Bayesian generalized fused lasso modeling via NEG distribution

doi: 10.1080/03610926.2018.1489056

Bayesian generalized fused lasso modeling via NEG distribution

Kaito Shimamura[1], Masao Ueki[2],
Shuichi Kawano[3] and Sadanori Konishi[4]

[1]*NTT Advanced Technology Corporation.*

[2]*Statistical Genetics Team, RIKEN Center for Advanced Intelligence Project.*

[3]*Graduate School of Informatics and Engineering, The University of Electro-Communications, 1-5-1, Chofugaoka, Chofu-shi, Tokyo 182-8585, Japan.*
skawano@uec.ac.jp (corresponding author)

[4]*Department of Mathematics, Faculty of Science and Engineering, Chuo University.*

Abstract: The fused lasso penalizes a loss function by the L_1 norm for both the regression coefficients and their successive differences to encourage sparsity of both. In this paper, we propose a Bayesian generalized fused lasso modeling based on a normal-exponential-gamma (NEG) prior distribution. The NEG prior is assumed into the difference of successive regression coefficients. The proposed method enables us to construct a more versatile sparse model than the ordinary fused lasso using a flexible regularization term. Simulation studies and real data analyses show that the proposed method has superior performance to the ordinary fused lasso.

Key Words and Phrases: Bayesian lasso, Hierarchical Bayesian model, Normal-exponential-gamma distribution, Markov chain Monte Carlo.

Short title: Bayesian generalized fused lasso via NEG

1 Introduction

With the advanced computer systems and progress in instrumentation technologies, the extremely high-dimensional data are being observed and recorded in biology, genomics, and many other fields of science. For such data, the usual methods of separating model estimation and evaluation are ineffectual for constructing an optimal model, and thus effective techniques are required to construct a statistical model with high reliability and prediction. This created a need for work on modeling and has led to the proposal of various regularization methods with an L_1 penalty term, in addition to the sum of squared errors or log-likelihood functions. A distinctive feature of the L_1 regularization methods is their capability for simultaneous model estimation and variable selection.

Lasso proposed by Tibshirani (1996) is the most fundamental tool, which imposes the sum of absolute values (L_1 norms) of the regression coefficients as a constraint on the sum of squared errors. Tibshirani et al. (2005) also proposed the fused lasso for the

analysis of data where the predictor variables are in some sense ordered. The fused lasso can be used for sparse modeling both for regression coefficients and for their successive differences, and it has become the focus of increasing interest as a useful technique in life sciences, image processing, and many other fields (see, e.g., Friedman et al. (2007) and Tibshirani and Wang (2008)). In the L_1 type of regularization, however, the L_1 norm constraint is non-differentiable at zero and no closed-form solution is available. Various estimation algorithms for lasso have therefore been developed such as the least angle regression (LARS) algorithm of Efron et al. (2004) and the coordinate descent algorithm of Friedman et al. (2007).

Tibshirani (1996) demonstrated that the lasso estimates can be interpreted as a posterior mode estimation when the regression parameters have independent and identical Laplace (double-exponential) priors. Park and Casella (2008) suggested Gibbs sampling for the lasso with a Laplace prior in a hierarchical model. Kyung et al. (2010) proposed a Bayesian fused lasso by interpreting the fused lasso in a Bayesian framework, assuming a product of the Laplace distribution in the prior of the regression coefficient vector. It might be, however, pointed out that the methods which encourage sparsity between neighboring variables via the L_1 norm such as the fused lasso and Bayesian fused lasso may have a substantial bias in their estimates, because the ordinary methods impose a large penalty for differences between regression coefficients that belong to different groups. As a result, the group difference is not contrasted, and then it may incur inaccuracy of prediction.

In order to overcome these issues, we propose a Bayesian sparse fused lasso and a Bayesian sparse generalized fused lasso based on the normal-exponential-gamma (NEG) prior distribution. The NEG penalty allows construction of highly versatile sparse models, because it has spike at zero and more extreme flatness in its tail than does the lasso penalty (Griffin and Brown 2005; Hoggart et al. 2008). Using a NEG prior to the difference of successive regression coefficients, our Bayesian sparse modeling can yield clearly different estimates for parameters in different groups and improves prediction accuracy. For parameter estimation, we present a simple implementation of the Gibbs sampling for the proposed Bayesian sparse model, by exploiting the hierarchical representation of the NEG prior analogous to that of the Laplace prior in Bayesian lasso (Park and Casella 2008). A drawback of Gibbs sampling in sparse Bayesian modeling is that the random numbers hinder producing exact sparse solutions (e.g. in estimate by posterior mode). To overcome the limitation, we develop an algorithm, called a sparse fused algorithm, which can produce exact sparse solutions from Gibbs sampling. We also investigate a model selection criterion for evaluating the estimated models.

The rest of this paper is organized as follows. Section 2 devotes the L_1 norm regularization. In Section 3, we describe the Bayesian sparse modeling which formulates the sparse estimation in a Bayesian framework. In Section 4, we propose a Bayesian sparse

modeling having higher versatility than the fused lasso using the NEG distribution. Monte Carlo simulations and real data analysis are conducted to examine the performance of our proposed procedure and to compare it with existing methods in Section 5. Concluding remarks are given in Section 6.

2 L_1 norm regularization

In this section, we describe the L_1 norm regularization, where the sum of absolute values of regression coefficients is imposed in a penalty term. In particular, we describe the lasso, fused lasso, and generalized fused lasso.

2.1 Regularized likelihood method

Suppose that we have observed data $\{(y_i, \boldsymbol{x}_i); i = 1, 2, \ldots, n\}$ for response variable y and p-dimensional predictor variables $\boldsymbol{x} = (x_1, x_2, \ldots, x_p)^T$. Without loss of generality, the response is centered around the mean and the predictors are standardized:

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

We consider the following linear regression model without the intercept:

$$\boldsymbol{y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon}, \tag{1}$$

where $\boldsymbol{y} = (y_1, y_2, \ldots, y_n)^T$ is the n-dimensional vector of observed values for the response variable, $X = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)^T$ is the $n \times p$ design matrix, $\boldsymbol{\beta}$ is the p-dimensional regression coefficient vector, and $\boldsymbol{\epsilon}$ is the n-dimensional error vector distributed as $\mathrm{N}_n(\mathbf{0}_n, \sigma^2 I_n)$. Then, the likelihood function is given by

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

where

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

Hereafter, we denote the probability density function $f(y_i|\boldsymbol{x}_i; \boldsymbol{\beta}, \sigma^2)$ as $f(y_i|\boldsymbol{\beta}, \sigma^2)$ for simplicity.

A regularization method imposes a constraint condition for $\boldsymbol{\beta}$ with a penalty function $P(\boldsymbol{\beta})$ (> 0) on the maximization of the loss function such as a log-likelihood function

$\log f(\boldsymbol{y}|\boldsymbol{\beta}, \sigma^2)$. We consider the following constrained optimization problem:

$$\max_{\boldsymbol{\beta}} \sum_{i=1}^{n} \log f(y_i|\boldsymbol{\beta}, \sigma^2), \quad \text{subject to } P(\boldsymbol{\beta}) \leq t, \tag{3}$$

where t (≥ 0) is a constant. The above optimization problem is equivalent to the maximization of the following objective function,

$$\sum_{i=1}^{n} \log f(y_i|\boldsymbol{\beta}, \sigma^2) - p_\gamma(\boldsymbol{\beta}), \tag{4}$$

where $p_\gamma(\boldsymbol{\beta})$ (> 0) is a penalty function corresponding to the constraint $P(\boldsymbol{\beta}) \leq t$ and γ (> 0) is a tuning parameter to control the degree of penalties, called the regularization parameter. When $p_\gamma(\boldsymbol{\beta}) = \gamma \|\boldsymbol{\beta}\|_2^2$, the optimization problem (4) reduces to the ridge regression problem proposed by Hoerl and Kennard (1970). The ridge regression improves the prediction performance, but it cannot produce zero values for regression coefficients.

2.2 Lasso

When $p_\gamma(\boldsymbol{\beta}) = \gamma \sum_{j=1}^{p} |\beta_j|$, the optimization problem (4) reduces to the lasso problem by Tibshirani (1996):

$$\hat{\boldsymbol{\beta}} = \arg\max_{\boldsymbol{\beta}} \left\{ \log f(\boldsymbol{y}|\boldsymbol{\beta}, \sigma^2) - \gamma \sum_{j=1}^{p} |\beta_j| \right\}. \tag{5}$$

In contrast to the shrinkage of regression coefficients toward zero that occurs in ridge regression, the lasso results in exactly zero estimates for some of the coefficients. The regularization parameter γ controls the overall model sparsity (that is, the model with exactly zero values for the coefficients) and shrinkage of the regression coefficients. A larger value of the regularization parameter produces sparser models.

