Japanese Journal of Statistics and Data Science (2023) 6:705–727
https://doi.org/10.1007/s42081-023-00213-2

ORIGINAL PAPER

Bayesian fused lasso modeling via horseshoe prior

Yuko Kakikawa[1] · **Kaito Shimamura**[2] · **Shuichi Kawano**[3]

Received: 10 February 2023 / Revised: 18 July 2023 / Accepted: 26 July 2023 /
Published online: 21 August 2023

Abstract
Bayesian fused lasso is one of the sparse Bayesian methods, which shrinks both regression coefficients and their successive differences simultaneously. In this paper, we propose a Bayesian fused lasso modeling via horseshoe prior. By assuming a horseshoe prior on the difference of successive regression coefficients, the proposed method enables us to prevent over-shrinkage of those differences. We also propose a Bayesian nearly hexagonal operator for regression with shrinkage and equality selection with horseshoe prior, which imposes priors on all combinations of differences of regression coefficients. Simulation studies and an application to real data show that the proposed method gives better performance than existing methods.

Keywords Fusion of coefficients · Hierarchical Bayesian model · Horseshoe prior · Markov chain Monte Carlo

1 Introduction

Recently, a wide variety of data have come to be used in statistical analysis. Especially, the analysis of high-dimensional data, such as image data and financial data is taking on added significance. To handle these data, it is important to perform variable selection and variable fusion, which correspond to extracting relevant variables and capturing the group structure of data, respectively. To this end, sparse regularization methods such as lasso (Tibshirani, 1996), fused lasso (Tibshirani et al., 2005), and a hexagonal operator for regression with shrinkage and equality selection (HORSES) (Jang et al.,

✉ Yuko Kakikawa
kakikawa.yuko@ism.ac.jp

1 Department of Statistical Science, Graduate University of Advanced Studies (SOKENDAI), 10-3 Midori-cho, Tachikawa, Tokyo 190-8562, Japan

2 NTT Advanced Technology Corporation, Tokyo Opera City Tower, 3-20-2, Nishi-shinjuku, Shinjuku-ku, Tokyo 163-1436, Japan

3 Faculty of Mathematics, Kyushu University, 744 Motooka, Nishi-ku, Fukuoka 819-0395, Japan

2015) have been proposed. These methods allow us to execute variable selection and variable fusion by the estimation of regression coefficients.

Meanwhile, many Bayesian approaches to these regularization methods, in which priors on regression coefficients correspond to regularization terms, have also been proposed. For example, Park and Casella (2008) proposed Bayesian lasso, which shrinks regression coefficients by assuming that they follow a Laplace distribution. Furthermore, Park and Casella (2008) developed a Gibbs sampling using a hierarchical expression of the Laplace distribution. Castillo et al. (2015) proposed Bayesian lasso with a quantification of uncertainty. Kyung et al. (2010) expanded Bayesian lasso by assuming Laplace distributions not only on regression coefficients but also on their successive differences, which is called Bayesian fused lasso.

A Laplace prior tends to shrink its targets, such as regression coefficients and their successive differences, too much. To overcome this problem, the Student-t prior and the normal-exponential-gamma (NEG) distribution (Griffin and Brown, 2005), which have heavier tails than a Laplace prior, have also been used. Song and Cheng (2020) proposed using a Student-t prior to construct Bayesian fusion models. Shimamura et al. (2019) proposed Bayesian fused lasso based on the hierarchical expression of an NEG prior. In addition, a horseshoe prior (Carvalho et al., 2010) is also often used instead of a Laplace prior. A horseshoe prior has an infinite spike at zero and a Cauchy-like tail, which leads to simultaneous weak shrinkage on non-zero elements and strong shrinkage on exactly zero ones. Makalic and Schmidt (2015) introduced a linear regression model in which a horseshoe prior is assumed on the regression coefficients and developed a simple Gibbs sampler for it. There are also many studies for regression models with horseshoe prior (Bhattacharya et al. (2016); Johndrow et al. (2020); Lee et al. (2021)). However, the existing methods assume a horseshoe prior on only the regression coefficients.

In this paper, we propose Bayesian fused lasso modeling with horseshoe prior under the framework of linear regression models. To formulate the Bayesian model, we assume a Laplace prior on the regression coefficients and a horseshoe prior on their successive differences. We also propose Bayesian nearly HORSES (nHORSES) with horseshoe prior, where the horseshoe prior is assumed on every pair of differences of regression coefficients. We develop a Gibbs sampler for the parameters by using the hierarchical expression of the half-Cauchy prior (Wand et al., 2011) shown by Makalic and Schmidt (2015). Through the proposal, we focus on grouping variables which have the similar role in the prediction more flexibly than the existing methods.

We note that Banerjee (2022) proposed imposing a horseshoe prior on differences of coefficients. However, Banerjee (2022) used the model assumed in the one-dimensional fused lasso signal approximation in Friedman et al. (2007), which is a special case of a linear regression model. In addition, Banerjee (2022) did not perform variable selection, unlike our proposed method.

The remainder of the paper is organized as follows. Section 2 describes the Bayesian models and introduces sparse Bayesian modelings with horseshoe prior. In Sect. 3, we propose Bayesian fused lasso and Bayesian nHORSES with horseshoe prior, and then develop Gibbs samplings for them. Section 4 presents Monte Carlo simulations and an application to real data to compare our proposed method with existing methods. We conclude our paper in Sect. 5.

2 Sparse Bayesian linear regression modeling

In this section, we review Bayesian linear regression, Bayesian lasso, and Bayesian fused lasso. We also describe Bayesian linear regression via horseshoe prior.

2.1 Preliminaries

Let $\boldsymbol{y} = (y_1, y_2, \ldots, y_n)^T$ be an n-dimensional vector of the response variable and $\boldsymbol{X} = (\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n)^T$ be an $n \times p$ design matrix. A linear regression model is formulated as

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon},$$

where $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_p)^T$ is a p-dimensional regression coefficient vector and $\boldsymbol{\epsilon} = (\epsilon_1, \epsilon_2, \ldots, \epsilon_n)^T$ is an n-dimensional error vector that is distributed as $\mathrm{N}_n(\boldsymbol{0}_n, \sigma^2 I_n)$. Here $\boldsymbol{0}_n$ is an n-dimensional vector of zeros, σ^2 is an error variance, and I_n is an $n \times n$ identity matrix. Without loss of generality, we suppose that the response variable is centered and the explanatory variable is standardized as follows:

$$\sum_{i=1}^{n} y_i = 0, \quad \sum_{i=1}^{n} x_{ij} = 0, \quad \sum_{i=1}^{n} x_{ij}^2 = n, \quad (j = 1, 2, \ldots, p).$$

Then, the likelihood function is given by

$$f(\boldsymbol{y} \mid \boldsymbol{X}; \boldsymbol{\beta}, \sigma^2) = \prod_{i=1}^{n} f(y_i \mid \boldsymbol{x}_i; \boldsymbol{\beta}, \sigma^2),$$

where

$$f(y_i \mid \boldsymbol{x}_i; \boldsymbol{\beta}, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{ -\frac{(y_i - \boldsymbol{x}_i^T \boldsymbol{\beta})^2}{2\sigma^2} \right\}.$$

2.2 Bayesian lasso

Tibshirani (1996) proposed lasso, which performs parameter estimation and variable selection simultaneously in terms of frequentist. He also mentioned that the lasso solution is identical to a posterior mode obtained by imposing the Laplace distribution on the parameter vector $\boldsymbol{\beta}$ as its prior.

Based on the perspective of Tibshirani (1996), Park and Casella (2008) established a Bayesian estimation for lasso. The estimation is called Bayesian lasso. Bayesian lasso considers a conditional Laplace prior in the form:

$$\begin{aligned} \pi(\boldsymbol{\beta} \mid \sigma^2) &= (\sigma^2)^{-\frac{p}{2}} \prod_{j=1}^{p} \mathrm{Laplace}\left(\frac{\beta_j}{\sqrt{\sigma^2}} \,\middle|\, 0, \lambda \right) \\ &= \prod_{j=1}^{p} \frac{\lambda}{2\sqrt{\sigma^2}} \exp\left(-\frac{\lambda}{\sqrt{\sigma^2}} |\beta_j| \right), \end{aligned} \tag{1}$$

where λ (> 0) is a hyper-parameter. Conditioning $\boldsymbol{\beta}$ on σ^2 makes the posterior distribution unimodal (for example, see Appendix A in Park and Casella (2008)).

The prior distribution in (1) can be rewritten as

$$\frac{\lambda}{2\sqrt{\sigma^2}}\exp\left(-\frac{\lambda}{\sqrt{\sigma^2}}|\beta|\right)=\int_0^\infty \frac{1}{\sqrt{2\pi\sigma^2\tau^2}}\exp\left(-\frac{\beta^2}{2\sigma^2\tau^2}\right)\frac{\lambda^2}{2}\exp\left(-\frac{\lambda^2}{2}\tau^2\right)d\tau^2$$

by using a scale mixture of normals (Andrews and Mallows, 1974). This equation means that the Laplace distribution is represented as the convolution of the following two distributions:

$$\pi(\boldsymbol{\beta}\mid\sigma^2,\tau_1^2,\ldots,\tau_p^2)=\prod_{j=1}^p\frac{1}{\sqrt{2\pi\sigma^2\tau_j^2}}\exp\left(-\frac{\beta_j^2}{2\sigma^2\tau_j^2}\right),$$

$$\pi(\tau_1^2,\ldots,\tau_p^2)=\prod_{j=1}^p\frac{\lambda^2}{2}\exp\left(-\frac{\lambda^2}{2}\tau_j^2\right).$$

For the parameter σ^2, the improper prior distribution $\pi(\sigma^2)\propto 1/\sigma^2$ or any inverse gamma distribution for σ^2 is assumed. Based on the likelihood and the prior distributions, a Gibbs sampling for Bayesian lasso is developed. We omit the Gibbs samplers. For details, we refer the reader to Park and Casella (2008).

