Appendix: Bayesian Variable Selection for Skew Normal Models

Arnold van Wyk[a], Janet van Niekerk[a,b], Mohammad Arashi[c], Andriette Bekker[a]

[a]*Department of Statistics, University of Pretoria, Pretoria, South Africa,*
[b]*Statistics Program, CEMSE, King Abdullah University of Science and Technology, Thuwal, Saudi Arabia,*
[c]*Department of Statistics, Faculty of Mathematical Sciences, Ferdowsi University of Mashhad, Mashhad, Iran,*

Abstract

Variable selection is one of the most commonly faced problems in statistical analysis. In the frequentist paradigm, penalized regression methods such as L1 regularization and LASSO are used to induce sparsity in high-dimensional settings. In the Bayesian setting, sparsity can be induced by means of a two-component mixture prior with sufficient probability mass at zero. There has also been a recent development that uses global-local shrinkage priors for high-dimensional Bayesian variable selection. The Dirichlet-Laplace (DL) prior is a popular example of this, and has shown promising results compared to existing feature selection methods in the Bayesian framework. In this paper, we propose incorporating an asymmetrical component into the variable selection framework. This is showcased by incorporating a skew normal random error component into the Dirichlet-Laplace prior for linear regression. We also propose a framework for prior selection and hyperparameter tuning of the proposed model. The performance of the proposed model is assessed and compared with its symmetrical counterpart in both simulated and real-data examples, and is found to not only perform well, but is also able to identify certain non-zero signals due to the inclusion of skewness in the proposed model.

Keywords: Bayesian variable selection, Dirichlet-Laplace, High-dimensional, Penalized regression, Shrinkage prior, Skew normal.

1. Proof of Proposition 1:

Given $U \sim \text{SN}(0, 1, \lambda)$. Since Y is a simple linear transformation of U, its distribution again skew normal. Then, for $f(y, z) = f(y|z)\ f(z)$, we have that $f(z)$ represents a standard normal density, while:

$$(Y|Z = z) = (\mathbf{x}^\top \boldsymbol{\beta} + \sigma U \mid Z = z) = \begin{cases} \mathbf{x}^\top \boldsymbol{\beta} + \sigma W & \text{if} \quad z_i \geq 0 \\ 0 & \text{if} \quad z_i < 0. \end{cases} \tag{1}$$

Then, by using results on conditional normal density functions, we get the following:

$$(Y|Z = z) \sim \begin{cases} \text{N}\left(\mathbf{x}^\top \boldsymbol{\beta} + \sigma\delta z, (1 - \delta^2)\sigma^2\right) & \text{if} \quad z_i \geq 0 \\ 0 & \text{if} \quad z_i < 0 \end{cases} \tag{2}$$

2. Full Conditional Distributions

2.1. Proof of Theorem **??**

To define the likelihood, consider again $y_i \sim \text{SN}_c(x_i^\top \boldsymbol{\beta}, \sigma^2, \gamma_1), i = 1, \ldots, n$. Then, by using the result in (2), we can write the augmented likelihood function as if we have observed the latent variable t_i, $i = 1, \ldots, n$. The augmented likelihood function is thus given by:

$$\begin{aligned} L(\boldsymbol{\Omega}) &= \prod_{i=1}^{n} \text{N}\left(y_i; x_i^\top \beta + \frac{\sigma_\epsilon}{S_U}\delta t_i + \frac{\sigma_e}{S_U} E_U, \zeta^{-1}\right) \times \text{N}(t_i; 0, 1)\ I_{(0,\infty)}(t_i) \\ &= \left\{ \prod_{i=1}^{n} \zeta\ \exp\left\{ -\frac{1}{2}\zeta^2 \left(y_i - x_i^\top \beta - \frac{\sigma_e}{S_U}\delta t_i - \frac{\sigma_e}{S_U} E_U \right)^2 - \frac{1}{2} t_i^2 \right\}\ I_{(0,\infty)}(t_i) \right\} \end{aligned} \tag{3}$$