2.3 Fused lasso

Tibshirani et al. (2005) proposed the fused lasso for the sake of analyzing data whose predictor variables are in some sense ordered. The regularization procedure gives estimates by

$$\hat{\boldsymbol{\beta}} = \arg\max_{\boldsymbol{\beta}} \left\{ \log f(\boldsymbol{y}|\boldsymbol{\beta}, \sigma^2) - \lambda_1 \sum_{j=1}^{p} |\beta_j| - \lambda_2 \sum_{j=2}^{p} |\beta_j - \beta_{j-1}| \right\},$$

where λ_1 (> 0) and λ_2 (> 0) are regularization parameters. The λ_1 controls the degree of sparsity and λ_2 controls the degree of smoothing between successive differences. If $\lambda_2 = 0$, the fused lasso reduces to the lasso. In recent years, the fused lasso has become the focus

of increasing interest as a useful technique in genomic data analysis, image processing, and many other fields (see, e.g., Friedman et al. (2007), Tibshirani and Wang (2008)). The upper left panel of Fig. 3 shows the penalty

$$p_{\lambda_2}(\beta_j) = \lambda_2\Big(|\beta_j - \beta_{j-1}| + |\beta_{j+1} - \beta_j|\Big) \tag{6}$$

as a function of β_j, while we fix both β_{j-1} and β_{j+1}.

A general form of the generalized fused lasso is given by

$$\hat{\boldsymbol{\beta}} = \arg\max_{\boldsymbol{\beta}} \left\{ \log f(\boldsymbol{y}|\boldsymbol{\beta}, \sigma^2) - \lambda_1 \sum_{j=1}^{p} |\beta_j| - \lambda_2 \sum_{(j,k)\in E} |\beta_j - \beta_k| \right\},$$

where $E \subset \{(j,k); j,k = 1,\ldots,p\}$. It is important to determine the set E according to the subject of the analysis. Examples of the generalized fused lasso include hexagonal operator for regression with shrinkage and equality selection (HORSES; Jang et al. 2013), which is a regularization method that maximizes the objective function

$$\log f(\boldsymbol{y}|\boldsymbol{\beta}, \sigma^2) - \lambda_1 \sum_{j=1}^{p} |\beta_j| - \lambda_2 \sum_{j>k} |\beta_j - \beta_k|.$$

In HORSES, all combinations between two regression coefficients are used as a penalty. Although in the fused lasso, the predictors must be in some sense ordered, HORSES, on the other hand, does not require that condition.

One of useful applications of the fused lasso is the fused lasso signal approximator (FLSA; Friedman et al. 2007). The FLSA solves the optimization problem

$$\min_{\beta_1,\ldots,\beta_n} \left\{ \sum_{i=1}^{n} (y_i - \beta_i)^2 + \lambda_1 \sum_{i=1}^{n} |\beta_i| + \lambda_2 \sum_{i=2}^{n} |\beta_i - \beta_{i-1}| \right\}. \tag{7}$$

The FLSA corresponds to the case where $n = p$ and $X = I_n$ in the ordinary fused lasso. Tibshirani and Wang (2008) applied the FLSA to the analysis of comparative genomic hybridization (CGH) data.

3 Bayesian sparse modeling via Gibbs sampling

In this section, we describe the Bayesian lasso which formulates the lasso in a Bayesian framework. We consider the Bayesian sparse estimation with an NEG distribution as the prior distribution instead of the Laplace prior distribution. In addition, the Bayesian fused lasso is described to formulate the fused lasso in a Bayesian framework.

3.1 Bayesian lasso

The posterior distribution of coefficient vector $\boldsymbol{\beta}$ is given by

$$\pi(\boldsymbol{\beta}|\boldsymbol{y}) \propto f(\boldsymbol{y}|\boldsymbol{\beta}, \sigma^2)\pi(\boldsymbol{\beta}|\sigma^2)\pi(\sigma^2).$$

The coefficient vector $\boldsymbol{\beta}$ is estimated by the posterior mode for given data $\boldsymbol{y}$. Park and Casella (2008) used the Laplace prior on the coefficient vector $\boldsymbol{\beta}$:

$$\pi(\boldsymbol{\beta}|\sigma^2) = \prod_{j=1}^{p} \frac{\lambda}{2\sqrt{\sigma^2}} \exp\left(-\frac{\lambda}{\sqrt{\sigma^2}}|\beta_j|\right) \tag{8}$$

and the non-informative scale-invariant prior $\pi(\sigma^2) = 1/\sigma^2$ or inverse-gamma prior $\pi(\sigma^2) = \mathrm{IG}(\nu_0/2, \eta_0/2)$ on σ^2, where ν_0 (> 0) is a shape parameter and η_0 (> 0) is a scale parameter. An inverse-gamma probability density function is given by

$$\mathrm{IG}(x|\nu, \eta) = \frac{\eta^\nu}{\Gamma(\nu)} x^{-(\nu+1)} \exp\left(-\frac{\eta}{x}\right),$$

where $\Gamma(\cdot)$ is the gamma function. The hyper-parameter λ in (8) plays the same role as that of regularization parameter γ in (5). It controls the degree of sparsity of the coefficients estimated. In other words, the larger values of hyper-parameter λ get, the more numbers of zero regression coefficients increase. The smaller values of λ get, the less numbers of zero regression coefficients increase.

The Laplace distribution is represented by a scale mixture of normals (Andrews and Mallows 1974):

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

From this relationship, Park and Casella (2008) assumed the following priors:

$$\begin{aligned} \pi(\boldsymbol{\beta}|\sigma^2, \tau_1^2, \tau_2^2, \ldots, \tau_p^2) &= \prod_{j=1}^{p} \frac{1}{\sqrt{2\pi\sigma^2\tau_j^2}} \exp\left(-\frac{\beta_j^2}{2\sigma^2\tau_j^2}\right), \\ \pi(\tau_1^2, \tau_2^2, \ldots, \tau_p^2) &= \prod_{j=1}^{p} \frac{\lambda^2}{2} \exp\left(-\frac{\lambda^2}{2}\tau_j^2\right). \end{aligned}$$

As a result, it enables us to carry out Bayesian estimation by Gibbs sampling. Assuming an inverse-gamma prior $\mathrm{IG}(\nu_0/2, \eta_0/2)$ on σ^2:

$$\pi(\sigma^2) = \frac{(\eta_0/2)^{\nu_0/2}}{\Gamma(\nu_0/2)} (\sigma^2)^{-(\nu_0/2+1)} \exp\left(-\frac{\eta_0/2}{\sigma^2}\right),$$

the full-conditional posteriors on $\boldsymbol{\beta}, \sigma^2, \tau_1^2, \tau_2^2, \ldots, \tau_p^2$ are given by

$$\begin{aligned}
\boldsymbol{\beta}|\boldsymbol{y}, X, \sigma^2, \tau_1^2, \tau_2^2, \ldots, \tau_p^2 &\sim \mathrm{N}_p(A^{-1}X^T\boldsymbol{y},\ \sigma^2 A^{-1}), \\
&\quad A = X^TX + D_r^{-1}, \quad D_r = \mathrm{diag}(\tau_1^2,\ \tau_2^2,\ \ldots,\ \tau_p^2), \\
\sigma^2|\boldsymbol{y}, X, \boldsymbol{\beta}, \tau_1^2, \tau_2^2, \ldots, \tau_p^2 &\sim \mathrm{IG}\left(\frac{\nu_1}{2},\ \frac{\eta_1}{2}\right), \\
&\quad \nu_1 = n + p + \nu_0, \quad \eta_1 = \|\boldsymbol{y} - X\boldsymbol{\beta}\|_2^2 + \boldsymbol{\beta}^T D_r^{-1}\boldsymbol{\beta} + \eta_0, \\
\frac{1}{\tau_j^2}\middle| \beta_j, \sigma^2, \lambda &\sim \mathrm{IGauss}(\mu',\ \lambda'), \\
&\quad \mu' = \sqrt{\frac{\lambda^2\sigma^2}{\beta_j^2}}, \quad \lambda' = \lambda^2, \quad j = 1, 2, \ldots, p,
\end{aligned}$$

where $\mathrm{IGauss}(\mu, \lambda)$ denotes the inverse-Gaussian distribution with a density function

$$\sqrt{\frac{\lambda}{2\pi}}x^{-3/2}\exp\left\{-\frac{\lambda(x-\mu)^2}{2\mu^2 x}\right\} \quad (x > 0).$$