2.3 Bayesian fused lasso

The fused lasso (Tibshirani et al., 2005) encourages sparsity in both the coefficients and their successive differences. Kyung et al. (2010) proposed Bayesian fused lasso as a Bayesian counterpart to fused lasso. Bayesian fused lasso assumes a prior distribution for $\boldsymbol{\beta}$ of the following form:

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

where λ_1 and λ_2 are positive hyper-parameters. Similar to Bayesian lasso, a scale mixture of normals is applied. Then the prior distribution (2) is transformed into

$$\begin{aligned}\pi(\boldsymbol{\beta}\mid\sigma^2)\propto(\sigma^2)^{-\frac{2p-1}{2}}&\prod_{j=1}^p\int\frac{1}{\sqrt{2\pi\tau_j^2}}\exp\left(-\frac{\beta_j^2}{2\sigma^2\tau_j^2}\right)\frac{\lambda_1^2}{2}\exp\left(-\frac{\lambda_1^2}{2}\tau_j^2\right)d\tau_j^2\\&\times\prod_{j=2}^p\int\frac{1}{\sqrt{2\pi\tilde{\tau}_j^2}}\exp\left\{-\frac{(\beta_j-\beta_{j-1})^2}{2\sigma^2\tilde{\tau}_j^2}\right\}\frac{\lambda_2^2}{2}\exp\left(-\frac{\lambda_2^2}{2}\tilde{\tau}_j^2\right)d\tilde{\tau}_j^2.\end{aligned}$$

Using this hierarchical relationship, Kyung et al. (2010) developed a Gibbs sampling for Bayesian fused lasso.

To perform the fully Bayesian estimation, Kyung et al. (2010) further assumed the gamma distribution for the hyper-parameters λ_1 and λ_2 as

$$\lambda_1^2 \sim \mathrm{Ga}(r_1, \delta_1),$$
$$\lambda_2^2 \sim \mathrm{Ga}(r_2, \delta_2),$$

where r_1, r_2, δ_1, and δ_2 are positive hyper-parameters. Here, the probability density function of the gamma distribution is given by

$$\mathrm{Ga}(x \mid m, c) = \frac{c^m}{\Gamma(m)} x^{m-1} \exp(-cx), \quad (x \geq 0),$$

where m is a shape parameter, c is a rate parameter, both taking positive values, and $\Gamma(\cdot)$ is the gamma function. In Kyung et al. (2010), $r_1 = 1, r_2 = 1, \delta_1 = 10$, and $\delta_2 = 10$ are used because the gamma distribution is relatively flat with these parameter values. We omit the full conditional posteriors and the Gibbs samplers. For details, we refer the reader to Kyung et al. (2010).

Next, we explain HORSES by Jang et al. (2015). HORSES was proposed as an extension of fused lasso; HORSES imposes an L_1 penalty on all combinations of differences of regression coefficients. In the Bayesian framework, this corresponds to assuming a Laplace prior of the form:

$$\pi(\boldsymbol{\beta} \mid \sigma^2) \propto (\sigma^2)^{-\frac{p^2+p}{4}} \prod_{j=1}^{p} \mathrm{Laplace}\left(\frac{\beta_j}{\sqrt{\sigma^2}} \,\middle|\, 0, \lambda_1\right) \times \prod_{j>k}^{p} \mathrm{Laplace}\left(\frac{\beta_j - \beta_k}{\sqrt{\sigma^2}} \,\middle|\, 0, \lambda_2\right)$$

for regression coefficients $\boldsymbol{\beta}$. Note that HORSES is also known as generalized fused lasso (She, 2010).

The Laplace distribution shrinks all of the regression coefficients to the same extent. Shimamura et al. (2019) proposed Bayesian fused lasso and Bayesian nHORSES with NEG prior. This method assumes a Laplace prior on the regression coefficients and an NEG prior on their differences. Because an NEG prior has two properties, a spike at zero and extreme flatness of its tail, the method with an NEG prior has the advantage that exactly identical regression coefficients tend to be estimated as identical, while different regression coefficients tend to be estimated as different.

2.4 Bayesian linear regression model with horseshoe prior

Makalic and Schmidt (2015) proposed the following Bayesian linear regression model:

$$\begin{aligned} \boldsymbol{y} \mid \boldsymbol{X}, \boldsymbol{\beta}, \sigma^2 &\sim \mathrm{N}_n(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2 \boldsymbol{I}_n), \\ \beta_j \mid \lambda_j^2, \tau^2, \sigma^2 &\sim \mathrm{N}(0, \lambda_j^2 \tau^2 \sigma^2), \\ \sigma^2 &\sim \sigma^{-2} d\sigma^2, \\ \lambda_j &\sim \mathrm{C}^+(0, 1), \\ \tau &\sim \mathrm{C}^+(0, 1). \end{aligned} \tag{3}$$

Here, $\mathrm{C}^+(0, a)$ $(a > 0)$ is a half-Cauchy distribution, which has the following density function:

$$p(x) = \frac{2a}{\pi \left(x^2 + a^2\right)}, \quad (x > 0).$$

The hierarchies of priors in (3) represent the horseshoe prior proposed in Carvalho et al. (2010). In the model with horseshoe prior, the half-Cauchy prior distribution is assumed on hyper-parameters λ_j and τ. Hyper-parameter λ_j adjusts the level of local shrinkage for regression coefficient β_j, while hyper-parameter τ determines the degree of global shrinkage for all regression coefficients. Owing to having these two types of hyper-parameters, the horseshoe prior simultaneously enjoys a heavy tail and infinitely tall spike at zero. These properties induce exactly identical regression coefficients to tend to be estimated as identical, while different regression coefficients tend to be estimated as different.

To develop a Gibbs sampling for the parameters, Makalic and Schmidt (2015) used a hierarchical expression of the half-Cauchy distribution (Wand et al., 2011), which means that x follows $\mathrm{C}^+(0, A)$ when x^2 and a have the following priors:

$$x^2 \mid a \sim \mathrm{IG}\left(\frac{1}{2}, \frac{1}{a}\right), \quad a \sim \mathrm{IG}\left(\frac{1}{2}, \frac{1}{A^2}\right), \tag{4}$$

where A is a positive constant. Here, $\mathrm{IG}(x \mid \nu_0, \eta_0)$ is the inverse gamma distribution whose probability density function is given by

$$\mathrm{IG}(x \mid \nu_0, \eta_0) = \frac{\eta_0^{\nu_0}}{\Gamma(\nu_0)} x^{-\nu_0 - 1} \exp\left(-\frac{\eta_0}{x}\right), \quad (x > 0),$$

where ν_0, η_0 (> 0) are hyper-parameters. Using (4), the priors of the model (3) can be expressed as follows:

$$\begin{aligned} \boldsymbol{y} \mid \boldsymbol{X}, \boldsymbol{\beta}, \sigma^2 &\sim \mathrm{N}_n(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2 \boldsymbol{I}_n), \\ \beta_j \mid \lambda_j^2, \tau^2, \sigma^2 &\sim \mathrm{N}(0, \lambda_j^2 \tau^2 \sigma^2), \\ \sigma^2 &\sim \sigma^{-2} d\sigma^2, \\ \lambda_j^2 \mid \nu_j &\sim \mathrm{IG}\left(\frac{1}{2}, \frac{1}{\nu_j}\right), \\ \tau^2 \mid \xi &\sim \mathrm{IG}\left(\frac{1}{2}, \frac{1}{\xi}\right), \\ \nu_1, \ldots, \nu_p, \xi &\sim \mathrm{IG}\left(\frac{1}{2}, 1\right). \end{aligned}$$

We omit the full conditional posteriors and the Gibbs samplers. For details, we refer the reader to Makalic and Schmidt (2015) and Nalenz and Villani (2018).

3 Proposed method

In this section, we propose the Bayesian linear regression modeling, which assumes the horseshoe prior on successive differences of regression coefficients. We also extend this approach to HORSES.

3.1 Bayesian fused lasso with horseshoe prior

We propose assuming a Laplace prior on regression coefficients and a horseshoe prior on their successive differences as follows:

$$\pi(\boldsymbol{\beta} \mid \sigma^2) \propto (\sigma^2)^{-\frac{p}{2}} \prod_{j=1}^{p} \text{Laplace}\left(\frac{\beta_j}{\sqrt{\sigma^2}} \,\middle|\, \tilde{\lambda}_1\right) \times \int \left[\prod_{j=2}^{p} \int \frac{1}{\sqrt{2\pi\lambda_j^2\tilde{\tau}^2\sigma^2}} \exp\left\{-\frac{(\beta_j-\beta_{j-1})^2}{2\lambda_j^2\tilde{\tau}^2\sigma^2}\right\} \frac{2}{\pi(1+\lambda_j^2)} d\lambda_j^2\right] \frac{2}{\pi(1+\tilde{\tau}^2)} d\tilde{\tau}^2. \tag{5}$$

By assuming the prior (5), small differences between successive regression coefficients are largely shrunk, while large differences are not much shrunk. Note that we assumed not a horseshoe prior but a Laplace prior on regression coefficients. It is because the MCMC chain for a model only with horseshoe prior does not converge in almost all cases. This problem might be related to the geometric ergodicity of MCMC chain. Therefore, we adopt the combination of Laplace prior and horseshoe prior to obtain a stable estimation.