where $\zeta = \frac{S_U^2}{(1-\delta^2)\sigma_e^2}$. From the above specification, we are then able to define the joint prior distribution as follows:

$$
\begin{aligned}
\pi(\{\beta_j\}, \psi, \phi, \tau, \delta) &= \prod_{j=1}^{p} \pi(\beta_j)\pi(\psi_j)\pi(\phi_j)\pi(\tau)\pi(\delta) \\
&\propto \prod_{j=1}^{p} \left(\frac{1}{\sqrt{2\pi}\sqrt{\psi_j \phi_j^2 \tau^2}} \exp\left\{ -\frac{1}{2} \frac{(\beta_j - 0)^2}{\psi_j \phi_j^2 \tau^2} \right\} \right) \\
&\times \prod_{j=1}^{p} \left(\frac{1}{2} \exp\left\{ -\frac{1}{2} \psi_j \right\} \right) \times \left(\frac{1}{\mathrm{B}(\alpha)} \prod_{j=1}^{p} \phi_j^{\alpha - 1} \right) \\
&\times \left(\frac{\left(\frac{1}{2}\right)^{p\alpha}}{\Gamma(p\alpha)} \tau^{p\alpha - 1} \exp\left\{ -\frac{1}{2} \tau \right\} \right) \\
&\times \left(\left(\sigma_e^2\right)^{-\left(\frac{df_e}{2} + 1\right)} \exp\left\{ -\frac{S_e}{2\sigma_e^2} \right\} \right) \\
&\times \left(\left(\frac{1 - \delta}{2} \right)^{a_0 - 1} \left(1 - \frac{1 - \delta}{2} \right)^{b_0 - 1} \right) \qquad (4)
\end{aligned}
$$

Finally, using (3) and (4), we can define the joint posterior distribution

as follows:

$$
\begin{aligned}
\pi(\{\beta_j\}, \psi, \phi, \tau | y_i) &\propto \pi(\{\beta_j\}, \psi, \phi, \tau, \delta)\, L(\{\beta_j\}, \delta, \mathbf{y}, \mathbf{t}) \\
&= \zeta^{\frac{n}{2}} \ \exp\left\{ -\frac{1}{2}\zeta^2 \sum_{i=1}^{n} \left(y_i - x_i^\top \beta - \frac{\sigma_e}{S_U}\delta t_i - \frac{\sigma_e}{S_U} E_U \right)^2 \right\} \\
&\times \exp\left\{ -\frac{1}{2} \sum_{i=1}^{n} t_i^2 \right\} \ I_{(0,\infty)}(t_i) \\
&\times \prod_{j=1}^{p} \left(\frac{1}{\sqrt{2\pi}\sqrt{\psi_j \phi_j^2 \tau^2}} \exp\left\{ -\frac{1}{2} \frac{(\beta_j - 0)^2}{\psi_j \phi_j^2 \tau^2} \right\} \right) \\
&\times \prod_{j=1}^{p} \left(\frac{1}{2} \exp\left\{ -\frac{1}{2}\psi_j \right\} \right) \times \left(\frac{1}{\mathrm{B}\,(\alpha)} \prod_{j=1}^{p} \phi_j^{\alpha - 1} \right) \\
&\times \left(\frac{\left(\frac{1}{2}\right)^{p\alpha}}{\Gamma(p\alpha)} \tau^{p\alpha - 1} \exp\left\{ -\frac{1}{2}\tau \right\} \right) \\
&\times \left(\left(\sigma_e^2\right)^{-\left(\frac{df_e}{2}+1\right)} \exp\left\{ -\frac{S_e}{2\sigma_e^2} \right\} \right) \\
&\times \left(\left(\frac{1-\delta}{2} \right)^{a_0 - 1} \left(1 - \frac{1-\delta}{2} \right)^{b_0 - 1} \right)
\end{aligned}
$$

