Supplementary Materials for “Bayesian Ising Graphical Model for Variable Selection”

Zaili Fang, Inyoung Kim

1 Marginal Odds and Shrinkage Parameter

1.1 Marginal Odds π_j^b

To understand the different impact of three normal mixture priors of β_j on the marginal selection probability $p(\gamma_j = 1|b) = \pi_j^b/(1 + \pi_j^b)$, it is convenient to introduce the shrinkage coefficient, $\kappa_j = \left(\frac{\tau_j}{b^2\phi}\right) / \left(1 + \frac{\tau_j}{b^2\phi}\right)$ similar to Carvalho and Polson (2010); Polson and Scott (2011). Under the orthogonal design, the posterior mean and variance of β_j 's corresponding to $j \in \{j : \gamma_j = 1\}$ are

$$\begin{aligned} E(\beta_j|\mathbf{y}, \phi, b) &= [1 - E(\kappa_j|\mathbf{y}, \phi, b)]a_j, \\ \text{Var}(\beta_j|\mathbf{y}, \phi, b) &= [1 - E(\kappa_j|\mathbf{y}, \phi, b)]\phi^{-1}, \end{aligned} \quad (1)$$

where $a_j = \mathbf{x}_j^T \mathbf{y}$ is considered as the signal. The coefficient κ_j 's represent how much shrinkage is placed on the initial estimation of β_j 's. $\kappa_j \rightarrow 0$, yields no shrinkage, and $\kappa_j \rightarrow 1$ yields near-total shrinkage. With this definition of κ_j , it is easy to derive the density function of κ_j , $p(\kappa_j)$. In the following discussions, we assume $\phi = 1$. This is equivalent to say that $\mathbf{y}$, a_j 's and b are scaled by $\sqrt{\phi}$, respectively.

The marginal odds for $\gamma_j = 1$ given b is calculated as

$$\pi_j^b = \frac{\int \sum_{\gamma_{\bar{j}} \in \{0,1\}^{p-1}} p(\gamma_j = 1, \gamma_{\bar{j}}, \boldsymbol{\beta}, \boldsymbol{\tau}|\mathbf{y}, b) d\boldsymbol{\beta} d\boldsymbol{\tau}}{\int \sum_{\gamma_{\bar{j}} \in \{0,1\}^{p-1}} p(\gamma_j = 0, \gamma_{\bar{j}}, \boldsymbol{\beta}, \boldsymbol{\tau}|\mathbf{y}, b) d\boldsymbol{\beta} d\boldsymbol{\tau}}$$

with $\bar{j} = \{1, 2, \dots, p\} \setminus j$, where $p(\boldsymbol{\gamma}, \boldsymbol{\beta}, \boldsymbol{\tau}|\mathbf{y}, b) \propto p(\mathbf{y}|\boldsymbol{\gamma}, \boldsymbol{\beta})p(\boldsymbol{\beta}|\boldsymbol{\tau}, b)p(\boldsymbol{\tau})$ is the joint posterior distribution of $\boldsymbol{\gamma}$, $\boldsymbol{\beta}$ and $\boldsymbol{\tau}$. For orthogonal designs, it is easy to show the marginal odds is the expectation of the function $\pi_j = \kappa_j^{1/2} \exp[a_j^2(1 - \kappa_j)/2]$ on the probability measure of κ_j

$$\pi_j^b = \int \pi_j p(\kappa_j) d\kappa_j. \quad (2)$$

Obviously, if $\kappa_j \rightarrow 0$, then $\pi_j \rightarrow 0$, and if $\kappa_j \rightarrow 1$, then $\pi_j \rightarrow 1$. Similarly, if $b \rightarrow 0$, then $\pi_j^b \rightarrow 1$, and if $b \rightarrow \infty$, then $\pi_j^b \rightarrow 0$. See A.2 for the details to derive Expression (2).

The above expression shows the distribution of shrinkage $p(\kappa_j)$ determines the behavior of the marginal selection probability $p(\gamma_j = 1|b) = \pi_j^b/(1 + \pi_j^b)$. The density function forms of

$p(\beta_j|b)$, $p(\tau_j)$ and $p(\kappa_j)$ of three normal mixture prior cases are given in Table 1. Furthermore, Figure 1 compares the density function $p(\beta_j|b)$'s around zero and along the tails, and the density function $p(\kappa_j)$'s given different b 's.

1.2 Expression for π_j^b

There is no closed form of π_j^b for the three different priors. However, at least for the Laplace and horseshoe priors, π_j^b can be expressed by some special functions.

For Laplace prior,

$$\pi_j^b = \sqrt{\frac{\pi}{2}} b^{-1} \exp \left[\frac{(|a_j| - b^{-1})^2}{2} \right] \{ \Phi(|a_j| - b^{-1}) + \exp(2|a_j|b^{-1}) \Phi(-|a_j| - b^{-1}) \}, \quad (3)$$

where $\Phi(\cdot)$ is the CDF of the standardized normal distribution. This expression can be directly used to calculate the marginal selection probability given b in Figure 2.

For horseshoe prior,

$$\pi_j^b = \frac{1}{\pi} b^{-1} Be \left(1, \frac{1}{2} \right) \Phi_1 \left(\frac{1}{2}, 1, \frac{3}{2}, \frac{a_j^2}{2}, 1 - b^{-2} \right), \quad (4)$$

where $Be(\cdots)$ denotes the beta function, and $\Phi_1(\cdots)$ is the degenerate hypergeometric function of two variables (Gordy, 1998; Polson and Scott, 2010). Φ_1 can be evaluated by using a series of hypergeometric ${}_2F_1$ functions (Gordy, 1998).

The derivative of the above expressions is shown in A.3. The odds π_j^b of the Cauchy prior does not have an analytic form and cannot be represented by a known set of special functions. Hence we use the Monte Carlo approach to calculate π_j^b under the Cauchy prior.

1.3 Dynamic Properties of the Odds with Different Priors

By examining $p(\kappa_j)$'s in Figure 1 together with Figure 2, we see how b effects the selection probability profile by putting different weight on shrinkage. Figure 2 (a) shows the selection probability calculated as $\pi_j/(1 + \pi_j)$ against κ_j at different signal magnitudes where the curves are identical for all three normal mixture priors, while Figure 2 (b) plots the marginal selection probability $\pi_j^b/(1 + \pi_j^b)$ at different signal size a_j 's against b where three normal mixture priors demonstrate different behaviors because of $p(\kappa_j)$. In Figure 2 (c) we plot the selection probability profile with the baseline ($a_j = 0$) subtracted. It can be seen that for an orthogonal design, the larger the magnitude of the selection probability, the more markedly the true signal ($a_j \neq 0$) is distinguished from the baseline.

In Figure 2 (c), for small b ($b \leq 1$) and a given signal the selection probability of three priors descends in the order of : the Cauchy prior, the horseshoe prior, and the Laplace prior, which is consistent with the $p(\kappa_j)$ plots in Figure 1 at $b = 0.1$, where the density masses on the small κ_j region also exhibits the same order. In addition, since the Cauchy and the horseshoe priors put similar mass around the small κ_j side, their selection probabilities behave almost identically for small b , but Laplace prior puts almost zero weight on small shrinkage thus its selection probability drops dramatically as shown in Figure 2 (c).

On the other hand, the order of density mass at $b = 100$ for small κ_j becomes as: the horse shoe, Laplace and Cauchy prior (Figure 1). Again this is consistent with the selection probability order in Figure 2 (c) for large b ($b \geq 100$). At a moderate b , such as $b = 1$, all priors have substantial mass around small κ_j side, as shown in Figure 1, thus all priors behave similarly. This is confirmed by Figure 2 (c), where the selection probabilities for three priors seem similar around $b = 1$ for large signals.