3.2 Bayesian fused lasso

Kyung et al. (2010) proposed the Bayesian fused lasso by interpreting the fused lasso in a Bayesian framework. In the Bayesian fused lasso, the prior distribution of the regression coefficients $\boldsymbol{\beta}$ is defined as follows:

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

This can be expressed as a hierarchical representation of the Laplace distribution,

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

where

$$\Sigma_{\boldsymbol{\beta}}^{-1} = \left\{ \begin{array}{cccccc} \frac{1}{\tau_1^2}+\frac{1}{\widetilde{\tau}_2^2} & -\frac{1}{\widetilde{\tau}_2^2} & 0 & \cdots & 0 & 0 \\ -\frac{1}{\widetilde{\tau}_2^2} & \frac{1}{\tau_2^2}+\frac{1}{\widetilde{\tau}_2^2}+\frac{1}{\widetilde{\tau}_3^2} & -\frac{1}{\widetilde{\tau}_3^2} & \cdots & 0 & 0 \\ 0 & -\frac{1}{\widetilde{\tau}_3^2} & \frac{1}{\tau_3^2}+\frac{1}{\widetilde{\tau}_3^2}+\frac{1}{\widetilde{\tau}_4^2} & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & \frac{1}{\tau_{p-1}^2}+\frac{1}{\widetilde{\tau}_{p-1}^2}+\frac{1}{\widetilde{\tau}_p^2} & -\frac{1}{\widetilde{\tau}_p^2} \\ 0 & 0 & 0 & \cdots & -\frac{1}{\widetilde{\tau}_p^2} & \frac{1}{\tau_p^2}+\frac{1}{\widetilde{\tau}_p^2} \end{array} \right\}. \tag{9}$$

This formulation enables us to implement Gibbs sampler for $\boldsymbol{\beta}, \sigma^2, \tau_1^2, \tau_2^2, \ldots, \tau_p^2$ and $\widetilde{\tau}_2^2, \widetilde{\tau}_3^2, \ldots, \widetilde{\tau}_p^2$. The full-conditional distribution is then given by

$$\begin{aligned} &\boldsymbol{\beta}|\boldsymbol{y}, X, \sigma^2, \tau_1^2, \tau_2^2, \ldots, \tau_p^2, \widetilde{\tau}_2^2, \widetilde{\tau}_3^2, \ldots, \widetilde{\tau}_p^2 \\ &\qquad \sim \mathrm{N}_p\left((X^TX + \Sigma_{\boldsymbol{\beta}}^{-1})^{-1}X^T\boldsymbol{y},\ \sigma^2(X^TX + \Sigma_{\boldsymbol{\beta}}^{-1})^{-1}\right), \\ &\sigma^2|\boldsymbol{y}, X, \boldsymbol{\beta}, \tau_1^2, \tau_2^2, \ldots, \tau_p^2, \widetilde{\tau}_2^2, \widetilde{\tau}_3^2, \ldots, \widetilde{\tau}_p^2 \sim \mathrm{IG}\left(\nu_1/2,\ \eta_1/2\right), \\ &\qquad \nu_1 = n + 2p - 1 + \nu_0, \\ &\qquad \eta_1 = (\boldsymbol{y} - X\boldsymbol{\beta})^T(\boldsymbol{y} - X\boldsymbol{\beta}) + \boldsymbol{\beta}^T\Sigma_{\boldsymbol{\beta}}^{-1}\boldsymbol{\beta} + \eta_0, \\ &\frac{1}{\tau_j^2}|\beta_j, \sigma^2, \lambda_1 \sim \mathrm{IGauss}\left(\sqrt{\frac{\lambda_1^2\sigma^2}{\beta_j^2}},\ \lambda_1^2\right), \\ &\frac{1}{\widetilde{\tau}_j^2}|\beta_j, \beta_{j-1}, \sigma^2, \lambda_2 \sim \mathrm{IGauss}\left(\sqrt{\frac{\lambda_2^2\sigma^2}{(\beta_j - \beta_{j-1})^2}},\ \lambda_2^2\right), \end{aligned}$$

where an inverse-gamma prior distribution $\mathrm{IG}(\nu_0/2, \eta_0/2)$ is assumed for σ^2.

3.3 Lasso-type Bayesian sparse regression via NEG prior

Griffin and Brown (2005) proposed an NEG distribution as a prior distribution for the regression coefficients $\boldsymbol{\beta}$ which is more flexible with respect to sparsity than a Laplace distribution. As in Laplace distribution, the NEG distribution has a hierarchical representation which is useful to derive a simple Gibbs sampling algorithm as shown in the following. The NEG density function is given by

$$\mathrm{NEG}(\beta_j|\lambda, \gamma) = \kappa \exp\left(\frac{\beta_j^2}{4\gamma^2}\right) D_{-2\lambda-1}\left(\frac{|\beta_j|}{\gamma}\right), \tag{10}$$

where $\kappa = (2^\lambda\lambda)/(\gamma\sqrt{\pi})\Gamma(\lambda + 1/2)$ is a normalization constant, $D_{-2\lambda-1}$ is a parabolic cylinder function, and λ and γ are hyper-parameters with positive values that control the sparsity of the coefficients. The parabolic cylinder function is a solution of the

second-order linear ordinary differential equation

$$\frac{d^2w}{dz^2} - \left(\frac{z^2}{4} - \frac{1}{2} - a\right) w = 0,$$

and its integral representation is given by

$$D_{-2\lambda-1}\left(\frac{|\beta|}{\gamma}\right) = \frac{1}{\Gamma(2\lambda+1)} \exp\left(-\frac{\beta^2}{4\gamma^2}\right) \int_0^\infty w^{2\lambda} \exp\left(-\frac{1}{2}w^2 - \frac{|\beta|}{\gamma}w\right) dw.$$

Then, NEG density function can be expressed as a hierarchical representation

$$\begin{aligned} & \mathrm{NEG}\left(\beta_j|\lambda,\gamma\right) \\ = & \int\int \frac{1}{\sqrt{2\pi\tau_j^2}} \exp\left(-\frac{\beta_j^2}{2\tau_j^2}\right) \psi_j \exp\left(-\psi_j\tau_j^2\right) \frac{(\gamma^2)^\lambda}{\Gamma(\lambda)} \psi_j^{\lambda-1} \exp\left(-\gamma^2\psi_j\right) d\tau_j^2 d\psi_j \\ = & \int\int \mathrm{N}(\beta_j|0,\tau_j^2)\mathrm{EXP}(\tau_j^2|\psi_j)\mathrm{Ga}(\psi_j|\lambda,\gamma^2) d\tau_j^2 d\psi_j. \end{aligned}$$

The lasso-type Bayesian sparse estimation via an NEG distribution (Griffin and Brown 2011; Rockova and Lesaffre 2014) assumes the following NEG distribution instead of the Laplace distribution as a prior distribution for the regression coefficients $\boldsymbol{\beta}$,

$$\pi(\boldsymbol{\beta}|\sigma^2) = \prod_{j=1}^{p} \frac{1}{\sqrt{\sigma^2}} \mathrm{NEG}\left(\frac{\beta_j}{\sqrt{\sigma^2}}\Big|\lambda,\gamma\right).$$

By assuming the above prior distribution, it is possible to guarantee a unimodal posterior distribution (Rockova and Lesaffre 2014) and perform Bayesian estimation of the regression coefficient vector by Gibbs sampling in the same way as the Bayesian lasso. The full-conditional distributions of $\boldsymbol{\beta}, \sigma^2, 1/\tau_j^2$ and ψ_j $(j = 1, 2, \ldots p)$ are given by

$$\begin{aligned} \boldsymbol{\beta}|\boldsymbol{y}, X, \sigma^2, \tau_1^2, \tau_2^2, \ldots, \tau_p^2 &\sim \mathrm{N}_p(A^{-1}X^T\boldsymbol{y}, \sigma^2 A^{-1}), \\ & A = X^TX + D_r^{-1}, \quad D_r = \mathrm{diag}(\tau_1^2, \ldots, \tau_p^2), \\ \sigma^2|\boldsymbol{y}, X, \boldsymbol{\beta}, \tau_1^2, \tau_2^2, \ldots, \tau_p^2 &\sim \mathrm{IG}(\nu_1/2, \eta_1/2), \\ & \nu_1 = n + p + \nu_0, \quad \eta_1 = \|\boldsymbol{y} - X\boldsymbol{\beta}\|_2^2 + \boldsymbol{\beta}^T D_r^{-1}\boldsymbol{\beta} + \eta_0, \\ \frac{1}{\tau_j^2}|\beta_j, \sigma^2, \psi_j &\sim \mathrm{IGauss}(\mu', \lambda'),\ j = 1, 2, \ldots, p, \\ & \mu' = \sqrt{\frac{2\psi_j\sigma^2}{\beta_j^2}}, \quad \lambda' = 2\psi_j, \\ \psi_j|\tau_j^2, \lambda, \gamma &\sim \mathrm{Ga}(\lambda + 1, \tau_j^2 + \gamma^2),\ j = 1, 2, \ldots, p. \end{aligned}$$