Using a scale mixture of normals (Andrews and Mallows, 1974), the prior (5) can be expressed as follows:

$$\begin{aligned}\pi(\boldsymbol{\beta} \mid \sigma^2) \propto & \int \cdots \int (\sigma^2)^{-\frac{2p-1}{2}} (\tilde{\tau}^2)^{-\frac{p-1}{2}} \pi(\tilde{\tau}^2 \mid \xi) \\ & \times \pi(\xi) \prod_{j=1}^{p} (\tau_j^2)^{-\frac{1}{2}} \prod_{j=2}^{p} (\lambda_j^2)^{-\frac{1}{2}} \exp\left(-\frac{1}{2\sigma^2}\boldsymbol{\beta}^T \boldsymbol{B}^{-1}\boldsymbol{\beta}\right) \\ & \times \prod_{j=1}^{p} \pi(\tau_j^2) \prod_{j=2}^{p} \pi(\lambda_j^2 \mid \nu_j) \prod_{j=2}^{p} \pi(\nu_j) d\tilde{\tau}^2 d\xi \prod_{j=1}^{p} d\tau_j^2 \prod_{j=2}^{p} d\lambda_j^2 \prod_{j=2}^{p} d\nu_j. \end{aligned} \tag{6}$$

Here, the inverse of matrix $\boldsymbol{B}$ is represented by

$$\boldsymbol{B}^{-1} = \begin{pmatrix} \frac{1}{\tau_1^2} + \frac{1}{\lambda_2^2\tilde{\tau}^2} & -\frac{1}{\lambda_2^2\tilde{\tau}^2} & 0 & \cdots & 0 & 0 \\ -\frac{1}{\lambda_2^2\tilde{\tau}^2} & \frac{1}{\tau_2^2} + \frac{1}{\lambda_2^2\tilde{\tau}^2} + \frac{1}{\lambda_3^2\tilde{\tau}^2} & -\frac{1}{\lambda_3^2\tilde{\tau}^2} & \cdots & 0 & 0 \\ 0 & -\frac{1}{\lambda_3^2\tilde{\tau}^2} & \frac{1}{\tau_3^2} + \frac{1}{\lambda_3^2\tilde{\tau}^2} + \frac{1}{\lambda_4^2\tilde{\tau}^2} & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & \frac{1}{\tau_{p-1}^2} + \frac{1}{\lambda_{p-1}^2\tilde{\tau}^2} + \frac{1}{\lambda_p^2\tilde{\tau}^2} & -\frac{1}{\lambda_p^2\tilde{\tau}^2} \\ 0 & 0 & 0 & \cdots & -\frac{1}{\lambda_p^2\tilde{\tau}^2} & \frac{1}{\tau_p^2} + \frac{1}{\lambda_p^2\tilde{\tau}^2} \end{pmatrix}.$$

The detailed calculation of (6) is given in Appendix A. Therefore, the priors on $\boldsymbol{\beta}, \tau_1^2, \ldots, \tau_p^2, \tilde{\tau}^2, \lambda_2^2, \ldots, \lambda_p^2, \nu_2, \ldots, \nu_p, \xi$ are given by

$$\begin{aligned} \boldsymbol{\beta} \mid \sigma^2, \tau_1^2, \ldots, \tau_p^2, \tilde{\tau}^2, \lambda_2^2, \ldots, \lambda_p^2 &\sim \mathrm{N}_p(\boldsymbol{0}_p, \sigma^2\boldsymbol{B}), \\ \tau_j^2 &\sim \mathrm{EXP}\left(\frac{\tilde{\lambda}_1^2}{2}\right), \\ \tilde{\tau}^2 \mid \xi &\sim \mathrm{IG}\left(\frac{1}{2}, \frac{1}{\xi}\right), \\ \lambda_j^2 \mid \nu_j &\sim \mathrm{IG}\left(\frac{1}{2}, \frac{1}{\nu_j}\right), \quad (j = 2, \ldots, p), \\ \xi, \nu_j &\sim \mathrm{IG}\left(\frac{1}{2}, 1\right), \quad (j = 2, \ldots, p), \end{aligned}$$

where $\mathrm{EXP}(x \mid d)$ is an exponential prior with density function

$$\mathrm{EXP}(x \mid d) = d\exp(-dx), \quad (x \geq 0).$$

Here d is positive. In addition, we assume the priors on σ^2 and $\tilde{\lambda}_1^2$ as

$$\begin{aligned} \sigma^2 &\sim \mathrm{IG}\left(\frac{\nu_0}{2}, \frac{\eta_0}{2}\right), \\ \tilde{\lambda}_1^2 &\sim \mathrm{Ga}(r_1, \delta_1). \end{aligned} \tag{7}$$

By using the likelihood and the priors for the parameters, we can obtain the full conditional distributions as follows:

$$\begin{aligned} \boldsymbol{\beta} \mid \boldsymbol{y}, \boldsymbol{X}, \sigma^2, \tau_1^2, \ldots, \tau_p^2, \tilde{\tau}^2, \lambda_2^2, \ldots, \lambda_p^2 &\sim \mathrm{N}_p(\boldsymbol{A}^{-1}\boldsymbol{X}^T\boldsymbol{y}, \sigma^2\boldsymbol{A}^{-1}), \\ \boldsymbol{A} &= \boldsymbol{X}^T\boldsymbol{X} + \boldsymbol{B}^{-1}, \\ \sigma^2 \mid \boldsymbol{y}, \boldsymbol{X}, \boldsymbol{\beta}, \tau_1^2, \ldots, \tau_p^2, \tilde{\tau}^2, \lambda_2^2, \ldots, \lambda_p^2 &\sim \mathrm{IG}\left(\frac{n_1}{2}, \frac{s_1}{2}\right), \\ n_1 &= n + 2p - 1 + \nu_0, \\ s_1 &= (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^T(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) + \boldsymbol{\beta}^T\boldsymbol{B}^{-1}\boldsymbol{\beta} + \eta_0, \end{aligned}$$

$$\begin{aligned}
\frac{1}{\tau_j^2} \mid \beta_j, \sigma^2, \tilde{\lambda}_1^2 &\sim \mathrm{IGauss}\left(\sqrt{\frac{\sigma^2\tilde{\lambda}_1^2}{\beta_j^2}}, \tilde{\lambda}_1^2\right), \\
\tilde{\lambda}_1^2 \mid \tau_1^2, \ldots, \tau_p^2 &\sim \mathrm{Ga}\left(p + r_1, \frac{1}{2}\sum_{j=1}^{p}\tau_j^2 + \delta_1\right), \\
\tilde{\tau}^2 \mid \beta_1, \ldots, \beta_p, \sigma^2, \lambda_2^2, \ldots, \lambda_p^2, \xi &\sim \mathrm{IG}\left(\frac{p}{2}, \frac{1}{2\sigma^2}\sum_{j=2}^{p}\frac{(\beta_j - \beta_{j-1})^2}{\lambda_j^2} + \frac{1}{\xi}\right), \\
\lambda_j^2 \mid \beta_j, \beta_{j-1}, \sigma^2, \tilde{\tau}^2, \nu_j &\sim \mathrm{IG}\left(1, \frac{(\beta_j - \beta_{j-1})^2}{2\sigma^2\tilde{\tau}^2} + \frac{1}{\nu_j}\right), \\
\nu_j \mid \lambda_j^2 &\sim \mathrm{IG}\left(1, \frac{1}{\lambda_j^2} + 1\right), \\
\xi \mid \tilde{\tau}^2 &\sim \mathrm{IG}\left(1, \frac{1}{\tilde{\tau}^2} + 1\right).
\end{aligned}$$

By using the full conditional distributions, we can perform the Gibbs sampling.

3.2 Bayesian nHORSES with horseshoe prior

Next, we propose assuming a Laplace prior on the regression coefficients and a horseshoe prior on all combinations of their differences as follows:

$$\begin{aligned}
\pi(\boldsymbol{\beta} \mid \sigma^2, \tilde{\tau}^2) \propto (\sigma^2)^{-\frac{p}{2}} &\prod_{j=1}^{p} \mathrm{Laplace}\left(\frac{\beta_j}{\sqrt{\sigma^2}} \,\middle|\, \tilde{\lambda}_1\right) \\
&\times \prod_{j>k} \int \frac{1}{\sqrt{2\pi\lambda_{j,k}^2\tilde{\tau}^2\sigma^2}} \exp\left\{-\frac{(\beta_j - \beta_k)^2}{2\lambda_{j,k}^2\tilde{\tau}^2\sigma^2}\right\} \frac{2}{\pi(1+\lambda_{j,k}^2)} \frac{2}{\pi(1+\tilde{\tau}^2)} d\lambda_{j,k}^2.
\end{aligned} \tag{8}$$

Therefore, the priors on $\boldsymbol{\beta}$, $\tau_1^2, \ldots, \tau_p^2$, $\tilde{\tau}^2$, $\lambda_{1,2}^2, \ldots, \lambda_{p-1,p}^2$, $\nu_{1,2}, \ldots, \nu_{p-1,p}$, and ξ are given by

$$\begin{aligned}
\boldsymbol{\beta} \mid \sigma^2, \tau_1^2, \ldots, \tau_p^2, \tilde{\tau}^2, \lambda_{1,2}^2, \ldots, \lambda_{p-1,p}^2 &\sim \mathrm{N}_p(\mathbf{0}_p, \sigma^2\boldsymbol{B}), \\
\tau_j^2 &\sim \mathrm{EXP}\left(\frac{\tilde{\lambda}_1^2}{2}\right), \\
\lambda_{j,k}^2 \mid \nu_{j,k} &\sim \mathrm{IG}\left(\frac{1}{2}, \frac{1}{\nu_{j,k}}\right), \\
\nu_{j,k} &\sim \mathrm{IG}\left(\frac{1}{2}, 1\right),
\end{aligned}$$

where $\lambda_{j,k}^2 = \lambda_{k,j}^2$, $\nu_{j,k}^2 = \nu_{k,j}^2$ and the (i, j)-element of $\boldsymbol{B}^{-1}$ is represented as

$$\boldsymbol{B}^{-1}{}_{(i,j)} = \begin{cases} \dfrac{1}{\tau_j^2} + \dfrac{1}{\tilde{\tau}^2} \displaystyle\sum_{l \neq i} \frac{1}{\lambda_{i,l}^2}, & (i = j), \\ -\dfrac{1}{\lambda_{i,j}^2}, & (i \neq j). \end{cases}$$

