Supplementary Material: Variational Lasso for Regression Models

Larissa C. Alves, Ronaldo Dias, Helio S. Migon

S-1 Introduction

The main objective of this supplementary material is to present the full Bayesian Lasso via variational inference (VI or VB) for linear regression models. Some extensive simulation studies are proposed. We also present the necessary algebraic developments to facilitate the understanding of the main article. This supplementary material is organized as follows. In Section S-2 the full Bayesian lasso is introduced. The Jeffreys prior for both the penalisation and precision parameters are developed exploring Fisher decomposition. The full conditional posterior distribution, useful for both MCMC and VB, are fully described. All the details are shown in Subsection S-2.2 and S-2.3. In Section S-3 three simulations studies are described. The first one compares VB and MCMC computational time and accuracy and the other two exercises simulate data with high correlation and large sparsity with $p >> n$.

S-2 The full Bayesian Lasso

Following the hierarchical representation for the Laplace distributions, Park and Casella [2008] shows a Bayesian formulation of the Lasso regression model. The hierarchical model is defined as:

$$\begin{aligned}
\mathbf{y}|\mathbf{X},\boldsymbol{\beta},\phi &\sim N[\mathbf{X}\boldsymbol{\beta},\phi^{-1}\mathbf{I_n}] \\
\boldsymbol{\beta}|\phi,\boldsymbol{\tau} &\sim N[\mathbf{0},\phi^{-\mathbf{1}}\mathbf{D}_{\boldsymbol{\tau}}], \quad \boldsymbol{\tau}=(\tau_{\mathbf{1}},\ldots,\tau_{\mathbf{p}}) \\
\tau_j|\lambda &\sim Exp(\lambda) \quad \text{with} \quad j=1,\ldots,p,
\end{aligned}$$

where $\mathbf{D}_{\boldsymbol{\tau}} = diag(\tau_1,\ldots,\tau_p)$ and $\tau_j|\lambda$ are conditionally independent for all j. The model can be completed with the hyperparameters of the priors $\phi \sim Ga(a_0,b_0)$ and $\lambda \sim Ga(g_0,h_0)$. In Subsection S-2.1 we propose an independent Jeffreys prior for ϕ and λ to automate the Lasso, and this implies supposing a_0, b_0, g_0 and h_0 tending to zero.

Let $\boldsymbol{\theta}=(\boldsymbol{\beta},\phi,\boldsymbol{\tau},\lambda)$ be the vector of the parameters and the latent variables of the model. The posterior distribution is obtained as proportional to the model distribution times the prior distribution for the latent component and the parameters:

$$p(\boldsymbol{\theta}|\mathbf{y},\mathbf{X}) \propto p(\mathbf{y}|\mathbf{X},\boldsymbol{\beta},\phi)\ p(\boldsymbol{\beta}|\phi,\boldsymbol{\tau})\ p(\boldsymbol{\tau}|\lambda)\ p(\phi)\ p(\lambda).$$

For instance, the above joint posterior is often intractable. An almost obvious numerical approach, since the breakthrough paper of Gelfand and Smith [1990], is to use stochastic simulation.

S-2.1 Jeffreys prior using Fisher decomposition

In order to develop an automatic Bayesian Lasso procedure it is worth to introduce non informative priors for the hyperparameters involved. Following Fonseca et al. [2019] and exploring the conditional independence involved in the Lasso model, the Fisher information decomposition for Lasso follows as:

$$I_{\mathbf{y}}(\lambda) = I_{\boldsymbol{\tau}}(\lambda) - E_{\mathbf{y}}\left[I_{\boldsymbol{\beta},\boldsymbol{\tau}}(\lambda|\mathbf{y})\right], \tag{1}$$

where $I_{\boldsymbol{\beta},\boldsymbol{\tau}}(\lambda|\mathbf{y})$ is the information obtained from the full conditional distribution $p(\boldsymbol{\beta},\boldsymbol{\tau}|\mathbf{y},\lambda)$. We also are using the conditional independence described by the graph that represents the Bayesian Lasso model.

We will develop, in turn, each of the components in the expression (1). The quantity $I_{\boldsymbol{\tau}}(\lambda)$ is based on the independent marginal distribution of τ_j, leading directly to $I_{\boldsymbol{\tau}}(\lambda) = \frac{p}{\lambda^2}$.

In order to obtain $I_{\boldsymbol{\beta},\boldsymbol{\tau}}(\lambda|\mathbf{y})$, we take advantage of the known full conditional distribution of $(\boldsymbol{\beta},\boldsymbol{\tau}|\lambda,\mathbf{y})$ (see (2)). Since $(\boldsymbol{\beta}|\boldsymbol{\tau},\lambda,\mathbf{y})$ does not depend on λ, then it is easy to obtain $E_{\mathbf{y}}\left[I_{\boldsymbol{\beta},\boldsymbol{\tau}}(\lambda|\mathbf{y})\right] = \frac{p}{\lambda}$.

Then substituting in (1), it follows $I_{\mathbf{y}}(\lambda) = \frac{p}{\lambda^2} + \frac{p}{\lambda^2}$ and so the prior for λ is $p(\lambda) \propto \lambda^{-1}$. This result is similar to the one reported in Fonseca et al. [2019], using the Uniform Gamma mixture. It is well known that the Jeffreys prior of ϕ is proportional of ϕ^{-1}.

S-2.2 The MCMC formulation

Considering the model and the prior distribution already specified, we know that the posterior distribution in this case has an unknown form. Therefore, we can use the MCMC to obtain a sample of the posterior distribution through the full conditional distributions (Gibbs Sampler). Calculations of full conditionals are as follows:

$$\begin{aligned}
(\boldsymbol{\beta}|\mathbf{y},\mathbf{X},\boldsymbol{\theta}_{-\boldsymbol{\beta}}) &\sim N\left((\mathbf{X}^T\mathbf{X}+\mathbf{D}_{\boldsymbol{\tau}}^{-1})^{-1}\mathbf{X}^T\mathbf{y}, \frac{1}{\phi}(\mathbf{X}^T\mathbf{X}+\mathbf{D}_{\boldsymbol{\tau}}^{-1})^{-1}\right)\\
(\tau_j|\mathbf{y},\mathbf{X},\boldsymbol{\theta}_{-\tau_j}) &\sim GIG\left(\frac{1}{2}, 2\lambda, \beta_j^2\phi\right)\\
(\phi|\mathbf{y},\mathbf{X},\boldsymbol{\theta}_{-\phi}) &\sim Ga\left(\frac{n}{2}+\frac{p}{2}+a_0, b_0+\frac{1}{2}[(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^T(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})+\boldsymbol{\beta}^T\mathbf{D}_{\boldsymbol{\tau}}^{-1}\boldsymbol{\beta}]\right)\\
(\lambda|\mathbf{y},\mathbf{X},\boldsymbol{\theta}_{-\lambda}) &\sim Ga\left(g_0+p, h_0+\sum_{j=1}^{p}\tau_j\right)
\end{aligned} \tag{2}$$

where $\boldsymbol{\theta}_{-}$ stands for the entire vector $\boldsymbol{\theta}$ without the parameter followed by symbol "_", and *GIG* denotes the generalized inverse Gaussian distribution.

