Technical Document to accompany "Bayesian Approaches to Shrinkage and Sparse Estimation: A guide for applied econometricians"

Dimitris Korobilis
University of Glasgow

Kenichi Shimizu
University of Glasgow

A Inference with non-hierarchical natural conjugate and independent priors

In this section, we review non-hierarchical Bayesian estimation of simple regression models under natural conjugate and independent priors. Most of the shrinkage priors that we review in this paper have forms of either conjugate or independent prior, conditional on the parameters such as prior variances of the slope coefficients. Therefore, it is helpful to first review the conditional posterior distributions under the non-hierarchical priors.

Consider the simple linear regression model of the form

$$y_i = x_i\boldsymbol{\beta} + \varepsilon_i, \quad \varepsilon_i \sim N(0, \sigma^2), \quad i = 1, ..., n, \tag{A.1}$$

where $\boldsymbol{\beta}$ is a $p \times 1$ vector. We define $\boldsymbol{y} = (y_1, ..., y_n)'$, $\boldsymbol{X} = (x_1', ..., x_n')'$ and $\boldsymbol{\varepsilon} = (\varepsilon_1, ..., \varepsilon_n)'$, such that the stacked form of the regression model is

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{A.2}$$

where $\boldsymbol{\varepsilon} \sim N_n(\mathbf{0}_{n\times 1}, \sigma^2\boldsymbol{I}_n)$. In this section, we assume that the prior variances on $\boldsymbol{\beta}$ are fixed and will review posterior sampling under generic normal-inverse-gamma priors on $\boldsymbol{\beta}$ and σ^2. The prior of $\boldsymbol{\beta}$ can be defined either dependent or independent on σ^2. In both cases, assume an inverse gamma prior[1] on σ^2.

$$\sigma^2 \quad \sim \quad Inv - Gamma\,(a, b)\,, \tag{A.3}$$

where we use the parametrization so that if $x \sim Inv - Gamma\,(a, b)$, then it has density $p(x) = \frac{b^a}{\Gamma(a)}\left(\frac{1}{x}\right)^{a+1}\exp\left(-\frac{b}{x}\right)$.

[1]The conditional posteriors under the improper prior $\sigma \sim \frac{1}{\sigma^2}d\sigma^2$ are similar.

A.1 Natural conjugate prior

In the first case, the prior on $\boldsymbol{\beta}$ is defined conditional on σ^2. The hierarchical structure is summarized as follows.

$$\boldsymbol{\beta}|\sigma^2 \sim N_p\left(\boldsymbol{\mu}_\beta, \sigma^2 \boldsymbol{V}_\beta\right), \tag{A.4}$$

The conditional posteriors are of the form

$$\begin{aligned} \boldsymbol{\beta} \mid \bullet &\sim N_p\left(\boldsymbol{V} \times \left[\boldsymbol{X}'\boldsymbol{y} + \boldsymbol{V}_\beta^{-1}\boldsymbol{\mu}_\beta\right], \sigma^2 \boldsymbol{V}\right), & \text{(A.5)} \\ \sigma^2 \mid \bullet &\sim Inv-Gamma\left(a + \frac{n+p}{2}, b + \frac{1}{2}\left[(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) + (\boldsymbol{\beta} - \boldsymbol{\mu}_\beta)\boldsymbol{V}_\beta^{-1}(\boldsymbol{\beta} - \boldsymbol{\mu}_\beta)\right]\right) & \text{(A.6)} \end{aligned}$$

where $\boldsymbol{V} = (\boldsymbol{X}'\boldsymbol{X} + \boldsymbol{V}_\beta^{-1})^{-1}$ and $\bullet$ denotes data and all the parameters except for the parameter that is being updated.

Derivation

The joint prior is

$$\begin{aligned} p(\boldsymbol{\beta}, \sigma^2) &= (2\pi)^{-p/2}|\sigma^2\boldsymbol{V}_\beta|^{-1/2} exp\left[-\frac{1}{2\sigma^2}(\boldsymbol{\beta} - \boldsymbol{\mu}_\beta)'\boldsymbol{V}_\beta^{-1}(\boldsymbol{\beta} - \boldsymbol{\mu}_\beta)\right]\frac{b^a}{\Gamma(a)}\left(\frac{1}{\sigma^2}\right)^{a+1} exp\left(-\frac{b}{\sigma^2}\right) \\ &\propto \left(\frac{1}{\sigma^2}\right)^{a+p/2+1} exp\left[-\frac{1}{\sigma^2}\left\{b + \frac{1}{2}(\boldsymbol{\beta} - \boldsymbol{\mu}_\beta)'\boldsymbol{V}_\beta^{-1}(\boldsymbol{\beta} - \boldsymbol{\mu}_\beta)\right\}\right], \end{aligned}$$

where the proportionality sign is with respect to the parameters $(\boldsymbol{\beta}, \sigma^2)$. The likelihood is

$$p(\boldsymbol{y}|\boldsymbol{\beta}, \sigma^2)(2\pi)^{-n/2}\left(\frac{1}{\sigma^2}\right)^{n/2} exp\left[-\frac{1}{2\sigma^2}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})\right].$$

The posterior is

$$\begin{aligned} p(\boldsymbol{\beta}, \sigma^2|\boldsymbol{y}) &\propto p(\boldsymbol{y}|\boldsymbol{\beta}, \sigma^2)p(\boldsymbol{\beta}, \sigma^2) \\ &\propto \left(\frac{1}{\sigma^2}\right)^{a+\frac{p+n}{2}+1} exp\left[-\frac{1}{\sigma^2}\left\{b + \frac{1}{2}\left[(\boldsymbol{\beta} - \boldsymbol{\mu}_\beta)'\boldsymbol{V}_\beta^{-1}(\boldsymbol{\beta} - \boldsymbol{\mu}_\beta) + (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})\right]\right\}\right]. \end{aligned}$$

From the right-hand-side above, it is easy to see that the conditional posterior $p(\sigma^2|\boldsymbol{\beta}, \boldsymbol{y})$ is of the form (A.6).

To see (A.5), note that

$$\begin{aligned} (\boldsymbol{\beta} - \boldsymbol{\mu}_\beta)'\boldsymbol{V}_\beta^{-1}(\boldsymbol{\beta} - \boldsymbol{\mu}_\beta) + (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) &= \boldsymbol{\beta}'\boldsymbol{V}_\beta^{-1}\boldsymbol{\beta} - 2\boldsymbol{\beta}'\boldsymbol{V}_\beta^{-1}\boldsymbol{\mu}_\beta + \boldsymbol{\mu}_\beta'\boldsymbol{V}_\beta^{-1}\boldsymbol{\mu}_\beta \\ &\quad + \boldsymbol{y}'\boldsymbol{y} - 2\boldsymbol{\beta}'\boldsymbol{X}'\boldsymbol{y} + \boldsymbol{\beta}'\boldsymbol{X}'\boldsymbol{X}\boldsymbol{\beta} \\ &= \boldsymbol{\beta}'\left[\boldsymbol{V}_\beta^{-1} + \boldsymbol{X}\boldsymbol{X}\right]\boldsymbol{\beta} - 2\boldsymbol{\beta}'\left[\boldsymbol{V}_\beta^{-1}\boldsymbol{\mu}_\beta + \boldsymbol{X}'\boldsymbol{y}\right] + \left[\boldsymbol{\mu}_\beta\boldsymbol{V}_\beta^{-1}\boldsymbol{\mu}_\beta + \boldsymbol{y}'\boldsymbol{y}\right] \\ &= (\boldsymbol{\beta} - \boldsymbol{\mu}_*)'\boldsymbol{V}_*^{-1}(\boldsymbol{\beta} - \boldsymbol{\mu}_*) - \boldsymbol{\mu}_*'\boldsymbol{V}_*^{-1}\boldsymbol{\mu}_* + \left[\boldsymbol{\mu}_\beta'\boldsymbol{V}_\beta^{-1}\boldsymbol{\mu}_\beta + \boldsymbol{y}'\boldsymbol{y}\right], \end{aligned}$$

where we used the identity

$$\boldsymbol{u}'\boldsymbol{A}\boldsymbol{u} - 2\boldsymbol{\alpha}'\boldsymbol{u} = (\boldsymbol{u} - \boldsymbol{A}^{-1}\boldsymbol{\alpha})'\boldsymbol{A}(\boldsymbol{u} - \boldsymbol{A}^{-1}\boldsymbol{\alpha}) - \boldsymbol{\alpha}'\boldsymbol{A}^{-1}\boldsymbol{\alpha},$$

in the last equality with $\boldsymbol{u} = \boldsymbol{\beta}, \boldsymbol{A} = \boldsymbol{V}_\beta^{-1} + \boldsymbol{X}\boldsymbol{X}$, and $\boldsymbol{\alpha} = \boldsymbol{V}_\beta^{-1}\boldsymbol{\mu}_\beta + \boldsymbol{X}'\boldsymbol{y}$ and defined

$$\begin{aligned}
\boldsymbol{\mu}_* &= \boldsymbol{A}^{-1}\boldsymbol{\alpha} = \left[\boldsymbol{V}_\beta^{-1} + \boldsymbol{X}\boldsymbol{X}\right]^{-1}\left[\boldsymbol{V}_\beta^{-1}\boldsymbol{\mu}_\beta + \boldsymbol{X}'\boldsymbol{y}\right], \\
\boldsymbol{V}_* &= \boldsymbol{A}^{-1} = \left[\boldsymbol{V}_\beta^{-1} + \boldsymbol{X}\boldsymbol{X}\right]^{-1}.
\end{aligned}$$

Hence the posterior is

$$p(\boldsymbol{\beta}, \sigma^2|\boldsymbol{y}) \propto \left(\frac{1}{\sigma^2}\right)^{a_*+1} exp\left[-\frac{1}{\sigma^2}\left\{b_* + \frac{1}{2}(\boldsymbol{\beta} - \boldsymbol{\mu}_*)'\boldsymbol{V}_*^{-1}(\boldsymbol{\beta} - \boldsymbol{\mu}_*)\right\}\right],$$

where $a_* = a + n/2 + p/2$ and $b_* = b + \frac{1}{2}\left[\boldsymbol{\mu}_\beta'\boldsymbol{V}_\beta^{-1}\boldsymbol{\mu}_\beta + \boldsymbol{y}'\boldsymbol{y} - \boldsymbol{\mu}_*'\boldsymbol{V}_*^{-1}\boldsymbol{\mu}_*\right]$. Therefore, the conditional posterior for $\boldsymbol{\beta}$ is of the form (A.5).

A.2 Independent prior

In this case, $\boldsymbol{\beta}$ and σ^2 are *a priori* independent.

$$\boldsymbol{\beta} \sim N_p\left(\boldsymbol{\mu}_\beta, \boldsymbol{V}_\beta\right), \tag{A.7}$$

The conditional posteriors are of the form

$$\begin{aligned}
\boldsymbol{\beta} \mid \bullet &\sim N_p\left(\boldsymbol{V} \times \left[\boldsymbol{X}'\boldsymbol{y}/\sigma^2 + \boldsymbol{V}_\beta^{-1}\boldsymbol{\mu}_\beta\right], \boldsymbol{V}\right), && \text{(A.8)} \\
\sigma^2 \mid \bullet &\sim Inv - Gamma\left(a + \frac{n}{2}, b + \frac{1}{2}\left(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\right)'\left(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\right)\right), && \text{(A.9)}
\end{aligned}$$

where $\boldsymbol{V} = (\boldsymbol{X}'\boldsymbol{X}/\sigma^2 + \boldsymbol{V}_\beta^{-1})^{-1}$.