The NEG distribution can maintain flat tails with a large preponderance of the density

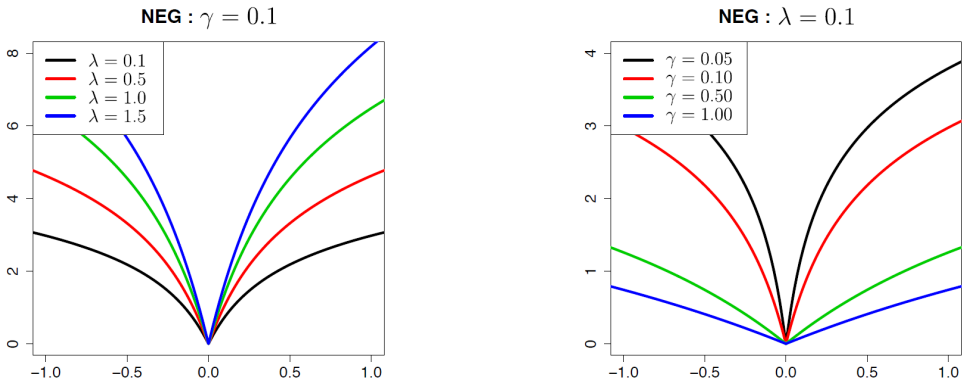


Figure 1: The NEG penalty function in Equation (13), $p_{\lambda,\gamma}(\beta) = \log \mathrm{NEG}(\beta|\lambda,\gamma) + C$. The left panel shows functions under varying λ at $\gamma = 0.1$, while the right panel shows those under varying γ at $\lambda = 0.1$.

around zero, making the resulting estimator more clear-cut. As both λ and γ increase such that $\xi = \sqrt{2\lambda}/\gamma$ remains a constant, the NEG distribution converges to the Laplace distribution with a parameter ξ. The NEG distribution is differentiable everywhere except at the point 0. First and second derivatives of the NEG density function at $\beta \neq 0$ are respectively given by

$$\frac{\partial}{\partial\beta}\mathrm{NEG}(\beta) = -\kappa\frac{2(\lambda+1/2)\mathrm{sign}(\beta)}{\gamma}\exp\left(\frac{\beta^2}{4\gamma^2}\right)D_{-(2\lambda+2)}\left(\frac{|\beta|}{\gamma}\right), \tag{11}$$

$$\frac{\partial^2}{\partial\beta^2}\mathrm{NEG}(\beta) = \kappa\frac{4(\lambda+1/2)(\lambda+1)}{\gamma^2}\exp\left(\frac{\beta^2}{4\gamma^2}\right)D_{-(2\lambda+3)}\left(\frac{|\beta|}{\gamma}\right). \tag{12}$$

Fig. 1 shows the NEG penalty function

$$p_{\lambda,\gamma}(\beta) = \log \mathrm{NEG}(\beta|\lambda,\gamma) + C, \tag{13}$$

when the hyper-parameters are varied, where C is a constant such that $p_{\lambda,\gamma}(\beta)$ takes zero value at $\widetilde{\beta} = \arg\min p_{\lambda,\gamma}(\beta)$. The hyper-parameters λ and γ affect the degree of sparsity of the solution: either a larger value of λ or a smaller value of γ produces sparser results. Setting an appropriate value of the hyper-parameters is an important problem. Rockova and Lesaffre (2014) summarized the properties of the NEG distribution. The most remarkable property is

$$\frac{\partial}{\partial\beta}\log\mathrm{NEG}\left(\beta\middle|\lambda,\gamma\right) = O\left(\frac{1}{|\beta|}\right) \quad \text{as} \quad |\beta| \to \infty,$$

which implies that the regression estimator is less biased for large $|\beta|$. The lasso estimator varies continuously, but is highly biased because of the strong constraint imposed

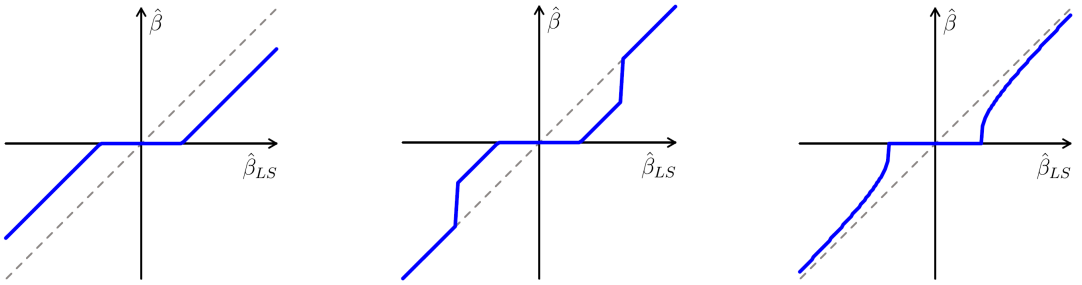


Figure 2: The relationship between the least-squares estimator and shrinkage estimator for lasso (left panel), SCAD (middle panel), and NEG (right panel). The dotted lines are the least-squares estimator $\hat{\beta}_{LS}$, while the solid lines are shrinkage estimators.

on nonzero estimates. It will be more clear by considering the univariate least-squares problem with a penalty term $p_\lambda(\beta)$,

$$\hat{\beta} = \arg\min_{\beta} \left\{ \frac{1}{2}(\hat{\beta}_{LS} - \beta)^2 + p_\gamma(\beta) \right\}, \tag{14}$$

where $\hat{\beta}_{LS}$ is the unpenalized least-squares estimate in univariate case. Fig. 2 shows $\hat{\beta}$ from the optimization problem (14) with lasso, smoothly-clipped absolute deviation (SCAD; Fan and Li 2001), and the NEG (13) penalties. The lasso has a large bias from $\hat{\beta}_{LS}$. SCAD has less biased for large $|\hat{\beta}_{LS}|$. This property is also true of the minimax concave penalty (MCP; Zhang 2010). The NEG penalty yields similar estimators to those from SCAD penalty, but the change is continuous in $\hat{\beta}_{LS}$.

4 Bayesian fused lasso modeling via NEG prior

4.1 Bayesian fused lasso via NEG prior

In this section, we propose a Bayesian sparse modeling with higher versatility than the fused lasso. The Bayesian fused lasso assumes two independent Laplace distributions as the prior distributions for the regression coefficients $\boldsymbol{\beta}$ and their successive differences. By replacing the Laplace distribution for the differences for the regression coefficients with the NEG distribution, we propose the prior distribution

$$\begin{aligned} \pi(\boldsymbol{\beta}|\sigma^2) &= (\sigma^2)^{-(2p-1)/2} \prod_{j=1}^{p} \text{Laplace}\left(\frac{\beta_j}{\sqrt{\sigma^2}} \middle| \lambda_1 \right) \\ &\quad \times \prod_{j=2}^{p} \text{NEG}\left(\frac{\beta_j - \beta_{j-1}}{\sqrt{\sigma^2}} \middle| \lambda_2, \gamma_2 \right), \end{aligned} \tag{15}$$

where $\lambda_1, \lambda_2, \gamma_2$ are hyper-parameters with positive values. Using the NEG distribution, compared to the Laplace distribution, the closer the difference between two regression coefficients is, the stronger the penalty becomes. Consequently, by adding the NEG penalty for the differences for regression coefficients, the truly identical regression coefficients tend to be estimated as identical, while the truly different regression coefficients tend to be estimated as different. Note that we adapt the NEG distribution only to the fused penalty, because imposing the NEG distribution to both penalties causes much computational cost.

The upper right panel of Fig. 3 shows the penalty function

$$\begin{aligned} p_{\lambda_2,\gamma_2}(\beta_j) &= \log \mathrm{NEG}(\beta_j - \beta_{j-1}|\lambda_2, \gamma_2) \\ &\quad + \log \mathrm{NEG}(\beta_{j+1} - \beta_j|\lambda_2, \gamma_2) + C, \end{aligned} \tag{16}$$

where C is a constant term such that the function $p_{\lambda_2,\gamma_2}(\beta_j)$ takes zero value at $\widetilde{\beta} = \arg\min p_{\lambda_2,\gamma_2}(\beta_j)$. When $\widetilde{\beta}$ satisfies an inequality $\beta_{j-1} \leq \widetilde{\beta} \leq \beta_{j+1}$, the fused lasso penalty $p_{\lambda_2}(\widetilde{\beta})$ always takes the minimum value, but the penalty of the proposed method does not always. The resulting estimator based on prior (15) tends to be identical to either β_{j-1} or β_{j+1}, and more contrasted result is obtained than the fused lasso penalty. This shows that the prior (15) is more flexible than that of the Bayesian fused lasso.

