |\beta_{0,j}|}{\sigma_0} + \frac{(n+3)p}{4} \quad (8)$$

by the basic properties of the inverse Gaussian distribution. Now note that $u \leq \delta u^2 + (4\delta)^{-1}$ for all $u \geq 0$ and $\delta > 0$. Applying this result to (8) for each j with $u = \lambda|\beta_{0,j}|/\sigma_0$ and $\delta = 4\gamma/(n+3)$ yields

$$E[V(\boldsymbol{\tau}_1, \sigma_1^2, \boldsymbol{\beta}_1)] \leq \gamma \sum_{j=1}^p \frac{\lambda^2 \beta_{0,j}^2}{\sigma_0^2} + \frac{n+3}{4} \left[\frac{(n+3)p}{16\gamma} \right] + \frac{(n+3)p}{4} = \gamma V(\boldsymbol{\tau}_0, \sigma_0^2, \boldsymbol{\beta}_0) + b,$$

establishing the drift condition.

To establish the associated minorization condition, observe that the density of $F_1(\sigma_0^2, \boldsymbol{\beta}_0)$ with respect to Lebesgue measure on $\mathbb{R}_+^p \times \mathbb{R}_+ \times \mathbb{R}^p$ is

$$\begin{aligned} f_1^{(\sigma_0^2, \boldsymbol{\beta}_0)}(\boldsymbol{\tau}, \sigma^2, \boldsymbol{\beta}) &= \frac{\lambda^p}{(2\pi)^{p/2}} \exp \left[\frac{\lambda \|\boldsymbol{\beta}_0\|_1}{\sigma_0} - \frac{\boldsymbol{\beta}_0^\top \mathbf{D}_\tau^{-1} \boldsymbol{\beta}_0}{2\sigma_0^2} - \frac{\lambda^2 \|\boldsymbol{\tau}\|_1}{2} \right] \prod_{j=1}^p \tau_j^{-1/2} \\ &\quad \times \frac{(C_\tau/2)^{(n-1)/2}}{\Gamma(\frac{n-1}{2}) \sigma^{n+1}} \exp \left(-\frac{C_\tau}{2\sigma^2} \right) \frac{\det(\mathbf{A}_\tau)^{1/2}}{(2\pi\sigma^2)^{p/2}} \exp \left[\frac{1}{2\sigma^2} \left\| \mathbf{A}_\tau^{1/2} (\boldsymbol{\beta} - \mathbf{A}_\tau^{-1} \mathbf{X}^\top \tilde{\mathbf{Y}}) \right\|_2^2 \right] \\ &= \frac{(\lambda/2\pi)^p}{2^{(n-1)/2} \Gamma(\frac{n-1}{2})} \frac{\det(\mathbf{A}_\tau \mathbf{D}_\tau^{-1})^{1/2} C_\tau^{(n-1)/2}}{\sigma^{n+p+1}} \exp \left(-\frac{\lambda^2 \|\boldsymbol{\tau}\|_1}{2} \right) \\ &\quad \times \exp \left(-\frac{\|\tilde{\mathbf{Y}} - \mathbf{X}\boldsymbol{\beta}\|_2^2}{2\sigma^2} - \frac{\boldsymbol{\beta}^\top \mathbf{D}_\tau^{-1} \boldsymbol{\beta}}{2\sigma^2} \right) \exp \left(\frac{\lambda \|\boldsymbol{\beta}_0\|_1}{\sigma_0} - \frac{\boldsymbol{\beta}_0^\top \mathbf{D}_\tau^{-1} \boldsymbol{\beta}_0}{2\sigma_0^2} \right). \end{aligned} \quad (9)$$