Derivation

The joint prior is

$$\begin{aligned}
p(\boldsymbol{\beta}, \sigma^2) &= (2\pi)^{-p/2}|\boldsymbol{V}_\beta|^{-1/2} exp\left[-\frac{1}{2}(\boldsymbol{\beta} - \boldsymbol{\mu}_\beta)'\boldsymbol{V}_\beta^{-1}(\boldsymbol{\beta} - \boldsymbol{\mu}_\beta)\right]\frac{b^a}{\Gamma(a)}\left(\frac{1}{\sigma^2}\right)^{a+1} exp\left(-\frac{b}{\sigma^2}\right) \\
&\propto \left(\frac{1}{\sigma^2}\right)^{a+1} exp\left[-\frac{1}{\sigma^2}\left\{b + \frac{1}{2}(\boldsymbol{\beta} - \boldsymbol{\mu}_\beta)'(\boldsymbol{V}_\beta/\sigma^2)^{-1}(\boldsymbol{\beta} - \boldsymbol{\mu}_\beta)\right\}\right].
\end{aligned}$$

The posterior is

$$
\begin{aligned}
p(\boldsymbol{\beta}, \sigma^2|\boldsymbol{y}) &\propto p(\boldsymbol{y}|\boldsymbol{\beta}, \sigma^2)p(\boldsymbol{\beta}, \sigma^2)\\
&\propto \left(\frac{1}{\sigma^2}\right)^{a+\frac{n}{2}+1} exp\left[-\frac{1}{\sigma^2}\left\{b+\frac{1}{2}\left[(\boldsymbol{\beta}-\boldsymbol{\mu}_\beta)'(\boldsymbol{V}_\beta/\sigma^2)^{-1}(\boldsymbol{\beta}-\boldsymbol{\mu}_\beta)+(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})\right]\right\}\right].
\end{aligned}
$$

To see (B.106), note that

$$
p(\sigma^2|\boldsymbol{\beta}, \boldsymbol{y}) \propto \left(\frac{1}{\sigma^2}\right)^{a+\frac{n}{2}+1} exp\left[-\frac{1}{\sigma^2}\left\{b+\frac{1}{2}(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})\right\}\right].
$$

To see (B.105), note that

$$
\begin{aligned}
&(\boldsymbol{\beta}-\boldsymbol{\mu}_\beta)'(\boldsymbol{V}_\beta/\sigma^2)^{-1}(\boldsymbol{\beta}-\boldsymbol{\mu}_\beta)+(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})\\
&=\boldsymbol{\beta}'(\boldsymbol{V}_\beta/\sigma^2)^{-1}\boldsymbol{\beta}-2\boldsymbol{\beta}'(\boldsymbol{V}_\beta/\sigma^2)^{-1}\boldsymbol{\mu}_\beta+\boldsymbol{\mu}_\beta'(\boldsymbol{V}_\beta/\sigma^2)^{-1}\boldsymbol{\mu}_\beta+\boldsymbol{y}'\boldsymbol{y}-2\boldsymbol{\beta}'\boldsymbol{X}'\boldsymbol{y}+\boldsymbol{\beta}'\boldsymbol{X}'\boldsymbol{X}\boldsymbol{\beta}\\
&=\boldsymbol{\beta}'\left[(\boldsymbol{V}_\beta/\sigma^2)^{-1}+\boldsymbol{X}\boldsymbol{X}\right]\boldsymbol{\beta}-2\boldsymbol{\beta}'\left[(\boldsymbol{V}_\beta/\sigma^2)^{-1}\boldsymbol{\mu}_\beta+\boldsymbol{X}'\boldsymbol{y}\right]+\left[\boldsymbol{\mu}_\beta(\boldsymbol{V}_\beta/\sigma^2)^{-1}\boldsymbol{\mu}_\beta+\boldsymbol{y}'\boldsymbol{y}\right]\\
&=(\boldsymbol{\beta}-\boldsymbol{\mu}_*)'\boldsymbol{V}_*^{-1}(\boldsymbol{\beta}-\boldsymbol{\mu}_*)-\boldsymbol{\mu}_*'\boldsymbol{V}_*^{-1}\boldsymbol{\mu}_*+\left[\boldsymbol{\mu}_\beta'(\boldsymbol{V}_\beta/\sigma^2)^{-1}\boldsymbol{\mu}_\beta+\boldsymbol{y}'\boldsymbol{y}\right],
\end{aligned}
$$

where

$$
\begin{aligned}
\boldsymbol{\mu}_* &= \left[(\boldsymbol{V}_\beta/\sigma^2)^{-1}+\boldsymbol{X}\boldsymbol{X}\right]^{-1}\left[(\boldsymbol{V}_\beta/\sigma^2)^{-1}\boldsymbol{\mu}_\beta+\boldsymbol{X}'\boldsymbol{y}\right]=\left[\boldsymbol{V}_\beta^{-1}+\boldsymbol{X}\boldsymbol{X}/\sigma^2\right]^{-1}\left[\boldsymbol{V}_\beta^{-1}\boldsymbol{\mu}_\beta+\boldsymbol{X}'\boldsymbol{y}/\sigma^2\right],\\
\boldsymbol{V}_* &= \left[(\boldsymbol{V}_\beta/\sigma^2)^{-1}+\boldsymbol{X}\boldsymbol{X}\right]^{-1}=\sigma^2\left[\boldsymbol{V}_\beta^{-1}+\boldsymbol{X}\boldsymbol{X}/\sigma^2\right]^{-1}.
\end{aligned}
$$

Hence the posterior is

$$
p(\boldsymbol{\beta}, \sigma^2|\boldsymbol{y}) \propto \left(\frac{1}{\sigma^2}\right)^{a_*+1} exp\left[-\frac{1}{\sigma^2}\left\{b_*+\frac{1}{2}(\boldsymbol{\beta}-\boldsymbol{\mu}_*)'\boldsymbol{V}_*^{-1}(\boldsymbol{\beta}-\boldsymbol{\mu}_*)\right\}\right],
$$

where $a_* = a + n/2$ and $b_* = b + \frac{1}{2}\left[\boldsymbol{\mu}_\beta'(\boldsymbol{V}_\beta/\sigma^2)^{-1}\boldsymbol{\mu}_\beta+\boldsymbol{y}'\boldsymbol{y}-\boldsymbol{\mu}_*'\boldsymbol{V}_*^{-1}\boldsymbol{\mu}_*\right]$.

B MCMC inference in linear regression model with hierarchical priors

We use the simple linear regression model of the form

$$y_i = x_i\boldsymbol{\beta} + \varepsilon_i, \quad \varepsilon_i \sim N(0, \sigma^2), \quad i = 1, ..., n, \tag{B.1}$$

where $\boldsymbol{\beta}$ is a $p \times 1$ vector. We define $\boldsymbol{y} = (y_1, ..., y_n)'$, $\boldsymbol{X} = (x_1', ..., x_n')'$ and $\boldsymbol{\varepsilon} = (\varepsilon_1, ..., \varepsilon_n)'$, such that the stacked form of the regression model is

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{B.2}$$

where $\boldsymbol{\varepsilon} \sim N_n(\boldsymbol{0}_{n\times 1}, \sigma^2 \boldsymbol{I}_n)$.

B.1 Normal-Jeffreys

The normal-Jeffreys hierarchical prior takes the form

$$\begin{aligned} \boldsymbol{\beta} | \{\tau_j^2\}_{j=1}^p, \sigma^2 &\sim N_p(\boldsymbol{0}, \sigma^2 \boldsymbol{D}), && \text{(B.3)} \\ \tau_j^2 &\sim \frac{1}{\tau_i^2}, \quad \text{for } j = 1, ..., p, && \text{(B.4)} \\ \sigma^2 &\sim \frac{1}{\sigma^2}, && \text{(B.5)} \end{aligned}$$

where $\boldsymbol{D} = diag(\tau_1^2, ..., \tau_p^2)$.

The conditional posteriors are of the form

$$\begin{aligned} \boldsymbol{\beta} \mid \bullet &\sim N_p\left(\boldsymbol{V} \times \boldsymbol{X}'\boldsymbol{y}, \sigma^2 \boldsymbol{V}\right), && \text{(B.6)} \\ \tau_j^2 \mid \bullet &\sim Inv-Gamma\left(\frac{1}{2}, \frac{\beta_j^2}{2\sigma^2}\right), \text{ for } j = 1, ..., p, && \text{(B.7)} \\ \sigma^2 \mid \bullet &\sim Inv-Gamma\left(\frac{n+2}{2}, \frac{\Psi + \boldsymbol{\beta}'\boldsymbol{D}^{-1}\boldsymbol{\beta}}{2}\right), && \text{(B.8)} \end{aligned}$$

where $\boldsymbol{V} = \left(\boldsymbol{X}'\boldsymbol{X} + \boldsymbol{D}^{-1}\right)^{-1}$ and $\Psi = (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})$.

B.2 Student-t shrinkage

The normal-inverse gamma prior is the scale mixture of normals representation of the fat-tailed Student-t distribution. This hierarchical prior, which is also called "sparse Bayesian Learning" prior in signal processing, takes the form

$$\begin{aligned}
\boldsymbol{\beta}|\{\tau_j^2\}_{j=1}^p, \sigma^2 &\sim N_p(\mathbf{0}, \sigma^2 \boldsymbol{D}), && \text{(B.9)}\\
\tau_j^2 &\sim Inv-Gamma\,(\rho, \xi)\,, \quad \text{for } j = 1, ..., p, && \text{(B.10)}\\
\sigma^2 &\sim \frac{1}{\sigma^2}, && \text{(B.11)}
\end{aligned}$$

where $\boldsymbol{D} = diag(\tau_1^2, ..., \tau_p^2)$.

The conditional posteriors are of the form

$$\begin{aligned}
\boldsymbol{\beta} \mid \bullet &\sim N_p\left(\boldsymbol{V} \times \boldsymbol{X}'\boldsymbol{y}, \sigma^2 \boldsymbol{V}\right), && \text{(B.12)}\\
\tau_j^2 \mid \bullet &\sim Inv-Gamma\left(\rho + \frac{1}{2}, \xi + \frac{\beta_j^2}{2\sigma^2}\right), \text{ for } j = 1, ..., p, && \text{(B.13)}\\
\sigma^2 \mid \bullet &\sim Inv-Gamma\left(\frac{n+p}{2}, \frac{\Psi + \boldsymbol{\beta}'\boldsymbol{D}^{-1}\boldsymbol{\beta}}{2}\right), && \text{(B.14)}
\end{aligned}$$

where $\boldsymbol{V} = \left(\boldsymbol{X}'\boldsymbol{X} + \boldsymbol{D}^{-1}\right)^{-1}$ and $\Psi = (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})$.

B.3 Bayesian LASSO

As noted first by Tibshirani (1996), the LASSO estimator

$$\hat{\boldsymbol{\beta}} = \arg\min_{\beta}\,(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})'\,(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) + \lambda_1 \sum_{j=1}^p |\beta_j|, \tag{B.15}$$

is equivalent to the posterior mode under the Laplace prior

$$\boldsymbol{\beta}|\sigma \sim \prod_{j=1}^p \frac{\lambda}{2\sqrt{\sigma^2}} e^{-\lambda|\beta_j|/\sqrt{\sigma^2}}, \tag{B.16}$$

which can be written as the following normal-exponential mixture

$$\boldsymbol{\beta}|\sigma \sim \prod_{j=1}^p \int_0^\infty \frac{1}{\sqrt{2\pi\sigma^2 s_j}} e^{\left(-\frac{\beta_j^2}{2\sigma^2 s_j}\right)} \frac{\lambda^2}{2} e^{-\frac{\lambda}{2s_j}} ds_j. \tag{B.17}$$

This is the mixture prior analyzed by Park and Casella (2008), which is by far the most popular form for the Bayesian LASSO. Hans (2009) provides an alternative formulation by means of the orthant-truncated normal distribution. A third possible formulation of the Laplace prior is the scale mixture of uniform distributions proposed by Mallick and Yi (2014). A related representation is that of a mixture of truncated normal distributions (see Alhamzawi and Ali, 2020).

B.3.1 Park and Casella (2008) algorithm

The Park and Casella (2008) Laplace prior takes the form

$$
\begin{aligned}
\boldsymbol{\beta}|\{\tau_j^2\}_{j=1}^p, \sigma^2 &\sim N_p(\mathbf{0}, \sigma^2 \boldsymbol{D}), && \text{(B.18)}\\
\tau_j^2|\lambda^2 &\sim Exponential\left(\frac{\lambda^2}{2}\right), \quad \text{for } j = 1, ..., p, && \text{(B.19)}\\
\lambda^2 &\sim Gamma(r, \delta), && \text{(B.20)}\\
\sigma^2 &\sim \frac{1}{\sigma^2}, && \text{(B.21)}
\end{aligned}
$$

where $\boldsymbol{D} = diag(\tau_1^2, ..., \tau_p^2)$.

The conditional posteriors are of the form

$$
\begin{aligned}
\boldsymbol{\beta} \mid \bullet &\sim N_p\left(\boldsymbol{V} \times \boldsymbol{X}'\boldsymbol{y}, \sigma^2 \boldsymbol{V}\right), && \text{(B.22)}\\
\frac{1}{\tau_j^2} \mid \bullet &\sim IG\left(\sqrt{\frac{\lambda^2\sigma^2}{\beta_j^2}}, \lambda^2\right), \quad \text{for } j = 1, ..., p, && \text{(B.23)}\\
\lambda^2 \mid \bullet &\sim Gamma\left(r + p, \frac{\sum_{j=1}^p \tau_j^2}{2} + \delta\right), && \text{(B.24)}\\
\sigma^2 \mid \bullet &\sim Inv - Gamma\left(\frac{n+p}{2}, \frac{\Psi + \boldsymbol{\beta}'\boldsymbol{D}^{-1}\boldsymbol{\beta}}{2}\right), && \text{(B.25)}
\end{aligned}
$$

where $\boldsymbol{V} = \left(\boldsymbol{X}'\boldsymbol{X} + \boldsymbol{D}^{-1}\right)^{-1}$, $\boldsymbol{D} = diag(\tau_1^2, ..., \tau_p^2)$, and $\Psi = (\boldsymbol{y} - \boldsymbol{X\beta})'(\boldsymbol{y} - \boldsymbol{X\beta})$.