By assuming an inverse gamma prior on σ^2 and a gamma prior on $\tilde{\lambda}_1^2$ in (7), the full conditional distributions are represented as

$$\begin{aligned} \boldsymbol{\beta} \mid \boldsymbol{y}, \boldsymbol{X}, \sigma^2, \tau_1^2, \ldots, \tau_p^2, \tilde{\tau}^2, \lambda_{1,2}^2, \ldots, \lambda_{p-1,p}^2 &\sim \mathrm{N}_p(\boldsymbol{A}^{-1}\boldsymbol{X}^T\boldsymbol{y}, \sigma^2\boldsymbol{A}^{-1}), \\ \boldsymbol{A} &= \boldsymbol{X}^T\boldsymbol{X} + \boldsymbol{B}^{-1}, \\ \sigma^2 \mid \boldsymbol{y}, \boldsymbol{X}, \boldsymbol{\beta}, \tau_1^2, \ldots, \tau_p^2, \tilde{\tau}^2, \lambda_{1,2}^2, \ldots, \lambda_{p-1,p}^2 &\sim \mathrm{IG}\left(\frac{n_1}{2}, \frac{s_1}{2}\right), \\ n_1 &= n + p(p+1)/2 + \nu_0, \\ s_1 &= (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^T(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) + \boldsymbol{\beta}^T\boldsymbol{B}^{-1}\boldsymbol{\beta} + \eta_0, \\ \frac{1}{\tau_j^2} \mid \beta_j, \sigma^2, \tilde{\lambda}_1^2 &\sim \mathrm{IGauss}\left(\sqrt{\frac{\sigma^2\tilde{\lambda}_1^2}{\beta_j^2}}, \tilde{\lambda}_1^2\right), \\ \tilde{\lambda}_1^2 \mid \tau_1^2, \ldots, \tau_p^2 &\sim \mathrm{Ga}\left(p + r_1, \frac{1}{2}\sum_{j=1}^{p}\tau_j^2 + \delta_1\right), \\ \lambda_{j,k}^2 \mid \beta_j, \beta_k, \sigma^2, \tilde{\tau}^2, \nu_{j,k} &\sim \mathrm{IG}\left(1, \frac{(\beta_j - \beta_k)^2}{2\sigma^2\tilde{\tau}^2} + \frac{1}{\nu_{j,k}}\right), \\ \nu_{j,k} \mid \lambda_{j,k}^2 &\sim \mathrm{IG}\left(1, \frac{1}{\lambda_{j,k}^2} + 1\right). \end{aligned}$$

By using the full conditional distributions, we can perform the Gibbs sampling for Bayesian nHORSES.

Note that the hyper-parameter $\tilde{\tau}^2$ in the prior (8) is treated as a tuning parameter. The value of the tuning parameter is selected by any model selection criterion such as the widely applicable information criterion (WAIC) (Watanabe, 2010).

4 Numerical studies

In this section, we compare the proposed method with existing methods through Monte Carlo simulations and show its effectiveness. In addition, we apply the proposed method to the Appalachian Mountains Soil Data (Bondell and Reich (2008); Jang et al. (2015)).

4.1 Monte carlo simulation

We conducted Monte Carlo simulations with artificial data generated from the true model:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta}^* + \boldsymbol{\epsilon},$$

where $\boldsymbol{\beta}^*$ is the p-dimensional true coefficient vector and the error vector $\boldsymbol{\epsilon}$ is distributed normally as $\mathrm{N}_n(\mathbf{0}_n, \sigma^2 \boldsymbol{I}_n)$. In addition, $\boldsymbol{x}_i$ $(i = 1, 2, \ldots, n)$ is distributed according to the multivariate normal distribution $\mathrm{N}_p(\mathbf{0}_p, \Sigma)$.

We considered the following settings:

Case 1: $\boldsymbol{\beta}^* = \boldsymbol{\beta}_1^*$ or $\boldsymbol{\beta}_2^*$, $\Sigma_{ii} = 1$, $\Sigma_{ij} = \rho$, $(i \neq j)$,
Case 2: $\boldsymbol{\beta}^* = \boldsymbol{\beta}_1^*$ or $\boldsymbol{\beta}_2^*$, $\Sigma_{ii} = 1$, $\Sigma_{ij} = 0.5^{|i-j|}$, $(i \neq j)$,
Case 3: $\boldsymbol{\beta}^* = \boldsymbol{\beta}_3^*$, $\Sigma_{ii} = 1$, $\Sigma_{ij} = 0.5$, $(i \neq j)$,
Case 4: $\boldsymbol{\beta}^* = \boldsymbol{\beta}_3^*$, $\Sigma_{ii} = 1$, $\Sigma_{ij} = 0.5^{|i-j|}$, $(i \neq j)$,
Case 5: $\boldsymbol{\beta}^* = \boldsymbol{\beta}_4^*$, $\Sigma_{ii} = 1$, $\Sigma_{ij} = \rho$, $(i \neq j)$,
Case 6: $\boldsymbol{\beta}^* = \boldsymbol{\beta}_5^*$, $\Sigma_{ii} = 1$, $\Sigma_{ij} = \rho$, $(i \neq j)$,

where Σ_{ij} is the (i, j)-th element of Σ. For each case, we considered $\sigma = 0.5, 1.5$. We simulated data with $\boldsymbol{\beta}_1^* = (\mathbf{0.0}_5^T, \mathbf{1.0}_5^T, \mathbf{0.0}_5^T, \mathbf{1.0}_5^T)^T$ and $\boldsymbol{\beta}_2^* = (\mathbf{0.0}_5^T, \mathbf{2.0}_5^T, \mathbf{0.0}_5^T, \mathbf{2.0}_5^T)^T$ for Cases 1 and 2. In addition, we considered $\boldsymbol{\beta}_3^* = (\mathbf{3.0}_5^T, \mathbf{-1.5}_5^T, \mathbf{1.0}_5^T, \mathbf{2.0}_5^T, \mathbf{0.0}_{30}^T)^T$, $\boldsymbol{\beta}_4^* = (\mathbf{3.0}_5^T, \mathbf{-1.5}_5^T, \mathbf{1.0}_5^T, \mathbf{2.0}_5^T, \mathbf{0.0}_{130}^T)^T$ and $\boldsymbol{\beta}_5^* = (\mathbf{3.0}_5^T, \mathbf{-1.5}_5^T, \mathbf{1.0}_5^T, \mathbf{2.0}_5^T, \mathbf{0.0}_{980}^T)^T$ for Cases 3, 4, 5, and 6. In Case 1, we considered $\rho = 0.5, 0.9$. In Cases 5 and 6, we considered $\rho = 0.5, 0.5^{|i-j|}$. We considered $n = 50$, $p = 20$ for Cases 1 and 2, $n = 30, 50$, $p = 50$ for Cases 3 and 4, $n = 50$, $p = 150$ for Case 5, and $n = 200$, $p = 1000$ for Case 6. Therefore, Cases 1 and 2 correspond to $n > p$ cases, whereas Cases 3, 4, 5, and 6 correspond to $n \leq p$ cases. For Cases 5 and 6, we used the fast sampling method from multivariate normal distribution based on Woodbury matrix identity (Hager (1989); Bhattacharya et al. (2016)). We simulated 100 datasets for each case. Cases 1 and 2 are according to example 1 in Shen and Huang (2010), whereas Cases 3, 4, 5, and 6 are, respectively, according to examples 2 and 3 in the same reference. For each Case 1 to Case 4, the Gibbs sampling was run with 5000 iterations, where we discarded the first 2000 iterations as burn-in. For Cases 5 and 6, the Gibbs sampling was run with 10,000 iterations with the first 7000 samples discarded as burn-in for BFL and BFNEG. For BFH, the Gibbs sampler was run in the same way except for the situation with $\rho = 0.5^{|i-j|}$ and $\sigma = 0.5$. In this situation, the Gibbs sampling was run with 13,000 iterations with the first 10,000 as burn-in.

We checked the convergence of MCMC chains by using Gelman–Rubin diagnostic. When the MCMC chain converges, the potential scale reduction factor $\hat{R}$ (Gelman and Rubin (1992); Brooks and Gelman (1998)) becomes close to 1.0. The factor $\hat{R}$ was computed by using its relationship with the effective sample size (Vats and Knudson, 2021). We used the package stableGR of the software **R** to compute $\hat{R}$. The package can be obtained from https://cran.r-project.org/web/packages/stableGR/index.html. As Vehtari et al. (2021) recommended, we checked whether $\hat{R}$ is below 1.01 or not. The maximum values of $\hat{R}$ in MCMC samples for 100 datasets for each regression

coefficients were lower than 1.01 (see Table S.1 in the supplementary material, for example.). Thus, we considered the MCMC chains are converged.

We compared Bayesian fused lasso with horseshoe prior (BFH) to Bayesian fused lasso (BFL) and Bayesian fused lasso with NEG prior (BFNEG). For BFNEG, the shape parameter in the gamma distribution in the NEG prior was set to 0.5 according to the simulation study in Griffin and Brown (2011), while the rate parameter was selected by WAIC in Cases 1 to 5. For Case 6, the rate parameter was set to 1.0, because it was too time-consuming to finish. The values of WAIC for the selected model are summarized in Table S.3 in the supplementary material.