Now suppose that $V(\boldsymbol{\tau}_0, \sigma_0^2, \boldsymbol{\beta}_0) \leq d$. Then $\lambda^2 \beta_{0,j}^2 / \sigma_0^2 \leq d$ for each j . Let $\boldsymbol{\xi} = d^{1/2} \lambda^{-1} \mathbf{1}_p$, and observe that

$$\exp \left(\frac{\lambda \|\boldsymbol{\beta}_0\|_1}{\sigma_0} - \frac{\boldsymbol{\beta}_0^\top \mathbf{D}_\tau^{-1} \boldsymbol{\beta}_0}{2\sigma_0^2} \right) \geq \exp \left(-\frac{\boldsymbol{\xi}^\top \mathbf{D}_\tau^{-1} \boldsymbol{\xi}}{2} \right) = \varepsilon \exp \left(\lambda \|\boldsymbol{\xi}\|_1 - \frac{\boldsymbol{\xi}^\top \mathbf{D}_\tau^{-1} \boldsymbol{\xi}}{2} \right),$$

noting that $\varepsilon = \exp(-\lambda \|\boldsymbol{\xi}\|_1)$. Combining this inequality with (9) yields

$$\begin{aligned} f_1^{(\sigma_0^2, \boldsymbol{\beta}_0)}(\boldsymbol{\tau}, \sigma^2, \boldsymbol{\beta}) &\geq \frac{(\lambda/2\pi)^p}{2^{(n-1)/2} \Gamma(\frac{n-1}{2})} \frac{\det(\mathbf{A}_\tau \mathbf{D}_\tau^{-1})^{1/2} C_\tau^{(n-1)/2}}{\sigma^{n+p+1}} \exp \left(-\frac{\lambda^2 \|\boldsymbol{\tau}\|_1}{2} \right) \\ &\quad \times \exp \left(-\frac{\|\tilde{\mathbf{Y}} - \mathbf{X}\boldsymbol{\beta}\|_2^2}{2\sigma^2} - \frac{\boldsymbol{\beta}^\top \mathbf{D}_\tau^{-1} \boldsymbol{\beta}}{2\sigma^2} \right) \varepsilon \exp \left(\lambda \|\boldsymbol{\xi}\|_1 - \frac{\boldsymbol{\xi}^\top \mathbf{D}_\tau^{-1} \boldsymbol{\xi}}{2} \right) \\ &= \varepsilon f_1^{(1, \boldsymbol{\xi})}(\boldsymbol{\tau}, \sigma^2, \boldsymbol{\beta}) \end{aligned}$$

for all $\boldsymbol{\tau}$, σ^2 , and $\boldsymbol{\beta}$. Note that $f_1^{(1, \boldsymbol{\xi})}(\boldsymbol{\tau}, \sigma^2, \boldsymbol{\beta})$ is the density of $F_1(1, \boldsymbol{\xi})$ with respect to Lebesgue measure on $\mathbb{R}_+^p \times \mathbb{R}_+ \times \mathbb{R}^p$, while $F_1(1, \boldsymbol{\xi})$ is the distribution of $(\boldsymbol{\tau}_1, \sigma_1^2, \boldsymbol{\beta}_1)$ if the chain is initialized at $\sigma^2 = 1$ and $\boldsymbol{\beta} = \boldsymbol{\xi}$. Thus, the minorization condition is established. \square

Note that for Theorem 2 to be useful, the bound must actually decrease with k . Thus, it is necessary to choose r small enough that $U^r / \alpha^{1-r} < 1$. Then for small enough r , the

bound is dominated by the term $(1 - \varepsilon)^{rk}$, which is approximately $(1 - r\varepsilon)^k$ for small r and ε . Now observe that

$$d > \frac{2b}{1 - \gamma} > 2b = \frac{(n + 3)p}{2} \left(1 + \frac{n + 3}{16\gamma} \right) > \frac{n^2 p}{32},$$

It follows that

$$1 - r\varepsilon = 1 - r \exp(-p d^{1/2}) > 1 - r \exp(-n p^{3/2}/32^{1/2}),$$

which tends to 1 exponentially fast as n or p tends to infinity. Thus, although Theorem 2 establishes that the fast Bayesian lasso Gibbs sampler is geometrically ergodic and provides a bound for the rate constant, it is not clear how sharp it is. Hence the bound may not be particularly informative in high-dimensional contexts. It is therefore essential to ascertain the actual behavior of the convergence rate of the fast Bayesian lasso in high-dimensional regimes, especially in comparison to the original Bayesian lasso. The numerical results of the following section provide valuable insights in this regard.