B.3.2 Hans (2009) algorithm

Before we proceed we need to define the notion of the normal orthant distribution, following Hans (2009). Let $\mathfrak{Z} = \{-1, 1\}^p$ represent the set of all 2^p possible vectors of length p whose elements are ± 1. For any realization $z \in \mathfrak{Z}$ define the orthant $\mathbb{O}_z \subset \mathbb{R}^p$. If $\boldsymbol{\beta} \in \mathbb{O}_z$, then $\beta_j \geq 0$ if $z = 1$ and $\beta_j < 0$ if $z = -1$. Then $\boldsymbol{\beta}$ follows the normal-orthant distribution with mean m and covariance S, which is of the form

$$
\boldsymbol{\beta} \sim N^{[z]}(\boldsymbol{m}, \boldsymbol{S}) \equiv \Phi(\boldsymbol{m}, \boldsymbol{S})\, N_p(\boldsymbol{m}, \boldsymbol{S})\, I(\in \mathbb{O}_z). \tag{B.26}
$$

The Hans (2009) prior takes the form

$$
\begin{aligned}
\boldsymbol{\beta}|\lambda, \sigma &\sim \left(\frac{\lambda}{2\sqrt{\sigma^2}}\right)^p \exp\left(-\lambda \sum_{j=1}^p |\beta_j|/\sqrt{\sigma^2}\right), && \text{(B.27)}\\
\lambda &\sim Gamma(r, \delta), && \text{(B.28)}\\
\sigma^2 &\sim \frac{1}{\sigma^2}, && \text{(B.29)}
\end{aligned}
$$

where the prior for $\boldsymbol{\beta}$ is an equivalent representation of the Laplace density in Equation (B.16).

The conditional posteriors are of the form

$$
\begin{aligned}
\beta_j|\beta_{-j},\lambda,\sigma^2,\boldsymbol{y} &\sim \phi_j N^{[+]}\left(\mu_j^+,\omega_{jj}^{-1}\right)+(1-\phi_j)N^{[-]}\left(\mu_j^-,\omega_{jj}^{-1}\right), && \text{(B.30)}\\
\lambda|\boldsymbol{y} &\sim Gamma\left(p+r,\frac{\sum_{j=1}^p|\beta|}{\sqrt{\sigma^2}}+\delta\right), && \text{(B.31)}\\
\sigma|\boldsymbol{\beta},\boldsymbol{y} &\propto (\sigma^2)^{-(\frac{n+p}{2}+1)}\exp\left(\frac{\Psi}{2\sigma^2}-\frac{\lambda\sum_{j=1}^p|\beta|}{\sqrt{\sigma^2}}\right), && \text{(B.32)}
\end{aligned}
$$

where:

- $N^{[-]}$ and $N^{[+]}$ correspond to the $N^{[z]}$ distribution for $z=-1$ and $z=1$, respectively,
- $\mu_j^+=\widehat{\beta}_j^{OLS}+\left\{\sum_{i=1,i\neq j}^p\left(\widehat{\beta}_i^{OLS}-\beta_i\right)(\omega_{ij}/\omega_{jj})\right\}+\left(-\frac{\lambda}{\sqrt{\sigma^2\omega_{jj}}}\right)$,
- ω_{ij} is the ij element of the matrix $\Omega=\Sigma^{-1}=\left(\sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}\right)^{-1}$,
- $\phi_j=\frac{\Phi\left(\frac{\mu_j^+}{\sqrt{\omega_{jj}}}\right)/N\left(0|\mu_j^+,\omega_{jj}^{-1}\right)}{\Phi\left(\frac{\mu_j^+}{\sqrt{\omega_{jj}}}\right)/N\left(0|\mu_j^+,\omega_{jj}^{-1}\right)+\Phi\left(-\frac{\mu_j^-}{\sqrt{\omega_{jj}}}\right)/N\left(0|\mu_j^-,\omega_{jj}^{-1}\right)}$, and
- $\Psi=(\boldsymbol{y}-\boldsymbol{X\beta})'(\boldsymbol{y}-\boldsymbol{X\beta})$.

Notice that the conditional posterior of σ^2 does not belong to a standard form we can sample from. Hans (2009) proposes a simple accept/reject algorithm in order to obtain samples from σ^2. The posterior of σ^2 simplifies to the standard Inv-Gamma form, if we consider a Laplace prior for $\boldsymbol{\beta}$ that is independent of σ, i.e. the prior $\boldsymbol{\beta}|\lambda\sim\left(\frac{\lambda}{2}\right)^p\exp\left(-\lambda\sum_{j=1}^p|\beta_j|\right)$. Finally, notice that sampling of β_j conditional on β_{-j} (i.e. all elements of the vector $\boldsymbol{\beta}$ other than the j-th) becomes very inefficient when predictors $\boldsymbol{X}$ are correlated. Hans (2009) proposes to use an alternative Gibbs sampler algorithm that orthogonalizes predictors, which comes at the cost of increased computational complexity (due to the rotations of data and parameters involved when orthogonalizing the predictors).

B.3.3 Mallick and Yi (2014) algorithm

The Mallick and Yi (2014) Laplace prior takes the form

$$
\begin{aligned}
\boldsymbol{\beta}|\{\tau_j^2\}_{j=1}^p,\sigma^2 &\sim \prod_{j=1}^p Uniform\left(-\sqrt{\sigma^2}\tau_j,\sqrt{\sigma^2}\tau_j\right), && \text{(B.33)}\\
\tau_j|\lambda &\sim Gamma\,(2,\lambda), \quad \text{for } j=1,...,p, && \text{(B.34)}\\
\lambda &\sim Gamma\,(r,\delta), && \text{(B.35)}\\
\sigma^2 &\sim \frac{1}{\sigma^2}. && \text{(B.36)}
\end{aligned}
$$

The conditional posteriors are of the form

$$\boldsymbol{\beta} \mid \bullet \sim N_p\left(\widehat{\boldsymbol{\beta}}_{OLS}, \sigma^2\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\right)\prod_{j=1}^{p} I\left(|\beta_j| < \sqrt{\sigma^2}\tau_j\right), \tag{B.37}$$

$$\tau_j \mid \bullet \sim Exponential\,(\lambda)\, I\left(\tau_j > \frac{|\beta_j|}{\sqrt{\sigma^2}}\right), \quad \text{for } j = 1, ..., p, \tag{B.38}$$

$$\lambda \sim Gamma\left(r + 2p, \delta + \sum_{j=1}^{p}|\beta_j|\right), \tag{B.39}$$

$$\left.\frac{1}{\sigma^2}\right| \bullet \sim Gamma\left(\frac{n-1+p}{2}, \frac{\Psi}{2}\right) I\left(\sigma^2 < \frac{1}{\max_j\left(\beta_j^2/\tau_j^2\right)}\right), \tag{B.40}$$

where $I(\bullet)$ is the indicator function and $\Psi = (\boldsymbol{y} - \boldsymbol{X\beta})'(\boldsymbol{y} - \boldsymbol{X\beta})$. Because of the truncation of the conditional posteriors, Mallick and Yi (2014) suggest the following sampling steps:

1. Generate first τ_j from the truncated Exponential distribution in Equation (B.38): Sample a $\tau_j^\star \sim Exponential(\lambda)$, and then set $\tau_j = \tau_j^\star + \frac{|\beta_j|}{\sqrt{\sigma^2}}$.
2. Sample β from the truncated normal distribution in Equation (B.37).
3. Sample λ from the Gamma distribution in Equation (B.39).
4. Generate σ^2 from the right truncated Gamma distribution in Equation (B.40): Use simple accept/reject sampling, that is, sample $\frac{1}{\sigma^{2\star}}$ from $Gamma\left(\frac{n-1+p}{2}, \frac{\Psi}{2}\right)$ until the condition $\sigma^{2\star} < \frac{1}{\max_j\left(\beta_j^2/\tau_j^2\right)}$ is met. If it is, set $\sigma = \frac{1}{\sigma^{2\star}}$.

B.4 Bayesian Adaptive LASSO

Fan and Li (2001) showed that the LASSO can perform automatic variable selection but it produces biased estimates for the larger coefficients. Thus, they argued that the oracle properties do not hold for the LASSO. To obtain the oracle property, Zou (2006) introduced the adaptive LASSO estimator as

$$\hat{\boldsymbol{\beta}} = \arg\min_{\beta}\,(\boldsymbol{y} - \boldsymbol{X\beta})'\,(\boldsymbol{y} - \boldsymbol{X\beta}) + \sum_{j=1}^{p}\lambda_j|\beta_j|, \tag{B.41}$$

with the weight vector $\lambda_j = \lambda|\hat{\beta}_j|^{-r}$ for $j = 1, ..., p$ where $\hat{\beta}_j$ is a $\sqrt{n}$ consistent estimator such as the least squares estimator. The adaptive LASSO enjoys the oracle property and it leads to a near-minimax-optimal estimator.

Alhamzawi and Ali (2018) proposed Bayesian adaptive LASSO. They show that a Laplace density can be written as a exponential scale mixture of truncated normal distribution i.e.

$$\frac{\lambda_j}{2\sqrt{\sigma^2}}e^{-\lambda|\beta_j|/\sqrt{\sigma^2}} = \int_0^\infty \int_{u_j > \sqrt{\lambda_j^2/\sigma^2}|\beta_j|} \frac{1}{\sqrt{2\pi\sigma^2 s_j}} e^{\left(-\frac{\beta_j^2}{2\sigma^2 s_j}\right)} e^{\left(-\frac{u_j}{2}\right)} \frac{\lambda_j^2}{8} e^{\left(-\frac{\lambda_j^2 s_j}{8}\right)} du_j ds_j$$
$$= \int_0^\infty \int_{u_j > \sqrt{\lambda_j^2/\sigma^2}|\beta_j|} N(\beta_j; 0, \sigma^2 s_j) Exponential\left(u_j; \frac{1}{2}\right) Exponential\left(s_j; \frac{\lambda_j^2}{8}\right) du_j ds_j.$$

Based on this fact, they propose the following conditional prior for Bayesian adaptive LASSO

$$\beta_j|\sigma^2, \lambda_j^2, s_j \sim N(0, \sigma^2 s_j) I\left(|\beta_j| < \sqrt{\sigma^2/\lambda_j^2} u_j\right) \quad j = 1, ..., p, \tag{B.42}$$
$$s_j|\lambda_j^2 \sim Exponential\left(\frac{\lambda_j^2}{8}\right) \quad j = 1, ..., p, \tag{B.43}$$
$$u_j \sim Exponential\left(\frac{1}{2}\right) \quad j = 1, ..., p, \tag{B.44}$$
$$\lambda_j^2 \sim Gamma(a, b) \quad j = 1, ..., p, \tag{B.45}$$
$$\sigma^2 \sim \sigma^{-2} d\sigma^2. \tag{B.46}$$

The conditional posteriors are of the form

$$\boldsymbol{\beta} \mid \bullet \sim N_p\left(\boldsymbol{V} \times \boldsymbol{X}\boldsymbol{y}, \sigma^2 \boldsymbol{V}\right) \prod_{j=1}^p I\left(|\beta_j| < \sqrt{\sigma^2/\lambda_j^2} u_j\right), \tag{B.47}$$
$$\sigma^2 \mid \bullet \sim Inv-Gamma\left(a^*, b^*\right) I\left(\sigma^2 > max_j \left\{\frac{\lambda_j^2 \beta_j^2}{u_j^2}\right\}\right), \tag{B.48}$$
$$s_j^{-1} \mid \bullet \sim IG\left(\sqrt{\frac{\sigma^2\lambda_j^2}{4\beta_j^2}}, \frac{\lambda_j^2}{4}\right) \quad j = 1, ..., p, \tag{B.49}$$
$$p(u_j \mid \bullet) \propto Exponential\left(\frac{1}{2}\right) I\left(u_j > \sqrt{\frac{\lambda_j^2}{\sigma^2}}|\beta_j|\right) \quad j = 1, ..., p, \tag{B.50}$$
$$p(\lambda_j^2 \mid \bullet) \propto Gamma\left(a + p, b + \frac{s_j}{8}\right) I\left(\lambda_j^2 < \frac{\sigma^2 u_j^2}{\beta_j^2}\right) \quad j = 1, ..., p, \tag{B.51}$$

where $\boldsymbol{V} = (\boldsymbol{X}'\boldsymbol{X} + \boldsymbol{S}^{-1})^{-1}$ with $\boldsymbol{S} = diag(s_1, ..., s_p)$, $a^* = \frac{n-1+p}{2}$, and $b^* = \frac{1}{2}\left[(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) + \sum_{j=1}^p \frac{\beta_j^2}{s_j}\right]$.

