STATS 620 Project

A Simulation Study on High Dimensional Shrinkage Feature Selection Using MCMC Methods

Benjamin Osafo Agyare[1]

[1]Department of Statistics, University of Michigan, Ann Arbor.

Abstract

Linear regression sometimes present us with the problem of high dimensionality –especially when the covariates, p is far greater than the sample size, n. A frequentist approach to tackling such problems include adopting shrinkage methods via penalization. However, penalization methods present yet another challenge of quantifying parameter uncertainties. Bayesian approach gives us the power to quantity parameters via estimating the posterior distribution for such parameters using Markov Chain Monte Carlo (MCMC) techniques. With such high dimensional shrinkage methods, we need super fast MCMC algorithms that are efficient and computationally relative inexpensive. In this simulation study, we compare and show that the two-Block Gibbs samplers (2BG) is a more efficient state of the art MCMC algorithm relative to the three-Block Gibbs samplers (3BG) method in estimating the posterior distributions of two commonly used Bayesian shrinkage models, viz: the Bayesian Lasso (BL) and the Spike-and-Slab shrinkage priors. Our criteria for evaluation include the one-lag autocorrelation and the average effective sample size per second, N_{eff}/T. Consequently, we apply these methods on the protein expression genetics data from the National cancer Institute.

Keywords: *Gibbs sampler; Bayesian Lasso; Spike-and-Slab; Geometric ergodicity, parallel computing*

1. Introduction

In linear regression, we model a response variable, y using potential predictors, $x_1, x_2, x_3, \ldots, x_p$ via the regression coefficients, $\boldsymbol{\beta} = (\beta_1, \beta_2, \beta_3, ..., \beta_p)$. Penalization is a technique used to handle situations where $p >> n$, which, depending on the choice of the choice of the penalty function, some β_j's would be shrunk to zero. Sparse estimates of β_j's which typically yield stable predictions for such situations present us with yet the problem of parameter uncertainty quantification. A solve to this challenge is to model such problems using Bayesian techniques.

Particularly, recent works in Bayesian shrinkage modeling use shrinkage priors to address this using two popular methods viz: the *spike-and-slab* priors such as Laplace-Zero mixtures (Johnstone and Silverman, 2004) which assign mixtures of the degenerate distribution at zero and the *purely continuous* prior distributions such as the Horseshoe distribution (Carvalho *et al.*, 2010) as alternatives to quantify parameter uncertainties. Moreover, Park and Casella (2008) developed a Bayesian lasso approach to the frequentist lasso objective following Tibshirani (1996)'s work which identified that the lasso objective could be interpreted as the posterior under a certain Bayesian model with an independent Laplace prior on the coefficients.

However, as with most Bayesian problems, there are no closed form expression for posterior quantification of the Bayesian Shrinkage models. An approach to approximating the posterior

distribution is via Variational Bayes algorithms such as mean field variational Bayes (MFVB) methodology. While this is a fast algorithm, Neville *et al.* (2014) points that the MFVB algorithm can perform quite poorly due to posterior dependence among auxiliary variables. Markov chain Monte Carlo (MCMC), an indispensable tool in contemporary Bayesian inference provides us with alternative methods that can be used to approximate the posterior distribution and hence make inferences on the parameters since the shrinkage priors of these models capitalize on hierarchical representations. Utilizing the power of MCMC, Park and Casella (2008) provided the three-step Gibbs sampler, otherwise known as the three-Block Gibbs sampler (3BG) algorithm to explore the posterior distribution for arbitrary n and p. While Khare and Hobert (2013) proved that the 3BG is geometrically ergodic for arbitrary values of n and p, it tends to suffer from slow convergence especially if p/n is large enough due to high correlation between components of the different blocks, especially that between the regression coefficients in one block and the variance parameters in another(Rajaratnam and Sparks, 2015). Owing to the deterioration in the convergence properties of the 3BG in high-dimesional settings, Rajaratnam *et al.* (2019) proposed an efficient version known as the two-Block Gibbs sampler (2BG) algorithm to overcome the sluggish convergence issues of the former. Theoretical convergence properties can be found in the aforementioned paper if it interests the reader.