5 Numerical Comparison

The upper bound in Theorem 2 above for the fast Bayesian lasso and the upper bound of Khare and Hobert (2013) for the original Bayesian lasso are both derived using the method originally proposed by Rosenthal (1995). As discussed by Rajaratnam and Sparks (2015), such bounds typically are not sharp in high-dimensional settings. Numerical investigations can therefore provide a more practically relevant comparison of the convergence rates of these two Gibbs samplers. As a proxy for the actual rate of convergence of the chain, we consider the autocorrelation in the marginal σ_k^2 chain. Note that this autocorrelation is exactly equal to the geometric convergence rate in the simplified case of standard Bayesian regression (Rajaratnam and Sparks, 2015) and is a lower bound for the true geometric convergence rate in general (Liu et al., 1994). Figure 1 plots the autocorrelation for the original and fast Bayesian lasso Gibbs samplers versus p for various values of n . (Similar plots under varying sparsity and multicollinearity may be found in the Supplementary Material.) The left side of Figure 2 plots the same autocorrelation versus $\log(p/n)$ for a wider variety of values of n and p . It is apparent that this autocorrelation for the fast Bayesian lasso is bounded away from 1 for all n and p . The center and right side of Figure 2 show dimensional autocorrelation function (DAF) surface plots (see Rajaratnam and Sparks, 2015) for the original (left) and fast (right) Bayesian lasso Gibbs samplers. (See the Supplementary Material for details of the generation of the various numerical quantities that were used in the execution of these chains.) It is clear that the autocorrelation tends to 1 as $p/n \rightarrow \infty$ for the original Bayesian lasso but remains bounded away from 1 for the fast Bayesian lasso.

6 Discussion

The Gibbs samplers associated with the standard Bayesian lasso and the new fast Bayesian lasso are geometrically ergodic. This property serves as a basic safeguard for the rapidity

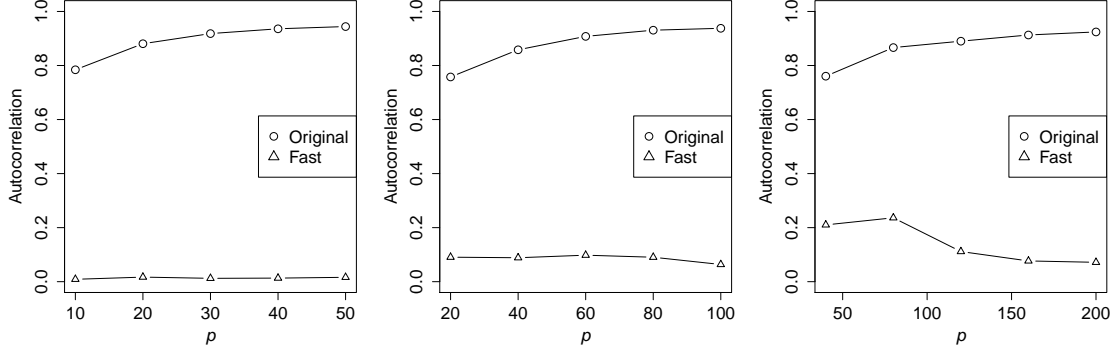


Figure 1: Autocorrelation of σ_k^2 versus p for the original and fast Bayesian lasso Gibbs samplers with $n = 5$ (top left), $n = 10$ (top right), and $n = 20$ (bottom left). See the Supplementary Material for details of the generation of the numerical quantities used in the execution of these chains.