B.5 Bayesian Fused LASSO

In some applications, there might be a meaningful order among the covariates (e.g. time). The original LASSO ignores such ordering. To compensate the ordering limitations of the LASSO, the fused LASSO was introduced. It penalizes the L_1-norm of both the coefficients and their differences:

$$\hat{\boldsymbol{\beta}} = \arg\min_{\beta} \left(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\right)' \left(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\right) + \lambda_1 \sum_{j=1}^{p} |\beta_j| + \lambda_2 \sum_{j=2}^{p} |\beta_j - \beta_{j-1}|. \tag{B.52}$$

Kyung et al. (2010) proposed Bayesian group LASSO with the following conditional prior.

$$p\left(\boldsymbol{\beta}|\sigma^2\right) \propto \exp\left(-\frac{\lambda_1}{\sigma}\sum_{j=1}^{p}|\beta_j| - \frac{\lambda_2}{\sigma}\sum_{j=2}^{p}|\beta_j - \beta_{j-1}|\right), \tag{B.53}$$

$$\sigma^2 \sim \sigma^{-2} d\sigma^2. \tag{B.54}$$

where the conditional prior is equivalent to the following gamma mixture of normals prior.

$$\boldsymbol{\beta}|\{\tau_j^2\}_{j=1}^{p}, \{\omega_j^2\}_{j=1}^{p-1}, \sigma^2 \sim N_p(\mathbf{0}, \sigma^2\boldsymbol{\Sigma}_\beta), \tag{B.55}$$

$$\tau_j^2 \sim \frac{\lambda_1^2}{2} e^{-\lambda_1 \tau_j^2/2} d\tau_j^2 \text{ for } , j = 1, ..., p, \tag{B.56}$$

$$\omega_j^2 \sim \frac{\lambda_2^2}{2} e^{-\lambda_2 \omega_j^2/2} d\omega_j^2 \text{ for } , j = 1, ..., p-1, \tag{B.57}$$

where $\tau_1^2, ..., \tau_p^2$ and $\omega_1^2, ..., \omega_{p-1}^2$ are mutually independent, and $\boldsymbol{\Sigma}_\beta$ is a tridiagonal matrix with

$$\text{Main diagonal} = \left\{\frac{1}{\tau_i^2} + \frac{1}{\omega_{i-1}^2} + \frac{1}{\omega_i^2}, i = 1, ..., p\right\}, \tag{B.58}$$

$$\text{Off diagonals} = \left\{-\frac{1}{\omega_i^2}, i = 1, ..., p-1\right\}, \tag{B.59}$$

where $1/\omega_0^2 = 1/\omega_p^2 = 0$.

The conditional posteriors are of the form

$$\boldsymbol{\beta} \mid \bullet \sim N_p\left(\boldsymbol{V} \times \boldsymbol{X}\boldsymbol{y}, \sigma^2\boldsymbol{V}\right), \tag{B.60}$$

$$1/\tau_j^2 \mid \bullet \sim IG\left(\left(\frac{\lambda_1^2\sigma^2}{\beta_j^2}\right)^{1/2}, \lambda_1^2\right) 1(1/\tau_j^2 > 0), j = 1, ..., p, \tag{B.61}$$

$$1/\omega_j^2 \mid \bullet \sim IG\left(\left(\frac{\lambda_2^2\sigma^2}{(\beta_{j+1} - \beta_j)^2}\right)^{1/2}, \lambda_2^2\right) 1(1/\omega_j^2 > 0), j = 1, ..., p-1, \tag{B.62}$$

$$\sigma^2 \mid \bullet \sim Inv - Gamma\left(a^*, b^*\right), \tag{B.63}$$

where $\boldsymbol{V} = (\boldsymbol{X}'\boldsymbol{X} + \boldsymbol{\Sigma}_\beta^{-1})^{-1}$, $a^* = \frac{n-1+p}{2}$, and $b^* = \frac{1}{2}\left[\left(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\right)'\left(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\right) + \boldsymbol{\beta}\boldsymbol{\Sigma}_\beta^{-1}\boldsymbol{\beta}\right]$.

When we place $Gamma(r, \delta)$ priors on λ_1 and λ_2, the conditional posteriors are

$$\lambda_1^2 \mid \bullet \sim Gamma\left(p + r, \frac{1}{2}\sum_{j=1}^{p} \tau_j^2 + \delta\right), \tag{B.64}$$

$$\lambda_2^2 \mid \bullet \sim Gamma\left(p - 1 + r, \frac{1}{2}\sum_{j=1}^{p-1} \omega_j^2 + \delta\right). \tag{B.65}$$

B.6 Bayesian Group LASSO

If there is a group of covariates among which the pairwise correlation is high (e.g. dummy variables), the LASSO tends to select only individual variables from the group. The group LASSO takes such group structure into account:

$$\hat{\boldsymbol{\beta}} = \arg\min_{\beta} \left(\boldsymbol{y} - \sum_{k=1}^{K} \boldsymbol{X}_k \boldsymbol{\beta}_k\right)' \left(\boldsymbol{y} - \sum_{k=1}^{K} \boldsymbol{X}_k \boldsymbol{\beta}_k\right) + \lambda \sum_{j=k}^{K} ||\boldsymbol{\beta}_k||_{G_k}, \tag{B.66}$$

where K is the number of groups, $\boldsymbol{\beta}_k$ is the vector of β's in the group k, and $||\boldsymbol{\beta}||_{G_k} = \left(\boldsymbol{\beta}' \boldsymbol{G}_k \boldsymbol{\beta}\right)^{1/2}$ with positive definite matrices $\boldsymbol{G}_k$'s. Typically, $\boldsymbol{G}_k = \boldsymbol{I}_{m_k}$ where m_k is the number of variables in group k.

Kyung et al. (2010) proposed Bayesian group LASSO with the following conditional prior.

$$p\left(\boldsymbol{\beta}|\sigma^2\right) \propto \exp\left(-\frac{\lambda}{\sigma}\sum_{j=k}^{K} ||\boldsymbol{\beta}_k||_{G_k}\right), \tag{B.67}$$

$$\sigma^2 \sim \sigma^{-2} d\sigma^2, \tag{B.68}$$

where the conditional prior is equivalent to the following gamma mixture of normals prior.

$$\boldsymbol{\beta}_{G_k}|\tau_k^2, \sigma^2 \sim N_{m_k}(\boldsymbol{0}, \sigma^2 \tau_k^2 \boldsymbol{I}_{m_k}), \tag{B.69}$$

$$\tau_k^2|\sigma^2 \sim Gamma\left(\frac{m_k + 1}{2}, \frac{\lambda^2}{2}\right) \text{ for } k = 1, ..., K. \tag{B.70}$$

The conditional posteriors are of the form

$$\boldsymbol{\beta}_{G_k}|\boldsymbol{\beta}_{-G_k}, \sigma^2, \tau_1^2, ..., \tau_K^2, \lambda, \boldsymbol{y} \sim N_p\left(\boldsymbol{V}_k \times \boldsymbol{X}_k'\left(\boldsymbol{y} - \frac{1}{2}\sum_{k' \neq k} \boldsymbol{X}_{k'} \boldsymbol{\beta}_{G_{k'}}\right), \sigma^2 \boldsymbol{V}_k\right), \tag{B.71}$$

$$1/\tau_k^2 \mid \bullet \sim IG\left(\left(\frac{\lambda^2 \sigma^2}{||\boldsymbol{\beta}_{G_k}||^2}\right)^{1/2}, \lambda^2\right) 1(1/\tau_k^2 > 0), \text{ for } k = 1, ..., K, \tag{B.72}$$

$$\sigma^2 \mid \bullet \sim Inv - Gamma\left(\frac{n - 1 + p}{2}, \frac{1}{2}||\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}||^2 + \frac{1}{2}\sum_{j=k}^{K} \frac{1}{\tau_k^2}||\boldsymbol{\beta}_{G_k}||^2\right), \tag{B.73}$$

where $\boldsymbol{\beta}_{-G_k} = \left(\boldsymbol{\beta}_{G_1}, ..., \boldsymbol{\beta}_{G_{k-1}}, \boldsymbol{\beta}_{G_{k+1}}, ..., \boldsymbol{\beta}_{G_K}\right)$ and $\boldsymbol{V}_k = (\boldsymbol{X}_k' \boldsymbol{X}_k + \tau_k^{-2} \boldsymbol{I}_{m_k})^{-1}$.

When we place a $Gamma(r, \delta)$ prior on λ, the posterior conditional on λ is

$$\lambda^2 \mid \bullet \sim Gamma\left(\frac{p+K}{2} + r, \frac{1}{2}\sum_{k=1}^{K} \tau_k^2 + \delta\right). \tag{B.74}$$

B.7 Bayesian Elastic Net

Here again we have various alternative algorithms. We look into the algorithm of Kyung et al. (2010) and the algorithm of Li and Lin (2010), but we can also mention here the algorithm of Hans (2011) that is based on the algorithm of Hans (2009) we examined for the Bayesian LASSO.

The elastic net combines the benefits of ridge regression (ℓ_2 penalization) and the LASSO (ℓ_1 penalization). The Bayesian prior that provides the solution to the elastic net estimation problem is of the form

$$\boldsymbol{\beta}|\sigma^2 \sim \exp\left\{-\frac{1}{2\sigma^2}\left(\lambda_1 \sum_{j=1}^{p} |\beta_j| + \lambda_2 \sum_{j=1}^{p} \beta_j^2\right)\right\}. \tag{B.75}$$

Li and Lin (2010) start from this prior and derive a mixture approximation and a Gibbs sampler that has the minor disadvantage that requires an accept-reject algorithm for obtaining samples from the conditional posterior of σ^2 (similar to the sampler of Hans (2009) for the LASSO). The formulation of the elastic net prior in Kyung et al. (2010) is slightly different to the one above, but they manage to derive a slightly different mixture representation and a slightly more straightforward Gibbs sampler.

B.7.1 Li and Lin (2010) algorithm

The Li and Lin (2010) prior takes the form

$$\boldsymbol{\beta}|\{\tau_j^2\}_{j=1}^p, \lambda_2, \sigma^2 \sim N_p\left(\mathbf{0}, \frac{\sigma^2}{\lambda_2}\boldsymbol{D}_\tau\right), \tag{B.76}$$

$$\tau_j^2|\sigma^2 \sim TG_{(1,\infty)}\left(\frac{1}{2}, \frac{8\lambda_2\sigma^2}{\lambda_1^2}\right), \quad \text{for } j = 1, ..., p, \tag{B.77}$$

$$\sigma^2 \sim \frac{1}{\sigma^2}, \tag{B.78}$$

where $TG_{(1,\infty)}$ is the Gamma distribution truncated to the support $(1, \infty)$, and $\boldsymbol{D}_\tau = diag\left(\frac{\tau_1^2 - 1}{\tau_1^2}, ..., \frac{\tau_p^2 - 1}{\tau_p^2}\right)$. Notice that λ_1, λ_2 do not have their own prior distributions, that is, they are not considered to be random variables in this algorithm. Instead, Li and Lin (2010) suggest to use empirical Bayes methods to calibrate these two parameters.

The conditional posteriors are of the form

$$\boldsymbol{\beta} \mid \bullet \sim N_p\left(\boldsymbol{V} \times \boldsymbol{X}'\boldsymbol{y}, \sigma^2\boldsymbol{V}\right), \tag{B.79}$$

$$\tau_j^2 - 1 \mid \bullet \sim GIG\left(\frac{1}{2}, \frac{\lambda_1}{4\lambda_2\sigma^2}, \frac{\lambda_2\beta_j^2}{\sigma^2}\right), \quad \text{for } j = 1, ..., p, \tag{B.80}$$

$$p(\sigma^2 \mid \bullet\,) \propto \left(\frac{1}{\sigma^2}\right)^{\frac{n}{2}+p+1} \left\{\Gamma_U\left(\frac{1}{2}, \frac{\lambda_1^2}{8\lambda_2\sigma^2}\right)\right\} \tag{B.81}$$

$$\exp\left[-\frac{1}{2\sigma^2}\left\{\Psi + \lambda_2\sum_{j=1}^{p}\frac{\tau_j^2}{\tau_j^2-1}\beta_j^2 + \frac{\lambda_1^2}{4\lambda_2}\sum_{j=1}^{p}\tau_j^2\right\}\right], \tag{B.82}$$

where $\boldsymbol{V} = \left(\boldsymbol{X}'\boldsymbol{X} + \lambda_2\boldsymbol{D}_\tau^{-1}\right)^{-1}$, $\boldsymbol{D}_\tau^{-1} = diag\left(\frac{\tau_1^2}{\tau_1^2-1}, ..., \frac{\tau_p^2}{\tau_p^2-1}\right)$, and $\Psi = (\boldsymbol{y} - \boldsymbol{X\beta})'(\boldsymbol{y} - \boldsymbol{X\beta})$. $\Gamma_U(\bullet)$ is the upper incomplete gamma function. *GIG* is the three parameter Generalized Inverse Gaussian distribution.[2] The conditional posterior distribution of σ^2 does not belong to a known density we can sample from. Therefore, for each Monte Carlo iteration we sample the first two parameters directly from their conditional posteriors but we sample σ^2 indirectly from its conditional posterior using an accept/reject step.