A full-conditional distribution is obtained for each of the prior distributions, enabling Bayesian estimation by Gibbs sampling. The prior (15) can be expressed as a hierarchical representation

$$\begin{aligned} \pi(\boldsymbol{\beta}|\sigma^2) &= (\sigma^2)^{-(2p-1)/2} \prod_{j=1}^{p} \mathrm{Laplace}\left(\frac{\beta_j}{\sqrt{\sigma^2}}\middle|\lambda_1\right) \prod_{j=2}^{p} \mathrm{NEG}\left(\frac{\beta_j - \beta_{j-1}}{\sqrt{\sigma^2}}\middle|\lambda_2, \gamma_2\right) \\ &= \int \cdots \int \prod_{j=1}^{p} \frac{1}{\sqrt{2\pi\sigma^2\tau_j^2}} \exp\left(-\frac{\beta_j^2}{2\sigma^2\tau_j^2}\right) \prod_{j=1}^{p} \frac{\lambda_1^2}{2} \exp\left(-\frac{\lambda_1^2\tau_j^2}{2}\right) \\ &\quad \times \prod_{j=2}^{p} \frac{1}{\sqrt{2\pi\sigma^2\widetilde{\tau}_j^2}} \exp\left\{-\frac{(\beta_j - \beta_{j-1})^2}{2\sigma^2\widetilde{\tau}_j^2}\right\} \prod_{j=2}^{p} \psi_j \exp\left(-\psi_j\widetilde{\tau}_j^2\right) \\ &\quad \times \prod_{j=2}^{p} \frac{(\gamma_2^2)^{\lambda_2}}{\Gamma(\lambda_2)} \psi_j^{\lambda_2 - 1} \exp(-\gamma_2^2\psi_j) \prod_{j=1}^{p} d\tau_j^2 \prod_{j=2}^{p} d\widetilde{\tau}_j^2 \prod_{j=2}^{p} d\psi_j. \end{aligned}$$

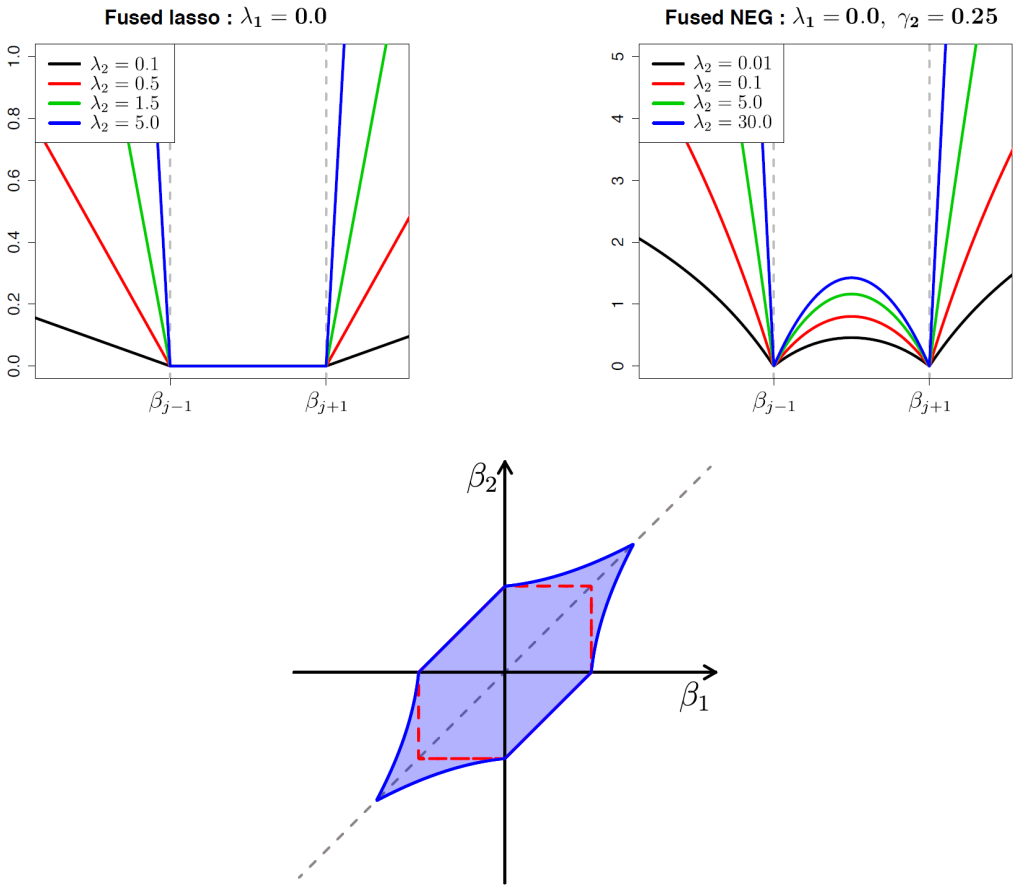


Figure 3: Upper left panel: The function (6), $p_{\lambda_2}(\beta_j) = \lambda_2\Big(|\beta_j - \beta_{j-1}| + |\beta_{j+1} - \beta_j|\Big)$, where β_{j-1} and β_{j+1} are fixed. Upper right panel: The function (16), $p_{\lambda_2,\gamma_2}(\beta_j) = \log \mathrm{NEG}(\beta_j - \beta_{j-1}|\lambda_2, \gamma_2) + \log \mathrm{NEG}(\beta_{j+1} - \beta_j|\lambda_2, \gamma_2) + C$, where β_{j-1} and β_{j+1} are fixed. Lower panel: A constraint region of fused lasso via NEG penalty (shaded region). The red dotted line indicates fused lasso.

Therefore, the priors on $\boldsymbol{\beta}, \tau_1^2, \tau_2^2, \ldots, \tau_p^2, \widetilde{\tau}_2^2, \widetilde{\tau}_3^2, \ldots, \widetilde{\tau}_p^2, \psi_2, \psi_3, \ldots, \psi_p$ are

$$\begin{aligned}
\boldsymbol{\beta}|\sigma^2, \tau_1^2, \tau_2^2, \ldots, \tau_p^2, \widetilde{\tau}_2^2, \widetilde{\tau}_3^2, \ldots, \widetilde{\tau}_p^2 &\sim \mathrm{N}_p(\mathbf{0}_p, \sigma^2 \Sigma_{\boldsymbol{\beta}}), \\
\tau_j^2 &\sim \mathrm{EXP}(\lambda_1^2/2), \\
\widetilde{\tau}_j^2|\psi_j &\sim \mathrm{EXP}(\psi_j), \\
\psi_j &\sim \mathrm{Ga}(\lambda_2, \gamma_2^2),
\end{aligned}$$

where $\Sigma_{\boldsymbol{\beta}}$ is given by the formula (9). Hence the full-conditional distributions of parameters are given by

$$\begin{aligned}
\boldsymbol{\beta}|\boldsymbol{y}, X, \sigma^2, \tau_1^2, \ldots, \tau_p^2, \widetilde{\tau}_2^2, \ldots, \widetilde{\tau}_p^2, \psi_2, \ldots, \psi_p &\sim \mathrm{N}_p\left(A^{-1}X^T\boldsymbol{y},\ \sigma^2 A^{-1}\right), \\
A &= X^TX + \Sigma_{\boldsymbol{\beta}}^{-1}, \\
\sigma^2|\boldsymbol{y}, X, \boldsymbol{\beta}, \tau_1^2, \ldots, \tau_p^2, \widetilde{\tau}_2^2, \ldots, \widetilde{\tau}_p^2, \psi_2, \ldots, \psi_p &\sim \mathrm{IG}\left(\nu_1/2,\ \eta_1/2\right), \\
\nu_1 &= n + 2p - 1 + \nu_0, \\
\eta_1 &= (\boldsymbol{y} - X\boldsymbol{\beta})^T(\boldsymbol{y} - X\boldsymbol{\beta}) + \boldsymbol{\beta}^T\Sigma_{\boldsymbol{\beta}}^{-1}\boldsymbol{\beta} + \eta_0, \\
\frac{1}{\tau_j^2}|\beta_j, \sigma^2, \lambda_1 &\sim \mathrm{IGauss}\left(\sqrt{\frac{\lambda_1^2\sigma^2}{\beta_j^2}}, \lambda_1^2\right), \\
\frac{1}{\widetilde{\tau}_j^2}|\beta_j, \beta_{j-1}, \sigma^2, \psi_j &\sim \mathrm{IGauss}\left(\sqrt{\frac{2\sigma^2\psi_j}{(\beta_j - \beta_{j-1})^2}},\ 2\psi_j\right), \\
\psi_j|\widetilde{\tau}_j^2, \lambda_2, \gamma_2 &\sim \mathrm{Ga}\left(\lambda_2 + 1, \widetilde{\tau}_j^2 + \gamma_2^2\right). \qquad (17)
\end{aligned}$$

4.2 Bayesian generalized fused lasso via NEG prior

The generalized fused lasso is given by the optimization problem

$$\max_{\beta_1, \ldots, \beta_p} \left\{ -\sum_{i=1}^{n}(y_i - \beta_i)^2 - \lambda_1 \sum_{j=1}^{p}|\beta_j| - \lambda_2 \sum_{(k,l)\in E}|\beta_k - \beta_l| \right\}. \qquad (18)$$

Various problems are included under this framework by changing the set E. In this section, we consider using the NEG distribution for the generalized fused lasso.