Once the likelihood, joint prior, and joint posterior has been defined, what follows is to derive the full conditional posteriors for each of the model components, such that we are able to formulate the Gibbs sampling steps necessary for posterior computation. As such, the full conditional posteriors for the relevant parameters will now be derived sequentially as they appear in (4:

- **Posterior of β**:
  The full conditional posterior distribution of β:

$$\begin{aligned}\pi(\beta_j|-) &\propto \prod_{i=1}^{n} \exp\left\{-\frac{1}{2}\zeta^2\left(y_i - \mathbf{x}_i^\top\boldsymbol{\beta} - \frac{\sigma_\epsilon}{S_U}(\delta t_i - E_U)\right)^2\right\} \\ &\quad\times \exp\left\{-\frac{1}{2}\frac{\theta_j^2}{\psi_j\phi_j^2\tau^2}\right\} \\ &= \mathrm{N}(y_i^*|x_{ij}\beta_j, \zeta^{-2})\,\mathrm{N}(\beta_j|0, \psi_j\phi_j^2\tau^2),\end{aligned}$$

  where $y_i^* = y_i - \mathbf{x}_{i,-j}^\top\boldsymbol{\beta}_{-j} - \frac{\sigma_\epsilon}{S_U}(\delta t_i - E_U)$, $\mathbf{x}_{i,-j}^\top$ denotes the i-th row of $\boldsymbol{X}$ with the j-th column removed, and the notation $\boldsymbol{\beta_{-j}}$ follows the same formulation. The right hand side of the above equation has the same kernel as a normal distribution such that

$$\pi(\beta_j|-) \sim \mathrm{N}\left(\mu_j, \sigma_j^2\right)$$

  with

$$\mu_j = \frac{\sum_{i=1}^n x_{ij}y_i^*}{\sum_{i=1}^n x_{ij}^2 + \frac{(1-\delta^2)\sigma_\epsilon^2}{S_U^2\psi_j\phi_j^2\tau^2}} \tag{5}$$

  and

$$\sigma_j^2 = \left(\sum_{i=1}^n x_{ij}^2 + \frac{(1-\delta^2)\sigma_\epsilon^2}{S_U^2\psi_j\phi_j^2\tau^2}\right)^{-1}\sum_{i=1}^n x_{ij}y_i^* \tag{6}$$

- **Posterior of t**:
  The full conditional posterior distribution of t:

$$\begin{aligned}\pi\left(t_i|-\right) &\propto \exp\left\{-\frac{1}{2(1-\delta^2)}\left(\left(\frac{y_i - x_{i.}\beta_j}{\sigma_e}\right)S_U^2 + E_U - \delta t_i\right)^2\right\} \\ &\quad\times \exp\left\{-\frac{1}{2}t_i^2\right\} I_{(0,\infty)}(t_i) \\ &= \exp\left\{-\frac{1}{2(1-\delta^2)}\left(t_i - \delta\left(\left(\frac{y_i - x_{i.}\beta_j}{\sigma_e}\right)S_U^2 + E_U\right)\right)^2\right\} I_{(0,\infty)}(t_i)\end{aligned}$$

  which corresponds with the kernel of a truncated normal distribution with location parameter:

$$\delta\left(\left(\frac{y_i - x_{i.}\beta_j}{\sigma_e}\right)S_U^2 + E_U\right), \tag{7}$$

scale parameter: $1-\delta^2$, lower truncation bound 0 and upper truncation bound ∞.

- **Posterior of $\boldsymbol{\psi}$**: The full conditional posterior distribution of ψ:

$$\begin{aligned}
\pi\left(\frac{1}{\psi_j}|\phi_j,\tau,\beta_j\right) &\propto \exp\left(-\frac{q}{2}\psi_j\right)\exp\left(-\frac{1}{2}\frac{\beta_j^2}{\psi_j\phi_j^2\tau^2}\right) \\
&= \exp\left(-\frac{1}{2}\frac{\psi_j\beta_j^2}{\phi_j^2\tau^2}\left(q\frac{\phi_j^2\tau^2}{\beta_j^2}+\frac{1}{\psi_j^2}\right)\right) \\
&= \exp\left(-\frac{1}{2}\frac{\psi_j\beta_j^2}{\phi_j^2\tau^2}\left(\frac{1}{\psi_j^2}+\sqrt{q}\frac{2}{\psi_j}\frac{\phi_j\tau}{\beta_j}-\sqrt{q}\frac{2}{\psi_j}\frac{\phi_j\tau}{\beta_j}+q\frac{\phi_j^2\tau^2}{\beta_j^2}\right)\right) \\
&= \exp\left(-\frac{1}{2}\frac{\psi_j\beta_j^2}{\phi_j^2\tau^2}\left(\frac{1}{\psi_j^2}-\sqrt{q}\frac{2}{\psi_j}\frac{\phi_j\tau}{\beta_j}+q\frac{\phi_j^2\tau^2}{\beta_j^2}\right)-\sqrt{q}\frac{\beta_j}{\phi_j\tau}\right) \\
&\propto \exp\left(-\frac{1}{2}\frac{\psi_j\beta_j^2}{\phi_j^2\tau^2}\left(\frac{1}{\psi_j^2}-\sqrt{q}\frac{2}{\psi_j}\frac{\phi_j\tau}{\beta_j}+q\frac{\phi_j^2\tau^2}{\beta_j^2}\right)\right) \\
&= \exp\left(-\frac{1}{2}\frac{\sqrt{q}\left(\frac{1}{\psi_j}-\sqrt{q}\frac{\phi_j\tau}{\beta_j}\right)^2}{\frac{1}{\psi_j}\sqrt{q}\frac{\phi_j^2\tau^2}{\beta_j^2}}\right)
\end{aligned}$$

That is we sample $\hat{\psi}_j = \frac{1}{\psi_j}$ from an inverse Gaussian distribution iG $\left(\sqrt{q}\frac{\phi_j\tau}{|\beta_j|},\sqrt{q}\right)$ and set $\psi_j = \frac{1}{\hat{\psi}_j}$. We set β_j to $|\beta_j|$ as the mean must be greater than zero and β_j is the only term in the mean that can cause the mean to be less than zero.

In terms of the use of the constant q, I found including it this way leads to a cleaner derivation of the inverse Gaussian posterior and gives easy control i.t.o varying the value of q. The only thing that has changed from before, is that instead of assuming an Exp $\left(\frac{1}{2}\right)$ prior for ψ, we assume an Exp $\left(\frac{q}{2}\right)$ prior.

What this then allows us to do is either increase the value of q to add more weight to values closer to 0 and this increase the effect of local shrinkage, or decrease the value of q for the opposite effect.

- **Posterior of τ**: The full conditional posterior distribution of τ:

$$\begin{aligned}\pi\left(\tau|\phi,\beta\right) &\propto \tau^{p\alpha-1}\exp\left\{-\frac{1}{2}\tau\right\}\prod_{j=1}^{p}\tau^{-1}\exp\left\{-\frac{|\beta_j|}{\phi_j\tau}\right\}\\ &= \tau^{(p\alpha-p)-1}\exp\left\{-\frac{1}{2}\left(\tau+\frac{1}{\tau}\left(2\sum_{j=1}^{p}\frac{|\beta_j|}{\phi_j}\right)\right)\right\}\end{aligned}$$

  That is we sample τ from a generalized inverse Gaussian distribution $\text{GIG}\left(p\alpha-p,1,2\sum_{j=1}^{p}\frac{|\beta_j|}{\phi_j}\right)$.