B.7.2 Kyung et al. (2010) algorithm

The Kyung et al. (2010) prior takes the form

$$\boldsymbol{\beta}|\{\tau_j^2\}_{j=1}^p, \lambda_2, \sigma^2 \sim N_p(\boldsymbol{0}, \sigma^2\boldsymbol{D}_{\tau,\lambda_2}), \tag{B.83}$$

$$\tau_j^2|\lambda^2 \sim Exponential\left(\frac{\lambda_1^2}{2}\right), \quad \text{for } j = 1, ..., p, \tag{B.84}$$

$$\lambda_1^2 \sim Gamma(r_1, \delta_1), \tag{B.85}$$

$$\lambda_2 \sim Gamma(r_2, \delta_2), \tag{B.86}$$

$$\sigma^2 \sim \frac{1}{\sigma^2}, \tag{B.87}$$

where $\boldsymbol{D}_{\tau,\lambda_2} = diag((\tau_1^{-2} + \lambda_2)^{-1}, ..., (\tau_p^{-2} + \lambda_2))^{-1})$.

[2] CRAN has several implementations in R of random number generators that allow sampling from the *GIG* distribution. As of the time of writing of this document, Mathworks does not provide a built-in function for MATLAB that allows to generate from this distribution, but external contributions do exist.

The conditional posteriors are of the form

$$
\begin{aligned}
\boldsymbol{\beta} \mid \bullet &\sim N_p\left(\boldsymbol{V} \times \boldsymbol{X}'\boldsymbol{y}, \sigma^2 \boldsymbol{V}\right), && \text{(B.88)}\\
\frac{1}{\tau_j^2} \mid \bullet &\sim IG\left(\sqrt{\frac{\lambda_1^2\sigma^2}{\beta_j^2}}, \lambda_1^2\right) I(1/\tau_j^2 > 0), \quad \text{for } j = 1, ..., p, && \text{(B.89)}\\
\lambda_1^2 \mid \bullet &\sim Gamma\left(r_1 + p, \frac{\sum_{j=1}^p \tau_j^2}{2} + \delta_1\right), && \text{(B.90)}\\
\lambda_2 \mid \bullet &\sim Gamma\left(r_2 + \frac{p}{2}, \frac{\sum_{j=1}^p \beta_j^2}{2\sigma^2} + \delta_2\right), && \text{(B.91)}\\
\sigma^2 \mid \bullet &\sim Inv - Gamma\left(\frac{n - 1 + p}{2}, \frac{\Psi + \boldsymbol{\beta}'\boldsymbol{D}_{\tau,\lambda_2}^{-1}\boldsymbol{\beta}}{2}\right), && \text{(B.92)}
\end{aligned}
$$

where $\boldsymbol{V} = \left(\boldsymbol{X}'\boldsymbol{X} + \boldsymbol{D}_{\tau,\lambda_2}^{-1}\right)^{-1}$, $\boldsymbol{D}_{\tau,\lambda_2}^{-1} = diag((\tau_1^{-2} + \lambda_2), ..., (\tau_p^{-2} + \lambda_2)))$, and $\Psi = (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})$.

B.8 Generalized Double Pareto

Armagan et al. (2013) propose the following Generalized Double Pareto (GDP) prior on $\boldsymbol{\beta}$

$$
\boldsymbol{\beta}|\sigma \sim \prod_{j=1}^{p} \frac{1}{2\sigma\delta/r}\left(1 + \frac{1}{r}\frac{|\beta_j|}{\sigma\delta/r}\right)^{-(r+1)}. \tag{B.93}
$$

This distribution can be represented using the familiar, from the Bayesian LASSO, normal-exponential-gamma mixture, see subsubsection B.3.1. The only difference is that, while the exponential component has the same rate parameter for all $j = 1, ..., p$, in the representation of the GDP mixture this parameter is adaptive.

The Generalized Double Pareto prior takes the form

$$
\begin{aligned}
\boldsymbol{\beta}|\{\tau_j\}_{j=1}^p, \sigma^2 &\sim N_p\left(\boldsymbol{0}, \sigma^2\boldsymbol{D}\right), && \text{(B.94)}\\
\tau_j^2|\lambda_j &\sim Exponential\left(\frac{\lambda_j^2}{2}\right), \quad \text{for } j = 1, ..., p, && \text{(B.95)}\\
\lambda_j &\sim Gamma(r, \delta), \quad \text{for } j = 1, ..., p, && \text{(B.96)}\\
\sigma^2 &\sim \frac{1}{\sigma^2}, && \text{(B.97)}
\end{aligned}
$$

where $\boldsymbol{D} = diag(\tau_1^2, ..., \tau_p^2)$.

The conditional posteriors are of the form

$$\begin{aligned}
\boldsymbol{\beta} \mid \bullet &\sim N_p\left(\boldsymbol{V} \times \boldsymbol{X}'\boldsymbol{y}, \sigma^2 \boldsymbol{V}\right), && \text{(B.98)} \\
\frac{1}{\tau_j^2} \,\Bigg|\, \bullet &\sim IG\left(\sqrt{\frac{\lambda_j^2 \sigma^2}{\beta_j^2}}, \lambda^2\right), \quad \text{for } j = 1, ..., p, && \text{(B.99)} \\
\lambda_j^2 \mid \bullet &\sim Gamma\left(r + 1, \sqrt{\frac{\beta_j^2}{\sigma^2}} + \delta\right), && \text{(B.100)} \\
\sigma^2 \mid \bullet &\sim Inv - Gamma\left(\frac{n - 1 + p}{2}, \frac{\Psi + \boldsymbol{\beta}'\boldsymbol{D}^{-1}\boldsymbol{\beta}}{2}\right), && \text{(B.101)}
\end{aligned}$$

where $\boldsymbol{V} = \left(\boldsymbol{X}'\boldsymbol{X} + \boldsymbol{D}^{-1}\right)^{-1}$, $\boldsymbol{D} = diag(\tau_1^2, ..., \tau_p^2)$, and $\Psi = (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})$.

B.9 Normal-Gamma

The normal-gamma prior of Griffin and Brown (2010) takes the form

$$\begin{aligned}
\boldsymbol{\beta} | \{\tau_j\}_{j=1}^p &\sim N(\boldsymbol{0}, \boldsymbol{D}), && \text{(B.102)} \\
\tau | \lambda, \gamma^2 &\sim Gamma\left(\lambda, \frac{1}{2\gamma^2}\right), && \text{(B.103)} \\
\sigma^2 &\sim \frac{1}{\sigma^2}, && \text{(B.104)}
\end{aligned}$$

where $\boldsymbol{D} = diag(\tau_1^2, ..., \tau_p^2)$.

The conditional posteriors $\boldsymbol{\beta}$ and σ^2 are of the usual form

$$\begin{aligned}
\boldsymbol{\beta} \mid \bullet &\sim N_p\left(\boldsymbol{V} \times \boldsymbol{X}'\boldsymbol{y}/\sigma^2, \boldsymbol{V}\right), && \text{(B.105)} \\
\sigma^2 \mid \bullet &\sim Inv - Gamma\left(\frac{n}{2}, \frac{1}{2}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})\right), && \text{(B.106)}
\end{aligned}$$

where $\boldsymbol{V} = (\boldsymbol{X}'\boldsymbol{X}/\sigma^2 + \boldsymbol{D}^{-1})^{-1}$.

The parameters $\tau_1, ..., \tau_p$ can be updated in a block since the full conditional distributions of $\tau_1, ..., \tau_p$ are independent. The full conditional distribution of τ_j follows a generalized inverse Gaussian distribution

$$\tau_j \mid \bullet \sim GIG(\lambda - 0.5, 1/\gamma^2, \beta_j^2), \quad j = 1, ..., p. \qquad \text{(B.107)}$$

B.10 Multiplicative Gamma process

Suppose we have the factor model

$$\begin{aligned}
\boldsymbol{X}_t &= \boldsymbol{\Lambda}\boldsymbol{F}_t + \boldsymbol{\epsilon}_t, && \text{(B.108)} \\
\boldsymbol{\epsilon}_t &\sim N_n(\boldsymbol{0}, \boldsymbol{\Sigma}), t = 1, ..., T, && \text{(B.109)}
\end{aligned}$$

where $\boldsymbol{X}_t$ is a $n \times 1$ vector, $\boldsymbol{\Lambda}$ is a $n \times k$ matrix of factor loadings, $\boldsymbol{F}_t$ is a $k \times 1$ vector, and $\boldsymbol{\Sigma} = diag(\Sigma_{11}, ..., \Sigma_{nn})$.

Bhattacharya and Dunson (2011) proposed a novel multiplicative gamma process prior on the factor loadings that shrinks more aggressively columns of $\boldsymbol{\Lambda}$ that correspond to a higher number of factors. They call their approach the sparse infinite factor model, as it allows to specify a maximum number of factors and the prior is able to determine zero and non-zero loadings, as well as the number of factors. The gamma process prior for the loadings matrix is of the following "global-local shrinkage" form

$$\begin{aligned}
\Lambda_{ij}|\phi_{ij}, \tau_j &\sim N(0, \phi_{ij}^{-1}\tau_j^{-1}), && \text{(B.110)}\\
\phi_{ij} &\sim Gamma(v/2, v/2), && \text{(B.111)}\\
\tau_j &= \prod_{l=1}^{j} \delta_l, \quad j = 1, ..., k, && \text{(B.112)}\\
\delta_1 &\sim Gamma(a_1, 1), && \text{(B.113)}\\
\delta_l &\sim Gamma(a_2, 1), \quad l \geq 2, && \text{(B.114)}\\
\Sigma_{ii} &\sim Inv - Gamma(a_0, b_0), i = 1, ..., n. && \text{(B.115)}
\end{aligned}$$

While the local shrinkage parameter is the same for each element of $\boldsymbol{\Lambda}$, the global shrinkage parameter τ_j is shrinking more aggressively as the index j increases, where $j = 1, ..., k$ indexes the number of factors. This is because τ_j is a j-dimensional product of gamma distributions.

Let $\boldsymbol{X}^{(i)}$ be the ith column of the $n \times k$ matrix $\boldsymbol{X}$ $\boldsymbol{\Lambda}_i'$ be the ith row of $\boldsymbol{\Lambda}$. The conditional posterior distributions are

$$\begin{aligned}
\boldsymbol{\Lambda}_i \mid \bullet &\sim N_k\left(\boldsymbol{V}_{L_i}\left(\boldsymbol{F}'\Sigma_{ii}^{-1}\boldsymbol{X}^{(i)}\right), \boldsymbol{V}_{L_i}\right) \quad i = 1, ..., n, && \text{(B.116)}\\
\boldsymbol{F}_t \mid \bullet &\sim N_k\left(\boldsymbol{V}_F\left(\boldsymbol{\Lambda}'\boldsymbol{\Sigma}^{-1}\boldsymbol{X}_t\right), \boldsymbol{V}_F\right) \quad t = 1, ..., T, && \text{(B.117)}\\
\phi_{ij} \mid \bullet &\sim Gamma\left(\frac{v+1}{2}, \frac{v + \tau_j \Lambda_{ij}^2}{2}\right) \quad i = 1, ..., n, j = 1, ..., k, && \text{(B.118)}\\
\tau_\ell^{(j)} &= \prod_{t=1, t\neq j}^{\ell} \delta_t \quad j = 1, ..., k, && \text{(B.119)}\\
\delta_1 \mid \bullet &\sim Gamma\left(a_1 + 0.5nk, 1 + 0.5\sum_{\ell=1}^{k}\tau_\ell^{(1)}\sum_{i=1}^{n}\phi_{i\ell}\Lambda_{i\ell}^2\right), && \text{(B.120)}\\
\delta_j \mid \bullet &\sim Gamma\left(a_2 + 0.5n(k-j+1), 1 + 0.5\sum_{\ell=j}^{k}\tau_\ell^{(j)}\sum_{i=1}^{n}\phi_{i\ell}\Lambda_{i\ell}^2\right), \quad j \geq 2, && \text{(B.121)}\\
\Sigma_{ii} \mid \bullet &\sim Inv - Gamma\left(a_0 + n/2, b_0 + SSE_i\right), \quad i = 1, ..., n, && \text{(B.122)}
\end{aligned}$$

where $\boldsymbol{V}_{L_i} = (\boldsymbol{D}_i^{-1} + \Sigma_{ii}^{-1}\boldsymbol{F}'\boldsymbol{F})^{-1}$, $\boldsymbol{D}_i^{-1} = diag(\phi_{i1}\tau_1, ..., \phi_{ik}\tau_k)$, $\boldsymbol{V}_F = (I + \boldsymbol{\Lambda}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda})^{-1}$, and $SSE_i = (\boldsymbol{X}^{(i)} - \boldsymbol{F}\boldsymbol{\Lambda}_i)'(\boldsymbol{X}^{(i)} - \boldsymbol{F}\boldsymbol{\Lambda}_i)$.