4.2.1 2d fused lasso

The 2d fused lasso is a useful application of the generalized fused lasso. The purpose of this method is the denoising of image data. The gray scale of $p_1 \times p_2$ pixel in the image data corresponds to each $y_{i,j}$ $(i = 1, \ldots, p_1,\ j = 1, \ldots, p_2)$. We consider the following

optimization problem:

$$\max_{\beta_{1,1},\ldots,\beta_{p_1,p_2}} \left\{ -\sum_{i=1}^{p_1}\sum_{j=1}^{p_2}(y_{i,j}-\beta_{i,j})^2 - \lambda_1 \sum_{i=1}^{p_1}\sum_{j=1}^{p_2}|\beta_{i,j}| \right.$$
$$\left. -\lambda_2 \sum_{i=1}^{p_1}\sum_{j=2}^{p_2}|\beta_{i,j}-\beta_{i,j-1}| - \lambda_2 \sum_{i=2}^{p_1}\sum_{j=1}^{p_2}|\beta_{i,j}-\beta_{i-1,j}| \right\}. \tag{19}$$

The estimated value of parameter β_{ij} corresponds to the denoised image.

Next, we formulate the 2d fused lasso in a Bayesian framework. For the following discussions, we use the notations

$$\begin{aligned}
\boldsymbol{y} &= (y_{1,1},\ldots,y_{1,p_2},y_{2,1},\ldots,y_{2,p_2},\ldots,y_{p_1,1},\ldots,y_{p_1,p_2})^T \\
&= (y_1,y_2,\cdots,y_p)^T, \\
\boldsymbol{\beta} &= (\beta_{1,1},\ldots,\beta_{1,p_2},\beta_{2,1},\ldots,\beta_{2,p_2},\ldots,\beta_{p_1,1},\ldots,\beta_{p_1,p_2})^T \\
&= (\beta_1,\beta_2,\cdots,\beta_p)^T,
\end{aligned}$$

where $p = p_1 \times p_2$. The likelihood function and prior distribution on $\boldsymbol{\beta}$ are, respectively,

$$f(\boldsymbol{y}|\boldsymbol{\beta},\sigma^2) = \mathrm{N}_p(\boldsymbol{\beta},\sigma^2 I_p), \tag{20}$$

$$\pi(\boldsymbol{\beta}|\sigma^2) \propto (\sigma^2)^{-(3p-p_1-p_2)/2} \prod_{j=1}^{p} \frac{\lambda_1}{2} \exp\left(-\frac{\lambda_1}{\sigma}|\beta_j|\right) \times \prod_{j\in\Omega_1} \mathrm{NEG}(\beta_j-\beta_{j-1}|\lambda_2,\gamma_2) \prod_{j\in\Omega_2} \mathrm{NEG}(\beta_j-\beta_{j-p_2}|\lambda_2,\gamma_2), \tag{21}$$

where $\Omega_1 = \{1,2,\ldots,p\}\backslash\{1,p_2+1,\ldots,(p_1-1)p_2+1\}$, $\Omega_2 = \{p_2+1,p_2+2,\ldots,p\}$. The

prior (21) can be expressed as a hierarchical representation

$$
\begin{aligned}
\pi(\boldsymbol{\beta}|\sigma^2) \;=\; & \int\cdots\int \prod_{j=1}^{p} \frac{1}{\sqrt{2\pi\sigma^2\tau_j^2}} \exp\left(-\frac{\beta_j^2}{2\sigma^2\tau_j^2}\right) \prod_{j=1}^{p} \frac{\lambda_1^2}{2} \exp\left(-\frac{\lambda_1^2\tau_j^2}{2}\right) \\
& \times \prod_{j\in\Omega_1} \frac{1}{\sqrt{2\pi\sigma^2\widetilde{\tau}_{j-1,j}^2}} \exp\left\{-\frac{(\beta_j-\beta_{j-1})^2}{2\sigma^2\widetilde{\tau}_{j-1,j}^2}\right\} \\
& \times \prod_{j\in\Omega_1} \psi_{j-1,j} \exp\left(-\psi_{j-1,j}\widetilde{\tau}_{j-1,j}^2\right) \\
& \times \prod_{j\in\Omega_1} \frac{(\gamma_2^2)^{\lambda_2}}{\Gamma(\lambda_2)} \psi_{j-1,j}^{\lambda_2-1} \exp(-\gamma_2^2\psi_{j-1,j}) \\
& \times \prod_{j\in\Omega_2} \frac{1}{\sqrt{2\pi\sigma^2\widetilde{\tau}_{j-p_2,j}^2}} \exp\left\{-\frac{(\beta_j-\beta_{j-p_2})^2}{2\sigma^2\widetilde{\tau}_{j-p_2,j}^2}\right\} \\
& \times \prod_{j\in\Omega_2} \psi_{j-p_2,j} \exp\left(-\psi_{j-p_2,j}\widetilde{\tau}_{j-p_2,j}^2\right) \\
& \times \prod_{j\in\Omega_2} \frac{(\gamma_2^2)^{\lambda_2}}{\Gamma(\lambda_2)} \psi_{j-p_2,j}^{\lambda_2-1} \exp\left(-\gamma_2^2\psi_{j-p_2,j}\right) \\
& \times \prod_{j=1}^{p} d\tau_j^2 \prod_{j\in\Omega_1} d\widetilde{\tau}_{j-1,j}^2 \prod_{j\in\Omega_1} d\psi_{j-1,j} \prod_{j\in\Omega_2} d\widetilde{\tau}_{j-p_2,j}^2 \prod_{j\in\Omega_2} d\psi_{j-p_2,j}.
\end{aligned}
$$

The full-conditional distribution is then obtained by replacing $\Sigma_{\boldsymbol{\beta}}^{-1}$ by the following expression in the fused lasso-type Bayesian modeling via the NEG distribution in Equation (17):

$$
(\Sigma_{\boldsymbol{\beta}}^{-1})_{(i,j)} = \begin{cases} \frac{1}{\tau_i^2} + \frac{1}{\widetilde{\tau}_{i-1,j}^2} + \frac{1}{\widetilde{\tau}_{i-p_2,j}^2} + \frac{1}{\widetilde{\tau}_{i,j+1}^2} + \frac{1}{\widetilde{\tau}_{i,j+p_2}^2}, & (i=j), \\ -\frac{1}{\widetilde{\tau}_{i,j}^2}, & (j\in\{i+1,\ i+p_2, i-1, i-p_2\}), \\ 0, & (\text{otherwise}), \end{cases}
$$

where $(\Sigma_{\boldsymbol{\beta}}^{-1})_{(i,j)}$ is the (i,j)-element of $\Sigma_{\boldsymbol{\beta}}^{-1}$ and $1/\widetilde{\tau}_{i,j}^2 = 1/\widetilde{\tau}_{j,i}^2$, $1/\widetilde{\tau}_{j'-1,j'}^2 = 0$ for $j' \in \{1,\ldots,p\}\setminus\Omega_1$, while $1/\widetilde{\tau}_{j'-p_2,j'}^2 = 0$ for $j' \in \{1,\ldots,p\}\setminus\Omega_2$.