In conclusion, the analysis shows that more weight on large shrinkage is not as important as on small shrinkage for the purpose of distinguishing the signals. Therefore, we find that the horseshoe prior is superior to other two, even though the marginal prior $p(\beta_j|b)$ of the horseshoe prior is not as heavy-tailed as the Cauchy prior for small b . The horseshoe prior demonstrates that for a wide range of b , $p(\kappa_j)$ maintains substantial mass on the small shrinkage side of κ_j . On the other hand, the Laplace prior has almost zero mass around small κ_j when b is small, and Cauchy prior has all mass absorbed to $\kappa_j = 0$ when b is very large, each diminishing its respective performance for the particularly b values.

2 Full Conditional Distributions

2.1 BIGM

The full conditional distribution for β is

$$[\beta_c | \mathbf{y}, \gamma, \beta_{\bar{c}}, \phi] \sim \begin{cases} N(\boldsymbol{\mu}_c, \Sigma_c) & \text{if } \gamma_c = \mathbf{1} \\ N(\mathbf{0}, D_c^{-1}) & \text{if } \gamma_c = \mathbf{0}. \end{cases} \quad (5)$$

Here we use “ c ” as a general subscript to stand for a subset of the index $\{1, \dots, p\}$. We use $\bar{c}$ to present the complementary index set of c . In above expression, D_c is a $|c| \times |c|$ diagonal matrix with $\tau_j/b^2, j \in c$ as the diagonal elements, where $|c|$ refers for the cardinality of c . Σ_c and $\boldsymbol{\mu}_c$ are expressed as

$$\begin{aligned} \Sigma_c &= (\phi X_c^T X_c + D_c)^{-1}, \\ \boldsymbol{\mu}_c &= \phi \Sigma_c X_c^T (\mathbf{y} - X_{\gamma_{\bar{c}}} \beta_{\bar{c}}). \end{aligned} \quad (6)$$

With some abuse of notations here, X_c stands for the sub-matrix of X corresponding the predictors in c . Similarly $X_{\gamma_{\bar{c}}}$ is the sub-matrix of $X_{\gamma} = (\gamma_1 \mathbf{x}_1, \dots, \gamma_p \mathbf{x}_p)$ corresponding to predictors indexed by elements of $\bar{c}$.

The full conditional distribution for τ_j in the Cauchy and Laplace priors are

$$[\tau_j | \beta_j, b, \phi] \propto \tau_j^{-3/2} \exp \left[-\frac{(\tau_j - b/|\beta_j|)^2}{2\tau_j b^2 / \beta_j^2} \right], \quad \text{Laplace prior} \quad (7)$$

$$[\tau_j | \beta_j, b, \phi] \propto \exp \left[-\frac{1}{2}(\beta_j^2/b^2 + 1)\tau_j \right], \quad \text{Cauchy prior} \quad (8)$$

and the full condition distribution for τ_j in the horseshoe prior is obtained by

$$\begin{aligned} [u_j|\beta_j, b, \phi, v_j] &\propto \exp \left[-\frac{1}{2} \left(\frac{\beta_j^2}{b^2 v_j} + 1 \right) u_j \right], \\ [v_j|\beta_j, b, \phi, u_j] &\propto v_j^{-1} \exp \left[-\frac{1}{2} \left(\frac{\beta_j^2 u_j}{b^2} \frac{1}{v_j} + v_j \right) \right], \\ \tau_j &= u_j/v_j. \end{aligned} \quad \text{Horseshoe prior} \quad (9)$$

(7) is an inverse Gaussian distribution $ING(b/|\beta_j|, 1)$ with mean $b/|\beta_j|$ and shape parameter 1. (8) is an exponential distribution or Gamma distribution $G(1, (\beta_j^2/b^2 + 1)/2)$. The Gibbs sampler for the horseshoe prior is implemented by using the redundant multiplicative reparameterization technique similar to that found in Gelman (2006). If we reparameterize τ_j as $\tau_j = u_j/v_j$ where u_j and v_j are independently identically distributed with prior $G(1/2, 1/2)$, then a priori $[\tau_j^{1/2}] \sim C^+(0, 1)$, and the prior for β_j is the horseshoe prior. In (9), the full conditional distributions for u_j and v_j are a Gamma distribution $G[1, (\beta_j^2/(b^2 v_j) + 1)/2]$ and the generalized inverse Gaussian distribution $GING(0, \frac{\beta_j^2 u_j}{b^2}, 1)$, respectively.

The full conditional distribution of ϕ becomes a gamma distribution,

$$[\phi|\mathbf{y}, \boldsymbol{\beta}, \boldsymbol{\gamma}] \sim G\left(\frac{n}{2}, \frac{1}{2}\|\mathbf{y} - X_{\boldsymbol{\gamma}}\boldsymbol{\beta}\|^2\right). \quad (10)$$

2.2 SSVS

In order to have a fair comparison with BIGM, we consider a similar hierarchical structure for Bayesian SSVS model in this paper:

$$\begin{aligned} [\mathbf{y}|\boldsymbol{\beta}, \phi] &\sim N(X\boldsymbol{\beta}, \phi^{-1}I), \\ [\beta_j|\gamma_j, \tau_j, b, d] &\sim \gamma_j N(0, b^2 \tau_j^{-1}) + (1 - \gamma_j) N(0, b^2 d \tau_j^{-1}), \\ [\tau_j] &\sim p(\tau_j), \\ [\phi] &\sim \phi^{-1}, \end{aligned} \quad (11)$$

where $p(\tau_j)$ is a similar prior distribution to BIGM (Gamma, Inverse Gamma and half Cauchy), and d is a small positive number so that if $\gamma_j = 0$ then $\beta_j \sim N(0, b^2 d \tau_j^{-1})$, a normal distribution with variance $b^2 \tau_j^{-1}$ shrunk by d .

Based on the above hierarchical model, the full conditional distribution of $\boldsymbol{\beta}$ is

$$[\boldsymbol{\beta}|\mathbf{y}, \boldsymbol{\gamma}, \phi] \sim N(\boldsymbol{\mu}, \Sigma), \quad (12)$$

where

$$\begin{aligned} \Sigma &= (\phi X^T X + S_{\boldsymbol{\gamma}})^{-1}, \\ \boldsymbol{\mu} &= \phi \Sigma X^T \mathbf{y}. \end{aligned} \quad (13)$$

and $S_{\boldsymbol{\gamma}}$ is a $p \times p$ diagonal matrix with $\frac{\tau_j}{b^2 d^{(1-\gamma_j)}}$ on the diagonal.

Because the prior of β_j is equivalent to $[\beta_j^*|\gamma_j, \tau_j, b, d] \sim N(0, b^2 \tau_j^{-1})$ by taking transformation $\beta_j^{*2} = \beta_j^2/d^{(1-\gamma_j)}$, the full condition distributions for τ_j 's can be derived the same as (7), (8) and (9) given the same prior for τ_j except that β_j 's are replaced with β_j^* .

The full condition distribution for γ_j 's is different from BIGM now, but it is still considered as a graphic model. Notice that $\frac{1}{d^{1-\gamma_j}} = \gamma_j + (1 - \gamma_j)\frac{1}{d}$, thus $[\beta_j|\gamma_j, \tau_j, b, d] \propto \exp\left[-\frac{1}{2}\beta_j^2 \frac{\tau_j}{b^2 d^{1-\gamma_j}}\right] \left(\frac{\tau_j}{b^2 d^{1-\gamma_j}}\right)^{1/2} \propto \exp\left[-\frac{1}{2}\beta_j^2 \frac{\tau_j}{b^2} \left(\gamma_j + (1 - \gamma_j)\frac{1}{d}\right) - \frac{1}{2}(1 - \gamma_j) \log d\right] \left(\frac{\tau_j}{b^2}\right)^{1/2}$. Given the same noninformative prior for γ_j the full condition distribution of γ_j for the SSVS model is then expressed as a similar graphic model:

$$[\gamma|\beta, \tau, b, d] \propto \exp(\gamma^T J \gamma + \mathbf{h}^T \gamma), \quad (14)$$

except where J and $\mathbf{h}^T$ consist of different elements: J is a $p \times p$ zero matrix so $\gamma^T J \gamma$ is actually equal to zero, and the j th element of the vector $\mathbf{h}$ is $h_j = \frac{1}{2} \left[\left(\frac{1}{d} - 1\right) \frac{\beta_j^2 \tau_j}{b^2} + \log d \right]$. With this expression for the full conditional distribution of γ_j , the single-site/cluster algorithm of updating γ introduced for BIGM is applied directly.