- **Posterior of $\boldsymbol{\phi}$**

  Integrating out τ we get the following joint posterior for for ϕ:

$$\begin{aligned}\pi(\phi_1,\ldots,\phi_{p-1}|\beta) \propto &\left\{\prod_{j=1}^{p}\frac{1}{\phi_j}\phi_j^{\alpha-1}\right\}\int_{\tau=0}^{\infty}\exp\left\{-\frac{\tau}{2}\right\}\tau^{p\alpha-p-1}\\ &\times\exp\left\{\sum_{j=1}^{p}\frac{|\beta_j|}{\phi_j\tau}\right\}d\tau. \qquad (8)\end{aligned}$$

  The next step involves a result from the theory of normalized random measures. Suppose $T_1,\ldots,T_p$ are independent random variables with T_j having a density f_j on $(0,\infty)$. Let $\phi_j=\frac{T_j}{T}$ where $T=\sum_{j=1}^{p}T_j$. Then, the joint density of f of $(\phi_1,\ldots,\phi_{p-1})$ has the form

$$f(\phi_1,\ldots,\phi_{p-1}) = \int_{t=0}^{\infty}t^{p-1}\prod_{j=1}^{p}f_j(\phi_j t)dt, \qquad (9)$$

  where $\phi_p=1-\sum_{j=1}^{p-1}\phi_j$. Setting $f_j(x)\propto\frac{1}{x^{\delta}}\exp\left\{-\frac{|\beta_j|}{x}\right\}\exp\left\{-\frac{1}{2}\right\}$ in (9) we get

$$\begin{aligned}f(\phi_1,\ldots,\phi_{p-1}) = &\left\{\prod_{j=1}^{p}\frac{1}{\phi_j^{\delta}}\right\}\int_{t=0}^{\infty}\exp\left\{-\frac{\phi_j t}{2}\right\}t^{p-1-p\delta}\\ &\times\exp\left\{-\sum_{j=1}^{p}\frac{|\beta_j|}{\phi_j t}\right\}. \qquad (10)\end{aligned}$$

The goal is to equate the expression (10) with the one in expression (8). By comparing the exponents of ϕ_j we get $\delta = 2 - \alpha$. Next we see that $p - 1 - p\delta = p\alpha - p - 1$ is also satisfied for $\delta = 2 - \alpha$. The final piece of the puzzle is observing that f_j corresponds to a giG$(\alpha - 1, 1, 2|\beta_j|)$ distribution.

- **Posterior of δ**:
  The full conditional posterior distribution of δ:

$$\pi(\delta|-) \propto \prod_{i=1}^{n} \frac{S_U}{\sqrt{1-\delta^2}} \exp\left\{ -\frac{S_U^2}{2(1-\delta^2)\sigma_e^2} \left(y_i - x_i^\top \beta - \frac{\sigma_e}{S_U}\delta t_i - \frac{\sigma_e}{S_U} E_U \right)^2 \right\}$$
$$\times \left(\frac{1-\delta}{2}\right)^{a_0-1} \left(1 - \frac{1-\delta}{2}\right)^{b_0-1} I_{(-1,1)}(\delta)$$

The above kernel for δ is quite complex and does not correspond to any known univariate density function, so we must obtain the samples some other way such as Metropolis Hastings or some other MCMC technique. The idea for the following implementation can be found in [1]. We will consider using Fisher's transformation [2] of δ, which is defined as follows:

$$\upsilon = \frac{1}{2}\log\left(\frac{1+\delta}{1-\delta}\right) = \tanh^{-1}(\delta) \tag{11}$$

such that υ has support in $\mathbb{R}$. We then obtain the density function of υ using the transformation method showcased in Casella and Berger [3, Chapter 2]. This gives us the following:

$$p(\upsilon|else) \propto p(\delta|else) \times \text{sech}^2(\upsilon). \tag{12}$$

In the Random Walk Metropolis Algorithm we generate values for υ by choosing a proposal transition kernel that will add noise to the current state. Assuming a value of $\upsilon = \upsilon_k$, the idea is to update the value so that at the next iteration we have $\upsilon = \upsilon_{k+1}$. The steps of the algorithm are thus given as:

1. Sample υ where $\upsilon = \upsilon_k + Z$, $Z \sim \text{N}(0, \nu^2)$.

2. Sample u, $U \sim Uniform(0,1)$.
3. If $u < \frac{p(\upsilon|else)}{p(\upsilon_k|else)}$, then $\upsilon_{k+1} = \upsilon$, else $\upsilon_{k+1} = \upsilon_k$.