S-2.3 The variational approximation applied to Lasso

In order to obtain a scalable inference procedure, we introduce an alternative methodology.

The joint distribution of the observations, latent components and parameters can easily be followed from Figure S1 which in turn summarizes the model.

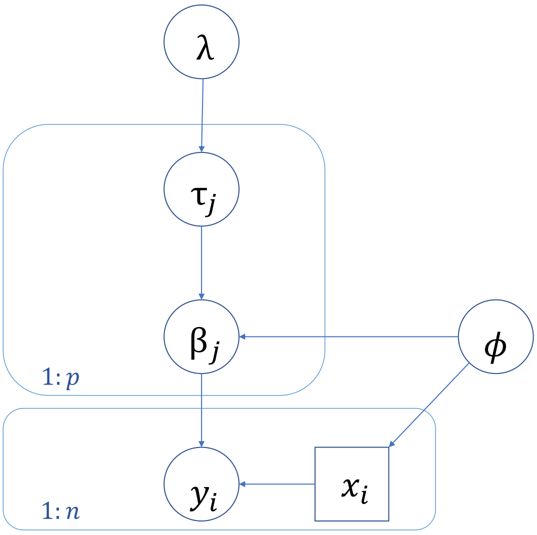

Figure S1: Directed acyclic graph

It is worth remembering the expression of the *mean field* posterior approximation for the latent components and parameters which is given by:

$$\log(q(\boldsymbol{\theta})) = \log(q_1(\boldsymbol{\beta}, \phi)) + \log(q_2(\boldsymbol{\tau})) + \log(q_3(\lambda)).$$

After quoting Blei et al. [2017] the optimal $q_l(\boldsymbol{\theta}_l)$ is proportional to the exponential of the log of the full conditional distribution that is calculated in (2) as the following:

$$q_l^*(\boldsymbol{\theta}_l) \propto \exp\{E_{-l}[\log p(\boldsymbol{\theta}_l | \boldsymbol{\theta}_{-l}, \mathbf{y})]\}, \quad l = 1, 2, 3.$$

In the first step, the variational distribution for $\boldsymbol{\beta}$ and ϕ, that maximizes the variational bound $\mathcal{L}(q)$ while holding $q_2(\boldsymbol{\tau})$ and $q_3(\lambda)$ fixed, is given by

$$\begin{aligned}
\log q_1^*(\boldsymbol{\beta},\phi) &= \log(p(\mathbf{y}|\boldsymbol{\beta},\phi)) + E_{\boldsymbol{\tau}}[\log(p(\boldsymbol{\beta},\phi|\boldsymbol{\tau}))] + const \\
&= \log\left[(2\pi)^{-n/2}|\phi^{-1}\mathbf{I_n}|^{-1/2}\exp\left\{-\frac{1}{2}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^T(\phi\mathbf{I_n})(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})\right\}\right] + \\
&\quad +E_{\boldsymbol{\tau}}\left\{\log\left[(2\pi)^{-p/2}|\phi^{-1}\mathbf{D}_{\boldsymbol{\tau}}|^{-1/2}\exp\left\{-\frac{1}{2}\boldsymbol{\beta}^T\phi\mathbf{D}_{\boldsymbol{\tau}}^{-1}\boldsymbol{\beta}\right\}\right]\right\} + \\
&\quad +E_{\boldsymbol{\tau}}\{\log[\phi^{a_0-1}\exp\{-\phi b_0\}]\} + const \\
&= \left(\frac{n}{2}+\frac{p}{2}+a_0-1\right)\log\phi + \\
&\quad -\frac{\phi}{2}\{\boldsymbol{\beta}^T[E_{\boldsymbol{\tau}}(\mathbf{D}_{\boldsymbol{\tau}}^{-1}) + \mathbf{X}^T\mathbf{X}]\boldsymbol{\beta} + \mathbf{y}^T\mathbf{y} - 2\mathbf{y}^T\mathbf{X}\boldsymbol{\beta} + 2b_0\} + const \\
&= \log N(\boldsymbol{\beta}|\mathbf{m}_{\boldsymbol{\beta}},\phi^{-1}\mathbf{C}_{\boldsymbol{\beta}}) \times Ga(\phi|a_\phi,b_\phi)
\end{aligned}$$

It is easy to see that this is a normal-gamma distribution with parameters:

$$\mathbf{C}_{\boldsymbol{\beta}}^{-1} = E_{\boldsymbol{\tau}}(\mathbf{D}_{\boldsymbol{\tau}}^{-1}) + \mathbf{X}^T\mathbf{X}, \quad \text{and} \quad \mathbf{m}_{\boldsymbol{\beta}} = \mathbf{C}_{\boldsymbol{\beta}}\mathbf{X}^T\mathbf{y},$$
$$a_\phi = a_0 + n/2, \quad \text{and} \quad b_\phi = b_0 + \frac{1}{2}(\mathbf{y}^T\mathbf{y} - \mathbf{m}_{\boldsymbol{\beta}}^T\mathbf{C}_{\boldsymbol{\beta}}^{-1}\mathbf{m}_{\boldsymbol{\beta}}).$$

Next, the variational distribution of $\boldsymbol{\tau}$, that maximizes the variational bound $\mathcal{L}(q)$ while holding $q_3(\lambda)$ fixed is given by

$\log q_2^*(\tau_j) =$

$$\begin{aligned}
&= E_\lambda[\log(p(\tau_j|\lambda))] + E_{\beta,\phi}[\log(p(\beta_j,\phi|\tau_j))] + const \\
&= E_\lambda\{\log[\exp\{-\lambda\tau_j\}]\} + E_{\beta,\phi}\left\{\log\left[(\phi^{-1}\tau_j)^{-1/2}\exp\left\{-\frac{\phi}{2\tau_j}\beta_j^2\right\}\right]\right\} + const \\
&= -\frac{1}{2}\log\tau_j - \frac{1}{2}\left(2E_\lambda[\lambda]\tau_j + \frac{1}{\tau_j}E_{\beta,\phi}[\phi\beta_j^2]\right) + const \\
&= \log GIG(\tau_j|c_\tau,d_\tau,f_{\tau_j})
\end{aligned}$$

with GIG being generalized inverse Gaussian distribution, where

$$c_\tau = \frac{1}{2}\ ;\ d_\tau = 2E_\lambda[\lambda]\ ;\ f_{\tau_j} = E_{\beta,\phi}[\phi\beta_j^2].$$