To evaluate the accuracy of the estimation of regression coefficients, we used mean squared error (MSE):

$$\mathrm{MSE} = \frac{1}{100}\sum_{k=1}^{100}(\hat{\boldsymbol{\beta}}^{(k)} - \boldsymbol{\beta}^*)^T(\hat{\boldsymbol{\beta}}^{(k)} - \boldsymbol{\beta}^*),$$

where $\hat{\boldsymbol{\beta}}^{(k)} = (\hat{\beta}_1^{(k)}, \ldots, \hat{\beta}_p^{(k)})^T$ is the regression coefficient vector estimated from the k-th dataset. We also computed $\mathrm{MSE}_{\mathrm{diff}}$, expressed as

$$\mathrm{MSE}_{\mathrm{diff}} = \frac{1}{100}\sum_{k=1}^{100}(\hat{\boldsymbol{\beta}}_{\mathrm{diff}}^{(k)} - \boldsymbol{\beta}_{\mathrm{diff}}^*)^T(\hat{\boldsymbol{\beta}}_{\mathrm{diff}}^{(k)} - \boldsymbol{\beta}_{\mathrm{diff}}^*),$$

where $\boldsymbol{\beta}_{\mathrm{diff}}^*$ is a vector of the non-zero differences of the true regression coefficients and $\hat{\boldsymbol{\beta}}_{\mathrm{diff}}^{(k)}$ is the estimated value of $\boldsymbol{\beta}_{\mathrm{diff}}^*$ from the k-th dataset. $\mathrm{MSE}_{\mathrm{diff}}$ is an index to assess how close the differences of estimated regression coefficients which are not zero are to the true differences. For example, regression coefficients for Case 1 are given by $\boldsymbol{\beta}^* = (\mathbf{0.0}_5^T, \mathbf{2.0}_5^T, \mathbf{0.0}_5^T, \mathbf{2.0}_5^T)^T$ and the non-zero successive differences are between the 5th and 6th, 10th and 11th, and 15th and 16th elements of $\boldsymbol{\beta}^*$. Then, $\mathrm{MSE}_{\mathrm{diff}}$ is calculated as follows:

$$\begin{aligned}\mathrm{MSE}_{\mathrm{diff}} =& \frac{1}{100}\sum_{k=1}^{100}\Bigg[\left\{(\hat{\beta}_6^{(k)} - \hat{\beta}_5^{(k)}) - (\beta_6^* - \beta_5^*)\right\}^2 \\ &+ \left\{(\hat{\beta}_{11}^{(k)} - \hat{\beta}_{10}^{(k)}) - (\beta_{11}^* - \beta_{10}^*)\right\}^2 + \left\{(\hat{\beta}_{16}^{(k)} - \hat{\beta}_{15}^{(k)}) - (\beta_{16}^* - \beta_{15}^*)\right\}^2\Bigg].\end{aligned}$$

In addition, we computed prediction squared error

$$\mathrm{PSE} = \frac{1}{100}\sum_{k=1}^{100}(\hat{\boldsymbol{\beta}}^{(k)} - \boldsymbol{\beta}^*)^T\Sigma(\hat{\boldsymbol{\beta}}^{(k)} - \boldsymbol{\beta}^*)$$

to evaluate the accuracy of prediction.

The results are summarized in Tables 1, 2, 3, 4, 5, 6, and 7. For Cases 1 to 5, the proposed method BFH shows the smallest MSEs and PSEs in all situations. This

Table 1 MSE, MSE_{diff}, and PSE for Case 1 and $\rho = 0.5$

		$\sigma = 0.5$			$\sigma = 1.5$		
		MSE (sd)	MSE_{diff} (sd)	PSE (sd)	MSE (sd)	MSE_{diff} (sd)	PSE (sd)
β_1^*	BFL	0.298	0.110	0.377	1.682	0.612	1.135
		(0.120)	(0.084)	(0.316)	(0.671)	(0.462)	(0.558)
	BFNEG	0.229	0.095	0.341	1.284	**0.604**	0.932
		(0.112)	(0.073)	(0.317)	(0.535)	(0.403)	(0.518)
	BFH	**0.169**	**0.073**	**0.310**	**0.990**	0.684	**0.778**
		(0.099)	(0.057)	(0.313)	(0.520)	(0.434)	(0.490)
β_2^*	BFL	0.632	0.217	1.199	2.107	0.802	2.009
		(0.318)	(0.171)	(1.210)	(0.819)	(0.618)	(1.380)
	BFNEG	0.531	0.187	1.145	1.620	0.758	1.759
		(0.321)	(0.154)	(1.211)	(0.730)	(0.574)	(1.380)
	BFH	**0.459**	**0.165**	**1.107**	**1.134**	**0.602**	**1.504**
		(0.310)	(0.139)	(1.207)	(0.627)	(0.488)	(1.343)

Bold font indicates smallest value among BFL, BFNEG, and BFH. Figures in parentheses give the estimated standard deviation

Table 2 MSE, MSE_{diff}, and PSE for Case 1 and $\rho = 0.9$

		$\sigma = 0.5$			$\sigma = 1.5$		
		MSE (sd)	MSE_{diff} (sd)	PSE (sd)	MSE (sd)	MSE_{diff} (sd)	PSE (sd)
β_1^*	BFL	0.835	**0.386**	1.042	4.636	1.573	1.467
		(0.332)	(0.256)	(1.285)	(1.969)	(1.070)	(1.405)
	BFNEG	0.673	0.415	1.027	3.366	**1.452**	1.342
		(0.321)	(0.273)	(1.285)	(1.656)	(0.751)	(1.407)
	*BFH	**0.594**	0.407	**1.019**	**3.003**	1.936	**1.311**
		(0.322)	(0.288)	(1.286)	(1.118)	(0.587)	(1.406)
β_2^*	BFL	1.277	0.536	3.950	6.273	**2.692**	4.490
		(0.639)	(0.370)	(5.092)	(2.509)	(1.712)	(5.216)
	BFNEG	0.947	0.403	3.917	5.090	3.041	4.376
		(0.623)	(0.301)	(5.094)	(2.257)	(1.676)	(5.216)
	BFH	**0.826**	**0.344**	**3.904**	**4.917**	3.620	**4.360**
		(0.607)	(0.263)	(5.096)	(2.402)	(1.963)	(5.217)

Bold font indicates smallest value among BFL, BFNEG, and BFH. Figures in parentheses give the estimated standard deviation

Table 3 MSE, MSE_{diff}, and PSE for Case 2

		$\sigma = 0.5$			$\sigma = 1.5$		
		MSE (sd)	MSE_{diff} (sd)	PSE (sd)	MSE (sd)	MSE_{diff} (sd)	PSE (sd)
$\boldsymbol{\beta}_1^*$	BFL	0.246	0.113	0.232	1.286	0.616	1.023
		(0.102)	(0.098)	(0.116)	(0.535)	(0.479)	(0.418)
	BFNEG	0.182	0.091	0.199	0.918	0.557	0.829
		(0.094)	(0.085)	(0.115)	(0.396)	(0.386)	(0.357)
	BFH	**0.127**	**0.063**	**0.166**	**0.577**	**0.469**	**0.620**
		(0.077)	(0.069)	(0.107)	(0.304)	(0.334)	(0.317)
$\boldsymbol{\beta}_2^*$	BFL	0.555	0.219	0.610	1.675	0.824	1.446
		(0.258)	(0.197)	(0.362)	(0.691)	(0.688)	(0.640)
	BFNEG	0.457	0.182	0.560	1.222	0.717	1.210
		(0.253)	(0.174)	(0.363)	(0.604)	(0.616)	(0.625)
	BFH	**0.394**	**0.155**	**0.523**	**0.766**	**0.473**	**0.931**
		(0.240)	(0.159)	(0.357)	(0.457)	(0.495)	(0.548)

Bold font indicates smallest value among BFL, BFNEG, and BFH. Figures in parentheses give the estimated standard deviation

Table 4 MSE, MSE_{diff}, and PSE for Case 3

		$\sigma = 0.5$			$\sigma = 1.5$		
		MSE (sd)	MSE_{diff} (sd)	PSE (sd)	MSE (sd)	MSE_{diff} (sd)	PSE (sd)
$n = 30$	BFL	4.412	2.411	3.988	13.605	5.228	8.672
		(2.513)	(2.250)	(3.270)	(6.302)	(3.857)	(4.509)
	BFNEG	2.685	1.112	3.145	12.804	4.591	8.266
		(1.444)	(1.446)	(3.122)	(6.522)	(3.982)	(4.561)
	BFH	**1.800**	**0.868**	**2.683**	**6.936**	**3.429**	**5.290**
		(1.142)	(1.040)	(3.010)	(4.487)	(3.599)	(3.668)
$n = 50$	BFL	1.897	0.522	1.978	8.840	2.152	5.494
		(0.630)	(0.423)	(1.510)	(3.237)	(1.563)	(2.075)
	BFNEG	1.972	0.478	2.019	10.454	2.270	6.318
		(0.620)	(0.394)	(1.496)	(4.260)	(1.699)	(2.402)
	BFH	**0.836**	**0.301**	**1.429**	**2.397**	**1.076**	**2.234**
		(0.475)	(0.290)	(1.507)	(1.205)	(0.759)	(1.535)

Bold font indicates smallest value among BFL, BFNEG, and BFH. Figures in parentheses give the estimated standard deviation