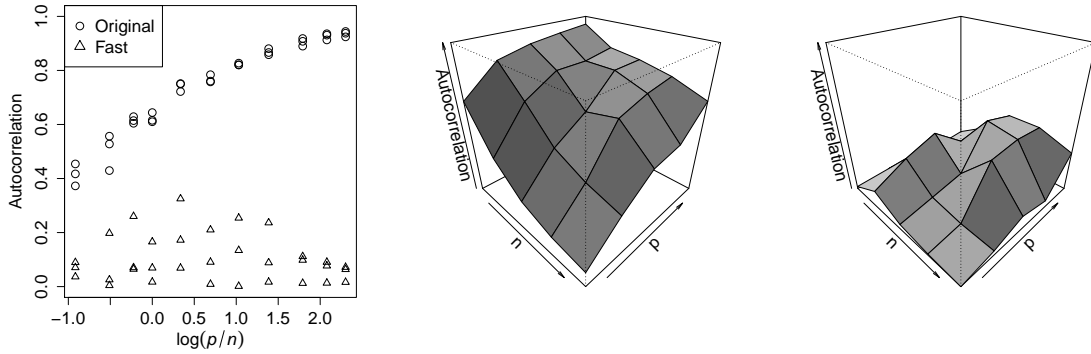


Figure 2: Autocorrelation of σ_k^2 versus $\log(p/n)$ for the original and fast Bayesian lasso Gibbs samplers with various values of n and p (left). Dimensional autocorrelation function (DACF) surface plots for the σ_k^2 chain relative to n and p for the original (center) and fast (right) Bayesian lasso Gibbs samplers. See the Supplementary Material for details of the generation of the numerical quantities used in the execution of these chains.

of MCMC convergence. Since the associated upper bounds are not sharp, it is important to understand the precise nature of the convergence via numerical simulations, especially in various n and p regimes. Our work above demonstrates that the geometric rate constant of the fast Bayesian lasso improves as the dimension increases and is thus ideally suited to ultra-high-dimensional applications. On the other hand, the standard Bayesian lasso encounters convergence problems, especially in the very setting for which the lasso is designed. Furthermore, it should also be noted that the ability to combine the draws of β and σ^2 into a single block can be extended beyond the Bayesian lasso posterior. The same blocking process may be applied to the Gibbs sampler for any Bayesian regression model that can be expressed in the form of (1) and (4), i.e., any Bayesian regression model in which the prior is a scale mixture of normal distributions. The blocked versions of any such Gibbs samplers are likely to exhibit the same improved convergence behavior in high dimensions as is seen in the Bayesian lasso (see Kyung et al., 2010, for a variety of such lasso-type Bayesian regression models and associated Gibbs samplers).

Acknowledgment

The authors were supported in part by the US National Science Foundation under grants DMS-CMG-1025465, AGS-1003823, DMS-1106642, and DMS-CAREER-1352656, and by the US Air Force Office of Scientific Research grant award FA9550-13-1-0043.

Supplementary material

Supplementary material comprises proofs of Lemmas 1 and 3, additional numerical results, and details of the execution of the numerical results in Section 5.

References

- JONES, G. L. and HOBERT, J. P. (2001). Honest exploration of intractable probability distributions via Markov chain Monte Carlo. *Statistical Science*, **16** 312–334.
- KHARE, K. and HOBERT, J. P. (2013). Geometric ergodicity of the Bayesian lasso. *Electronic Journal of Statistics*, **7** 2150–2163.
- KYUNG, M., GILL, J., GHOSH, M. and CASELLA, G. (2010). Penalized regression, standard errors, and Bayesian lassos. *Bayesian Analysis*, **5** 369–412.
- 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.
- PARK, T. and CASELLA, G. (2008). The Bayesian lasso. *Journal of the American Statistical Association*, **103** 681–686.

- 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. <http://arxiv.org/abs/1508.00947>. Tech. rep., Stanford University.
- ROSENTHAL, J. S. (1995). Minorization conditions and convergence rates for Markov chain Monte Carlo. *Journal of the American Statistical Association*, **90** 558–566.
- TIBSHIRANI, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B*, **58** 267–288.