Although the full conditional distribution of ϕ is not dependent on γ , it has similar form:

$$[\phi|\mathbf{y}, \beta] \sim G\left(\frac{n}{2}, \frac{1}{2}\|\mathbf{y} - X\beta\|^2\right). \quad (15)$$

2.3 g -prior

Similarly, in order to have a fair comparison with the g -prior model, we consider the following hierarchical structure using the KM model while the prior of β has the form of Zellner's g -prior:

$$\begin{aligned} [\mathbf{y}|\beta, \gamma, \phi] &\sim N(X_\gamma \beta, \phi^{-1} I), \\ [\beta|\phi, b] &\sim N\left(\mathbf{0}, \frac{b^2}{\phi}(X^T X)^{-1}\right), \\ [\phi] &\sim \phi^{-1}. \end{aligned} \quad (16)$$

The full conditional distribution of β is derived as

$$[\beta|\mathbf{y}, \gamma, \phi] \sim N(\boldsymbol{\mu}, \Sigma), \quad (17)$$

where

$$\begin{aligned} \Sigma &= \left(\phi X_\gamma^T X_\gamma + \frac{\phi}{b^2} X^T X \right)^{-1}, \\ \boldsymbol{\mu} &= \phi \Sigma X_\gamma^T \mathbf{y}, \end{aligned} \quad (18)$$

and the full condition distribution of ϕ is

$$[\phi|\mathbf{y}, \beta, \gamma] \sim G\left[\frac{n+p}{2}, \frac{1}{2}\left(\|\mathbf{y} - X_\gamma \beta\|^2 + \frac{\beta^T X^T X \beta}{b^2}\right)\right]. \quad (19)$$

Because the distribution of $\mathbf{y}$ is exactly the same as BIGM, the full conditional distribution of γ_j 's are then exactly the same as BIGM and the algorithms to update γ keeps the same for g -prior model too.

3 Additional Simulation Results

3.1 Case One: Comparison of Three Priors

This simulation study will examine a simple linear regression model with a general form

$$y = \sum_{j \in S} \beta_j x_j + \epsilon, \quad (20)$$

where $S = \{2, 3, 5, 10\}$, $n = 50$ is the sample size, $p = 100$, $x_j \sim N(0, 1)$, $j = 1, \dots, p$ and $\epsilon \sim N(0, 1)$. Particularly, we will consider one large signal set and one small signal set: $\{\beta_2, \beta_3, \beta_5, \beta_{10}\} = \{-4, 2, -1, 2.5\}$ and $\{-0.9, 0.7, -0.6, 0.8\}$.

In this simulation, we performed the single-site updating with a total of 6000 iterations for each settings, discarding the first 2000 iterations as burn-in, then calculated the average γ_j 's over the remaining $N = 4000$ iterations as the marginal selection probabilities. Figure 5 plots the marginal selection probability of all variables against the global shrinkage parameter b . For the large signal setting, as shown in the upper row of Figure 5, the horseshoe and Cauchy priors perform similarly and show the robustness of large signals. The horseshoe prior shows even more robustness than Cauchy prior on the large b region. Both priors have a wide window of b in which the true signals are well separated from the noise signals. On the other hand, Laplace prior does not exhibit such robustness for large signals: as the shrinkage parameter $b \rightarrow 0$, the selection probability for the true signals drops to 0.5 very fast. Around $b = 0.1$, all signal probabilities reach the 0.5 line under Laplace prior. In the large b region, the Laplace prior seems perform a little better than the Cauchy prior. We can also see that the conclusions drawn from the simulation about the performance of the three priors agrees with the exact calculation of the marginal selection probabilities from Figure 2.

The bottom row of Figure 5 displays the simulation results for small signals. In general, the window for true signals maintaining a high selection probability narrows for all three priors. The selection probability of the true signal for all priors starts to drop to 0.5 around $b = 0.1$, and around $b = 1000$, it drops to 0. However, the drop rate on both ends of b is different for three priors, resulting in different widths for the working widow of b . The horseshoe prior has the widest window, while the Laplace prior has the narrowest. Again, the conclusion we make is exactly the same as from the calculation in Figure 2.

We may note some other observations from Figure 5, especially for small signals. For example, we see that around $b = 10$, the selection probability of true signals first drops a little and then increases. This happens right above the peak of the selection probability of the noise, which indicates a potential interaction between the noise and the true signals. This is easy to understand since when sample size is small or collinearity exists, the correlation between a true signal and noise is large. Expression (2) becomes to $\pi_j^b = \int \pi_j \xi_j p(\kappa_j) d\kappa_j$ (See A.2). The parameter ξ_j represent the interaction among signals and distorts the profile curves.

Based on this first case study, we find that the horseshoe prior has the largest working window of b . Hence the rest of the paper employs only horseshoe prior unless otherwise stated.

3.2 Case Three: Bayesian Sparse Additive Model (BSAM)

This is the same simulation of Equation (7) and (8) in the main paper but with additional results.

Figure 6 (a-b) show us how the selection probabilities changes for a range of b with $t = 0$ and $t = 1$ for the simulation setting $p = 80$. Note that in Figure 6 (a), one true signal is buried in the noise until $b = 1$, while the same signal is always mixed with noise in Figure 6 (b). As shown in Figure 6, when $b \approx 26$ all false signals go to 0 and we achieve the largest gap between signals and the noise.

One feature of our BIGM is its ability to simultaneously select the important variables and estimate the selected function components. We show the four true functions and one noise function, the corresponding estimated functions, and the selection probability for all nodes in the case where $b = 26$ and $p = 80$ for $t = 0$ and $t = 1$ in Figure 7 and Figure 8 respectively. Based on the LS basis, the function components estimated are always centered, so we also center the true functions. In this simulation run, the four true nodes are selected exactly, and their estimated functions are calculated by $\hat{f}_j = Z_j E(\beta_j | \gamma_j = 1)$, where the expectation is based on the $N = 4000$ iterations, and the 95% credible intervals are plotted as well.

As shown in Figure 7 and 8, for both $t = 0$ and $t = 1$ the estimated functions are very close to the true functions. Note that for the noise function, f_{11} , the selection probability is close to zero, and thus the estimated function is calculated by $\hat{f}_j = Z_j E(\beta_j)$, an expectation over all N iterations. This is why we see a very wide credible interval for f_{11} , because when $\gamma_j = 0$ the posterior of β_j is a multivariate normal with large variance. Also note that to have a better view of the selection probability, we reordered the nodes such that the 4 true nodes are (2, 20, 50, 70).

4 Additional Real Data Analysis: Ozone Data

As an illustration of BSAM implemented by the BIGM, we consider the ozone data analyzed by Lin and Zhang (2006). The ozone data are available in the R packages `cosso` and `gss`. The Ozone dataset consists of eight explanatory meteorological variables that were recorded in the Los Angeles area for 330 days in 1976. The response variable of interest is the daily maximum hourly average ozone concentration. The eight explanatory variables are Height (Hgt), Wind Speed (WS), Humidity (Hum), Temperature (Temp), Inversion Base Height (InvHt), Pressure (Press), Inversion Base Temperature (InvTp), and Visibility (Vis). All predictors were standardized and the response was log-transformed to normalize its distribution.