In this project, we perform a simulation study comparing the computational efficiencies between the 3BG and 2BG using the *spike-and-slab* priors and Bayesian lasso model following Rajaratnam *et al.* (2019). Computing efficiency in this case is measured as the effective sample size per second (N_{eff}/T). We also assess the mixing rates in terms of lag-one autocorrelations ρ_1, hence having ρ_1 closer to 0 implies better mixing rate (Rajaratnam and Sparks, 2015). We consequently apply this method on the well-known NCI-60 cancer cell panel from the National cancer Institute.

The rest of the article is organized as follows: Section 2 provides an overview of the Bayesian skrinkage frameowork for regression as well as a review of the *spike-and-slab* priors and the Bayesian lasso and the 2BG algorithms that explore the posterior distributions. In Sections 3 and 4, we perform simulation studies and real data analysis to empirically compare the two algorithms under study. We close the project with concluding remarks in Section 5. The codes for this project are found at https://github.com/bosafoagyare/2-BGS.

2. Methods

2.1. Bayesian Shrinkage Models

Let $Y \in \mathbb{R}^n$ be the response vector, X be the $n \times p$ design matrix of standardized covariates, $\boldsymbol{\beta} \in \mathbb{R}^p$ be the vector of regression coefficients, $\sigma^2 > 0$ be the residual variance and $\mu \in \mathbb{R}$ be an unknown intercept in a regression problem. This model is represented as:

$$\boldsymbol{Y} \mid \boldsymbol{\beta}, \sigma^2 \sim \mathrm{N}_n \left(\mu \mathbf{1}_n + \boldsymbol{X}\boldsymbol{\beta}, \sigma^2 \boldsymbol{I}_n\right) \tag{1}$$

We more especially focus on situations where the number of covariates p is much larger than the sample size n. In the Bayesian framework where sparsity is desired, the *spike-and-slab* priors which mixes a normal density with a spike at zero and another normal density which is flat near zero as well as the alternative *purely continuous* skrinkage priors have been explored in shrinking the regression coefficients toward zero. Customarily, many Bayesian methods in high-dimesional regression follow a set-up where the prior density is specified as:

$$\boldsymbol{\beta} \mid \sigma^2, \boldsymbol{\tau} \sim \mathrm{N}_p \left(\mathbf{0}_p, \sigma^2 \boldsymbol{D}_{\boldsymbol{\tau}}\right), \quad \boldsymbol{\tau} \sim \pi(\boldsymbol{\tau}), \tag{2}$$

where $\pi(\boldsymbol{\tau})$ is a prior on $\tau = (\tau_1, \ldots, \tau_p)$. Now, further assume that the prior on σ^2 and μ is once again the improper prior $\pi\left(\sigma^2, \mu\right) = 1/\sigma^2$ and that this prior is independent of the prior on τ. It

follows that integrating out μ after combining 1 and 2 yields the following conditional distributions:

$$\begin{aligned}
&\boldsymbol{\tau} \mid \boldsymbol{\beta}, \sigma^2, \boldsymbol{Y} \sim \pi\left(\boldsymbol{\tau} \mid \boldsymbol{\beta}, \sigma^2, \boldsymbol{Y}\right), \\
&\sigma^2 \mid \boldsymbol{\beta}, \boldsymbol{\tau}, \boldsymbol{Y} \sim \text{ Inverse-Gamma} \\
&\left[(n+p-1)/2, \|\tilde{\boldsymbol{Y}} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2/2 + \boldsymbol{\beta}^{\mathrm{T}} \boldsymbol{D}_{\boldsymbol{\tau}}^{-1} \boldsymbol{\beta}/2\right], \\
&\boldsymbol{\beta} \mid \sigma^2, \boldsymbol{\tau}, \boldsymbol{Y} \sim \mathrm{N}_p\left(\boldsymbol{A}_{\boldsymbol{\tau}}^{-1} \boldsymbol{X}^{\mathrm{T}} \tilde{\boldsymbol{Y}}, \sigma^2 \boldsymbol{A}_{\boldsymbol{\tau}}^{-1}\right),
\end{aligned} \tag{3}$$