Table 5 MSE, MSE_{diff}, and PSE for Case 4

		$\sigma = 0.5$			$\sigma = 1.5$		
		MSE (sd)	MSE_{diff} (sd)	PSE (sd)	MSE (sd)	MSE_{diff} (sd)	PSE (sd)
$n = 30$	BFL	2.716	1.873	3.159	8.483	4.430	8.965
		(1.110)	(1.432)	(1.375)	(3.569)	(3.279)	(4.058)
	BFNEG	1.921	0.869	2.423	8.464	4.117	8.601
		(0.799)	(0.818)	(1.010)	(4.101)	(3.942)	(3.775)
	BFH	**1.274**	**0.604**	**1.853**	**4.007**	**2.563**	**4.657**
		(0.627)	(0.528)	(0.884)	(2.991)	(3.360)	(2.439)
$n = 50$	BFL	1.530	0.482	1.587	6.366	2.097	5.340
		(0.531)	(0.373)	(0.662)	(2.595)	(1.599)	(1.966)
	BFNEG	1.613	0.464	1.638	7.737	2.304	6.186
		(0.544)	(0.371)	(0.666)	(3.476)	(1.856)	(2.508)
	BFH	**0.677**	**0.229**	**1.012**	**1.280**	**0.660**	**1.808**
		(0.345)	(0.195)	(0.559)	(0.752)	(0.613)	(0.881)

Bold font indicates smallest value among BFL, BFNEG, and BFH. Figures in parentheses give the estimated standard deviation

Table 6 MSE, MSE_{diff}, and PSE for Case 5

		$\sigma = 0.5$			$\sigma = 1.5$		
		MSE (sd)	MSE_{diff} (sd)	PSE (sd)	MSE (sd)	MSE_{diff} (sd)	PSE (sd)
$\rho = 0.5$	BFL	3.315	1.869	2.535	10.221	4.653	6.072
		(1.704)	(1.111)	(1.609)	(4.065)	(2.528)	(2.408)
	BFNEG	1.335	0.377	1.534	7.739	2.221	4.830
		(0.530)	(0.325)	(1.377)	(3.202)	(1.899)	(2.104)
	BFH	**0.889**	**0.272**	**1.347**	**6.926**	**1.957**	**4.437**
		(0.544)	(0.254)	(1.421)	(3.341)	(1.759)	(2.191)
$\rho = 0.5^{\lvert i-j\rvert}$	BFL	2.038	1.791	2.679	5.257	3.915	6.939
		(0.870)	(0.976)	(1.273)	(1.979)	(2.253)	(2.750)
	BFNEG	0.997	0.369	1.495	4.191	1.795	5.600
		(0.416)	(0.324)	(0.669)	(1.849)	(1.742)	(2.154)
	BFH	**0.788**	**0.276**	**1.176**	**4.083**	**1.664**	**5.400**
		(0.423)	(0.239)	(0.642)	(2.000)	(1.651)	(2.271)

Bold font indicates smallest value among BFL, BFNEG, and BFH. Figures in parentheses give the estimated standard deviation

Table 7 MSE, MSE_{diff}, and PSE for Case 6

		$\sigma = 0.5$			$\sigma = 1.5$		
		MSE (sd)	MSE_{diff} (sd)	PSE (sd)	MSE (sd)	MSE_{diff} (sd)	PSE (sd)
$\rho = 0.5$	BFL	0.536	0.209	0.527	2.734	0.771	1.648
		(0.207)	(0.144)	(0.507)	(0.486)	(0.391)	(0.567)
	BFNEG	**0.468**	0.095	**0.492**	**3.410**	**0.399**	**1.990**
		(0.169)	(0.087)	(0.514)	(0.556)	(0.282)	(0.601)
	BFH	0.605	**0.092**	0.562	4.461	0.417	2.516
		(0.169)	(0.081)	(0.507)	(0.744)	(0.310)	(0.648)
$\rho = 0.5^{\lvert i-j \rvert}$	BFL	0.337	0.213	0.517	**1.426**	0.669	**2.334**
		(0.093)	(0.102)	(0.142)	(0.254)	(0.330)	(0.358)
	BFNEG	**0.288**	**0.077**	**0.449**	1.553	**0.322**	2.362
		(0.088)	(0.059)	(0.121)	(0.316)	(0.250)	(0.355)
	BFH	0.369	0.080	0.524	2.161	0.338	2.919
		(0.091)	(0.061)	(0.123)	(0.423)	(0.283)	(0.441)

Bold font indicates smallest value among BFL, BFNEG, and BFH. Figures in parentheses give the estimated standard deviation

indicates that BFH outperformed other methods when $n > p$ and but also when $n \leq p$. In addition, BFH gives smaller MSE_{diff}s than BFL in almost all situations. The reason is why BFH does not shrink non-zero differences of regression coefficients too much compared to BFL. Furthermore, BFH gives smaller values of MSE_{diff} in 21 situations, out of the 24 situations, in comparison to BFNEG. These results show that BFH gives a closer estimation to the true regression coefficients. For Case 6, BFNEG provides smaller MSEs and PSEs than BFH and is competitive with BFH in terms of MSE_{diff}s. Similar to BFNEG, BFL gives smaller MSEs and PSEs than BFH. However, BFH outperforms BFL in terms of MSE_{diff}s in all situations.

We next conducted simulations for Bayesian HORSES methods. The following settings were considered:

Case 7: $\boldsymbol{\beta}^* = (3.0, -1.5, 1.0, 2.0, \mathbf{0.0}_4^T, 3.0, -1.5, 1.0, 2.0, \mathbf{0.0}_4^T)^T$,

$$\Sigma_{ii} = 1,\ \Sigma_{ij} = \begin{cases} 0.7 & (\beta_i \neq 0 \text{ and } \beta_j \neq 0,\ \beta_i = 0 \text{ and } \beta_j = 0) \\ 0.2 & \text{otherwise} \end{cases} \quad (i \neq j),$$

Case 8: $\boldsymbol{\beta}^* = (3.0, \mathbf{0.0}_2^T, -1.5, \mathbf{0.0}_2^T, 3.0, \mathbf{0.0}_2^T, -1.5)^T$,

$$\Sigma_{ii} = 1,\ \Sigma_{ij} = \begin{cases} 0.7 & (\beta_i \neq 0 \text{ and } \beta_j \neq 0) \\ 0 & \text{otherwise} \end{cases} \quad (i \neq j),$$

Case 7 means that variables with non-zero coefficients are highly correlated; this is true of those with zero coefficients, while Case 8 does that variables with non-zero coefficients are highly correlated. For each case, we considered $\sigma = 0.5, 1.5$. The sample size and the number of parameters were, respectively, set by $n = 40, 80$, $p = 16$ for Case 7 and $n = 30, 50$, $p = 10$ for Case 8. We compared Bayesian nHORSES with horseshoe prior (BHH) to Bayesian HORSES (BH) and Bayesian nHORSES with NEG prior (BHNEG). We chose the hyper-parameters λ_2^2 for BH,

Table 8 MSE and PSE for Case 7

		$\sigma = 0.5$		$\sigma = 1.5$	
		MSE (sd)	PSE (sd)	MSE (sd)	PSE (sd)
$n = 40$	BH	0.995	0.879	**3.964**	**1.891**
		(0.626)	(0.865)	(1.827)	(1.162)
	BHNEG	0.910	0.844	4.217	1.955
		(0.513)	(0.836)	(1.867)	(1.141)
	BHH	**0.815**	**0.808**	4.075	1.916
		(0.531)	(0.839)	(2.086)	(1.199)
$n = 80$	BH	0.386	0.389	1.571	0.812
		(0.234)	(0.355)	(0.698)	(0.435)
	BHNEG	0.395	0.390	1.641	0.831
		(0.224)	(0.353)	(0.711)	(0.431)
	BHH	**0.348**	**0.376**	**1.465**	**0.780**
		(0.225)	(0.356)	(0.786)	(0.455)

Bold font indicates smallest value among BH, BHNEG, and BHH. Figures in parentheses give the estimated standard deviation

$\tilde{\tau}^2$ for BHH and the rate parameter in the gamma distribution in the NEG prior for BHNEG by WAIC. The values of WAIC for the selected model are summarized in Table S.4 in the supplementary material. We set the shape parameter in the gamma distribution in the NEG prior for BHNEG as 0.5.

We measured the accuracy of the estimation of regression coefficients by MSE and performance for prediction by PSE. We only considered MSE and PSE for the Bayesian HORSES methods, because the structure of the coefficient vector $\boldsymbol{\beta}^*$ is too complicated to assess the performance of capturing groups of variables by an index such as MSE_{diff}.

The results in Cases 7 and 8 are summarized in Tables 8 and 9, respectively. Table 8 shows that BHH gives smaller MSEs and PSEs in many situations. BH achieves the smallest MSE and PSE in $n = 40$ and $\sigma = 1.5$ and PSE in $n = 80$ and $\sigma = 0.5$. Table 9 shows that BHH outperforms BH and BHNEG in all situations in terms of MSEs and PSEs. From these results, we believe that BHH might be a useful method for analyzing the complex structure treated by HORSES.

4.2 Application

We applied Bayesian nHORSES with horseshoe prior in Section 3.2 to the Appalachian Mountains Soil Data, which was analyzed in Bondell and Reich (2008) and Jang et al. (2015). This dataset is available from https://blogs.unimelb.edu.au/howard-bondell/#tab25 and was used for showing the relationship between soil characteristics and rich cove forest diversity. The dataset was collected at twenty 500 m^2 plots in the Appalachian Mountains. Forest diversity, which is represented as the number of different plant species, is used for a response variable and 15 soil characteristics in 20

Table 9 MSE and PSE for Case 8

		$\sigma = 0.5$		$\sigma = 1.5$	
		MSE (sd)	PSE (sd)	MSE (sd)	PSE (sd)
$n = 30$	BH	0.906	0.485	4.029	1.752
		(0.829)	(0.406)	(2.530)	(0.945)
	BHNEG	0.520	0.355	2.448	1.293
		(0.500)	(0.338)	(1.619)	(0.719)
	BHH	**0.489**	**0.338**	**2.363**	**1.221**
		(0.472)	(0.332)	(1.603)	(0.706)
$n = 50$	BH	0.456	0.263	2.178	0.976
		(0.387)	(0.257)	(1.469)	(0.588)
	BHNEG	0.293	0.214	1.281	0.724
		(0.250)	(0.248)	(0.924)	(0.483)
	BHH	**0.279**	**0.204**	**1.192**	**0.654**
		(0.248)	(0.247)	(0.947)	(0.483)

Bold font indicates smallest value among BH, BHNEG, and BHH. Figures in parentheses give the estimated standard deviation

plots are used as explanatory variables. The data are the average of five equally spaced measurements in each plot. We standardized the dataset before the analysis.