Therefore,

$$\log q_2^*(\boldsymbol{\tau}) = \log\prod_{j=1}^{p} GIG(\tau_j|c_\tau,d_\tau,f_{\tau_j}).$$

Finally, we will identify the variational distribution of λ:

$$\begin{aligned}\log q_3^*(\lambda) &= \log(p(\lambda)) + E_{\boldsymbol{\tau}}[\log(p(\boldsymbol{\tau}|\lambda))] + const \\ &= \log[\lambda^{g_0-1}\exp\{-h_0\lambda\}] + E_{\boldsymbol{\tau}}\left(\log\left[\prod_{j=1}^{p}\lambda\exp\{-\tau_j\lambda\}\right]\right) + const \\ &= (g_0+p-1)\log\lambda - \lambda[h_0 + \sum_{j=1}^{p}E_\tau(\tau_j)] + const \\ &= \log Ga(\lambda|g_\lambda, h_\lambda)\end{aligned}$$

which is a gamma distribution with parameters

$$g_\lambda = g_0 + p\ ;\ h_\lambda = h_0 + \sum_{j=1}^{p}E_\tau(\tau_j).$$

The expected values involved in the definition of the above variational distributions are computed as follows (see Jørgensen [1982] for details).

It is worth pointing out that if $X \sim GIG(p,a,b)$, then its density is

$$f(x|p,a,b) = \frac{(\frac{a}{b})^{\frac{p}{2}}}{2\kappa_p(\sqrt{ab})}\, x_{p-1}\ \exp\{-(ax+b/x)/2\}, \quad x>0,$$

where $\kappa_p(\cdot)$ is a modified Bessel function of the second kind, with

$$\begin{aligned}E[X] &= \sqrt{\frac{b}{a}}\frac{\kappa_{p+1}(\sqrt{ab})}{\kappa_p(\sqrt{ab})} \quad \text{and} \quad Var[X] = \left(\frac{b}{a}\right)\left[\frac{\kappa_{p+2}(\sqrt{ab})}{\kappa_p(\sqrt{ab})} - \left(\frac{\kappa_{p+1}(\sqrt{ab})}{\kappa_p(\sqrt{ab})}\right)^2\right] \\ E[X^{-1}] &= \sqrt{\frac{a}{b}}\frac{\kappa_{p+1}(\sqrt{ab})}{\kappa_p(\sqrt{ab})} - \frac{2p}{b}\end{aligned}$$

Therefore,

$$E_{\boldsymbol{\tau}}[\mathbf{D}_{\boldsymbol{\tau}}^{-1}] = diag(E_\tau(\tau_1^{-1}),\ldots,E_\tau(\tau_p^{-1})),$$

$$E_\tau(\tau_j^{-1}) = \frac{\sqrt{d_\tau}\kappa_{c_\tau+1}(\sqrt{d_\tau f_{\tau_j}})}{\sqrt{f_{\tau_j}}\kappa_{c_\tau}(\sqrt{d_\tau f_{\tau_j}})} - \frac{2c_\tau}{f_{\tau_j}}, \qquad E_\tau(\tau_j) = \frac{\sqrt{f_{\tau_j}}\kappa_{c_\tau+1}(\sqrt{d_\tau f_{\tau_j}})}{\sqrt{d_\tau}\kappa_{c_\tau}(\sqrt{d_\tau f_{\tau_j}})},$$

$$E_\lambda(\lambda) = \frac{g_\lambda}{h_\lambda}.$$

For the calculus of $E_{\beta,\phi}[\phi\beta_j^2]$, let $x|y \sim N(\mu_x, y^{-1}\sigma_x)$ and $y \sim Ga(a,b)$, so

$$E[X^2Y] = E[E(X^2Y|Y)] = E[Y(E^2(X|Y) + Var(X|Y))] = \mu_x^2E(Y) + \sigma_x = \frac{a\mu_x^2}{b} + \sigma_x.$$

Thus,

$$E_{\beta,\phi}[\phi\beta_j^2] = m_{\beta_j}^2 a_\phi / b_\phi + (C_\beta)_{jj}.$$

Evidence Lower Bound

The evidence lower bound (ELBO) for this model consists of:

$$\begin{aligned} \mathcal{L}(q) \quad = \quad & E_{\boldsymbol{\beta},\phi}(\log p(\mathbf{y}|\mathbf{X},\boldsymbol{\beta},\phi)) + E_{\boldsymbol{\beta},\phi,\boldsymbol{\tau}}(\log p(\boldsymbol{\beta},\phi|\boldsymbol{\tau})) + E_{\boldsymbol{\tau},\lambda}(\log p(\boldsymbol{\tau}|\lambda)) + \\ & + E_\lambda(\log p(\lambda)) - E_{\boldsymbol{\beta},\phi}(\log q_1(\boldsymbol{\beta},\phi)) - E_{\boldsymbol{\tau},\lambda}(\log q_2(\boldsymbol{\tau})) - E_\lambda(\log q_3(\lambda)) \end{aligned}$$

Each of the above terms are evaluated as function of the variational parameters, as follows:

$$\begin{aligned}
E_{\boldsymbol{\beta},\phi}(\log p(\mathbf{y}|\mathbf{X},\boldsymbol{\beta},\phi)) &= \frac{n}{2}(\psi(a_\phi) - \log b_\phi - \log 2\pi) + \\
&\quad -\frac{1}{2}\left[\frac{a_\phi}{b_\phi}(\mathbf{y} - \mathbf{X}\mathbf{m}_{\boldsymbol{\beta}})^T(\mathbf{y} - \mathbf{X}\mathbf{m}_{\boldsymbol{\beta}}) + tr(\mathbf{X}^T\mathbf{X}\mathbf{C}_{\boldsymbol{\beta}})\right] \\
E_{\boldsymbol{\beta},\phi,\boldsymbol{\tau}}(\log p(\boldsymbol{\beta},\phi|\boldsymbol{\tau})) &= \frac{p}{2}(\psi(a_\phi) - \log b_\phi - \log 2\pi) + (a_0 - 1)(\psi(a_\phi) - \log b_\phi) + \\
&\quad -b_0\frac{a_\phi}{b_\phi} + \frac{1}{2}\sum_{j=1}^{p} E_\tau(\log \tau_j) + \\
&\quad -\frac{1}{2}\sum_{j=1}^{p}\left[m_{\beta_j}\frac{a_\phi}{b_\phi} + (C_\beta)_{jj}\right] E_\tau\left(\frac{1}{\tau_j}\right) \\
E_{\boldsymbol{\tau},\lambda}(\log p(\boldsymbol{\tau}|\lambda)) &= p(\psi(g_\lambda) - \log h_\lambda) - \frac{g_\lambda}{h_\lambda}\sum_{j=1}^{p} E_\tau(\tau_j) \\
E_\lambda(\log p(\lambda)) &= g_0 \log h_0 - \log\Gamma(g_0) + (g_0 - 1)(\psi(g_\lambda) - \log h_\lambda) - h_0\frac{g_\lambda}{h_\lambda} \\
E_{\boldsymbol{\beta},\phi}(\log q_1(\boldsymbol{\beta},\phi)) &= \frac{p}{2}(\psi(a_\phi) - \log b_\phi - \log 2\pi) - \frac{1}{2}\log|\mathbf{C}_{\boldsymbol{\beta}}| + a_\phi \log b_\phi + \\
&\quad - \log\Gamma(a_\phi) + (a_\phi - 1)(\psi(a_\phi) - \log b_\phi) - a_\phi \\
E_{\boldsymbol{\tau},\lambda}(\log q_2(\boldsymbol{\tau})) &= \sum_{j=1}^{p}\left[\frac{c_\tau}{2}\log\frac{d_\tau}{f_{\tau_j}} - \log 2 - \log \kappa_{c_\tau}(\sqrt{d_\tau f_{\tau_j}}) + \right. \\
&\quad \left. + (c_\tau - 1)E_\tau(\log\tau_j) - \frac{d_\tau}{2}E_\tau(\tau_j) + \frac{f_{\tau_j}}{2}E_\tau\left(\frac{1}{\tau_j}\right)\right] \\
E_\lambda(\log q_3(\lambda)) &= -\log\Gamma(g_\lambda) + (g_\lambda - 1)\psi(g_\lambda) + \log h_\lambda - g_\lambda
\end{aligned}$$

The second order Taylor expansion for $\log \tau_j$ at $E(\tau_j)$ is used to obtain the approximation for its expected value: $E(\log\tau_j) \approx \log E(\tau_j) - \frac{Var(\tau_j)}{2E^2(\tau_j)}$ where the mean and the variance of τ_j are are specified above.

Note that the variational bound depends on the quantities $\mathbf{m}_{\boldsymbol{\beta}}$, $\mathbf{C}_{\boldsymbol{\beta}}$, b_ϕ, d_τ, f_{τ_j} e h_λ. The algorithm updates these quantities in each iteration. The ELBO is maximized and reaches a plateau with stabilization of those quantities. Convergence can be achieved by analyzing changes to ELBO in consecutive iterations or by analyzing the quantities on which it depends.

We end this section by showing the predictive distribution. Let $\mathbf{y}^o$ e $\mathbf{y}^p$ be the observed and the predicted vectors, respectively. Finally, let $p(\boldsymbol{\beta}, \phi|\mathbf{y}^o)$ be its variational component. Then, after some algebraic calculations, we have a Student's t-distribution (St) as follows:

$$\begin{aligned} p(\mathbf{y}^p|\mathbf{y}^o, \mathbf{X}^p) &= \int\int p(\mathbf{y}^p|\boldsymbol{\beta}, \phi)p(\boldsymbol{\beta}, \phi|\mathbf{y}^o)d\boldsymbol{\beta}d\phi \approx \int\int p(\mathbf{y}^p|\boldsymbol{\beta}, \phi)q_1(\boldsymbol{\beta}, \phi)d\boldsymbol{\beta}d\phi \\ &= St\left(\mathbf{y}^p|\mathbf{X}^T\mathbf{m}_{\boldsymbol{\beta}}, (1+\mathbf{X}^T\mathbf{C}_{\boldsymbol{\beta}}\mathbf{X})\frac{b_\phi}{(a_\phi - 1)}, 2a_\phi\right), \end{aligned}$$

where $q_1(\boldsymbol{\beta}, \phi)$ is the variational approximation of the posterior distribution, a normal-gamma distribution.

S-3 Simulations Studies

This section proposes 3 exercises with artificial data. The inference procedure assumes Jeffreys prior for ϕ and λ and a vague prior for $\boldsymbol{\beta} \sim N(\mathbf{1}, 100\ \mathbf{I_{p+1}})$.

For MCMC, 15,000 iterations were necessary to achieve convergence. The first 5,000 iterations were discarded as the burn in process and one observation was taken for every ten observations to remove autocorrelation, ending with a sample size of 10000. These quantities were obtained by using the criterion found in Lewis and Raftery [1997], that provides the number of iterations needed to guarantee the convergence in the Gibbs Sampler. The variational inference (or variational bayes - VB) algorithm is repeated until the changes in $\mathbf{m}_{\boldsymbol{\beta}}$, $\mathbf{C}_{\boldsymbol{\beta}}$, b_ϕ, d_τ, f_{τ_j} and h_λ between two consecutive iterations are less than 0.1%.

When used, the classic procedure was implemented using the R software glmnet package, which in turn applies 5-fold cross-validation to estimate the penalty parameter λ.

Considering the linear regression models, the goal of these exercises is twofold. Firstly to compare the estimation methods VB and MCMC (eventually we also use the classic Lasso in the comparison). Secondly, we wish to compare the credible interval (CI), scaled neighborhood (SN) and posterior ratio (PR) selection criteria (see Section 3.2 in the principal article for definitions). Moreover, different sparsity scenarios, variations in the sample size, different correlations between explanatory variables and different values for the accuracy of the model are considered.

Specifically, exercise 1 aims to estimate Lasso hyperparameters via VB and MCMC. Only one data set is simulated from which the real values of all parameters and hyperparameters of the model are known. VB presents results similar to MCMC and computational time 14 times shorter.

Exercise 2 is a simulation study with 100 replicates that presents a lesser sparsity structure. Variations in the sample size and in the correlation among the explanatory variables are considered. Again, VB and MCMC present similar and superior results to the classic Lasso. When the CI, SN and PR selection criteria are compared, PR gives the best results, with high proportions of exclusion for coefficients that are zero and low exclusion proportions for coefficients that are different from zero.