where $\boldsymbol{A}_{\boldsymbol{\tau}} = \boldsymbol{X}^{\mathrm{T}}\boldsymbol{X} + \boldsymbol{D}_{\tau}^{-1}$ and $D_{\boldsymbol{\tau}} = \mathrm{Diag}\left(\tau_1, \tau_2, \ldots, \tau_p\right)$. So long as it is possible to draw from $\pi\left(\boldsymbol{\tau} \mid \boldsymbol{\beta}, \sigma^2, \boldsymbol{Y}\right)$, one can use the three conditionals above to construct a useful 3BG to draw from the joint posterior $\pi\left(\boldsymbol{\beta}, \sigma^2 \mid \boldsymbol{Y}\right)$. The one-step transition density $\hat{k}$ with respect to Lebesgue measure on $\mathbb{R}^p \times \mathbb{R}_+$is given by

$$\begin{aligned}
\hat{k}\left[\left(\boldsymbol{\beta}_0, \sigma_0^2\right), \left(\boldsymbol{\beta}_1, \sigma_1^2\right)\right] = \int_{\mathbb{R}_+^p} &\pi\left(\sigma_1^2 \mid \boldsymbol{\beta}_1, \boldsymbol{\tau}, \boldsymbol{Y}\right) \pi\left(\boldsymbol{\beta}_1 \mid \boldsymbol{\tau}, \sigma_0^2, \boldsymbol{Y}\right) \\
&\times \pi\left(\boldsymbol{\tau} \mid \boldsymbol{\beta}_0, \sigma_0^2, \boldsymbol{Y}\right) d\boldsymbol{\tau}.
\end{aligned} \tag{4}$$

2.1.1. *The Spike-and-Slab* Prior

Under the *spike-and-slab* prior framework, we assign independent discrete priors to τ_j such that each assign probability w_j to the point $\kappa_j \zeta_j$ and probability $1 - w_j$ to the point ζ_j, where $\zeta_j > 0$ is small, $\kappa_j > 0$ is large, and $w_j \in (0, 1)$. This yields the conditional posterior distribution of $\boldsymbol{\tau} \mid (\boldsymbol{\beta}, \sigma^2, \boldsymbol{Y})$ which is a product of independent discrete distributions that each assign probability $\tilde{w}_j$ to the point $\kappa_j \zeta_j$ and probability $1 - \tilde{w}_j$ to the point ζ_j (Rajaratnam *et al.*, 2019), where

$$\tilde{w}_j = \left\{1 + \frac{(1 - w_j)\sqrt{\kappa_j}}{w_j} \exp\left[-\frac{\beta_j^2}{2\sigma^2}\left(\frac{\kappa_j - 1}{\kappa_j \zeta_j}\right)\right]\right\}^{-1} \tag{5}$$

In our study, we treat $\tilde{w}_j, \kappa_j, \zeta_j$ as constants whose values are chosen to meet the constraints in the analyses that follow.

2.1.2. The Bayesian Lasso

Here, we assign independent Exponential $\left(\lambda^2/2\right)$ priors to τ_j where $\left(\lambda^2/2\right)$ is the rate parameter of the exponential distribution. It follows that the marginal prior of $\boldsymbol{\beta}$ (given σ^2) assigns independent Laplace densities to each component. Hence, the conditional posterior distribution of $\boldsymbol{\tau} \mid (\boldsymbol{\beta}, \sigma^2, \boldsymbol{Y})$ assigns independent inverse Gaussian distributions to each $1/\tau_j$, which makes it straightforward to sample from. The full framework is given by Park and Casella (2008) under the 3BG algorithm.

2.2. Fast MCMC for Bayesian Skrinkage Models

Rajaratnam *et al.* (2019) provide and prove the following lemma which provides a 2BG alternative algorithm to overcome the sluggish convergence properties of the 3BG algorithm.

Lemma 1. For the Bayesian model in 2, $\sigma^2 \mid \boldsymbol{\tau}, \boldsymbol{Y}$ has the inverse gamma distribution with shape parameter $(n-1)/2$ and scale parameter $\tilde{\boldsymbol{Y}}^{\mathrm{T}}\left(\boldsymbol{I}_n - \boldsymbol{X}\boldsymbol{A}_{\boldsymbol{\tau}}^{-1}\boldsymbol{X}^{\mathrm{T}}\right)\tilde{\boldsymbol{Y}}/2$.