We applied the Bayesian graph model described in Section 4.3 of the main paper with $M_j = 6$ for all predictors. A total of 20000 iterations were employed using the single-site algorithm, and half of them were discarded as burn in. By quickly examining a small number of b , we saw the selection probability profile curves are all well defined (not shown here). We conclude that it is appropriate to choose a modest shrinkage, $b = 1.6$, in this case such that all selected predictors reach their highest selection probability. At this value of b , only two predictors have selection probability less than 0.5.

The estimated results for β_j 's are summarized in Figure 9, where the additive function components, $\hat{f}_j = Z_j E(\beta_j | \gamma_j = 1)$'s, are plotted with an enveloping 95% credible band. The marginal selection probability, $P = p(\gamma_j | \mathbf{y})$, for each variable is labeled in each plot. Because the smallest P is 0.14, we decide that there are enough iterations for all $\gamma_j = 1$ to estimate all $\hat{f}_j$. We can therefore identify three groups of the variables. The first group, which includes Temp, Press, InvHt and Vis, has $P \geq 0.88$. The second group includes InvTp and Hgt with $P = 0.69$ and 0.58 . The last group contains Hum and WS with $P < 0.5$.

From a variable selection standpoint, we would certainly want to choose all variables in the first group, and the last group of variables would be excluded, since their selection probabilities are very close to the baseline. Because of the small number of variables admitted and the near independence of each variable, we can consider the second group as true variables with small signals.

Regarding function components estimation, the selection probability is consistent with signal estimation. As shown in Figure 9, the first group of variables has their credible interval bands only cover a small part of the zero line. The third group of variables has its credible interval band almost completely centered on the zero line. While the credible intervals of the second group covers the zero line entirely, the zero line is only on the edge of the credible interval.

The summary statistics for all precision parameters and the intercept are reported in Table 1, which shows how the posterior means of these parameters adapt to the data.

Appendix A

A.1 Proof of Theorem 1

In general, in the Markov chain of MH algorithm, the move from current state γ_c^0 to the proposed state γ_c^* in the cluster has the transition probability, $P(\gamma_c^0 \rightarrow \gamma_c^*)$, which satisfies the detailed balance condition

$$\frac{P(\gamma_c^0 \rightarrow \gamma_c^*)}{P(\gamma_c^* \rightarrow \gamma_c^0)} = \frac{p(\gamma_c^* | \mathbf{y}, \beta, \phi)}{p(\gamma_c^0 | \mathbf{y}, \beta, \phi)} = \exp \{ - [U(\gamma_c^*) - U(\gamma_c^0)] \}. \quad (\text{A.1.1})$$

The transition probability can be broken down into two parts:

$$P(\gamma_c^0 \rightarrow \gamma_c^*) = g(\gamma_c^0 \rightarrow \gamma_c^*) A(\gamma_c^0 \rightarrow \gamma_c^*),$$

where $g(\cdot)$ is the selection probability, which is the probability given γ_c^0 that the new target state generated, and $A(\cdot)$ is the acceptance ratio. Thus

$$\frac{g(\gamma_c^0 \rightarrow \gamma_c^*) A(\gamma_c^0 \rightarrow \gamma_c^*)}{g(\gamma_c^* \rightarrow \gamma_c^0) A(\gamma_c^* \rightarrow \gamma_c^0)} = \exp \{ - [U(\gamma_c^*) - U(\gamma_c^0)] \}. \quad (\text{A.1.2})$$

Now we consider the move $\gamma_c^0 \rightarrow \gamma_c^*$, starting with a particular cluster c and then adding the others to it in a particular order. Consider also the reverse move, which takes us back to γ_c^0 from γ_c^* , starting with exactly the same cluster (except the state in the cluster is flipped), and adding the others to it in exactly the same way as in the forward move. The probability of

choosing the cluster (if the cluster is the seed node) is exactly the same in the two directions, as is the probability of adding each node to the cluster. The only difference between the two directions is the probability of “breaking” bonds around the edge of the cluster. Because the cluster couples with all $j \in \bar{c}$, for both directions, there are $|\bar{c}|$ bonds which have to be broken in order to flip the cluster. These broken bonds represent the affinity between the cluster and the spins which were not added to the cluster by the algorithm. We represent the probability of not adding such a node in forward move as $1 - p_{a,j}^0, j \in \bar{c}$ and in backward move as $1 - p_{a,j}^*, j \in \bar{c}$. Thus the probability of not adding all of them, which is proportional to the selection probability $g(\gamma_c^0 \rightarrow \gamma_c^*)$ for the forward move, is $\prod_{j \in \bar{c}} (1 - p_{a,j}^0)$. In the reverse move then the probability of doing it is $\prod_{j \in \bar{c}} (1 - p_{a,j}^*)$. The condition of detailed balance, Equation (A.1.1), along with Equation (A.1.2), then tells us that

$$\frac{g(\gamma_c^0 \rightarrow \gamma_c^*) A(\gamma_c^0 \rightarrow \gamma_c^*)}{g(\gamma_c^* \rightarrow \gamma_c^0) A(\gamma_c^* \rightarrow \gamma_c^0)} = \prod_{j \in \bar{c}} \left(\frac{1 - p_{a,j}^0}{1 - p_{a,j}^*} \right) \cdot \frac{A(\gamma_c^0 \rightarrow \gamma_c^*)}{A(\gamma_c^* \rightarrow \gamma_c^0)} = \exp \{ - [U(\gamma_c^*) - U(\gamma_c^0)] \}. \quad (\text{A.1.3})$$

Note that the energy change $U(\gamma_c^*) - U(\gamma_c^0)$ is only determined by the bonds (the coupling between c and $\bar{c}$) and coupling of c with the external field $\mathbf{h}^*$, i.e.,

$$U(\gamma_c^*) - U(\gamma_c^0) = \sum_{j \in \bar{c}} (-1)^{\gamma_j} \left(\sum_{k \in c_0} J_{jk} - \sum_{l \in c_1} J_{jl} \right) + \sum_{j \in c_1} h_j^* - \sum_{j \in c_0} h_j^*. \quad (\text{A.1.4})$$

The first part of right hand side of Equation (A.1.4) can be decomposed as

$$\lambda \sum_{j \in \bar{c}} (-1)^{\gamma_j} \left(\sum_{k \in c_0} J_{jk} - \sum_{l \in c_1} J_{jl} \right) + (1 - \lambda) \sum_{j \in \bar{c}} (-1)^{\gamma_j} \left(\sum_{k \in c_0} J_{jk} - \sum_{l \in c_1} J_{jl} \right).$$

With the probability of adding a node $j \in \bar{c}$ to the cluster, $p_{a,j}$, defined as (??),

$$\frac{1 - p_{a,j}^0}{1 - p_{a,j}^*} = \exp \left\{ -\lambda (-1)^{\gamma_j} \left(\sum_{k \in c_0} J_{jk} - \sum_{l \in c_1} J_{jl} \right) \right\}.$$

Substituting above equation into Expression (A.1.3) and rearranging, we derive the acceptance ratio for the moves in the two directions as

$$\frac{A(\gamma_c^0 \rightarrow \gamma_c^*)}{A(\gamma_c^* \rightarrow \gamma_c^0)} = \exp \left[(1 - \lambda) \sum_{j \in \bar{c}} (-1)^{\gamma_j} \left(\sum_{k \in c_1} J_{jk} - \sum_{l \in c_0} J_{jl} \right) + \sum_{j \in c_0} h_j^* - \sum_{j \in c_1} h_j^* \right],$$

and the acceptance probability for move from γ_c^0 to γ_c^* is

$$\alpha(\gamma_c^0 \rightarrow \gamma_c^*) = \min \left\{ \frac{A(\gamma_c^0 \rightarrow \gamma_c^*)}{A(\gamma_c^* \rightarrow \gamma_c^0)}, 1 \right\}.$$

As well as satisfying the detailed balance, the algorithm also guarantees the ergodicity by the fact that there is always a finite chance that any spin will be chosen as the sole member of cluster of one, which is then flipped. The appropriate succession of such moves will get us from any state to any other in a finite time as ergodicity requires.