Exercise 3 is designed for scenarios with 100 replicates and with greater sparsity when compared to exercise 2. This exercise takes into account cases where $n < p$ and different values for the model's precision. The results are similar of those obtained in simulation 2.

S-3.1 Simulation 1: MCMC vs. VB

The purpose of simulation 1 is to compare the MCMC and VB methods to curve fitting and computational time. For this study we considered $n = 100$, $p = 10$ and each column of the matrix $\mathbf{X}$ was generated from a distribution $N(\mathbf{0}, \mathbf{I_n})$. For the parameters, were taken $\phi = 0.4$, $\lambda = 5$ and $\tau_j|\lambda \sim Exp(\lambda), \quad \forall j$. The regression coefficients and observations were generated considering the Lasso regression model.

Table S1 shows us a posterior summary of the model parameters. There, one can see the mean and the standard deviation of the approximate posterior obtained by using VB. Also, the posterior mean, the posterior standard deviation via MCMC and the true value of the parameters. Note that the point estimates obtained by the VB are close to those obtained by the MCMC. In addition, for both methods, the results are close to the real values with small standard deviation. This same conclusion can be seen in Figure S2.

In fact, Figure S2 exhibits a graphical comparison between MCMC and VB. The histogram represents the sample of the posterior distribution obtained via MCMC and the curve in red the approximate posterior density obtained by the VB. The green dot indicates the true value of the parameters. Note that the curves approximated by the VB are close to the histograms and both centered on the actual values. The remaining parameters τ_j show similar results.

Table S1: Posterior summary.

Parameters	Real	Mean VB	Sd VB	Mean MCMC	Sd MCMC
β_1	0.463	0.557	0.132	0.558	0.139
β_2	0.116	-0.046	0.136	-0.048	0.138
β_3	-1.251	-1.316	0.153	-1.315	0.160
β_4	0.250	0.396	0.161	0.383	0.171
β_5	-0.319	-0.078	0.137	-0.079	0.142
β_6	0.826	0.844	0.140	0.845	0.148
β_7	-0.036	0.091	0.142	0.096	0.149
β_8	0.144	0.074	0.136	0.081	0.143
β_9	0.064	-0.090	0.131	-0.081	0.134
β_{10}	-0.298	-0.370	0.150	-0.369	0.163
ϕ	0.4	0.473	0.066	0.473	0.070
τ_1	0.340	0.233	0.188	0.265	0.326
τ_2	0.088	0.137	0.159	0.162	0.250
τ_3	0.148	0.401	0.231	0.453	0.398
τ_4	0.470	0.201	0.179	0.234	0.359
τ_5	0.048	0.140	0.160	0.152	0.209
τ_6	0.162	0.296	0.205	0.345	0.333
τ_7	0.120	0.142	0.161	0.164	0.266
τ_8	0.069	0.139	0.160	0.173	0.365
τ_9	0.027	0.140	0.161	0.148	0.244
τ_{10}	0.275	0.194	0.177	0.212	0.278
λ	5	4.745	1.493	5.600	3.776

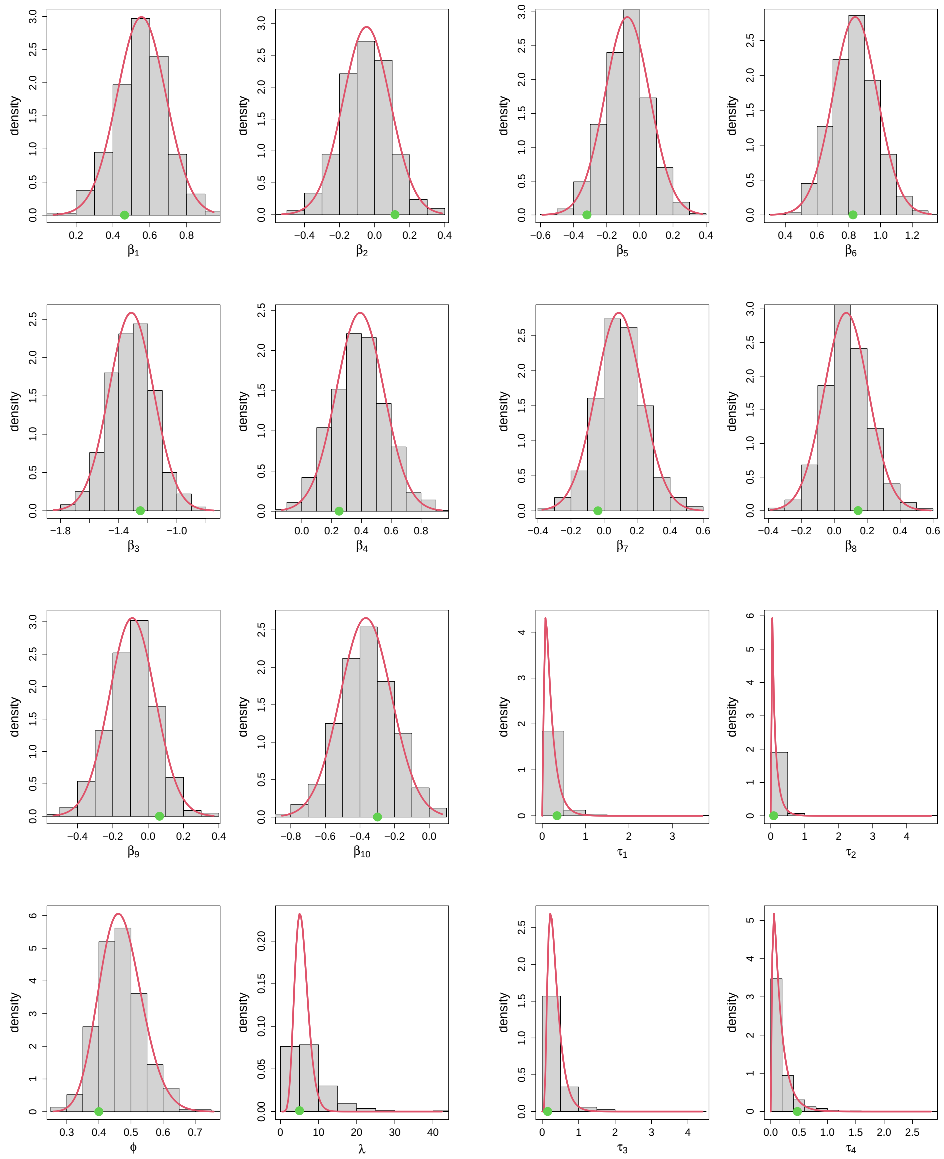

Figure S2: Comparison MCMC (histogram) versus VB (solid line). The dot marks the actual value of the parameter used to generate the data.