The lemma above leads to the construction of a novel 2BG sampler for generating samples from the joint posterior density of $\left(\boldsymbol{\beta}, \sigma^2\right)$, and which is equally tractable as the original 3BG sampler. This algorithm alternates between drawing $\left(\boldsymbol{\beta}, \sigma^2\right) \mid \boldsymbol{\tau}$ and $\boldsymbol{\tau} \mid \left(\boldsymbol{\beta}, \sigma^2\right)$ where, $\left(\boldsymbol{\beta}, \sigma^2\right) \mid \boldsymbol{\tau}$ may be drawn by first drawing $\sigma^2 \mid \boldsymbol{\tau}$ and then drawing $\boldsymbol{\beta} \mid \sigma^2, \boldsymbol{\tau}$. In other words, the 2BG sampler may be constructed by replacing the draw of $\sigma^2 \mid \boldsymbol{\beta}, \boldsymbol{\tau}, \boldsymbol{Y}$ in 3 with a draw of $\sigma^2 \mid \boldsymbol{\tau}$ as given by lemma

1. In cyclical fashion, the algorithm is summarized as follows:

$$\begin{aligned}
&\boldsymbol{\tau} \mid \boldsymbol{\beta}, \sigma^2, \boldsymbol{Y} \sim \pi\left(\boldsymbol{\tau} \mid \boldsymbol{\beta}, \sigma^2, \boldsymbol{Y}\right) \\
&\left(\boldsymbol{\beta}, \sigma^2\right) \mid \boldsymbol{\tau}, \boldsymbol{Y} \sim \\
&\times \begin{cases} \sigma^2 \mid \boldsymbol{\tau}, \boldsymbol{Y} \sim \text{ Inverse-Gamma } \left[(n-1)/2, \tilde{\boldsymbol{Y}}^{\mathrm{T}}\left(\boldsymbol{I}_n - \boldsymbol{X}\boldsymbol{A}_{\boldsymbol{\tau}}^{-1}\boldsymbol{X}^{\mathrm{T}}\right)\tilde{\boldsymbol{Y}}/2\right] \\ \qquad\qquad\qquad\qquad \boldsymbol{\beta} \mid \sigma^2, \boldsymbol{\tau}, \boldsymbol{Y} \sim N_p\left(\boldsymbol{A}_{\boldsymbol{\tau}}^{-1}\boldsymbol{X}^{\mathrm{T}}\tilde{\boldsymbol{Y}}, \sigma^2\boldsymbol{A}_{\boldsymbol{\tau}}^{-1}\right), \end{cases}
\end{aligned}$$

3. Simulation studies

We study the computational efficiency between the 2 and 3BG samplers using simulation studies. All studies were performed on a Windows 11 Pro PC with 16.0 GB RAM, 8 cores and Intel (R) Core(TM) i7-8650U @ 1.90GHz processor.

The data is simulated using the following model:

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta}_* + \boldsymbol{\epsilon}, \tag{6}$$

where $\boldsymbol{\epsilon} \in \mathbb{R}^{n\times 1}$ is a vector of standard normal variables and $\boldsymbol{\beta}_* \in \mathbb{R}^{p\times 1}$ is a vector of the true regression coefficients. For the *spike-and-slab* model, we perform two sets of simulations at $n = \{50, 100\}$ while we do $n = 75$ for a single set of simulation for the Bayesian lasso model. We set p such that $\frac{p}{n} = \{0.5, 0.6, 0.7, 0.8, 0.9, 1, 2, 3, 4, 5\}$. Hence we use 2-block simulations with each (n, p) combination giving 10 datasets each for the *spike-and-slab*, and 1-10 block simulation with each (n, p) combination giving 10 datasets for the Bayesian models respectively. For each dataset, the rows of the $n \times p$ design matrix, $\boldsymbol{X}$ are independently drawn, each from the $p - dim$ standard multivariate normal distribution after which the columns are standardized to have mean zero and squared Euclidean norm n. For each p we experiment with, $\boldsymbol{\beta}_*$ is such that its first $p/5$ elements are nonzero, and are drawn independently from the t_2 distribution.