Supplementary Material

Proofs

of Lemma 1. Integrating out β from the joint posterior $\pi(\beta, \sigma^2, \tau \mid \mathbf{Y})$ yields

$$\begin{aligned}
 \pi(\sigma^2, \tau \mid \mathbf{Y}) &= \int \pi(\beta, \sigma^2, \tau \mid \mathbf{Y}) d\beta \\
 &= \frac{(\lambda^2/2)^p}{(2\pi\sigma^2)^{(n+p+1)/2}} \exp\left(-\frac{\lambda^2}{2} \sum_{j=1}^p \tau_j\right) \\
 &\quad \times \int \exp\left[-\frac{1}{2\sigma^2} \left(\|\tilde{\mathbf{Y}} - \mathbf{X}\beta\|_2^2 + \|\mathbf{D}_\tau^{-1/2}\beta\|_2^2\right)\right] d\beta \\
 &= \frac{(\lambda^2/2)^p}{(2\pi\sigma^2)^{(n+p+1)/2}} \exp\left[-\frac{\lambda^2\|\tau\|_1}{2} - \frac{1}{2\sigma^2} \tilde{\mathbf{Y}}^\top (\mathbf{I}_n - \mathbf{X}\mathbf{A}_\tau^{-1}\mathbf{X}^\top) \tilde{\mathbf{Y}}\right] \\
 &\quad \times \int \exp\left[-\frac{1}{2\sigma^2} \left\|\mathbf{A}_\tau^{1/2}(\beta - \mathbf{A}_\tau^{-1}\mathbf{X}^\top \tilde{\mathbf{Y}})\right\|_2^2\right] d\beta \\
 &= \frac{(\lambda^2/2)^p}{(2\pi\sigma^2)^{(n+p+1)/2} \det(\mathbf{A}_\tau)^{1/2}} \exp\left[-\frac{\lambda^2\|\tau\|_1}{2} - \frac{\tilde{\mathbf{Y}}^\top (\mathbf{I}_n - \mathbf{X}\mathbf{A}_\tau^{-1}\mathbf{X}^\top) \tilde{\mathbf{Y}}}{2\sigma^2}\right].
 \end{aligned}$$

Then it is clear that

$$\pi(\sigma^2 \mid \tau, \mathbf{Y}) \propto \frac{1}{(\sigma^2)^{(n+p+1)/2}} \exp\left[-\frac{\tilde{\mathbf{Y}}^\top (\mathbf{I}_n - \mathbf{X}\mathbf{A}_\tau^{-1}\mathbf{X}^\top) \tilde{\mathbf{Y}}}{2\sigma^2}\right],$$

and the result follows immediately. \square

of Lemma 3. Let $\mathbf{X} = \mathbf{U}\mathbf{\Omega}\mathbf{V}^\top$ be a singular value decomposition of \mathbf{X} , where $\omega_1, \dots, \omega_{\min\{n,p\}}$ are the the singular values, and let $\tau_{\max} = \max_{1 \leq j \leq p} \tau_j$. Then