Since MCMC and VB present similar results, it is worth to point out the main difference between these estimation methods, which is computational time. For exercise 1, the computational time of the VB was 0.19 seconds while that of the MCMC was 10.15 seconds. In the following exercises these computational times become even more discrepant as we will be dealing with simulations with replicates.

S-3.2 Simulation 2: High correlation

In this exercise a simulation was developed considering 100 replicates $p = 8$, $\boldsymbol{\beta} = (3, 1.5, 0, 0, 2, 0, 0, 0)^T$ and the design matrix is generated from a multivariate normal distribution with zero mean, variance 1 and two different correlation structures between x_i e x_j: 0 e $0.7^{|i-j|}$, $\forall i$ e j. Let's consider $\phi = 1/9$ and 3 nested scenarios varying the sample size with $\{n_T, n_V\} = \{20, 10\}, \{100, 50\}$ e $\{200, 100\}$, where n_T e n_V denote the size of the training set and the size of the validation set, respectively. Therefore, we have a total of 6 different scenarios. Note that the explanatory variables are standardized to have mean 0 and variance 1. The Table S2 summarize the scenarios of simulation 2.

Table S2: Simulation 2 with 100 replicates, $p = 8$ explanatory variables and the vector of coefficients $\boldsymbol{\beta} = (3, 1.5, 0, 0, 2, 0, 0, 0)^T$.

Simulation	n_T	n_V	$cov(X_i, X_j)$
S2.1	20	10	0
S2.2	100	50	–
S2.3	200	100	–
S2.4	20	10	$0.7^{\|i-j\|}$
S2.5	100	50	–
S2.6	200	100	–

The comparison of the MCMC and VB methods is our main objective in this simulation, however, frequentist Lasso is also considered through the glmnet package of the R software. For the frequentist Lasso, a 5-fold cross-validation is used to select the parameter λ. In addition, different variable selection criteria will be compared as described in Subsection 3.2 of the principal article: credible interval (CI), scaled neighborhood (SN) and ratio of posteriors (PR).

In order to compare Lasso's predictive power from the different estimation techniques, MCMC, VB and frequentist Lasso, the mean absolute error (MAE) was calculated for each replicate of the validation set using the following expression:

$$MAE = \frac{1}{n_V} \sum_{i=1}^{n_V} |y_i^P - y_i^V| \qquad (3)$$

where y_i^P are the predicted values in the validation set, obtained from the fitted model after the selection of the coefficients. y_i^V are the observed values in the validation set and n_V is the size of the validation set. Note that MCMC generates a sample of the predictive distribution from each iteration of the method. Then, y_i^P is obtained as follows:

$$p(y_i^P|\mathbf{y}) = \frac{1}{AM} \sum_{j=1}^{AM} p(y_i^P|\boldsymbol{\theta}^{(j)}),$$

where AM is MCMC number of iterations and $\boldsymbol{\theta}$ is the vector of coefficients.

Figure S3 shows the box-plots of the mean absolute errors for each of the six proposed scenarios. As the sample size increases, we observe a smaller difference between the three estimation methods. When the sample is small, similar results are obtained between MCMC and VB. These have the median MAE and the lowest dispersion when compared to the frequentist Lasso. Next, we will detail the performance of the selection criteria for each β_j.

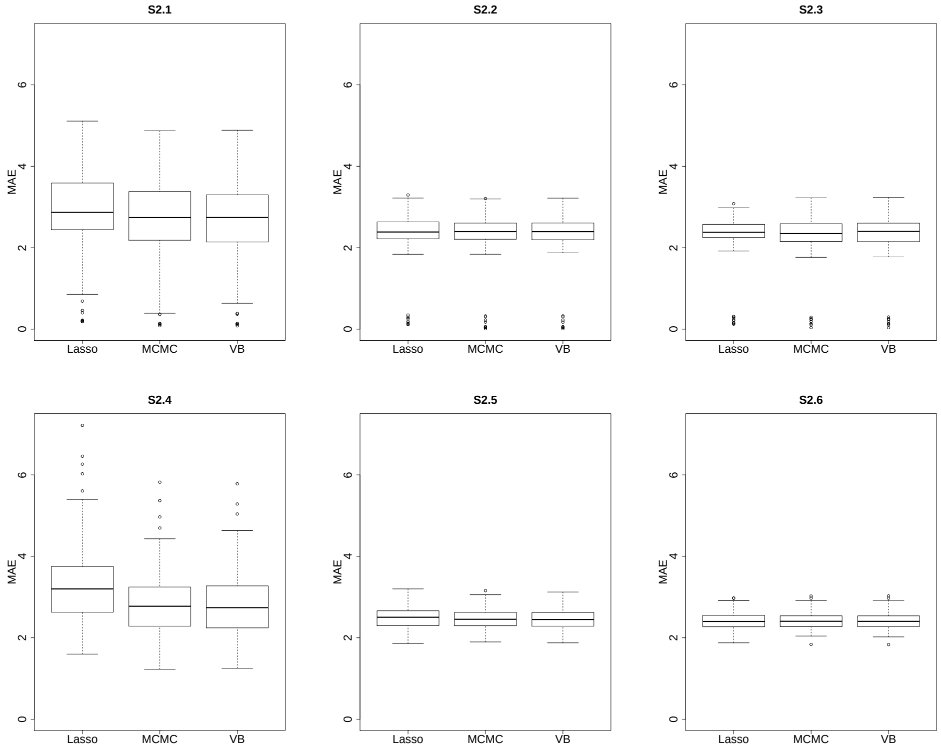

Figure S3: Mean absolute error (MAE) using (S-3.2) for the 6 scenarios of the simulation 2. Estimation methods: MCMC, VB and frequentist Lasso.

The Table S3 shows the frequency of times that the predictor x_j, $j = 1, \ldots, 8$ was excluded in the 100 replicates, considering the three variables selection methods and all six scenarios built in Simulation 2. We present the proportions only for the VB because so far its results are similar to those of the MCMC. Note that for this simulation exercise the PR presents the best results in all scenarios, with a greater proportion of exclusion when the actual values of β_j are zero and a small proportion when the β 's are different from zero. In addition, it is noted that as the sample size increases, the three criteria tend to correctly choose coefficients that are zero and the coefficients that are different from zero.