For all studies, we run 6 Markov chains under a *parallel computing* architecture using the *foreach* package in the R programming language. Each chain is allowed to run at 15,000 iterations after which the first 10% are chopped-off as burn-in. For all chains in all simulation, we set $\boldsymbol{\beta} = \mathbf{1}_p$ and $\sigma_0^2 = 0$ as initial values. The regularization parameter λ was set to $\lambda = 1$. For the *spike-and-slab* model, we consistently set the hyperparameters of the priors $\tilde{w}_j = 1/2, \kappa = 100$ and $\zeta = 1/100$. To assess computational efficiency, we estimate the average lag-*1* autocorrelation, ρ_1 of the σ^2-marginal of the chain under stationarity (post burn-in). The reason for using this metric as well as the σ^2 parameter rather than the $\boldsymbol{\beta}$ is discussed by Rajaratnam and Sparks (2015). Also, we estimate the average effective sample size per second, N_{eff}/T, where N_{eff}, the effective sample size (computed from the R package *coda* Plummer *et al.* (2006))is defined by:

$$N_{\text{eff}} = \frac{N}{1 + 2\sum_{k=1}^{\infty}\rho_k}. \tag{7}$$

3.1. Results for the *Spike-and-slab* model

The left panel of Figure 1 shows the empirical average lag-one autocorrelation of the σ^2 component of the six MCMC chains for the *spike-and-slab* model at $n = 50$. It is indeed clear that the new 2BG exhibits smaller average autocorrelations across all (n, p) combinations than the 3BG counterpart. This implies that the 2BG has better mixing rate and hence is more efficient computationally than the 3BG algorithm. We can see that the gap widens even more as the dimension of the covariates, p gets bigger and bigger relative to the sample size, n. The right panel shows the average effective sample size per second, N_{eff}/T, in the base-10 log scale of each sampler. It can be observed that the 2BG produces many effective sample than the 3BG sampler and this becomes even more as p grows relative to n, hence we pay smaller computational cost per chain for using the 2BG sampler which makes it an ideal algorithm for high-dimesional problems. Similar story can be told when we scale up the sample size $n = 100$ as depicted in 2.

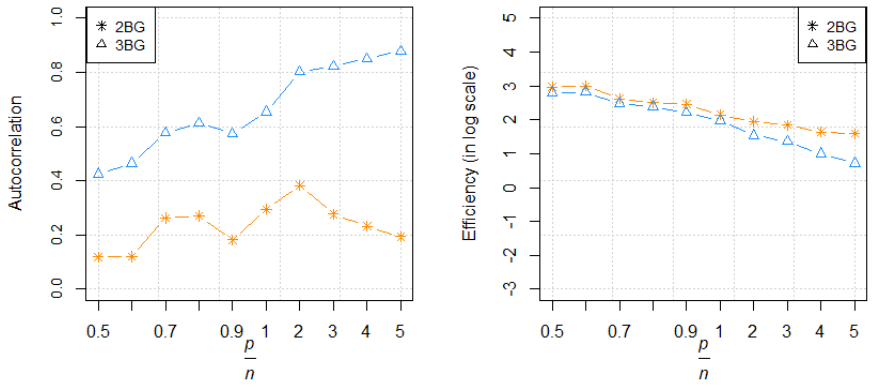

Figure 1. Empirical lag-one autocorrelation at $n = 50$ (left) and the average effective sample size per second in base-10 log (right) of the σ^2 component of the four MCMC chains for the *spike-and-slab* model.

3.2. Results for the Bayesian lasso model

We make similar analysis as Section 3.1 under the Bayesian lasso model. The left panel of Figure 3 shows the empirical average lag-one autocorrelation of the σ^2 component of the six MCMC chains at $n = 75$. Similarly, we observe that the new 2BG exhibits smaller average autocorrelations across all (n, p) combinations than the 3BG counterpart under both "n-small, p-large" and "n-large, p-small" regimes. The closest average lag-one autocorrelation of the two algorithms happens when $n = p = 100$. In terms of the average effective sample size per second, N_{eff}/T shown on the right panel, it is no surprise that the 2BG produces much more effective samples than the 3BG counterpart, demonstrating its computational dominance over the latter.