$$\begin{aligned}
 \frac{\|\mathbf{A}_\tau^{-1}\mathbf{X}^\top \tilde{\mathbf{Y}}\|_2^2}{C_\tau} &= \frac{\tilde{\mathbf{Y}}^\top \mathbf{U}\mathbf{\Omega}\mathbf{V}^\top (\mathbf{V}\mathbf{\Omega}^\top \mathbf{\Omega}\mathbf{V}^\top + \mathbf{D}_\tau^{-1})^{-2} \mathbf{V}\mathbf{\Omega}^\top \mathbf{U}^\top \tilde{\mathbf{Y}}}{\tilde{\mathbf{Y}}^\top \tilde{\mathbf{Y}} - \tilde{\mathbf{Y}}^\top \mathbf{U}\mathbf{\Omega}\mathbf{V}^\top (\mathbf{V}\mathbf{\Omega}^\top \mathbf{\Omega}\mathbf{V}^\top + \mathbf{D}_\tau^{-1})^{-1} \mathbf{V}\mathbf{\Omega}^\top \mathbf{U}^\top \tilde{\mathbf{Y}}} \\
 &\leq \frac{\tilde{\mathbf{Y}}^\top \mathbf{U}\mathbf{\Omega}\mathbf{V}^\top (\mathbf{V}\mathbf{\Omega}^\top \mathbf{\Omega}\mathbf{V}^\top + \tau_{\max}^{-1} \mathbf{I}_p)^{-2} \mathbf{V}\mathbf{\Omega}^\top \mathbf{U}^\top \tilde{\mathbf{Y}}}{\tilde{\mathbf{Y}}^\top \tilde{\mathbf{Y}} - \tilde{\mathbf{Y}}^\top \mathbf{U}\mathbf{\Omega}\mathbf{V}^\top (\mathbf{V}\mathbf{\Omega}^\top \mathbf{\Omega}\mathbf{V}^\top + \tau_{\max}^{-1} \mathbf{I}_p)^{-1} \mathbf{V}\mathbf{\Omega}^\top \mathbf{U}^\top \tilde{\mathbf{Y}}} \\
 &= \frac{\tilde{\mathbf{Y}}^\top \mathbf{U}\mathbf{\Omega}(\mathbf{\Omega}^\top \mathbf{\Omega} + \tau_{\max}^{-1} \mathbf{I}_p)^{-2} \mathbf{\Omega}^\top \mathbf{U}^\top \tilde{\mathbf{Y}}}{\tilde{\mathbf{Y}}^\top \mathbf{U}\mathbf{U}^\top \tilde{\mathbf{Y}} - \tilde{\mathbf{Y}}^\top \mathbf{U}\mathbf{\Omega}(\mathbf{\Omega}^\top \mathbf{\Omega} + \tau_{\max}^{-1} \mathbf{I}_p)^{-1} \mathbf{\Omega}^\top \mathbf{U}^\top \tilde{\mathbf{Y}}} = \frac{\tilde{\mathbf{Y}}_\star^\top \mathbf{G} \tilde{\mathbf{Y}}_\star}{\tilde{\mathbf{Y}}_\star^\top \mathbf{H} \tilde{\mathbf{Y}}_\star},
 \end{aligned}$$

where $\tilde{\mathbf{Y}}_\star = \mathbf{U}^\top \tilde{\mathbf{Y}}$ and

$$\mathbf{G} = \text{Diag}\left[\frac{\omega_1^2}{(\omega_1^2 + \tau_{\max}^{-1})^2}, \dots, \frac{\omega_n^2}{(\omega_n^2 + \tau_{\max}^{-1})^2}\right], \quad \mathbf{H} = \text{Diag}\left(\frac{\tau_{\max}^{-1}}{\omega_1^2 + \tau_{\max}^{-1}}, \dots, \frac{\tau_{\max}^{-1}}{\omega_n^2 + \tau_{\max}^{-1}}\right),$$

taking $\omega_i = 0$ for all $i > p$ if $n > p$. Then it is clear that

$$\begin{aligned} \frac{\left\| \mathbf{A}_\tau^{-1} \mathbf{X}^\top \tilde{\mathbf{Y}} \right\|_2^2}{C_\tau} &\leq \max_{1 \leq i \leq n} \left[\frac{\omega_i^2}{(\omega_i^2 + \tau_{\max}^{-1})^2} \left(\frac{\tau_{\max}^{-1}}{\omega_i^2 + \tau_{\max}^{-1}} \right)^{-1} \right] \\ &\leq \max_{1 \leq i \leq n} \left[\frac{1}{4\tau_{\max}^{-1}} \left(\frac{\tau_{\max}^{-1}}{\omega_i^2 + \tau_{\max}^{-1}} \right)^{-1} \right] \leq \frac{\tau_{\max}}{4} \leq \frac{\|\boldsymbol{\tau}\|_1}{4}, \end{aligned}$$

noting for the second inequality that $a/(a+b)^2 \leq 1/(4b)$ for all $a \geq 0$ and $b > 0$. \square

Additional numerical results

Figure 3 is similar to the left-hand side of Figure 2 under settings of high multicollinearity (left), low sparsity (center), and both high multicollinearity and low sparsity (right). (See the following section for details.) It is clear that the behavior seen in Figure 1 is also seen in other settings of multicollinearity and sparsity.