A.2 Derive of Expression (2)

First some extra notations are introduced here. Since $\mathbf{x}_j$'s are standardized, $C = X^T X = [\mathbf{c}_1, \dots, \mathbf{c}_p]$ is the correlation matrix of $\mathbf{x}_j$'s and $\mathbf{c}_j$'s stands for the vector of the correlation between $\mathbf{x}_j$ with all predictors. For the orthogonal data set, $C = I_n$ or $\mathbf{x}_i^T \mathbf{x}_j = \delta_{ij}$ and $\mathbf{x}_j^T \boldsymbol{\epsilon} = 0, i, j = 1, \dots, p$. The projection of $\mathbf{y}$ on $\mathbf{x}_j$ can be expressed as $\mathbf{a} = X^T \mathbf{y} = (\mathbf{x}_1^T \mathbf{y}, \dots, \mathbf{x}_p^T \mathbf{y})^T = (a_1, \dots, a_p)^T$, which are the estimates of the signals of the β_j 's under an orthogonal design. We may also need the notation $\mathbf{c}_{\gamma_{\bar{j}}} = \gamma_{\bar{j}} \circ \{\mathbf{c}_j\}_{\bar{j}}$, where "o" stands for the pointwise product of two vectors, " $\bar{j}$ " stands for the j th element removed for corresponding vectors and matrices,

Then define $\mathbf{y}^* = \mathbf{y} - \sum_{k \neq j} \gamma_k \mathbf{x}_k \beta_k$, and integrate out β_j and $\boldsymbol{\beta}_{\bar{j}}$ separately in following expression

$$\begin{aligned} p(\boldsymbol{\gamma}|\mathbf{y}, \boldsymbol{\tau}, b) &\propto \int p(\mathbf{y}|\boldsymbol{\gamma}, \boldsymbol{\beta}) p(\boldsymbol{\beta}|\boldsymbol{\tau}, b) d\boldsymbol{\beta} \\ &\propto \int \exp \left[-\frac{1}{2} (\mathbf{y}^* - \gamma_j \mathbf{x}_j \beta_j)^2 \right] p(\beta_j|\tau_j, b) d\beta_j \prod_{k \neq j} p(\beta_k|\tau_k, b) d\beta_k \\ &\propto \exp \left[\frac{\gamma_j}{2} a_j^2 \left(\gamma_j + \frac{\tau_j}{b^2} \right)^{-1} \right] \left(\frac{\tau_j/b^2}{\gamma_j + \tau_j/b^2} \right)^{1/2} \cdot \xi(\gamma_j, \kappa_j, \boldsymbol{\gamma}_{\bar{j}}, \boldsymbol{\tau}_{\bar{j}}), \end{aligned} \quad (\text{A.2.1})$$

where $\xi(\gamma_j, \kappa_j, \boldsymbol{\gamma}_{\bar{j}}, \boldsymbol{\tau}_{\bar{j}})$ is calculated by integrating out $\boldsymbol{\beta}_{\bar{j}}$:

$$\begin{aligned} \xi(\gamma_j, \kappa_j, \boldsymbol{\gamma}_{\bar{j}}, \boldsymbol{\tau}_{\bar{j}}) &\propto \int \exp \left[\frac{\gamma_j}{2} \left(\gamma_j + \frac{\tau_j}{b^2} \right)^{-1} \left(\sum_{k \neq j; l \neq j} \beta_k \gamma_k \mathbf{x}_k^T \mathbf{x}_j \mathbf{x}_j^T \mathbf{x}_l \beta_l \gamma_l - 2 \sum_{k \neq j} \mathbf{x}_k^T \mathbf{x}_j \mathbf{x}_j^T \mathbf{y} \beta_k \gamma_k \right) \right] \\ &\quad \times \exp \left(-\frac{\mathbf{y}^{*T} \mathbf{y}^*}{2} \right) \prod_{k \neq j} p(\beta_k|\tau_k, b) d\beta_k \\ &\propto \int \exp \left[\frac{\gamma_j}{2} \left(\gamma_j + \frac{\tau_j}{b^2} \right)^{-1} \left(\boldsymbol{\beta}_{\bar{j}}^T (\mathbf{c}_{\gamma_{\bar{j}}} \mathbf{c}_{\gamma_{\bar{j}}}^T) \boldsymbol{\beta}_{\bar{j}} - 2 a_j \mathbf{c}_{\gamma_{\bar{j}}}^T \boldsymbol{\beta}_{\bar{j}} \right) \right] \\ &\quad \times \exp \left[-\frac{1}{2} \left(\mathbf{y} - X_{\gamma_{\bar{j}}} \boldsymbol{\beta}_{\bar{j}} \right)^2 \right] p(\boldsymbol{\beta}_{\bar{j}}|\boldsymbol{\tau}_{\bar{j}}, b) d\boldsymbol{\beta}_{\bar{j}} \\ &\propto \exp \left[\frac{1}{2} \left(\mathbf{a}_{\bar{j}} - (1 - \kappa_j^{\gamma_j}) a_j \mathbf{c}_{\gamma_{\bar{j}}} \right)^T \Omega_j^{-1} \left(\mathbf{a}_{\bar{j}} - (1 - \kappa_j^{\gamma_j}) a_j \mathbf{c}_{\gamma_{\bar{j}}} \right) \right] |\Omega_j|^{1/2} |D_{\bar{j}}|^{1/2}, \end{aligned} \quad (\text{A.2.2})$$

where $\Omega_j = [D_{\bar{j}} + X_{\gamma_{\bar{j}}}^T X_{\gamma_{\bar{j}}} - (1 - \kappa_j^{\gamma_j}) (\mathbf{c}_{\gamma_{\bar{j}}} \mathbf{c}_{\gamma_{\bar{j}}}^T)]^{-1}$ and is easy to show it is positive definite. We also used the identity $\gamma_j / (\gamma_j + \frac{\tau_j}{b^2}) = (1 - \kappa_j^{\gamma_j})$. Note if $\gamma_j = 0$, $\xi(\dots)$ does not depend on κ_j or τ_j .

Then by definition and above expressions,

$$\begin{aligned}
\pi_j^b &= \frac{\int \sum_{\gamma_{\bar{j}} \in \{0,1\}^{p-1}} p(\gamma_j = 1, \gamma_{\bar{j}} | \mathbf{y}, \boldsymbol{\tau}, b) p(\boldsymbol{\tau}) d\boldsymbol{\tau}}{\int \sum_{\gamma_{\bar{j}} \in \{0,1\}^{p-1}} p(\gamma_j = 0, \gamma_{\bar{j}} | \mathbf{y}, \boldsymbol{\tau}, b) p(\boldsymbol{\tau}) d\boldsymbol{\tau}} \\
&= \int \pi_j \cdot \frac{\int \sum_{\gamma_{\bar{j}} \in \{0,1\}^{p-1}} \xi(\gamma_j = 1, \kappa_j, \gamma_{\bar{j}}, \boldsymbol{\tau}_{\bar{j}}) p(\boldsymbol{\tau}_{\bar{j}}) d\boldsymbol{\tau}_{\bar{j}}}{\int \sum_{\gamma_{\bar{j}} \in \{0,1\}^{p-1}} \xi(\gamma_j = 0, \kappa_j, \gamma_{\bar{j}}, \boldsymbol{\tau}_{\bar{j}}) p(\boldsymbol{\tau}_{\bar{j}}) d\boldsymbol{\tau}_{\bar{j}}} \cdot p(\tau_j) d\tau_j \\
&= \int \pi_j \xi_j p(\kappa_j) d\kappa_j
\end{aligned} \tag{A.2.3}$$

with $\xi_j = \frac{\int \sum_{\gamma_{\bar{j}} \in \{0,1\}^{p-1}} \xi(\gamma_j=1, \kappa_j, \gamma_{\bar{j}}, \boldsymbol{\tau}_{\bar{j}}) p(\boldsymbol{\tau}_{\bar{j}}) d\boldsymbol{\tau}_{\bar{j}}}{\int \sum_{\gamma_{\bar{j}} \in \{0,1\}^{p-1}} \xi(\gamma_j=0, \kappa_j, \gamma_{\bar{j}}, \boldsymbol{\tau}_{\bar{j}}) p(\boldsymbol{\tau}_{\bar{j}}) d\boldsymbol{\tau}_{\bar{j}}}$ and $\bar{j} = \{1, 2, \dots, p\} \setminus j$. It is easy to show that $\xi_j = 1$ for orthogonal design since $\mathbf{c}_{\gamma_{\bar{j}}} = \mathbf{0}$ for all j .