4. Application to Real Data

We now apply both the Bayesian lasso and te *spike-and-slab* models to a real dataset. We use the well-known NCI-60 cancer cell panel from the National cancer Institute. This dataset comprises protein expressions for a specific protein selected as the response variable, and the gene expressions of the 100 genes that have the highest (robustly estimated) correlations with the response variable which are screened as candidate predictors. This pre-processing is done using the R *robustHD* package (Alfons, 2021). Hence, we have $n = 59$ samples and $p = 100$ covariates. Each covariate was further standardized to have mean zero and squared Euclidean norm n. We set the hyperparameters of the priors $\tilde{w}_j = 1/2, \kappa = 100$ and $\zeta = 1/200$ with $\lambda = 0.5$. For both shrinkage models, we run 18,000 chains and set the first 10% aside as burn-in.

As we can see from Figures 4 and 5, the chains for both the *spike-and-slab* and Bayesian lasso models have attained stationarity and hence sufficient for estimating the posterior distribution. Further, from Table 1, we can see that the lag-one autocorrelation for the 2BG model for both the *spike-and-slab* and Bayesian lasso models (0.387 and 0.161 respectively) are much smaller than the 3BG counterparts and hence mixes better. Moreover, we can see that the 2BG sampler produces about twice as many effective samples than the 3BG sampler for the *spike-and-slab* model, whereas it produces about five-times as many for the Bayesian lasso which is highly efficient and computationally much cheaper.

5. Discussion

In this study, we have used both simulation study and real data analysis to demonstrate the computational prowess of the two-block Gibbs sampler algorithm, specifically using the *spike-and-slab* and Bayesian lasso models. The 2BG has proven to be a faster and more efficient algorithm

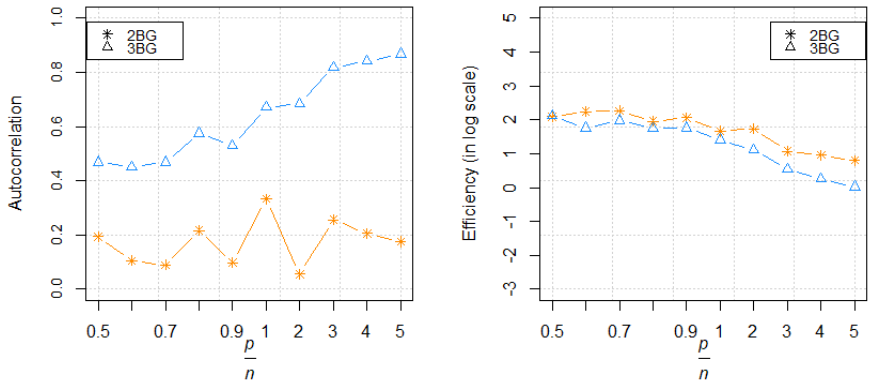

Figure 2. Empirical lag-one autocorrelation at $n = 100$ (left) and the average effective sample size per second in base-10 log (right) of the σ^2 component of the four MCMC chains for the *spike-and-slab* model

			Autocorrelation		N_{eff}	
Dataset	n	p	2GB	3GB	2BG	3BG
Gene (*spike-and-slab*)	59	100	0.387	0.774	3,263	1,639
Gene (*Bayesian lasso*)	59	100	0.161	0.700	10,921	2,856

Table 1. Lag-one autocorrelation and effective sample size per second for the σ^2-component of 2BG and 3BG of the *spike-and-slab* and Bayesian Lasso models as applied to the protein gene dataset

relative to the 3BG counterpart. This presents researchers and analysts edge to leverage the Bayesian framework in high-dimensional data modeling, especially as it presents the possibility of quantifying parameter uncertainties for inference via the Bayesian credible intervals while leaving little to worry about computational costs and complexities.