From simulations 1 and 2, one may notice that the approximations of the VB are as good as the results obtained by the MCMC. Nevertheless, the gain in computational time provided by VB (approximately 0.02 seconds on average in the simulation S2.6, for example) is far superior than MCMC (approximately 16 seconds on average in the simulation S2.6, for example). In addition, we saw that PR is a variable selection criterion that presents superior results when compared with CI and SN. In the following subsection we show the performance of the VB estimation method and the PR selection criterion for a more complex numerical experiment with greater sparsity.

Table S3: Comparison of the three methods on variable selection accuracy using VB for the six scenarios (the frequency of exclusions for the predictor x_j, $j = 1, \ldots, 8$) with $\boldsymbol{\beta} = (3, 1.5, 0, 0, 2, 0, 0, 0)^T$.

Simulation	Method	β_1	β_2	β_3	β_4	β_5	β_6	β_7	β_8
S2.1	VB + CI	0.01	0.09	0.64	0.57	0.13	0.66	0.65	0.62
	VB + SN	0.02	0.15	0.73	0.71	0.20	0.77	0.75	0.71
	VB + PR	0.00	0.09	0.88	0.70	0.13	0.82	0.75	0.92
S2.2	VB + CI	0.00	0.00	0.47	0.51	0.00	0.57	0.67	0.56
	VB + SN	0.00	0.00	0.70	0.62	0.00	0.71	0.78	0.72
	VB + PR	0.00	0.00	0.73	0.78	0.00	0.73	0.84	0.72
S2.3	VB + CI	0.00	0.00	0.50	0.47	0.00	0.67	0.60	0.58
	VB + SN	0.00	0.00	0.71	0.62	0.00	0.71	0.73	0.76
	VB + PR	0.00	0.00	0.72	0.73	0.00	0.77	0.80	0.77
S2.4	VB + CI	0.02	0.09	0.53	0.53	0.19	0.71	0.60	0.64
	VB + SN	0.06	0.11	0.70	0.62	0.24	0.75	0.73	0.74
	VB + PR	0.02	0.09	0.65	0.80	0.18	0.85	0.81	0.88
S2.5	VB + CI	0.00	0.00	0.57	0.49	0.00	0.60	0.51	0.57
	VB + SN	0.00	0.00	0.73	0.68	0.00	0.75	0.70	0.75
	VB + PR	0.00	0.00	0.78	0.77	0.00	0.76	0.80	0.82
S2.6	VB + CI	0.00	0.00	0.55	0.47	0.00	0.60	0.46	0.57
	VB + SN	0.00	0.00	0.70	0.66	0.00	0.77	0.64	0.75
	VB + PR	0.00	0.00	0.77	0.75	0.00	0.79	0.72	0.77

S-3.3 Simulation 3: High sparsity with small n and large p

In this simulation we consider a situation with sparsity given by $p = 40$ and $\boldsymbol{\beta} = (\mathbf{0}^{\mathbf{T}}, \mathbf{3}^{\mathbf{T}}, \mathbf{0}^{\mathbf{T}}, \mathbf{3}^{\mathbf{T}})^T$, where $\mathbf{0}$ and $\mathbf{3}$ are vectors of dimension 10 and each of their entries are 0 and 3 respectively. The design matrix $\mathbf{X}$ is generated from a multivariate normal distribution with mean zero, variance 1 and the correlation between the columns x_i e x_j is equal to 0.5, $\forall i \neq j$. We analyze 4 different scenarios by varying the sample size and the precision parameter ϕ. The simulated data were analyzed as follows, $\{n_T, n_V\} = \{20, 10\}$ e $\{200, 100\}$ where n_T e n_V are the size of the training set and the size of the validation set respectively. In addition, we set the precision parameter as $\phi = 1/9$ and $\phi = 1/225$. For each scenario we consider 100 replicates. Table S4 summarizes all the scenarios considered in this simulation exercise 3. It is worth mentioning that in scenarios S3.1 and S3.3 we have $n < p$.

Table S4: Scenarios in Simulation 3

Simulation	n_T	n_V	ϕ
S3.1	20	10	$1/9$
S3.2	200	100	-
S3.3	20	10	$1/225$
S3.4	200	100	-

Similarly to simulation 2, the MAE was calculated for each replicate as a predictive measure. Figure S4 shows the box-plots of each scenario for MCMC, VB and Lasso. One may see that the MCMC and VB present similar and superior results to the Lasso when the sample size is small. As the sample increases the results become similar in the 3 approaches.

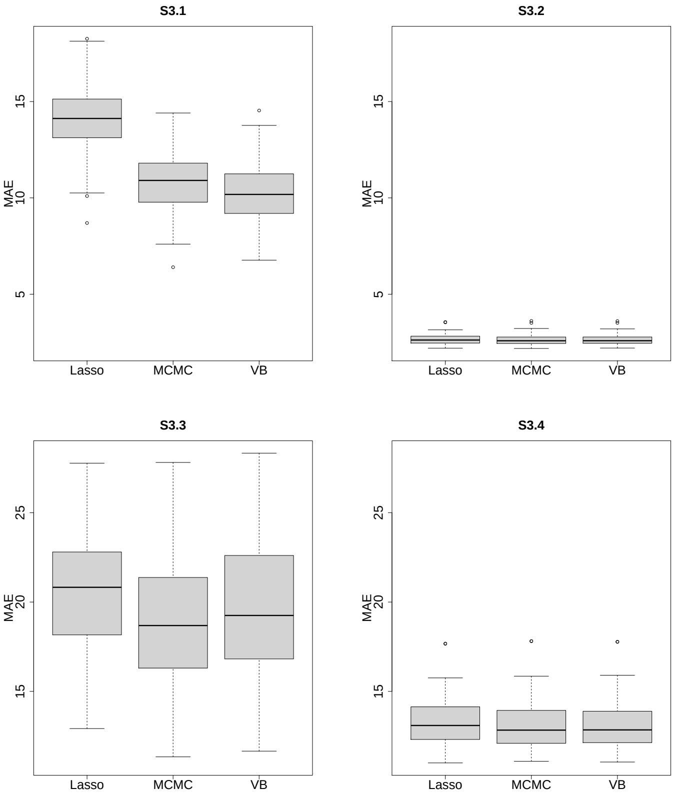

Figure S4: Mean absolute error (MAE) obtained by using (S-3.2) for the 4 scenarios in simulation 3, comparing the estimation methods MCMC, VB and Lasso.