Obviously, $\pi_j \rightarrow 1$ as $\kappa_j \rightarrow 1$ and $\pi_j \rightarrow 0$ as $\kappa_j \rightarrow 0$. For π_j^b , it is more convenient using $\pi_j^b = \int \pi_j \xi_j p(\tau_j) d\tau_j$, where π_j and ξ_j are measurable functions indexed by b . Both π_j and ξ_j are bounded by some positive number for all b . When $b \rightarrow 0$, $\lim_{b \rightarrow 0} \pi_j \rightarrow 1$ and $\lim_{b \rightarrow 0} \xi_j \rightarrow 1$, thus according to Lebesgue's Dominated Convergence Theorem (DCT), $\lim_{b \rightarrow 0} \pi_j^b = \lim_{b \rightarrow 0} \int \pi_j \xi_j p(\tau_j) d\tau_j = \int \lim_{b \rightarrow 0} (\pi_j \xi_j) p(\tau_j) d\tau_j = 1$. When $b \rightarrow \infty$, $\lim_{b \rightarrow \infty} \pi_j \rightarrow 0$ and $\lim_{b \rightarrow \infty} \xi_j$ equal to some finite number. Again the limit and integral commute by DCT, thus we have $\lim_{b \rightarrow \infty} \pi_j^b = \int \lim_{b \rightarrow \infty} (\pi_j \xi_j) p(\tau_j) d\tau_j = 0$.

A.3 The Calculation of π_j^b

For orthogonal designs, π_j^b with Laplace prior can be integrated out directly from Equation (2).

$$\begin{aligned}
\pi_j^b &= \int_0^1 \kappa_j^{\frac{1}{2}} \exp \left[\frac{a_j^2}{2} (1 - \kappa_j) \right] \cdot \frac{1}{2b^2} \kappa_j^{-2} \exp \left(-\frac{1 - \kappa_j}{2b^2 \kappa_j} \right) d\kappa_j \\
&= \frac{1}{2b^2} \left(\frac{2\pi}{\lambda} \right)^{1/2} \exp \left(\frac{1}{2b^2} + \frac{a_j^2}{2} - \frac{\sqrt{a_j^2}}{b} \right) \int_0^1 \left(\frac{\lambda}{2\pi} \right)^{1/2} \kappa_j^{-3/2} \exp \left[-\frac{\lambda(\kappa_j - \mu)^2}{2\mu^2 \kappa_j} \right] d\kappa_j,
\end{aligned} \tag{A.3.1}$$

where $\lambda = 1/b^2$ and $\mu = \sqrt{1/(b^2 a_j^2)}$, and the expression in the integral is the CDF of inverse Gaussian distribution. Borrowing the expression of the CDF of the inverse Gaussian, we then integrated out the integral to get expression (3).

π_j^b with horseshoe prior for orthogonal design can also be derived directly from Equation (2):

$$\begin{aligned}
\pi_j^b &= \int_0^1 \kappa_j^{\frac{1}{2}} \exp \left[\frac{a_j^2}{2} (1 - \kappa_j) \right] \cdot \frac{b}{\pi} \kappa_j^{-1/2} (1 - \kappa_j)^{-1/2} (1 - \kappa_j + b^2 \kappa_j)^{-1} d\kappa_j \\
&= \frac{1}{\pi b} \exp \left(\frac{a_j^2}{2} \right) \int_0^1 \kappa_j^{1-1} (1 - \kappa_j)^{1/2-1} [(1 - b^{-2}) \kappa_j + b^{-2}]^{-1} \exp \left(-\frac{a_j^2}{2} \kappa_j \right) d\kappa_j,
\end{aligned} \tag{A.3.2}$$

where the expression in the integral is the transformation of the hypergeometric inverted-beta distribution which was shown to be represented by degenerate hypergeometric functions (Gordy, 1998; Polson and Scott, 2010), thus we can follow Polson and Scott (2010) to express it as (4).

A.4 Lancaster and Šalkauskas Basis for Natural Cubic Spline

In this paper, we follow Chib and Greenberg (2010) to employ the cubic spline LS basis described by Lancaster and Šalkauskas (1986). Consider the j th function $f_j(x)$, and let $\boldsymbol{\nu}_j = (\nu_{1j}, \dots, \nu_{K_j j})$ be the set of $100 \times \frac{k-1}{K_j-1} \%$, $k = 1, \dots, K_j$ quantile of x_{ij} , $i = 1, \dots, n$. Thus $\nu_{1j} = \min_i(x_{ij})$ and $\nu_{K_j j} = \max_i(x_{ij})$. K_j denotes the number of knots for the spline functions. Then the cubic spline expansion of $f_j(x)$ is expressed as

$$\begin{aligned} f_j(x_{ij}) &= \sum_{k=1}^{K_j} [\Phi_{kj}(x_{ij})g_{kj} + \Psi_{kj}(x_{ij})s_{kj}] \\ &= \Phi_j(x_{ij})^T \mathbf{g}_j + \Psi_j(x_{ij})^T \mathbf{s}_j, \end{aligned} \quad (\text{A.4.1})$$

where $\mathbf{g}_j = (g_{1j}, \dots, g_{K_j j})^T$ and $\mathbf{s}_j = (s_{1j}, \dots, s_{K_j j})^T$ are the coefficients of this expression, $\Phi_j(x_{ij}) = [\Phi_{1j}(x_{ij}), \dots, \Phi_{K_j j}(x_{ij})]^T$ and $\Psi_j(x_{ij}) = [\Psi_{1j}(x_{ij}), \dots, \Psi_{K_j j}(x_{ij})]^T$ are two basis vectors, and the basis functions $\{\Phi_{kj}(x)\}_{k=1}^{K_j}$ and $\{\Psi_{kj}(x)\}_{k=1}^{K_j}$ are defined as

$$\begin{aligned} \Phi_{kj}(x) &\propto \begin{cases} 0, & x < \nu_{k-1,j} \\ -(2/h_{kj}^3)(x - \nu_{k-1,j})^2(x - \nu_{kj} - 0.5h_{kj}), & \nu_{k-1,j} \leq x < \nu_{kj} \\ (2/h_{k+1,j}^3)(x - \nu_{k+1,j})^2(x - \nu_{kj} + 0.5h_{k+1,j}), & \nu_{kj} \leq x < \nu_{k+1,j} \\ 0 & x \geq \nu_{k+1,j}, \end{cases} \\ \Psi_{kj}(x) &\propto \begin{cases} 0, & x < \nu_{k-1,j} \\ (1/h_{kj}^2)(x - \nu_{k-1,j})^2(x - \nu_{kj}), & \nu_{k-1,j} \leq x < \nu_{kj} \\ (1/h_{k+1,j}^2)(x - \nu_{k+1,j})^2(x - \nu_{kj}), & \nu_{kj} \leq x < \nu_{k+1,j} \\ 0 & x \geq \nu_{k+1,j}, \end{cases} \end{aligned} \quad (\text{A.4.2})$$