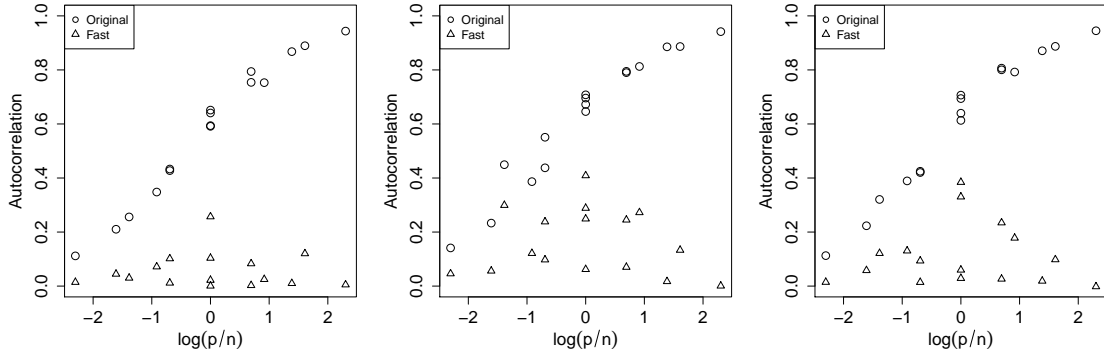


Figure 3: Autocorrelation of the σ_k^2 chain versus p/n for various values of n for the original and fast Bayesian lasso Gibbs samplers in settings of high multicollinearity (left), low sparsity (center), and both high multicollinearity and low sparsity (right). See the following section for details of the generation of the numerical quantities used in the execution of these chains.

Details of numerical results

Each plotted point in Figures 1 and 2 represents the average lag-one autocorrelation over 10 Gibbs sampling runs of 10,000 iterations each. For each of the 10 runs at each n and p setting, the np elements of the $n \times p$ covariate matrix \mathbf{X} were drawn as $N(0,1)$ random variables with all pairwise correlations equal to $1/5$. Also, for each run, the $n \times 1$ response vector \mathbf{Y} was generated based on the sparsity s as $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta}_* + \boldsymbol{\varepsilon}$, where $\boldsymbol{\beta}_*$ is a $p \times 1$ vector with its first $\lceil p/5 \rceil$ elements drawn as independent t_2 random variables and its remaining $p - \lceil p/5 \rceil$ elements set to zero, and where $\boldsymbol{\varepsilon}$ is an $n \times 1$ vector of independent t_4 random variables. The initial values were set as $\boldsymbol{\beta}_0 = \mathbf{1}_p$ and $\sigma_0^2 = 1$. The regularization parameter λ was set to $\lambda = 1$. In Figure 1, for each n , the values of p were $2n$, $4n$, $6n$, $8n$, and $10n$. The left side of Figure 2 includes the same results as in Figure 1 but also includes p equal

to $2n/5$, $3n/5$, $4n/5$, $7n/5$, and $14n/5$. The DACF surface plots in the center and right side of Figure 2 used each combination of $n, p \in \{5, 15, 25, 35, 45\}$.

The plots in Figure 3 were constructed similarly to the left-hand side of Figure 2, but with the values of n and p set to each combination of $n, p \in \{5, 10, 20, 50\}$, and with the following additional modifications. For the left-hand side and the right-hand side (but not the center), the pairwise correlation between each element of the design matrix was $4/5$ (rather than $1/5$). For the center and right-hand side (but not the left-hand side), the number of nonzero coefficients for the generation of the response vector was taken to be $\lceil 4p/5 \rceil$ (rather than $\lceil p/5 \rceil$).