We compared BHH to BH and BHNEG. For this application, we chose the hyper-parameters λ_2^2 for BH from five candidates between 10^{-4} and 10^{-2}, $\tilde{\tau}^2$ for BHH from five candidates between 10^4 and 10^6, and the rate parameter in the gamma distribution in the NEG prior for BHNEG from five candidates between 0.1 and 1. The values of the hyper-parameters were selected by WAIC. The values of WAIC for the selected model are summarized in Table S.5 in the supplementary material. We set the shape parameter in the gamma distribution in the NEG prior for BHNEG as 0.5.

We executed a leave-one-out cross-validation to assess the performance of the models. In each estimation, the Gibbs sampling was run with 10,000 iterations and 5000 iterations were discarded as burn-in. We also computed the maximum $\hat{R}$ from MCMC samples for 20 datasets in the leave-one-out cross-validation. The maximum values of 20 $\hat{R}$s for each regression coefficients were lower than 1.01 (see Table S.2 in the supplementary material.). Thus, we considered the MCMC chains are converged.

The mean values of cross-validation, CV, are summarized in Table 10. From Table 10, the value of CV for BHH is smaller than that for BH. BHNEG gives the smallest value of CV, but gives the largest value of standard deviation. On the other hand, BHH gives the smallest value of standard deviation and the second largest value of CV.

Table 11 shows the regression coefficients and its 95% credible intervals estimated by all 20 samples. The mean of estimated regression coefficients and 95% credible intervals are also described in Fig. 1. From Fig. 1, we observe that the range of the credible intervals of BH tends to be larger than ones of BHH and BHNEG. From Table 11 and Fig. 1, the regression coefficients of BHH and BHNEG for "base saturation", "sum cations", "CEC buffer", "calcium", and "potassium" are negative. Therefore,

Table 10 Results for the Appalachian Mountains Soil Data

	BH	BHNEG	BHH
CV	0.000685	0.000618	0.000652
(sd)	(0.000807)	(0.000852)	(0.000839)

Figures in parentheses give the estimated standard deviation

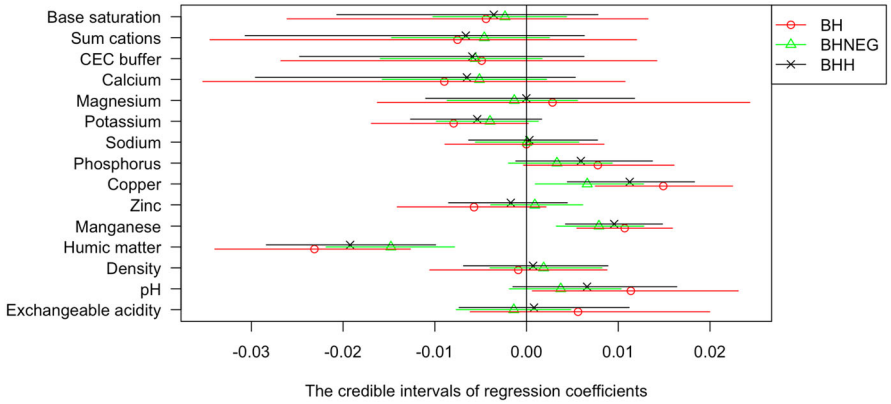

Fig. 1 95% credible intervals of estimated regression coefficients

those regression coefficients seem to form a same group, which is supported by the results in Jang et al. (2015). BHH provides positive coefficients for "phosphorus", "copper", and "exchangeableacidity" and negative for "zinc", while BHNEG negative for "exchangeableacidity" and positive for "zinc". The results by BHH are same as Jang et al. (2015). From the viewpoint of the range of credible intervals, we observe that BHNEG is the most stable in three and BHH captures the difference between variables.

5 Conclusion and discussion

We proposed Bayesian fused lasso modeling with horseshoe prior, and then developed the Gibbs sampler for the parameters by using a scale mixture of normals and a hierarchical expression of a half-Cauchy prior. In addition, we extended the method to the Bayesian nHORSES. Through numerical studies, we showed our proposed method is better than the existing methods in terms of prediction and estimation accuracy.

There are several studies about fused lasso modeling with regression coefficients on a general graph, which includes the fusion of all possible pairs of coefficients (Wang et al. (2016); Lee et al. (2021); Banerjee (2022)). Our proposed method can be also expanded as the model whose parameters exist on a general graph. Let $G = (V, E)$ be an arbitrary undirected graph, where V is the node set whose elements consists of indexes of explanatory variables in the design matrix $\boldsymbol{X}$ and E is the edge set that represents a relationship among the explanatory variables. Based on the graph, we

Table 11 Estimated regression coefficients for Appalachian Mountains Soil Data

	BH	BHNEG	BHH
Base saturation	−0.00442[−0.0261, 0.0132]	−0.00235 [−0.0102, 0.00436]	−0.00357 [−0.0207, 0.00779]
Sum cations	−0.00753 [−0.0345, 0.0120]	−0.00460 [−0.0147, 0.00249]	−0.00662 [−0.0307, 0.00630]
CEC buffer	−0.00489 [−0.0268, 0.0142]	−0.00560 [−0.0159, 0.00170]	−0.00592 [−0.0247, 0.00626]
Calcium	−0.00897 [−0.0353, 0.0107]	−0.00513 [−0.0157, 0.00218]	−0.00651 [−0.0296, 0.00532]
Magnesium	0.00283 [−0.0163, 0.0243]	−0.00134 [−0.00867, 0.00556]	−0.0000239 [−0.0110, 0.0118]
Potassium	−0.00795 [−0.0169, 0.000189]	−0.003986 [−0.00984, 0.00128]	−0.00538 [−0.0127, 0.00165]
Sodium	−0.0000158 [−0.00889, 0.00845]	0.0000555 [−0.00557, 0.00570]	0.000313 [−0.00631, 0.00774]
Phosphorus	0.00777 [−0.000337, 0.0161]	0.00331 [−0.00197, 0.00934]	0.00592 [−0.00118, 0.0137]
Copper	0.0149 [0.00750, 0.0225]	0.00661 [0.000958, 0.0128]	0.0113 [0.00445, 0.0183]
Zinc	−0.00573 [−0.0141, 0.00212]	0.000903 [−0.00390, 0.00612]	−0.00172 [−0.00850, 0.00445]
Manganese	0.0107 [0.00549, 0.0159]	0.00789 [0.00325, 0.0128]	0.00956 [0.00425, 0.0148]
Humic matter	−0.0231 [−0.0340, −0.0127]	−0.0148 [−0.0219, −0.00786]	−0.0192 [−0.0284, −0.00991]
Density	−0.000900 [−0.0105, 0.00877]	0.00186 [−0.00398, 0.00822]	0.000711 [−0.00687, 0.00888]
pH	0.0114 [0.000631, 0.0231]	0.00373 [−0.00187, 0.0103]	0.00660 [−0.00149, 0.0164]
Exchangeable acidity	0.00561 [−0.00614, 0.0200]	−0.00139 [−0.00766, 0.00482]	0.000815 [−0.00735, 0.0112]

Figures in parentheses give the credible interval

consider the following prior on regression coefficients:

$$\pi(\boldsymbol{\beta} \mid \sigma^2, \tilde{\tau}^2) \propto (\sigma^2)^{-\frac{p}{2}} \prod_{j=1}^{p} \text{Laplace}\left(\frac{\beta_j}{\sqrt{\sigma^2}} \,\middle|\, \tilde{\lambda}_1\right) \times \prod_{(j,k)\in E} \int \frac{1}{\sqrt{2\pi\lambda_{j,k}^2\tilde{\tau}^2\sigma^2}} \exp\left\{-\frac{(\beta_j-\beta_k)^2}{2\lambda_{j,k}^2\tilde{\tau}^2\sigma^2}\right\} \frac{2}{\pi(1+\lambda_{j,k}^2)} \frac{2}{\pi(1+\tilde{\tau}^2)} d\lambda_{j,k}^2.$$

By using this prior, we can construct a Bayesian nHORSES with horseshoe prior on a general graph. The sampling algorithm by MCMC may be built in a similar manner of the algorithm of Bayesian nHORSES with horseshoe prior.

In Bayesian nHORSES with horseshoe prior, we select the value of global shrinkage parameter $\tilde{\tau}^2$ by WAIC. It would be interesting to assume any proper prior on $\tilde{\tau}^2$. For high-dimensional data, our proposed method requires much computational burden. The reduction of computational time is thus necessary. For example, we might be able to make the sampling algorithm faster by using the approximation method of the horseshoe posterior (Johndrow et al., 2020). We leave these topics as future work.

Acknowledgements The authors thank the reviewers for their helpful comments and constructive suggestions. Y. K. was supported by JST, the establishment of university fellowships towards the creation of science technology innovation, Grant Number JPMJFS2136. S. K. was supported by JSPS KAKENHI Grant Numbers JP19K11854, JP23K11008, JP23H03352, and JP23H00809. The computational resource was provided by the Super Computer System, Human Genome Center, Institute of Medical Science, The University of Tokyo.

Data Availability We have used the real data of Bondell and Reich (2008) available online.