where $h_{kj} = \nu_{kj} - \nu_{k-1,j}$. Note that Φ_{1j}, Ψ_{1j} and $\Phi_{K_j j}, \Psi_{K_j j}$ are defined by last two lines and first two lines of above expressions respectively. $\mathbf{g}_j$ and $\mathbf{s}_j$ are interpreted the ordinate and slope of $f_j(x)$. Since $f_j(x)$ is a natural cubic splines with the second derivative equal to zero at two end points, and continuous derivative at knot points, both $\mathbf{g}_j$ and $\mathbf{s}_j$ are constrained (Lancaster and Šalkauskas, 1986) by $\mathbf{s}_j = A_j^{-1} C_j \mathbf{g}_j$, where

$$A_j = \begin{pmatrix} 2 & 1 & 0 & 0 & 0 & \dots & 0 & 0 & 0 \\ \omega_{2j} & 2 & \mu_{2j} & 0 & 0 & \dots & 0 & 0 & 0 \\ 0 & \omega_{2j} & 2 & \mu_{2j} & 0 & \dots & 0 & 0 & 0 \\ \vdots & \dots & \ddots & \ddots & \ddots & \dots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & 0 & \dots & \omega_{K_j-1,j} & 2 & \mu_{K_j-1,j} \\ 0 & 0 & 0 & 0 & 0 & \dots & 0 & 1 & 2 \end{pmatrix},$$

and

$$C_j = \begin{pmatrix} -\frac{1}{h_{2j}} & \frac{1}{h_{2j}} & 0 & 0 & \cdots & 0 & 0 & 0 \\ -\frac{\omega_{2j}}{h_{2j}} & \frac{\omega_{2j}}{h_{2j}} - \frac{\mu_{2j}}{h_{3j}} & \frac{\mu_{2j}}{h_{3j}} & 0 & \cdots & 0 & 0 & 0 \\ 0 & -\frac{\omega_{3j}}{h_{3j}} & \frac{\omega_{3j}}{h_{3j}} - \frac{\mu_{3j}}{h_{4j}} & \frac{\mu_{3j}}{h_{4j}} & \cdots & 0 & 0 & 0 \\ \vdots & \cdots & \ddots & \ddots & \cdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & -\frac{\omega_{K_j-1,j}}{h_{K_j-1}} & \frac{\omega_{K_j-1,j}}{h_{K_j-1}} - \frac{\mu_{K_j-1,j}}{h_{K_j}} & \frac{\mu_{K_j-1,j}}{h_{K_j}} \\ 0 & 0 & 0 & 0 & \cdots & 0 & -\frac{1}{h_{K_j}} & \frac{1}{h_{K_j}} \end{pmatrix},$$

where $\omega_{kj} = h_{kj}/(h_{kj} + h_{k+1,j})$ and $\mu_{kj} = 1 - \omega_{kj}$ for $k = 2, \dots, K_j$. With this constraints, $\mathbf{s}_j$ can be replaced from the function expression (A.4.1),

$$\begin{aligned} f_j(x_{ij}) &= [\Phi_j(x_{ij})^T + \Psi_j(x_{ij})^T A_j^{-1} C_j] \mathbf{g}_j \\ &= \mathbf{t}_j^T(x_{ij}) \mathbf{g}_j, \end{aligned} \quad (\text{A.4.3})$$

where $\mathbf{t}_j^T(x_{ij}) = (t_{1j}(x_{ij}), \dots, t_{K_j j}(x_{ij})) = \Phi_j(x_{ij})^T + \Psi_j(x_{ij})^T A_j^{-1} C_j$. Furthermore, consider the identifying constraints, $\sum_k g_{kj} = 0$, we can express $g_{1j} = -(g_{2j} + \dots + g_{K_j j})$, thus

$$f_j(x_{ij}) = \mathbf{t}_j^T(x_{ij}) \mathbf{g}_j = [t_{2j}(x_{ij}) - t_{1j}(x_{ij})]g_{2j} + \dots + [t_{K_j j}(x_{ij}) - t_{1j}(x_{ij})]g_{K_j j} = \mathbf{z}_j^{*T}(x_{ij}) \boldsymbol{\beta}_j,$$

where $\boldsymbol{\beta}_j = (g_{2j}, \dots, g_{K_j j})^T$ and we define matrix

$$Z_j^* = \begin{pmatrix} \mathbf{z}_j^{*T}(x_{1j}) \\ \vdots \\ \mathbf{z}_j^{*T}(x_{n_j}) \end{pmatrix}.$$

Now the j th nonparametric function expressed by the natural cubic spline basis is $f_j(\mathbf{x}_j) = Z_j^* \boldsymbol{\beta}_j$. In order to incorporate the assumption of a priori smoothness, Chib and Greenberg (2010) consider a prior distribution on $\boldsymbol{\beta}_j$'s as,

$$[\boldsymbol{\beta}_j | \sigma_{ej}^2, \sigma_{dj}^2] \sim N[\mathbf{0}, \Delta_j^{-1} T_j (\Delta_j^{-1})^T], \quad (\text{A.4.4})$$

where N is the $K_j - 1$ dimensional multivariate normal distribution, and

$$T_j = \begin{pmatrix} \sigma_{ej}^2 & 0 & 0 \\ 0 & \sigma_{dj}^2 I_{K_j-3} & 0 \\ 0 & 0 & \sigma_{ej}^2 \end{pmatrix},$$

where two variance components σ_{ej}^2 and σ_{dj}^2 are selected here because of the different normal assumptions for the differences of the ordinates and the differences of slopes. Δ_j is given by

$$\Delta_j = \begin{pmatrix} \frac{2}{h_{2j}} & \frac{1}{h_{2j}} & \frac{1}{h_{2j}} & \frac{1}{h_{2j}} & \frac{1}{h_{2j}} & \cdots & \frac{1}{h_{2j}} & \frac{1}{h_{2j}} \\ \frac{1}{h_{2j}} & -\left(\frac{1}{h_{2j}} + \frac{1}{h_{3j}}\right) & \frac{1}{h_{3j}} & 0 & 0 & \cdots & 0 & 0 \\ 0 & \frac{1}{h_{3j}} & -\left(\frac{1}{h_{3j}} + \frac{1}{h_{4j}}\right) & \frac{1}{h_{4j}} & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \cdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & 0 & \cdots & -\frac{1}{h_{K_j j}} & \frac{1}{h_{M_j j}} \end{pmatrix}.$$

So far the construction of function $f_j(\mathbf{x}_j)$ is exactly the same as Chib and Greenberg (2010). Note that $f_j(\mathbf{x}_j) = Z_j^* \boldsymbol{\beta}_j$ with the prior of $\boldsymbol{\beta}_j$ given by (A.4.4) is equivalent to have $f_j(\mathbf{x}_j) = Z_j^* \Delta_j^{-1} \boldsymbol{\beta}_j$ with $[\boldsymbol{\beta}_j] \sim N(\mathbf{0}, T_j)$. Henceforth, we define the final $n \times M_j$ basis matrix $Z_j = Z_j^* \Delta_j^{-1}$ such that $f_j(\mathbf{x}_j) = Z_j \boldsymbol{\beta}_j$, where $M_j = K_j - 1$. Define $\tau_{ej} = \sigma_{ej}^{-2}$ and $\tau_{bj} = \sigma_{bj}^{-2}$, and modify the one variance component prior algorithm in Section 4.3 of the main paper, then we can easily employ the LS basis into BSAM.

Appendix B Tables and Figures

Table 1: Summary of the Cauchy, Laplace and horseshoe priors for the marginal prior of β_j 's, corresponding to priors for $p(\tau_j)$ and the density functions of the shrinkage parameter κ_j .