B.11 Dirichlet-Laplace

The Dirichlet-Laplace prior of Bhattacharya et al. (2015), as analyzed in Zhang and Bondell (2018), takes the form

$$\begin{aligned}
\boldsymbol{\beta}|\{\tau_j\}_{j=1}^p, \{\psi_j\}_{j=1}^p, \lambda, \sigma^2 &\sim N_p\left(\mathbf{0}, \sigma^2 \boldsymbol{D}_{\lambda,\tau,\psi}\right), && \text{(B.123)}\\
\tau_j^2 &\sim Exponential(1/2), \quad \text{for } j=1,...,p, && \text{(B.124)}\\
\psi_j &\sim Dirichlet(\alpha), \quad \text{for } j=1,...,p, && \text{(B.125)}\\
\lambda &\sim Gamma(n\alpha, 1/2), && \text{(B.126)}\\
\sigma^2 &\sim \frac{1}{\sigma^2}, && \text{(B.127)}
\end{aligned}$$

where $\boldsymbol{D}_{\lambda,\tau,\psi} = diag(\lambda^2\tau_1^2\psi_1^2, ..., \lambda^2\tau_p^2\psi_p^2)$.

The conditional posteriors are of the form

$$\begin{aligned}
\boldsymbol{\beta} \mid \bullet &\sim N_p\left(\boldsymbol{V} \times \boldsymbol{X}'\boldsymbol{y}, \sigma^2\boldsymbol{V}\right), && \text{(B.128)}\\
\frac{1}{\tau_j^2} \,\Big|\, \bullet &\sim IG\left(\sqrt{\frac{\lambda^2\psi_j^2\sigma^2}{\beta_j^2}}, 1\right), \quad \text{for } j=1,...,p, && \text{(B.129)}\\
\lambda \mid \bullet &\sim GIG\left(2\frac{\sum_{j=1}^p |\beta_j|}{\psi_j\sigma}, 1, p(\alpha-1)\right), && \text{(B.130)}\\
T_j \mid \bullet &\sim GIG\left(2\sqrt{\frac{\beta_j^2}{\sigma^2}}, 1, \alpha-1\right), \quad \text{for } j=1,...,p, && \text{(B.131)}\\
\psi_j &= \frac{T_j}{\sum_{j=1}^p T_j}, \quad \text{for } j=1,...,p, && \text{(B.132)}\\
\sigma^2 \mid \bullet &\sim Inv-Gamma\left(\frac{n+p}{2}, \frac{\Psi + \boldsymbol{\beta}'\boldsymbol{D}_{\tau,\lambda,\psi}^{-1}\boldsymbol{\beta}}{2}\right), && \text{(B.133)}
\end{aligned}$$

where $\boldsymbol{V} = \left(\boldsymbol{X}'\boldsymbol{X} + \boldsymbol{D}_{\tau,\lambda,\psi}^{-1}\right)^{-1}$, $\boldsymbol{D}_{\tau,\lambda,\psi} = diag(\lambda^2\tau_1^2\psi_1^2, ..., \lambda^2\tau_p^2\psi_p^2)$, and $\Psi = (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})$.

B.12 Horseshoe

The horseshoe prior on a regression coefficient $\boldsymbol{\beta}$ takes the following hierarchical form

$$\begin{aligned}
\boldsymbol{\beta}|\{\lambda_j\}_{j=1}^p, \tau &\sim N\left(\mathbf{0}, \sigma^2\tau^2\boldsymbol{\Lambda}\right), && \text{(B.134)}\\
\lambda_j|\tau &\sim C^+(0,1), \quad \text{for } j=1,...,p, && \text{(B.135)}\\
\tau &\sim C^+(0,1), && \text{(B.136)}
\end{aligned}$$

where $\boldsymbol{\Lambda} = diag(\lambda_1^2, ..., \lambda_p^2)$, and $C^+(0, \alpha)$ is the half-Cauchy distribution on the positive reals with scale parameter α. That is, λ_j has conditional prior density

$$\lambda_j|\tau = \frac{2}{\pi\tau\left(1 + (\lambda_j/\tau)^2\right)}. \tag{B.137}$$

B.12.1 Makalic and Schmidt (2016) algorithm

Makalic and Schmidt (2016) note that the half-Cauchy distribution can be written as a mixture of inverse-Gamma distributions. In particular, if

$$x^2|z \sim Inv-Gamma(1/2, 1/z), \qquad z \sim Inv-Gamma(1/2, 1/\alpha^2), \tag{B.138}$$

then $x \sim C^+(0, \alpha)$. Therefore, the Makalic and Schmidt (2016) prior takes the form

$$\begin{aligned}
\boldsymbol{\beta}|\{\lambda_j\}_{j=1}^p, \tau, \sigma^2 &\sim N\left(\mathbf{0}, \sigma^2\tau^2\boldsymbol{\Lambda}\right), && \text{(B.139)}\\
\lambda_j^2|v_j &\sim Inv-Gamma(1/2, 1/v_j), \quad \text{for } j = 1, ..., p, && \text{(B.140)}\\
v_j &\sim Inv-Gamma(1/2, 1), \quad \text{for } j = 1, ..., p, && \text{(B.141)}\\
\tau^2|\xi &\sim Inv-Gamma(1/2, 1/\xi), && \text{(B.142)}\\
\xi &\sim Inv-Gamma(1/2, 1), && \text{(B.143)}\\
\sigma^2 &\sim \frac{1}{\sigma^2}, && \text{(B.144)}
\end{aligned}$$

where $\boldsymbol{\Lambda} = diag(\lambda_1^2, ..., \lambda_p^2)$.

The conditional posteriors are of the form

$$\begin{aligned}
\boldsymbol{\beta} \mid \bullet &\sim N_p\left(\boldsymbol{V} \times \boldsymbol{X}'\boldsymbol{y}, \sigma^2\boldsymbol{V}\right), && \text{(B.145)}\\
\lambda_j^2 \mid \bullet &\sim Inv-Gamma\left(1, \frac{1}{v_j} + \frac{\beta_j^2}{2\tau^2\sigma^2}\right), \quad \text{for } j = 1, ..., p, && \text{(B.146)}\\
v_j \mid \bullet &\sim Inv-Gamma\left(1, 1 + \frac{1}{\lambda_j^2}\right), \quad \text{for } j = 1, ..., p, && \text{(B.147)}\\
\tau^2 \mid \bullet &\sim Inv-Gamma\left(\frac{p+1}{2}, \frac{1}{\xi} + \frac{1}{2\sigma^2}\sum_{j=1}^p \frac{\beta_j^2}{\lambda_j^2}\right), && \text{(B.148)}\\
\xi \mid \bullet &\sim Inv-Gamma\left(1, 1 + \frac{1}{\tau^2}\right), && \text{(B.149)}\\
\sigma^2 \mid \bullet &\sim Inv-Gamma\left(\frac{n+p}{2}, \frac{\Psi + \boldsymbol{\beta}'\boldsymbol{D}^{-1}\boldsymbol{\beta}}{2}\right), && \text{(B.150)}
\end{aligned}$$

where $\boldsymbol{V} = \left(\boldsymbol{X}'\boldsymbol{X} + \boldsymbol{D}^{-1}\right)^{-1}$, $\boldsymbol{D} = diag(\tau^2\lambda_1^2, ..., \tau^2\lambda_p^2) = \tau^2\boldsymbol{\Lambda}$, and $\Psi = (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})$.

B.12.2 Slice sampler

Under the hosrshoe prior, we have

$$\begin{aligned}
\boldsymbol{\beta}|\{\lambda_j\}_{j=1}^p, \tau, \sigma^2 &\sim N\left(\mathbf{0}, \sigma^2\tau^2 diag(\lambda_1^2, ..., \lambda_p^2)\right), && \text{(B.151)}\\
\lambda_j &\sim C^+(0,1), \quad \text{for } j = 1, ..., p, && \text{(B.152)}\\
\tau &\sim C^+(0,1), && \text{(B.153)}\\
\sigma^2 &\sim \frac{1}{\sigma^2}. && \text{(B.154)}
\end{aligned}$$

The conditional posteriors are of the form

$$\begin{aligned}
\boldsymbol{\beta} \mid \bullet &\sim N_p\left(\boldsymbol{V} \times \boldsymbol{X}'\boldsymbol{y}, \sigma^2\boldsymbol{V}\right), && \text{(B.155)}\\
\sigma^2 \mid \bullet &\sim Inv-Gamma\left(\frac{n}{2} + \frac{p}{2}, \frac{1}{2}\left[(\boldsymbol{y} - \boldsymbol{X\beta})'(\boldsymbol{y} - \boldsymbol{X\beta}) + \boldsymbol{\beta}'\boldsymbol{D}^{-1}\boldsymbol{\beta}\right]\right), && \text{(B.156)}\\
p(\lambda_j \mid \bullet\,) &\propto \left(\frac{1}{\lambda_j^2}\right)^{1/2} \exp\left[-\frac{\beta_j}{2\sigma^2\tau^2}\frac{1}{\lambda_j^2}\right]\frac{1}{1+\lambda_j^2}d\lambda_j, \quad \text{for } j = 1, ..., p, && \text{(B.157)}\\
p(\tau \mid \bullet\,) &\propto \left(\frac{1}{\tau^2}\right)^{p/2} \exp\left[-\frac{1}{2\sigma^2}\sum_{j=1}^{p}\frac{\beta_j^2}{\lambda_j^2}\frac{1}{\tau^2}\right]\frac{1}{1+\tau^2}d\tau, && \text{(B.158)}
\end{aligned}$$

where $\boldsymbol{V} = (\boldsymbol{X}'\boldsymbol{X} + \boldsymbol{D}^{-1})^{-1}$ with $\boldsymbol{D} = diag(\tau^2\lambda_1^2, ..., \tau^2\lambda_p^2)$.

With a change of variable $\eta_j = \frac{1}{\lambda_j^2}$, it can be seen that

$$\eta_j|\boldsymbol{\beta}, \tau^2, \sigma^2 \propto \exp\left(-\mu_j\eta_j\right)\frac{1}{1+\eta_j}d\eta_j, \qquad \text{(B.159)}$$

where $\mu_j = \frac{\beta_j}{2\sigma^2\tau^2}$. The λ_j's are updated with a slice sampler:

1. Sample $u_j \sim Unif\left[0, \frac{1}{1+\eta_j}\right]$.
2. Sample $\eta_j|u_j \sim \exp\left(-\mu_j\eta_j\right) I\left(\eta_j < \frac{1-u_j}{u_j}\right)$ [3].
3. Set $\lambda_j = \eta_j^{-1/2}$.

Similarly, with a change of variable $\eta = \frac{1}{\tau^2}$, we have

$$\eta|\boldsymbol{\beta}, \{\lambda_j\}_{j=1}^p, \sigma^2, \boldsymbol{y} \propto \eta^{\frac{p+1}{2}-1}\exp\left(-\mu\eta\right)\frac{1}{1+\eta}d\eta, \qquad \text{(B.160)}$$

where $\mu = \frac{1}{2\sigma^2}\sum_{j=1}^p \frac{\beta_j^2}{\lambda_j^2}$. The τ can be updated in a similar fashion:

1. Sample $u \sim Unif\left[0, \frac{1}{1+\eta}\right]$.

[3] This is an exponential density with parameter μ_j^{-1} truncated on $\left(0, \frac{1-u_j}{u_j}\right)$.