4.2.2 HORSES

In the fused lasso, the predictors must be in some sense ordered. On the other hand, HORSES does not have such a requirement. In the HORSES, all pairwise differences of two regression coefficients are used as a penalty. The regularization method maximizes

the objective function

$$\log f(\boldsymbol{y}|\boldsymbol{\beta},\sigma^2) - \lambda_1 \sum_{j=1}^{p} |\beta_j| - \lambda_2 \sum_{j>k} |\beta_j - \beta_k|. \tag{22}$$

Next, we formulate HORSES in a Bayesian framework. The prior on $\boldsymbol{\beta}$ is assumed as

$$\pi(\boldsymbol{\beta}|\sigma^2) = (\sigma^2)^{-(p+p(p-1)/2)/2} \prod_{j=1}^{p} \text{Laplace}\left(\frac{\beta_j}{\sqrt{\sigma^2}}\Big|\lambda_1\right) \prod_{j>k} \text{NEG}\left(\frac{\beta_j - \beta_k}{\sqrt{\sigma^2}}\Big|\lambda_2, \gamma_2\right).$$

The full-conditional distribution is obtained by replacing the $p \times p$ matrix $\Sigma_{\boldsymbol{\beta}}$ in the fused lasso-type Bayesian modeling via an NEG distribution in (17) by

$$(\Sigma_{\boldsymbol{\beta}}^{-1})_{(i,j)} = \begin{cases} \dfrac{1}{\tau_i^2} + \displaystyle\sum_{j' \neq i} \frac{1}{\widetilde{\tau}_{i,j'}^2} & (i = j) \\ -\dfrac{1}{\widetilde{\tau}_{i,j}^2} & (\text{otherwise}) \end{cases},$$

where $(\Sigma_{\boldsymbol{\beta}}^{-1})_{(i,j)}$ is the (i,j)-element of $\Sigma_{\boldsymbol{\beta}}^{-1}$.

4.3 Computational algorithm for exact sparse solution

Since a posterior mode is estimated by random numbers, the Gibbs sampling does not produce exact zero estimates of the coefficients. The fused lasso has two purposes: sparse estimation of both the coefficients and differences between adjacent regression coefficients. To achieve these two purposes, we propose a sparse fused algorithm, which allows both regression coefficients and differences of regression coefficients to be exactly zero. Table 1 shows the proposed algorithm.

A detail of the algorithm is given as follows. Steps 1 and 2 are initialization. The index vector I stores information on groups of regression coefficients, where the same values indicate that they are in the same group. In addition, we assume that $\hat{\boldsymbol{\beta}}$ is the mode of sampled estimates. Step 3 updates three vectors $\widetilde{\boldsymbol{\beta}}^{(f)}, \widetilde{\boldsymbol{\beta}}^{(b)}, \widetilde{\boldsymbol{\beta}}^{(z)}$ that are used for comparison of posterior distribution. $\widetilde{\beta}_k^{(f)}$ is updated in the value of the group before the group of regression coefficients belonging to $\hat{\beta}_k$, while $\widetilde{\beta}_k^{(b)}$ is updated in that after the group of regression coefficients belonging to $\hat{\beta}_k$. We assign $\widetilde{\beta}_k^{(z)}$ into zero value. Step 3.1 computes values of the posterior distributions. In Step 3.2, estimated values are obtained from the magnitude of the posterior distribution. In addition, we update the estimate of regression coefficient and group information. Step 2 and Step 3 are repeated until convergence. By modifying this algorithm slightly, we can also construct an algorithm for the generalized fused lasso.

Table 1: Sparse fused algorithm

1. Let $\hat{\boldsymbol{\beta}} = (\hat{\beta}_1, \ldots, \hat{\beta}_p)^T$
   be a vector of estimates obtained from Gibbs sampling.
   $\hat{\boldsymbol{\beta}}$ is the mode of sampled estimates.
   $\boldsymbol{I} = (I_1, I_2, \ldots, I_p) \leftarrow (1, 2, \ldots, p)$
2. $\widetilde{\boldsymbol{\beta}} = (\widetilde{\beta}_1, \widetilde{\beta}_2, \ldots, \widetilde{\beta}_p)^T \leftarrow \hat{\boldsymbol{\beta}}$
   $\widetilde{\boldsymbol{\beta}}^{(f)} = (\widetilde{\beta}_1^{(f)}, \widetilde{\beta}_2^{(f)}, \ldots, \widetilde{\beta}_p^{(f)})^T \leftarrow \hat{\boldsymbol{\beta}}$
   $\widetilde{\boldsymbol{\beta}}^{(b)} = (\widetilde{\beta}_1^{(b)}, \widetilde{\beta}_2^{(b)}, \ldots, \widetilde{\beta}_p^{(b)})^T \leftarrow \hat{\boldsymbol{\beta}}$
   $\widetilde{\boldsymbol{\beta}}^{(z)} = (\widetilde{\beta}_1^{(z)}, \widetilde{\beta}_2^{(z)}, \ldots, \widetilde{\beta}_p^{(z)})^T \leftarrow \hat{\boldsymbol{\beta}}$
3. FOR $j = 1, \ldots, p$
   FOR $k = 1, \ldots, p$
   IF $I_k = j$ THEN
   IF $j \neq 1$ THEN
   SET $\widetilde{\beta}_k^{(f)} \leftarrow \hat{\beta}_{j-1}$
   END IF
   IF $j \neq p$ THEN
   SET $\widetilde{\beta}_k^{(b)} \leftarrow \hat{\beta}_{j+1}$
   END IF
   SET $\widetilde{\beta}_k^{(z)} \leftarrow 0$
   END IF
   END FOR

   3.1 $G = g(\tilde{\boldsymbol{\beta}},\ \hat{\boldsymbol{\xi}},\ \boldsymbol{y})$
   $G_{(f)} = g(\widetilde{\boldsymbol{\beta}}^{(f)},\ \hat{\boldsymbol{\xi}},\ \boldsymbol{y})$
   $G_{(b)} = g(\widetilde{\boldsymbol{\beta}}^{(b)},\ \hat{\boldsymbol{\xi}},\ \boldsymbol{y})$
   $G_{(z)} = g(\widetilde{\boldsymbol{\beta}}^{(z)},\ \hat{\boldsymbol{\xi}},\ \boldsymbol{y})$
   $\mathcal{G} = \{G,\ G_{(f)},\ G_{(b)},\ G_{(z)}\}$

   3.2 FOR $k = 1, \ldots, p$
   IF $I_k = j$ THEN
   CASE $\max\{\mathcal{G}\}$ OF
   G : $\hat{\beta}_k \leftarrow \widetilde{\beta}_j$
   IF $j \neq 1$ THEN
   $G_{(f)}$: $\hat{\beta}_k \leftarrow \widetilde{\beta}_{j-1}$
   $I_k \leftarrow j - 1$
   END IF
   IF $j \neq p$ THEN
   $G_{(b)}$: $\hat{\beta}_k \leftarrow \widetilde{\beta}_{j+1}$
   $I_k \leftarrow j + 1$
   END IF
   $G_{(z)}$: $\hat{\beta}_k \leftarrow 0$
   $I_k \leftarrow 0$
   END CASE
   END IF
   END FOR
   END FOR
4. Repeat Steps 2 and 3 until convergence and sparsified estimates are stored in $\hat{\boldsymbol{\beta}}$.

Here, $g(\boldsymbol{\beta},\ \boldsymbol{\xi},\ \boldsymbol{y}) = \log f(\boldsymbol{y}|\boldsymbol{\beta},\ \boldsymbol{\xi}) + \log \pi(\boldsymbol{\beta},\ \boldsymbol{\xi})$, $f(\boldsymbol{y}|\boldsymbol{\beta},\ \boldsymbol{\xi})$ is a likelihood function, $\pi(\boldsymbol{\beta},\ \boldsymbol{\xi})$ is a prior on $(\boldsymbol{\beta},\ \boldsymbol{\xi})$, and $\hat{\boldsymbol{\xi}}$ is an estimate of parameter vector $\boldsymbol{\xi}$ that consists of parameter vectors except for $\boldsymbol{\beta}$.

4.4 Model selection

A set of processes for selecting the optimal model using a model selection criterion such as Akaike information criterion (AIC; Akaike 1973) or Bayesian information criterion (BIC; Schwarz 1978) is effective for evaluating a regression model estimated by the maximum likelihood or least-squares method (see, e.g., Konishi and Kitagawa (2008)). However, when analyzing high-dimensional data, the traditional methods are not effective. Chen and Chen (2008) proposed an extended Bayesian information criterion (EBIC) to overcome the difficulties in model selection for small sample and high-dimensional data frequently encountered in genomic studies and image analysis.