Marginal prior	$p(\beta_j b)$	Prior for τ_j	Distribution for κ_j
Cauchy	$\pi b(\beta_j^2 + b^2)^{-1}$	$\tau_j^{-\frac{1}{2}} \exp(-\frac{\tau_j}{2})$	$\kappa_j^{-\frac{1}{2}} (1 - \kappa_j)^{-\frac{3}{2}} \exp(-\frac{b^2 \kappa_j}{2(1-\kappa_j)})$
Laplace	$(2b)^{-1} \exp(- \beta_j /b)$	$\tau_j^{-2} \exp(-\frac{2}{\tau_j})$	$\kappa_j^{-2} \exp(-\frac{(1-\kappa_j)}{2b^2 \kappa_j})$
Horseshoe	-	$\tau_j^{-\frac{1}{2}} (1 + \tau_j)^{-1}$	$\kappa_j^{-\frac{1}{2}} (1 - \kappa_j)^{-\frac{1}{2}} [1 - \kappa_j + b^2 \kappa_j]^{-1}$

Table 2: Parameter estimation of ozone data under Bayesian sparse additive model at $b = 1.6$. In this example we include the intercept term μ in model (7) of the main paper and assign a prior for μ : $[\mu] \sim N(0, \tau_\mu^{-1})$, and a Gamma prior for τ_μ : $[\tau_\mu] \sim G(4, 2)$. Note how the posterior means of these parameters adapt to the data. Especially interesting is that the τ_{ej} 's and τ_{dj} 's all start with the same prior, but the analyses give different posterior means. For the first group of variables, the posterior means for τ_{ej} 's and τ_{dj} 's are obviously different from their priors.

Parameter	Prior		Posterior				
	Mean	Std.Dev.	Mean	Std.Dev.	Median	Lower 2.5%	Upper 2.5%
τ_e^{Temp}	1.000	1.414	0.671	1.192	0.150	0.003	4.199
τ_e^{Press}	1.000	1.414	0.787	1.270	0.282	0.008	4.570
τ_e^{InvHt}	1.000	1.414	0.198	0.597	0.021	0.000	1.700
τ_e^{Vis}	1.000	1.414	1.316	1.517	0.805	0.021	5.485
τ_e^{InvTp}	1.000	1.414	1.010	1.365	0.500	0.010	4.884
τ_e^{Hgt}	1.000	1.414	1.062	1.392	0.547	0.005	5.032
τ_e^{Hum}	1.000	1.414	1.050	1.431	0.509	0.001	5.129
τ_e^{WS}	1.000	1.414	1.104	1.505	0.545	0.001	5.334
τ_d^{Temp}	1.000	1.414	0.226	0.724	0.026	0.002	2.212
τ_d^{Press}	1.000	1.414	0.033	0.047	0.021	0.002	0.125
τ_d^{InvHt}	1.000	1.414	0.889	1.352	0.349	0.004	4.731
τ_d^{Vis}	1.000	1.414	0.751	1.235	0.227	0.004	4.226
τ_d^{InvTp}	1.000	1.414	1.122	1.413	0.611	0.005	5.227
τ_d^{Hgt}	1.000	1.414	1.045	1.429	0.521	0.009	5.140
τ_d^{Hum}	1.000	1.414	1.061	1.474	0.513	0.001	5.048
τ_d^{WS}	1.000	1.414	0.949	1.383	0.404	0.000	4.985
ϕ	-	-	6.602	0.542	6.585	5.599	7.681
μ	0.000	$\sqrt{\tau_\mu}$	2.143	0.066	2.145	2.011	2.265
τ_μ	2.000	1.000	1.043	0.495	0.971	0.316	2.205

Figure 1: Marginal prior density functions of β_j and density functions of κ_j for different b .

Figure 2: Selection probability curves against κ_j (a). The curves of marginal selection probability against global shrinkage parameter b (b). Marginal selection probabilities with baseline subtracted (c). All plots are under orthogonal designs.

Figure 3: Diagram of variable selection as a random graph model with selected nodes (filled circles), excluded nodes (circles), edges of positive interaction (black lines), and edges of negative interaction (red lines). Independent variable selection: no interactions among nodes (a). General variable selection: a complete graph (b).

Figure 4: Diagram of the cluster algorithm. Forming the cluster (a-c). Flipping clustered nodes (d-e). Figure (a-c) demonstrate the clustering process. First we randomly select a seed node (in this diagram, we use node 8). Then we throw the bond to all neighbors of node 8, and find that node 5 is bonded to 8 with probability $p_{a,5}$, forming the cluster (the dashed line is turned into solid lines, which indicates that 5 is added to the cluster). We scan the remaining nodes again, but whether or not a new node should be added is determined by the bonding between the new node and node 5 and 8. For example in Figure (b), the bond between the new node 4 and the cluster is composed of the bonds 4-5 and 4-8. In Figure (c), after adding the last new node 1 to the cluster, we scan all the remaining nodes and find no new node added, which then terminates the clustering process. The cluster formed contains nodes $c = \{8, 5, 4, 2, 1\}$. To flip these nodes, we have to cut off the bonding of the cluster with all other nodes in $\bar{c}$ because in a complete graph the neighbors of a cluster are all the other nodes outside of the cluster. For example, the bond between the cluster and node 10 is demonstrated in Figure (d), where we can see the bonds between 10 and all nodes in the cluster should be cut to flip the cluster. Thus to completely flip the cluster, the bonds between all other nodes in $\bar{c}$ and the nodes in c should be cut. Similarly, in the reverse process to flip the cluster to its original configuration, as shown in Figure (e), all the bonds between cluster c and the nodes in $\bar{c}$ must be cut.

Figure 5: The profile curves of the selection probability of simulation model (20) with different priors for large signal setting (a-c), and small signal setting (d-f).

Figure 6: The profile curves of the selection probability of simulation model (7) in the main paper with $p = 80, n = 100$ for independent setting ($t=0$) (a), and correlated setting ($t=1$) (b).

Figure 7: True function f_j 's (blue dashed lines) and estimated function $\hat{f}_j$'s (blue solid lines) with 95% credible interval (red dashed lines) for the 4 true nodes (a-d) and a noise node (e) of a run of simulation model (7) of the main paper with independent setting $t = 0$ and $p = 80$. The marginal selection probability at $b = 26$ (f). Note we reordered the first 4 true nodes number to (2, 20, 50, 70) for a better view.

Figure 8: True function f_j 's (blue dashed lines) and estimated function $\hat{f}_j$'s (blue solid lines) with 95% credible interval (red dashed lines) for the 4 true nodes (a-d) and a noise node (e) of a run of simulation model (7) of the main paper with independent setting $t = 1$ and $p = 80$. The marginal selection probability at $b = 26$ (f). Note we reordered the first 4 true nodes number to (2, 20, 50, 70) for a better view.

Figure 9: Estimated function $\hat{f}_j$ (blue solid lines) with 95% credible interval (red dashed lines) for the 8 predictors of ozone data labeled by the marginal selection probability $P = p(\gamma_j = 1 | \mathbf{y})$ at $b = 1.6$.

References

- Carvalho, C. and Polson, N. G. (2010). The Horseshoe Estimator for Sparse Signals. *Biometrika*, **97**, 465-480.
- Chib, S. and Greenberg, E. (2010). Additive Cubic Spline Regression with Dirichlet Process Mixture Errors. *Journal of Econometrics*, **156**, 322-336.
- Gelman, A. (2006). Prior Distribution for Variance Parameters in Hierarchical Models. *Bayesian Analysis*, **1**, 515-533.
- Gordy, M. B. (1998). A generalization of Generalized Beta Distribution. In *Finance and Economics Discussion Series*. Board of Governors of the Federal Reserve System.
- Lancaster, P. and Šalkauskas, K. (1986). *Curve and Surface Fitting: an Introduction*. San Diego: Academic Press.
- Lin, Y. and Zhang, H. H. (2006). Component Selection and Smoothing in Multivariate Nonparametric Regression. *The Annals of Statistics*, **34**, 2272-2297.
- Polson, N. G. and Scott, J. G. (2010). On the Half-Cauchy Prior for a Global Scale Parameter. *Technical report*, University of Texas at Austin.
- Polson, N. G. and Scott, J. G. (2011). Shrink Globally, Act Locally: Sparse Bayesian Regularization and Prediction. In *Proceedings of the 9th Valencia World Meeting on Bayesian Statistics*. Oxford University Press, 501-538.