Once we have υ_{k+1}, simply compute the values of δ using $\delta = \tanh(\upsilon_{k+1})$. Also note that ν^2 can be modified to obtain an optimal acceptance rate. Ideally we want to tune it such that the acceptance rate is roughly 0.25 [4].

- **Posterior of σ_ϵ^2:**

The full conditional posterior distribution of σ_ϵ^2:

$$\begin{aligned}\pi(\sigma_\epsilon^2|-) &\propto \prod_{i=1}^{n} \frac{1}{\sigma_\epsilon} \exp\left\{-\frac{S_U^2}{2(1-\delta^2)\sigma_e^2}\left(y_i - x_i^\top\beta - \frac{\sigma_e}{S_U}\delta t_i - \frac{\sigma_\epsilon}{S_U}E_U\right)^2\right\} \\ &\quad \times \chi^{-2}(\sigma_\epsilon^2|S_e, df_e) \\ &= \exp\left\{-\frac{1}{2\sigma_\epsilon^2}\left[S_e + \frac{S_U^2}{(1-\delta^2)}\sum_{i=1}^{n}\left(y_i - x_i^\top\beta - \frac{\sigma_e}{S_U}\delta t_i - \frac{\sigma_e}{S_U}E_U\right)^2\right]\right\} \\ &\quad \times (\sigma_\epsilon^2)^{-\left(\frac{n+df_e}{2}+1\right)}\end{aligned}$$

As in the case of δ, we are again faced with a complex kernel that does not correspond to any known univariate density function. As such, we consider the idea in [5], whereby we sample σ_ϵ^2 using a Random Walk Metropolis Algorithm with a de-constraint transformation of the parameter. Thus, given that $\sigma_\epsilon^2 > 0$, we let $\xi = \log(\sigma_\epsilon^2)$ such that $\xi \in \mathbb{R}$. The density function of ξ is then obtained using the transformation detailed in (cite cassella and berger 2002), giving us the following representation:

$$p(\xi|\cdot) \propto p(\sigma^2|\cdot) \times \exp(\xi). \tag{13}$$

In the subsequent Random Walk Metropolis algorithm we generate values for ξ by choosing an appropriate proposal transition kernel to add noise to the current state of ξ. Let the current state of xi be ξ_k, so that the updated value in the iteration will be defined as ξ_{k+1}. The steps of the algorithm are thus given as:

1. Sample ξ where $\xi = \xi_k + Z$, with $Z \sim \mathrm{N}(0, \nu^2)$.
2. Sample u, where $U \sim \mathrm{Uniform}(0,1)$.

3. If $u < \frac{p(\xi|else)}{p(\xi_k|else)}$, then $\xi_{k+1} = \xi$, else $\xi_{k+1} = \xi_k$.

Once ξ_{k+1} is obtained, we again simply compute the values of σ_ϵ^2 using $\sigma_\epsilon^2 = \exp(\xi_{k+1})$. This algorithm is also tuned to an acceptance rate of roughly 0.25 [4].

References

[1] P. Pérez-Rodríguez, R. Acosta-Pech, S. Pérez-Elizalde, C. V. Cruz, J. S. Espinosa, J. Crossa, A bayesian genomic regression model with skew normal random errors, G3: Genes, Genomes, Genetics 8 (5) (2018) 1771–1785.

[2] R. A. Fisher, Frequency distribution of the values of the correlation coefficient in samples from an indefinitely large population, Biometrika 10 (4) (1915) 507–521.

[3] G. Casella, R. L. Berger, Statistical Inference, Cengage Learning, 2021.

[4] A. Gelman, W. R. Gilks, G. O. Roberts, Weak convergence and optimal scaling of random walk metropolis algorithms, The Annals of Applied Probability 7 (1) (1997) 110–120.

[5] K. Hea-Jung, Bayesian estimation for skew normal distributions using data augmentation, Journal of the Korean Statistical Society 12 (2) (2005) 323–333.