2. Sample $\eta|u \sim \eta^{\frac{p+1}{2}-1} \exp\left(-\mu\eta\right) I\left(\eta < \frac{1-u}{u}\right)$ [4].

3. Set $\tau = \eta^{-1/2}$.

B.12.3 Johndrow et al. (2020) algorithm

The horseshoe prior in Johndrow et al. (2020) has its original form

$$\begin{aligned}
\boldsymbol{\beta}|\{\lambda_j\}_{j=1}^{p}, \tau, \sigma^2 &\sim N\left(\mathbf{0}, \sigma^2\tau^2\boldsymbol{\Lambda}\right), & \text{(B.161)}\\
\lambda_j|\tau &\sim C^+(0,1), \quad \text{for } j = 1, ..., p, & \text{(B.162)}\\
\tau &\sim C^+(0,1), & \text{(B.163)}\\
\sigma^2 &\sim \frac{1}{\sigma^2}. & \text{(B.164)}
\end{aligned}$$

In order to improve the mixing of the global parameter τ^2, they propose a blocked Metropolis-within-Gibbs sampler where $(\boldsymbol{\beta}, \tau^2, \sigma)$ are updated in one block. The conditional posterior of τ^2 given $\boldsymbol{\lambda} = (\lambda_1^2, ..., \lambda_p^2)$ is

$$p(\tau^2|\boldsymbol{\lambda}, \boldsymbol{y}) \propto |\boldsymbol{M}|^{-1/2}\left(\frac{1}{2}\boldsymbol{y}'\boldsymbol{M}^{-1}\boldsymbol{y}\right)^{-\frac{n}{2}} \times \frac{\tau}{1+\frac{1}{\tau^2}}, \tag{B.165}$$

where $\boldsymbol{M} = I_n + \boldsymbol{X}\boldsymbol{D}\boldsymbol{X}'$. Their Metropolis-within-Gibbs algorithm is as follows

$$\begin{aligned}
p(\lambda_j^2\ |\tau^2, \boldsymbol{\beta}, \sigma^2) &\propto \frac{\lambda_j^2}{\lambda_j^2+1}\exp\left(-\frac{\beta_j}{2\sigma^2\tau^2}\frac{1}{\lambda_j^2}\right), \quad \text{for } j = 1, ..., p, & \text{(B.166)}\\
log(\tau^{-2*}) &\sim N\left(log(\tau^{-2}), s\right), \text{ accept } \tau^{2*} \text{ w.p. } \frac{p(\tau^{2*}|\boldsymbol{\lambda}, \boldsymbol{y})\tau^{2*}}{p(\tau^2|\boldsymbol{\lambda}, \boldsymbol{y})\tau^2}, & \text{(B.167)}\\
\sigma^2\ |\tau^2, \boldsymbol{\lambda}^2 &\sim Inv-Gamma\left(\frac{n}{2}, \frac{\boldsymbol{y}'\boldsymbol{M}^{-1}\boldsymbol{y}}{2}\right), & \text{(B.168)}\\
\boldsymbol{\beta}\ |\tau^2, \boldsymbol{\lambda}^2, \sigma^2 &\sim N_p\left(\boldsymbol{V}\times\boldsymbol{X}'\boldsymbol{y}, \sigma^2\boldsymbol{V}\right), & \text{(B.169)}
\end{aligned}$$

where $\boldsymbol{V} = \left(\boldsymbol{X}'\boldsymbol{X} + \boldsymbol{D}^{-1}\right)^{-1}$ and $\boldsymbol{D} = diag(\tau^2\lambda_1^2, ..., \tau^2\lambda_p^2)$.

The λ_j^2 can be updated via a slice sampler:

1. Sample $u \sim Unif\left[0, \frac{\lambda_j^2}{\lambda_j^2+1}\right]$.

2. Sample $\lambda_j^2|u \sim \exp\left(-\frac{\beta_j}{2\sigma^2\tau^2}\frac{1}{\lambda_j^2}\right) I\left(\frac{1-u}{u} > \frac{1}{\lambda_j^2}\right)$.

[4] This is a gamma density with the shape parameter $\frac{p+1}{2}$ and the scale parameter μ^{-1} truncated on $\left(0, \frac{1-u}{u}\right)$.

B.13 Generalized Beta mixtures of Gaussians

In their paper, Armagan et al. (2011) motivate the use of a three-parameter beta (TPB) distribution as a flexible class of shrinkage priors. The TPB distribution takes the form

$$p(x|a,b,\varphi) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \varphi^b x^{b-1} (1-x)^{a-1} \left[1 + (\varphi - 1)x\right]^{-(a+b)}, \tag{B.170}$$

for $0 < x < 1$, $a, b, \varphi > 0$. Proposition 1 in Armagan et al. (2011) shows that this distribution can either be written as normal-inverted beta mixture, or a normal-gamma-gamma mixture. The second choice gives a very straightforward Gibbs sampler scheme so we present an algorithm based on the normal-gamma-gamma representation of TPB.

The generalized beta mixtures of Gaussians prior takes the form

$$\begin{aligned}
\boldsymbol{\beta}|\{\tau_j^2\}_{j=1}^p, \sigma^2 &\sim N_p\left(0, \sigma^2 \boldsymbol{D}_\tau\right), && \text{(B.171)}\\
\tau_j^2|\lambda_j &\sim Gamma\left(a, \lambda_j\right), \quad \text{for } j = 1, ..., p, && \text{(B.172)}\\
\lambda_j|\varphi &\sim Gamma(b, \varphi), \quad \text{for } j = 1, ..., p, && \text{(B.173)}\\
\varphi &\sim Gamma\left(\frac{1}{2}, \omega\right), && \text{(B.174)}\\
\omega &\sim Gamma\left(\frac{1}{2}, 1\right), && \text{(B.175)}\\
\sigma^2 &\sim \frac{1}{\sigma^2}, && \text{(B.176)}
\end{aligned}$$

where $\boldsymbol{D}_\tau = diag(\tau_1^2, ..., \tau_p^2)$. Note that setting $a = b = 1/2$ we can obtain the horseshoe prior of Carvalho et al. (2010). For other choices we can recover popular cases of shrinkage priors.

The conditional posteriors are of the form

$$\begin{aligned}
\boldsymbol{\beta} \mid \bullet &\sim N_p\left(\boldsymbol{V} \times \boldsymbol{X}'\boldsymbol{y}, \sigma^2 \boldsymbol{V}\right), && \text{(B.177)}\\
\tau_j^2 \mid \bullet &\sim GIG\left(a - \frac{1}{2}, 2\lambda_j, \frac{\beta_j^2}{\sigma^2}\right), \quad \text{for } j = 1, ..., p, && \text{(B.178)}\\
\lambda_j \mid \bullet &\sim Gamma(a + b, \tau_j^2 + \varphi), \quad \text{for } j = 1, ..., p, && \text{(B.179)}\\
\varphi \mid \bullet &\sim Gamma\left(pb + \frac{1}{2}, \sum_{j=1}^p \lambda_j + \omega\right), && \text{(B.180)}\\
\omega \mid \bullet &\sim Gamma(1, \varphi + 1), && \text{(B.181)}\\
\sigma^2 \mid \bullet &\sim Gamma\left(\frac{n+p}{2}, \frac{\Psi + \boldsymbol{\beta}'\boldsymbol{D}_\tau^{-1}\boldsymbol{\beta}}{2}\right), && \text{(B.182)}
\end{aligned}$$

where $\boldsymbol{V} = \left(\boldsymbol{X}'\boldsymbol{X} + \boldsymbol{D}_\tau^{-1}\right)^{-1}$, $\boldsymbol{D}_\tau = diag(\tau_1^2, ..., \tau_p^2)$, and $\Psi = (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})$.

B.14 Spike and slab

B.14.1 Kuo and Mallick (1998) algorithm

Kuo and Mallick (1998) consider the following modified formulation of the regression problem.

$$\boldsymbol{y}|\boldsymbol{\beta},\boldsymbol{\gamma},\sigma^2 \quad\sim\quad N_p\left(\boldsymbol{X}\boldsymbol{\theta},\sigma^2\boldsymbol{I}\right), \tag{B.183}$$

where $\boldsymbol{X}=(\boldsymbol{x}_1,\ldots,\boldsymbol{x}_p)$ and $\boldsymbol{\theta}=(\beta_1\gamma_1,\ldots,\beta_p\gamma_p)'$ with $\gamma_j=1$ if $\boldsymbol{x}_j$ is included in the model and 0 otherwise.

The authors consider the following independent prior.

$$\begin{aligned}
\boldsymbol{\beta} &\sim N_p\left(\boldsymbol{0},\boldsymbol{D}\right), && \text{(B.184)}\\
\gamma_j &\sim Bernoulli(p_j),\ \text{for } j=1,\ldots,p, && \text{(B.185)}\\
\sigma^2 &\sim Inv-Gamma\left(a,b\right). && \text{(B.186)}
\end{aligned}$$

With $\boldsymbol{X}^*=(\gamma_1\boldsymbol{x}_1,...,\gamma_p\boldsymbol{x}_p)$, the conditional posteriors can be written as follows.

$$\begin{aligned}
\boldsymbol{\beta}\mid\bullet &\sim N_p\left(\boldsymbol{V}\times\boldsymbol{X}^{*'}\boldsymbol{y}/\sigma^2,\boldsymbol{V}\right), && \text{(B.187)}\\
\sigma^2\mid\bullet &\sim Inv-Gamma\left(a+\frac{n}{2},b+\frac{1}{2}\left(\boldsymbol{y}-\boldsymbol{X}^*\boldsymbol{\beta}\right)'\left(\boldsymbol{y}-\boldsymbol{X}^*\boldsymbol{\beta}\right)\right), && \text{(B.188)}\\
\gamma_j\mid\bullet &\sim Bernoulli\left(\frac{c_j}{c_j+d_j}\right), && \text{(B.189)}
\end{aligned}$$

where $\boldsymbol{\gamma}_{-j}=(\gamma_1,\ldots,\gamma_{j-1},\gamma_{j+1},\ldots,\gamma_p)$ and $\boldsymbol{V}=\left(\boldsymbol{X}^{*'}\boldsymbol{X}^*/\sigma^2+\boldsymbol{D}^{-1}\right)^{-1}$ and

$$\begin{aligned}
c_j &= p_j\exp\left[-\frac{1}{2\sigma^2}\left(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\theta}_j^*\right)'\left(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\theta}_j^*\right)\right], && \text{(B.190)}\\
d_j &= (1-p_j)\exp\left[-\frac{1}{2\sigma^2}\left(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\theta}_j^{**}\right)'\left(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\theta}_j^{**}\right)\right], && \text{(B.191)}
\end{aligned}$$

where $\boldsymbol{\theta}_j^*$ is $\boldsymbol{\theta}$ with the j-component replaced by β_j and $\boldsymbol{\theta}_j^{**}$ is $\boldsymbol{\theta}$ with the j-component replaced by 0. Note that the conditional posterior of γ_j depends on $\boldsymbol{\gamma}_{-j}$. In order to facilitate the mixing, it is preferred to update γ_j for $j=1,\ldots,p$ in random order.

Note that although the formulation above holds for a generic prior variance $\boldsymbol{V}_\beta$, but an important special case is when it is a diagonal matrix $\boldsymbol{V}_\beta=diag\left(\tau_1^2,\ldots,\tau_p^2\right)$. This is equivalent to assume a spike and slab prior on θ_j, which is a mixture of a point mass at 0 with probability $1-p_j$ and a normal density $N\left(\mu_{\beta,j},\tau_j^2\right)$ with probability p_j.

B.15 Stochastic search variable selection

Consider the following stochastic search variable selection prior with fixed values of the prior variances.

$$\begin{aligned}
\beta_j|\sigma^2, \gamma_j = 0 &\sim N\left(0, \sigma^2\tau_{0j}^2\right), && \text{(B.192)}\\
\beta_j|\sigma^2, \gamma_j = 1 &\sim N\left(0, \sigma^2\tau_{1j}^2\right), && \text{(B.193)}\\
P(\gamma_j = 1) &= \theta \text{ for } j = 1, \ldots, p, && \text{(B.194)}\\
\theta &\sim Beta(c, d), && \text{(B.195)}\\
\sigma^2 &\sim Inv - Gamma(a, b). && \text{(B.196)}
\end{aligned}$$

George and McCulloch (1993) use non-conjugate prior in (B.192) and (B.193).

(B.192) and (B.193) can be equivalently written as

$$\boldsymbol{\beta}|\sigma^2, \boldsymbol{\gamma}, \{\tau_{0j}^2, \tau_{1j}^2\}_{j=1}^{p} \sim N_p\left(\mathbf{0}, \sigma^2\boldsymbol{D}\right), \tag{B.197}$$

where $\boldsymbol{D}$ is a diagonal matrix with diagonal elements with $\{(1-\gamma_j)\tau_{0j}^2 + \gamma_j\tau_{1j}^2\}_{j=1}^{p}$

The conditional posteriors are of the form

$$\begin{aligned}
\boldsymbol{\beta} \mid \bullet &\sim N_p\left(\boldsymbol{V} \times \boldsymbol{X}'\boldsymbol{y}, \sigma^2\boldsymbol{V}\right), \text{ where } \boldsymbol{V} = (\boldsymbol{X}'\boldsymbol{X} + \boldsymbol{D}^{-1})^{-1}, && \text{(B.198)}\\
\sigma^2 \mid \bullet &\sim Inv - Gamma\left(a + \frac{n}{2} + \frac{p}{2}, b + \frac{1}{2}\left[(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) + \boldsymbol{\beta}'\boldsymbol{D}^{-1}\boldsymbol{\beta}\right]\right) && \text{(B.199)}\\
\gamma_j| \bullet &\sim Bernoulli\left(\frac{\phi\left(\beta_j|0, \sigma^2\tau_{1j}^2\right)\theta}{\phi\left(\beta_j|0, \sigma^2\tau_{1j}^2\right)\theta + \phi\left(\beta_j|0, \sigma^2\tau_{0j}^2\right)(1-\theta)}\right), \text{ for } j = 1, ..., p, && \text{(B.200)}\\
\theta| \bullet &\sim Beta\left(c + \sum_{j=1}^{p}\gamma_j, d + \sum_{j=1}^{p}(1-\gamma_j)\right), \text{ for } j = 1, ..., p, && \text{(B.201)}
\end{aligned}$$

where $\phi(x|m, v)$ is the normal density with mean m and variance v.