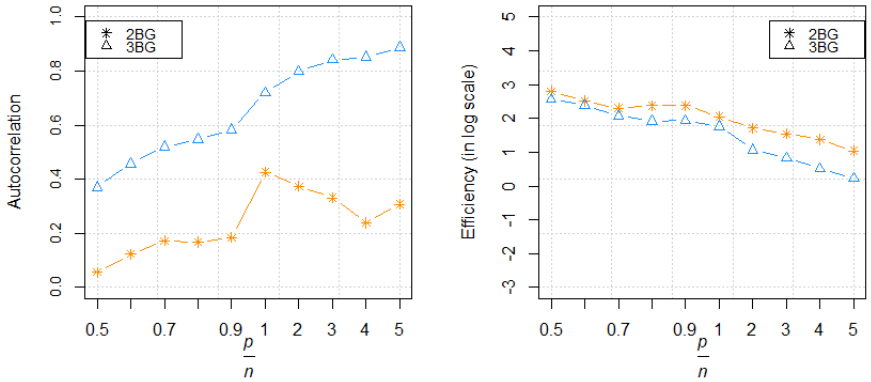

Figure 3. Empirical lag-one autocorrelation at $n = 75$ (left) and the average effective sample size per second in base-10 log (right) of the σ^2 component of the four MCMC chains for the Bayesian lasso model

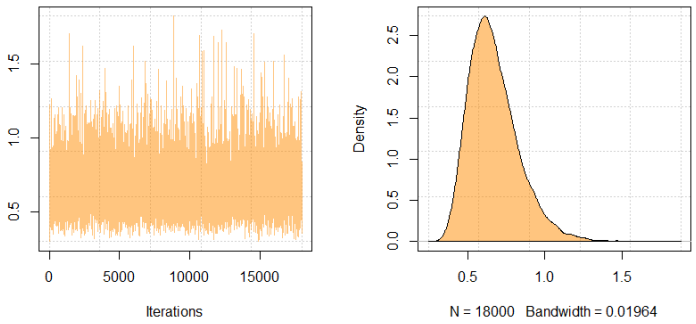

Figure 4. Trace-plot and density plot of the 2BG *spike-and-slab* model for the proteins data

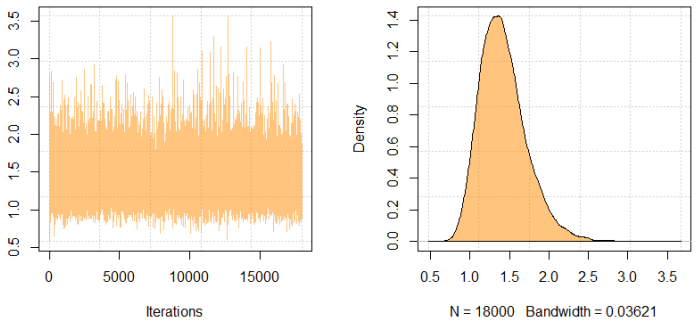

Figure 5. Trace-plot and density plot of the 2BG Bayesian lasso model for the proteins data

References

Alfons, A. (2021). Robusthd: an r package for robust regression with high-dimensional data. *Journal of Open Source Software*, **6**(67), 3786.

Carvalho, C. M. *et al.* (2010). The horseshoe estimator for sparse signals. *Biometrika*, **97**(2), 465–480.

Johnstone, I. M. and Silverman, B. W. (2004). Needles and straw in haystacks: Empirical bayes estimates of possibly sparse sequences.

Khare, K. and Hobert, J. P. (2013). Geometric ergodicity of the bayesian lasso.

Neville, S. E. *et al.* (2014). Mean field variational bayes for continuous sparse signal shrinkage: pitfalls and remedies.

Park, T. and Casella, G. (2008). The bayesian lasso. *Journal of the American Statistical Association*, **103**(482), 681–686.

Plummer, M. *et al.* (2006). Coda: convergence diagnosis and output analysis for mcmc. *R news*, **6**(1), 7–11.

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*.

Rajaratnam, B. *et al.* (2019). Uncertainty quantification for modern high-dimensional regression via scalable bayesian methods. *Journal of Computational and Graphical Statistics*, **28**(1), 174–184.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, **58**(1), 267–288.