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

A Detailed calculation of the formula (6)

The detailed calculation of rewriting (5) as (6) is as follows:

$$\pi(\boldsymbol{\beta} \mid \sigma^2) \propto (\sigma^2)^{-\frac{p}{2}} \prod_{j=1}^{p} \int \frac{1}{\sqrt{2\pi\tau_j^2}} \exp\left(-\frac{\beta_j^2}{2\sigma^2\tau_j^2}\right) \frac{\tilde{\lambda}_1^2}{2} \exp\left(-\frac{\tilde{\lambda}_1^2}{2}\tau_j^2\right) d\tau_j^2$$

$$\times \int \left[\prod_{j=2}^{p} \int \frac{1}{\sqrt{2\pi\lambda_j^2\tilde{\tau}^2\sigma^2}} \exp\left\{ -\frac{(\beta_j - \beta_{j-1})^2}{2\lambda_j^2\tilde{\tau}^2\sigma^2} \right\} \right.$$

$$\left. \times \int \frac{\left(\frac{1}{\nu_j}\right)^{\frac{1}{2}}}{\Gamma\left(\frac{1}{2}\right)} (\lambda_j^2)^{-\frac{3}{2}} \exp\left(-\frac{1}{\nu_j\lambda_j^2}\right) \frac{1}{\Gamma\left(\frac{1}{2}\right)} (\nu_j)^{-\frac{3}{2}} \exp\left(-\frac{1}{\nu_j}\right) d\nu_j d\lambda_j^2 \right]$$

$$\times \int \frac{(\frac{1}{\xi})^{\frac{1}{2}}}{\Gamma\left(\frac{1}{2}\right)} (\tilde{\tau}^2)^{-\frac{3}{2}} \exp\left(-\frac{1}{\xi\tilde{\tau}^2}\right) \frac{1}{\Gamma(\frac{1}{2})} (\xi)^{-\frac{3}{2}} \exp\left(-\frac{1}{\xi}\right) d\xi d\tilde{\tau}^2$$

$$\propto \int \cdots \int \prod_{j=1}^{p} \frac{1}{\sqrt{2\pi\sigma^2\tau_j^2}} \exp\left(-\frac{\beta_j^2}{2\sigma^2\tau_j^2}\right) \prod_{j=1}^{p} \frac{\tilde{\lambda}_1^2}{2} \exp\left(-\frac{\tilde{\lambda}_1^2}{2}\tau_j^2\right)$$

$$\times \prod_{j=2}^{p} \frac{1}{\sqrt{2\pi\lambda_j^2\tilde{\tau}^2\sigma^2}} \exp\left\{ -\frac{(\beta_j - \beta_{j-1})^2}{2\lambda_j^2\tilde{\tau}^2\sigma^2} \right\} \prod_{j=2}^{p} \frac{(\frac{1}{\nu_j})^{\frac{1}{2}}}{\Gamma(\frac{1}{2})} (\lambda_j^2)^{-\frac{3}{2}} \exp\left(-\frac{1}{\nu_j\lambda_j^2}\right)$$

$$\times \frac{(\frac{1}{\xi})^{\frac{1}{2}}}{\Gamma(\frac{1}{2})} (\tilde{\tau}^2)^{-\frac{3}{2}} \exp(-\frac{1}{\xi\tilde{\tau}^2}) \frac{1}{\Gamma(\frac{1}{2})} (\xi)^{-\frac{3}{2}} \exp\left(-\frac{1}{\xi}\right) \prod_{j=2}^{p} \frac{1}{\Gamma(\frac{1}{2})} (\nu_j)^{-\frac{3}{2}} \exp\left(-\frac{1}{\nu_j}\right)$$

$$\times d\tilde{\tau}^2 d\xi \prod_{j=1}^{p} d\tau_j^2 \prod_{j=2}^{p} d\lambda_j^2 \prod_{j=2}^{p} d\nu_j$$

$$\propto \int \cdots \int (\sigma^2)^{-\frac{2p-1}{2}} (\tilde{\tau}^2)^{-\frac{p-1}{2}} \pi(\tilde{\tau}^2 \mid \xi)\pi(\xi) \prod_{j=1}^{p} (\tau_j^2)^{-\frac{1}{2}} \prod_{j=2}^{p} (\lambda_j^2)^{-\frac{1}{2}} \exp\left(-\frac{1}{2\sigma^2}\boldsymbol{\beta}^T\boldsymbol{B}^{-1}\boldsymbol{\beta}\right)$$

$$\times \prod_{j=1}^{p} \pi(\tau_j^2) \prod_{j=2}^{p} \pi(\lambda_j^2 \mid \nu_j) \prod_{j=2}^{p} \pi(\nu_j) d\tilde{\tau}^2 d\xi \prod_{j=1}^{p} d\tau_j^2 \prod_{j=2}^{p} d\lambda_j^2 \prod_{j=2}^{p} d\nu_j.$$

References

Andrews, D. F., & Mallows, C. L. (1974). Scale mixtures of normal distributions. *Journal of the Royal Statistical Society: Series B, 36*(1), 99–102.

Banerjee, S. (2022). Horseshoe shrinkage methods for Bayesian fusion estimation. *Computational Statistics & Data Analysis, 174*, 107450.

Bhattacharya, A., Chakraborty, A., & Mallick, B. K. (2016). Fast sampling with Gaussian scale-mixture priors in high-dimensional regression. *Biometrika, 103*(4), 985–991.

Bondell, H. D., & Reich, B. J. (2008). Simultaneous regression shrinkage, variable selection, and supervised clustering of predictors with oscar. *Biometrics, 64*(1), 115–123.

Brooks, S. P., & Gelman, A. (1998). General methods for monitoring convergence of iterative simulations. *Journal of Computational and Graphical Statistics, 7*(4), 434–455.

Carvalho, C. M., Polson, N. G., & Scott, J. G. (2010). The horseshoe estimator for sparse signals. *Biometrika, 97*(2), 465–480.

Castillo, I., Schmidt-Hieber, J., & Van der Vaart, A. (2015). Bayesian linear regression with sparse priors. *The Annals of Statistics, 43*(5), 1986–2018.

Friedman, J., Hastie, T., Höfling, H., & Tibshirani, R. (2007). Pathwise coordinate optimization. *The Annals of Applied Statistics, 1*(2), 302–332.

Gelman, A., & Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science, 7*(4), 457–472.

Griffin, J., & Brown, P. (2005). Alternative prior distributions for variable selection with very many more variables than observations. University of Kent Technical Report

Griffin, J. E., & Brown, P. J. (2011). Bayesian hyper-lassos with non-convex penalization. *Australian & New Zealand Journal of Statistics, 53*(4), 423–442.
Hager, W. W. (1989). Updating the inverse of a matrix. *SIAM Review, 31*(2), 221–239.
Jang, W., Lim, J., Lazar, N. A., Loh, J. M., & Yu, D. (2015). Some properties of generalized fused lasso and its applications to high dimensional data. *Journal of the Korean Statistical Society, 44*(3), 352–365.
Johndrow, J., Orenstein, P., & Bhattacharya, A. (2020). Scalable Approximate MCMC Algorithms for the Horseshoe Prior. *Journal of Machine Learning Research, 21*(73), 1–61.
Kyung, M., Gill, J., Ghosh, M., & Casella, G. (2010). Penalized regression, standard errors, and Bayesian lassos. *Bayesian Analysis, 5*(2), 369–411.
Lee, C., Luo, Z. T., & Sang, H. (2021). T-LoHo: A Bayesian Regularization Model for Structured Sparsity and Smoothness on Graphs. *Advances in Neural Information Processing Systems, 34*, 598–609.
Makalic, E., & Schmidt, D. F. (2015). A simple sampler for the horseshoe estimator. *IEEE Signal Processing Letters, 23*(1), 179–182.
Nalenz, M., & Villani, M. (2018). Tree ensembles with rule structured horseshoe regularization. *The Annals of Applied Statistics, 12*(4), 2379–2408.
Park, T., & Casella, G. (2008). The Bayesian lasso. *Journal of the American Statistical Association, 103*(482), 681–686.
She, Y. (2010). Sparse regression with exact clustering. *Electronic Journal of Statistics, 4*, 1055–1096.
Shen, X., & Huang, H. C. (2010). Grouping pursuit through a regularization solution surface. *Journal of the American Statistical Association, 105*(490), 727–739.
Shimamura, K., Ueki, M., Kawano, S., & Konishi, S. (2019). Bayesian generalized fused lasso modeling via neg distribution. *Communications in Statistics-Theory and Methods, 48*(16), 4132–4153.
Song, Q., & Cheng, G. (2020). Bayesian fusion estimation via t shrinkage. *Sankhya A, 82*(2), 353–385.
Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B, 58*(1), 267–288.
Tibshirani, R., Saunders, M., Rosset, S., Zhu, J., & Knight, K. (2005). Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society: Series B, 67*(1), 91–108.
Vats, D., & Knudson, C. (2021). Revisiting the Gelman-Rubin diagnostic. *Statistical Science, 36*(4), 518–529.
Vehtari, A., Gelman, A., Simpson, D., Carpenter, B., & Bürkner, P. C. (2021). Rank-normalization, folding, and localization: an improved $\hat{R}$ for assessing convergence of MCMC (with discussion). *Bayesian Analysis, 16*(2), 667–718.
Wand, M. P., Ormerod, J. T., Padoan, S. A., & Frühwirth, R. (2011). Mean field variational Bayes for elaborate distributions. *Bayesian Analysis, 6*(4), 847–900.
Wang, Y. X., Sharpnack, J., Smola, A. J., & Tibshirani, R. J. (2016). Trend Filtering on Graphs. *Journal of Machine Learning Research, 17*(105), 1–41.
Watanabe, S. (2010). Equations of states in singular statistical estimation. *Neural Networks, 23*(1), 20–34.