Narisetty et al. (2018) propose to fix the value of the prior variance parameters as $\tau_{0j}^2 = \frac{\hat{\sigma}^2}{10n}$ and $\tau_{1j}^2 = \hat{\sigma}^2 \max\left(\frac{p^{2.1}}{100n}, \log(n)\right)$ where $\hat{\sigma}^2$ is the sample variance of y_i. The prior inclusion probability θ is chosen so that $Pr\left(\sum_{j=1}^{p}\gamma_j > K\right) = 0.1$ for $K = \max(10, \log(n))$.

B.16 Spike and slab LASSO

Consider the generic SSVS prior (B.192)-(B.196). Instead of fixing the prior variances τ_{0j} and τ_{1j}, one could place priors on them. A hierarchical Bayes version of the spike and slab LASSO prior in Ročková and George (2014) and Bai et al. (2021)[5] would correspond to placing two

[5]They propose an EM algorithm for estimation.

separate Laplace densities on the components i.e.

$$\tau_{0j}^2|\lambda_0^2 \sim Exponential\left(\frac{\lambda_0^2}{2}\right), \quad \text{for } j = 1, ..., p, \tag{B.202}$$

$$\tau_{1j}^2|\lambda_1^2 \sim Exponential\left(\frac{\lambda_1^2}{2}\right), \quad \text{for } j = 1, ..., p, \tag{B.203}$$

with $\lambda_0 \gg \lambda_1$ so that the density for $N(0, \sigma^2\tau_{0j}^2)$ is the "spike" and $N(0, \sigma^2\tau_{1j}^2)$ is the "slab".

The prior variances are updated according to

$$1/\tau_{0j}^2| \bullet \sim IG\left(\sqrt{\lambda_0^2\sigma^2/\beta_j^2}, \lambda_0^2\right), \quad \text{for } j = 1, ..., p, \tag{B.204}$$

$$1/\tau_{1j}^2| \bullet \sim IG\left(\sqrt{\lambda_1^2\sigma^2/\beta_j^2}, \lambda_1^2\right), \quad \text{for } j = 1, ..., p. \tag{B.205}$$

B.17 Semiparametric spike and slab

Dunson et al. (2008) allows for simultaneous selection of important predictors and soft clustering of predictors having similar impact on the variable of interest. This prior is a generalization of the typical "spike and slab" priors used for Bayesian variable selection and model averaging in the statistics literature. The coefficient $\boldsymbol{\beta}$ admit a prior of the form

$$\begin{aligned} \beta_j &\sim \pi\delta_0(\beta) + (1-\pi)G, \\ G &\sim DP(\alpha G_0), \\ G_0 &\sim N(0, \tau^2). \end{aligned}$$

G is a nonparametric density which follows a Dirichlet process with base measure G_0 and concentration parameter α. In this case the base measure G_0 is Gaussian with zero mean and variance τ^2, which is the typical conjugate prior distribution used on linear regression coefficients. Hence, this prior implies that each coefficient β_j will either be restricted to 0 with probability π, or with probability $1-\pi$ will come from a mixture of Gaussian densities. If it comes from a mixture of Gaussian densities, then due to a property of the Dirichlet process, β_j's in the same mixture component will share the same mean and the variance.

As an example, consider coefficients β_j, $j = 1, ..., 6$ with $(\beta_1, \beta_3) \sim N(0, 10^6)$, $(\beta_2, \beta_4) \sim N(0, 0.1)$, and $(\beta_5, \beta_6) \sim \delta_0$. In this case, (β_1, β_3) are clustered together and have a Gaussian prior with variance 10^6 which means that their posterior mean/median will be close to the least squares estimator. The second cluster consists of (β_2, β_4) which have prior variance 0.1, hence their posterior median will be equivalent to a ridge regression estimator. Finally, (β_5, β_6) are restricted to be zero.

Inference using the Gibbs sampler is straightforward, once we write the Dirichlet process prior using its stick-breaking representation, that is, an infinite sum of point mass functions.

The general form of the semiparametric spike and slab prior we use is of the form

$$\begin{aligned}
\beta_j &\sim \pi\delta_0(\beta) + (1-\pi)\, G, && \text{(B.206)}\\
G &\sim DP(\alpha G_0), && \text{(B.207)}\\
G_0 &\sim N\left(\underline{\mu}, \tau^2\right), && \text{(B.208)}\\
\tau^2 &\sim Inv - Gamma\left(\underline{a}_1, \underline{a}_2\right), && \text{(B.209)}\\
\alpha &\sim Gamma\left(\underline{\rho}_1, \underline{\rho}_2\right), && \text{(B.210)}\\
\pi &\sim Beta\left(\underline{c}, \underline{d}\right), && \text{(B.211)}\\
\sigma^2 &\sim \frac{1}{\sigma^2}, && \text{(B.212)}
\end{aligned}$$

where $\underline{\mu}, \underline{a}_1, \underline{a}_2, \underline{\rho}_1, \underline{\rho}_2, \underline{c}, \underline{d}$ are parameters to be chosen by the researcher. The usual stick breaking representation for β_j conditional on β_{-j} and marginalized over G is of the form

$$\left(\beta_j | \boldsymbol{\beta}_{-j}\right) \sim \frac{\alpha(1-\pi)}{\alpha + K - p_{\beta_1} - 1} N\left(\underline{\mu}, \tau^2\right) + \pi\delta_0(\beta) + \sum_{l=2}^{k_\beta} \frac{p_{\beta_l}(1-\pi)}{\alpha + K - p_{\beta_1} - 1} \delta_{\beta_l}(\beta), \qquad \text{(B.213)}$$

where k_β is the number of atoms in the above equation (number of mixture components plus the $\delta_\beta(0)$ component), and p_{β_n} is the number of elements of the vector β which which are equal to $\delta_{\beta_l}(\beta)$, $n = 1, 2, ..., k_\beta$, where it holds that $\delta_{\beta_1}(\beta) = \delta_0(\beta)$. Additionally, for notational convenience define the prior weights as

$$\begin{aligned}
w_0 &= \frac{\alpha(1-\pi)}{\alpha + K - p_{\beta_1} - 1},\\
w_1 &= \pi,\\
w_l &= \frac{p_{\beta_l}(1-\pi)}{\alpha + K - p_{\beta_1} - 1},\ l = 2, ..., k_\beta.
\end{aligned}$$

Gibbs sampling from the conditional posterior:

- Given k_β number of mixture components, sample $\boldsymbol{\theta} = \left(\boldsymbol{\theta}_1, ..., \boldsymbol{\theta}_{k_\beta}\right)$ from

  $$(\boldsymbol{\theta}|-) \sim N\left(\boldsymbol{E}_{\boldsymbol{\theta}}, \boldsymbol{V}_{\boldsymbol{\theta}}\right),$$

  with $\boldsymbol{E}_{\boldsymbol{\theta}} = \boldsymbol{V}_\theta\left(\boldsymbol{D}^{-1}\boldsymbol{M} + \sigma^{-2}\boldsymbol{X}_\pi'\boldsymbol{y}\right)$ and $\boldsymbol{V}_\beta = \left(\boldsymbol{D}^{-1} + \sigma^{-2}\boldsymbol{X}_\pi'\boldsymbol{X}_\pi\right)^{-1}$, where $\boldsymbol{D} = \tau^2\boldsymbol{I}_{k_\beta}$ and $\boldsymbol{M} = \mu\boldsymbol{1}_{k_\beta}$. Here $\boldsymbol{X}_\pi'$ denotes the matrix $\boldsymbol{X}$ with the columns corresponding to coefficients belonging to θ_1 being replaced with zeros (or equivalently, with these columns removed). Hence the remaining columns correspond to unrestricted coefficients which belong to one of the remaining $k_\beta - 1$ mixture components.

- Sample β_j conditional on β_{-j}, data, and other model parameters for $j = 1, ..., K$ from

$$\left(\beta_j | \boldsymbol{\beta}_{-j}, -\right) \sim \overline{w}_0 N\left(E_\beta, V_\beta\right) + \sum_{l=1}^{k_\beta} \overline{w}_l \boldsymbol{\theta}_l,$$

so that with probability $\overline{w}_l$ we assign β_j equal to the atom of mixture component l (i.e. $\beta_j = \theta_l$), while with probability $\overline{w}_0$ we assign β_j to a new $N\left(E_\beta, V_\beta\right)$ component. In the expression above it holds that

$$\begin{aligned} E_\beta &= V_\beta \left(\tau^{-2}\underline{\mu} + \sigma^{-2}\boldsymbol{X}'\boldsymbol{y}\right), \\ V_\beta &= \left(\tau^{-2} + \sigma^{-2}\boldsymbol{X}'\boldsymbol{X}\right)^{-1}, \end{aligned}$$

and that

$$\begin{aligned} \overline{w}_0 &\propto \frac{w_0 N\left(0; \underline{\mu}, \tau^2\right) \prod_{i=1}^n N\left(\widetilde{y}_i; 0, \sigma^2\right)}{N\left(0; E_\beta, V_\beta\right)}, \\ \overline{w}_l &\propto w_l N\left(0; \underline{\mu}, \tau^2\right) \prod_{i=1}^n N\left(\widetilde{y}_i; \boldsymbol{X}_{i,l}\boldsymbol{\theta}_l, \sigma^2\right), \ l = 1, ..., k_\beta, \end{aligned}$$

where $\widetilde{y}_i = y_i - \sum_{j' \neq j} X_{i,j'}\beta_{j'} = y_i - \left(\boldsymbol{X}_\pi\right)_i \boldsymbol{\theta} + X_{j',i}\beta_{j'}$ for $j, j\prime = 1, ..., K$, $\left(X_\pi\right)_i$ is the i-th observation of the matrix X_π constructed in step 1, and $N\left(a; b, c\right)$ denotes the normal density with mean b and variance c, evaluated at observation a.

- Introduce an indicator variable $S_\beta = l$ if the coefficient β_j belongs to cluster l, where $j = 1, ..., K$ and $l = 1, ..., k_\beta$, in which case it holds that $\beta_j = \theta_l$. In addition, set $S_\beta = 0$ if $\beta_j \neq \theta_l$, that is when β_j does not belong to a preassigned cluster and a new cluster is introduced for this coefficient. Then the conditional posterior of S_β is

$$\left(S_\beta | -\right) \sim Multinomial\left(0, 1, ..., k_\beta; \overline{w}_0, \overline{w}_1, ..., \overline{w}_{k_\beta}\right).$$

- Sample the restriction probability π from the coniditional distribution

$$\left(\pi | -\right) \sim Beta\left(\underline{c} + \sum_{j=1}^K I\left(S_\beta = 1\right), d + \sum_{j=1}^K I\left(S_\beta \neq 1\right)\right).$$

- Sample the latent variable η from the posterior conditional

$$\left(\eta | -\right) \sim Beta\left(a + 1, K - \sum_{j=1}^K I\left(S_\beta = 1\right)\right).$$

- Sample the Dirichlet process precision coefficient α from the conditional posterior

$$\begin{aligned} \left(\alpha | -\right) \sim\ & \pi_\eta Gamma\left(\underline{\rho}_1 + k_\beta - n_{S_\beta = 1}, \underline{\rho}_2 - \log\eta\right) + \\ & \left(1 - \pi_\eta\right) Gamma\left(\underline{\rho}_1 + k_\beta - n_{S_\beta = 1} - 1, \underline{\rho}_2 - \log\eta\right), \end{aligned}$$

where the weight π_η is given by

$$\frac{\pi_\eta}{1-\pi_\eta} = \frac{\underline{\rho}_1 + k_\beta - n_{S_\beta=1} - 1}{\left(K - \sum_{j=1}^{K} I\left(S_\beta = 1\right)\right)\left(\underline{\rho}_2 - \log \eta\right)},$$

and $n_{S_\beta=1} = 1$ if $\sum_{j=1}^{K} I\left(S_\beta = 1\right) > 0$, and it is 0 otherwise (i.e. when no coefficient β_j is restricted).

- Sample the variance τ^2 coefficient from the conditional density

$$\left(\tau^2|-\right) \sim iGamma\left(\underline{a}_1 + \frac{1}{2}\left(k_\beta - 1\right), \underline{a}_2^{-1} + \frac{1}{2}\sum_{l=2}^{k_\beta}\left(\boldsymbol{\theta}_l - \underline{\mu}\mathbf{1}\right)^2\right).$$