Figure S5 shows the proportions of exclusions (gray bars) and selections (black bars) for each of the 40 coefficients in the 100 replicates, when comparing the estimation methods, VB and Lasso. MCMC was omitted for presenting results similar to VB. In the Bayesian context, the selection criterion used in all scenarios was the PR. It is expected that the black bars will be larger when the true coefficients are different from zero and that the gray bars will be large when the true coefficients are equal to zero. The proportions of the errors are represented by the black bars when the coefficients are zero (type I error) and by the gray bars when the coefficients are different from zero (type II error).

Thus, it can be seen for $n < p$, both VB and Lasso do not have a good selection and exclusion performance, with a slight advantage of VB. On the other hand, as the sample increases, the VB presents good results, better than those presented by Lasso. Also note that when $n > p$ both VB and Lasso have the same type II error. However, for all coefficients, the type I error is considerably less in VB than in frequentist Lasso. Although the MCMC and VB present similar results, in terms of the computational time the VB presents 0.06 second on average and the MCMC 30 seconds on average, both in simulation S3.4.

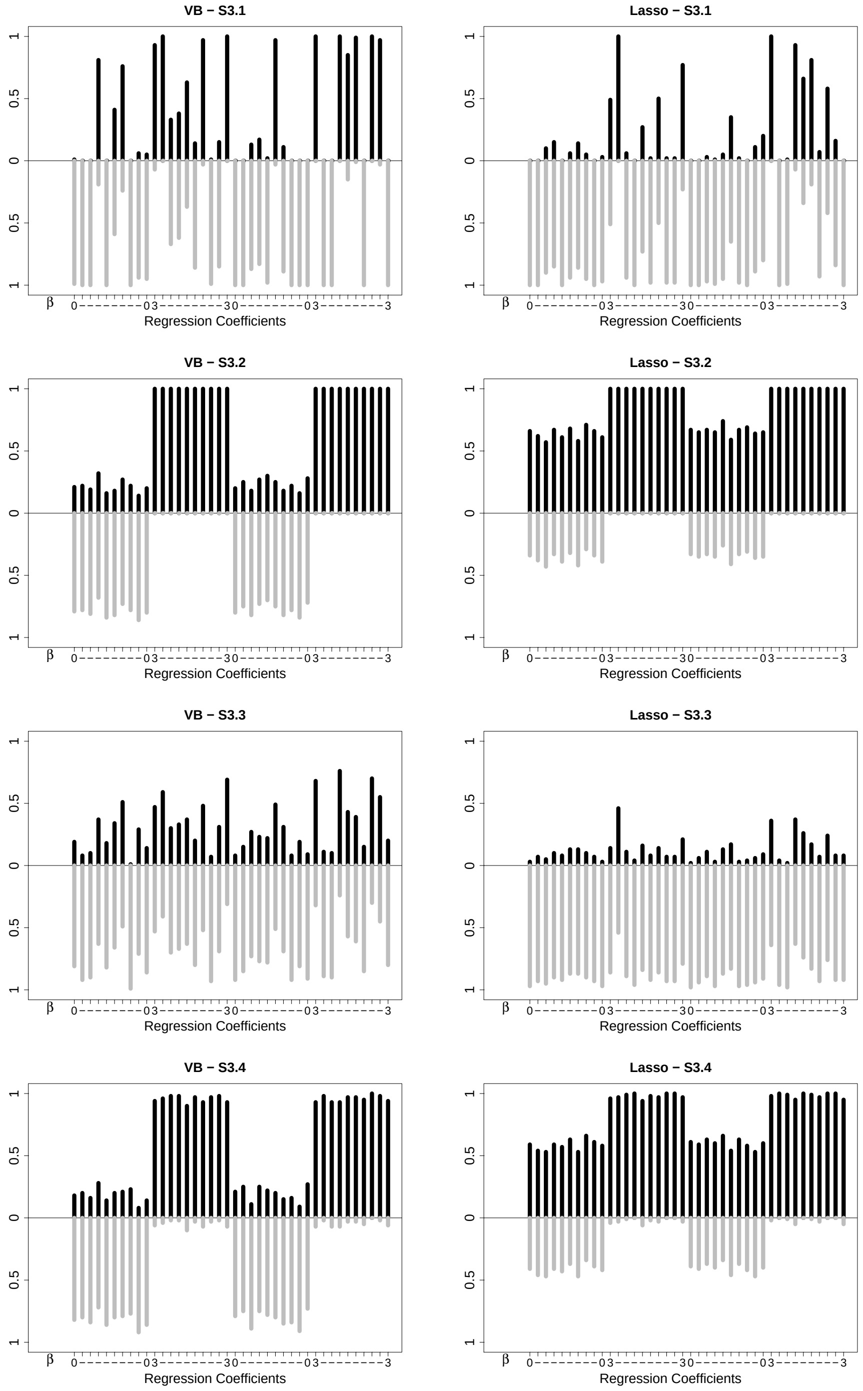

Figure S5: Proportion of selected (black) and excluded coefficients (gray) for the 4 scenarios in simulation 3 with the estimation methods VB (left column) and Lasso (right column).

References

D. M. Blei, A. Kucukelbir, and J. D. McAuliffe. Variational inference: a review for statisticians. *J. Amer. Statist. Assoc.*, 112(518):859–877, 2017. ISSN 0162-1459. doi: 10.1080/01621459.2017.1285773. URL https://doi.org/10.1080/01621459.2017.1285773.

T. C. O. Fonseca, H. S. Migon, and H. Mirandola. Reference bayesian analysis for hierarchical models. *Arxiv preprint*, 2019. URL https://arxiv.org/abs/1904.11609.

A. E. Gelfand and A. F. M. Smith. Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association*, 85(410):398–409, 1990. ISSN 01621459. URL http://www.jstor.org/stable/2289776.

B. Jørgensen. *Statistical properties of the generalized inverse Gaussian distribution*, volume 9 of *Lecture Notes in Statistics*. Springer-Verlag, New York-Berlin, 1982. ISBN 0-387-90665-7.

S. M. Lewis and A. E. Raftery. Estimating Bayes factors via posterior simulation with the Laplace-Metropolis estimator. *J. Amer. Statist. Assoc.*, 92(438):648–655, 1997. ISSN 0162-1459. doi: 10.2307/2965712. URL https://doi.org/10.2307/2965712.

T. Park and G. Casella. The Bayesian Lasso. *Journal of the American Statistical Association*, 103(482): 681–686, June 2008. ISSN 0162-1459, 1537-274X. doi: 10.1198/016214508000000337. URL http://www.tandfonline.com/doi/abs/10.1198/016214508000000337.