The basic idea of EBIC is as follows. Suppose that the likelihood function is $L_n(\boldsymbol{\theta}) = f(y|\boldsymbol{\theta}) = \prod_{i=1}^n f(y_i|x_i, \boldsymbol{\theta})$, where $\boldsymbol{\theta} \in \Theta \subset R^p$. A model M is a subset of $\{1, \ldots, p\}$. It indicates indexes of variables included in the model. For M included in the model space $\mathcal{M}$, the posterior of M is given by

$$p(M|Y) = \frac{m(Y|M)p(M)}{\sum_{M\in\mathcal{M}} m(Y|M)p(M)},$$

where $m(Y|M)$ is the marginal likelihood and $p(M)$ is the prior of M. The marginal likelihood is

$$m(Y|M) = \int f\big\{Y|\boldsymbol{\theta}(M)\big\}\pi\big\{\boldsymbol{\theta}(M)\big\}d\boldsymbol{\theta}(M),$$

where $\pi\{\boldsymbol{\theta}(M)\}$ is the prior of $\boldsymbol{\theta}(M)$ being the parameter $\boldsymbol{\theta}$ of the model M. By the Laplace approximation for integrals in the above quantity, we derive

$$-2\log m(Y|M) = -2\log L_n\{\hat{\boldsymbol{\theta}}(M)\} + \nu(M)\log n - 2p(M),$$

where $\hat{\boldsymbol{\theta}}(M)$ is the maximum likelihood estimator of $\boldsymbol{\theta}(M)$, $\nu(M)$ is the degrees of freedom of M. In addition, terms of smaller order than $O(1)$ with respect to the sample size n are ignored. The BIC (Schwarz 1978) approximates the posterior probability of a model by assuming that the prior is uniform over all models, and is of the form

$$\mathrm{BIC}(M) = -2\log L_n\big\{\hat{\boldsymbol{\theta}}(M)\big\} + \nu(M)\log n.$$

For the theoretical aspect and derivation of the BIC, we refer to Konishi et al. (2004); Konishi and Kitagawa (2008).

On the other hand, the EBIC considers the prior probability on a model M which takes the number of candidate models into consideration, rather assuming a uniform prior. Suppose that a model space $\mathcal{M}$ is partitioned into $\coprod_j \mathcal{M}_j$. The EBIC is then given

by, for $M \in \mathcal{M}_j$,

$$\mathrm{EBIC}(M) = -2 \log L_n\{\hat{\boldsymbol{\theta}}(M)\} + \nu(M) \log n + 2\eta \log \tau(\mathcal{M}_j),$$

where η $(0 < \eta < 1)$ is a tuning parameter and $\tau(\mathcal{M}_j)$ is a quantity which characterizes $\mathcal{M}_j$. Chen and Chen (2008) used $\tau(\mathcal{M}_j) = \binom{p}{j} = p!/\{(p-j)!j!\}$ for variable selection problem. For the fused lasso-type problem, Tibshirani et al. (2005) proposed the degrees of freedom

$$\mathrm{df}(\hat{\boldsymbol{\beta}}) = \#\left\{\text{nonzero coefficient blocks in } \hat{\boldsymbol{\beta}}\right\}.$$

It can be rewritten as

$$\mathrm{df}(\hat{\boldsymbol{\beta}}) = p - \#\left\{\hat{\beta}_j = 0\right\} - \#\left\{\hat{\beta}_j = \hat{\beta}_{j-1}; \hat{\beta}_j, \hat{\beta}_{j-1} \neq 0\right\}.$$

In this paper, we use $\mathrm{df}(\hat{\boldsymbol{y}})$ as the degrees of freedom $\nu(M)$ in the EBIC above and $\tau(\mathcal{M}_j) = \binom{p_g}{\mathrm{df}(\hat{\boldsymbol{\beta}})} = p_g!/[\{p_g - \mathrm{df}(\hat{\boldsymbol{\beta}})\}!\mathrm{df}(\hat{\boldsymbol{\beta}})!]$, where p_g is the number of coefficient blocks in $\hat{\boldsymbol{\beta}}$ including zero coefficients. We also use $\eta = 1 - \log n/(2 \log p)$ as recommended by Chen and Chen (2008). The values of the hyper-parameters are determined by minimizing the EBIC.

5 Numerical studies

5.1 Monte Carlo simulation

We simulated data from the model with n observations and p predictors:

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

where $\boldsymbol{\beta}^*$ is the p-dimensional true coefficient vector, $\boldsymbol{\epsilon}$ is an error vector distributed as $\mathrm{N}_n(\mathbf{0}_n, \sigma^2 I_n)$. In addition, $\boldsymbol{x}_i$ $(i = 1, 2, \ldots, n)$ was generated from a multivariate normal distribution with mean vector $\mathbf{0}_p$ and variance-covariance matrix Σ. We simulated 200 datasets with n observations. We considered the following three cases.

- Case 1: $n = 50, p = 20$, $\boldsymbol{\beta}^* = (\mathbf{0.0}_5^T, \mathbf{2.0}_5^T, \mathbf{0.0}_5^T, \mathbf{2.0}_5^T)^T$, $\sigma = 0.75$, $\Sigma_{ii} = 1$, and $\Sigma_{ij} = 0.5$ $(i \neq j)$, where Σ_{ij} is the (i, j)-element of Σ.
- Case 2: $n = 50, p = 50$, $\boldsymbol{\beta}^* = (\mathbf{0.0}_5^T, \mathbf{5.0}_3^T, \mathbf{0.0}_{15}^T, \mathbf{3.5}_7^T, \mathbf{0.0}_{10}^T, \mathbf{4.5}_5^T, \mathbf{0.0}_5^T)^T$, $\sigma = 0.75$, and $\Sigma = I_p$.
- Case 3: $n = 30, p = 50$, $\boldsymbol{\beta}^* = (\mathbf{3.0}_5^T, -\mathbf{1.5}_5^T, \mathbf{1.0}_5^T, \mathbf{2.0}_5^T, \mathbf{0.0}_{30}^T)^T$, $\sigma = 5.0$, and $\Sigma_{ij} = 0.5^{|i-j|}$.

These simulation settings are the same as those in Tibshirani et al. (2005) and Shen and Huang (2010). Cases 1 and 3 correspond to Examples 1 and 2 in Shen and Huang (2010), respectively, while Case 2 corresponds to Figure 3 in Tibshirani et al. (2005).

We denote the blocks of indexes which have distinctive regression coefficients by $B_1, B_2, \ldots, B_L \subset \{1, 2, \ldots, p\}$. For example, $L = 4$ in Case 1. For each generated dataset, the estimates were obtained by using 5,000 iterations of Gibbs sampler (after 2,000 burn-in iterations). Candidates of the hyper-parameters were set by

$$\lambda_{\min} \exp\{(\log \lambda_{\max} - \log \lambda_{\min}) \cdot (i/m)\} \tag{23}$$

for $i = 1, \ldots, m$. For the hyper-parameters λ_1 and λ_2, we set $m = 100$, $\lambda_{\min} = 10^{-4}$, and $\lambda_{\max} = 50$ for Cases 1 and 2 and $\lambda_{\max} = 100$ for Case 3 such that all coefficient parameters are zero. For the hyper-parameters γ_2, we set $m = 100$, $\lambda_{\min} = 0.1$, and $\lambda_{\max} = 2$.

We used the lasso, fused lasso, and Bayesian fused lasso as competitors. Regularization parameters were selected by the EBIC.

The performances were evaluated in terms of two accuracies: variable selection and prediction. For variable selection accuracy, we used three measures:

$$\begin{aligned}
P_{\mathrm{Z}} &= \frac{1}{200}\sum_{k=1}^{200} \frac{\#\{j : \beta_j^{(k)} = 0 \wedge \beta_j^* = 0\}}{\#\{j : \beta_j^* = 0\}}, \\
P_{\mathrm{NZ}} &= \frac{1}{200}\sum_{k=1}^{200} \frac{\#\{j : \beta_j^{(k)} \neq 0 \wedge \beta_j^* \neq 0\}}{\#\{j : \beta_j^* \neq 0\}}, \\
P_{\mathrm{B}} &= \frac{1}{200}\sum_{k=1}^{200} \frac{p - \sum_{l=1}^{L} N_l^{(k)}}{p - L},
\end{aligned}$$

where $\hat{\boldsymbol{\beta}}^{(k)} = (\hat{\beta}_1^{(k)}, \ldots, \hat{\beta}_p^{(k)})^T$ is the estimate of coefficient vector for the k-th dataset, and $N_l^{(k)}$ is the number of distinct regression coefficients $\{\hat{\beta}_j^{(k)} : j \in B_l\}$. P_{Z} indicates the accuracy of identifying truly zero coefficients. P_{NZ} indicates the accuracy of identifying truly nonzero coefficients. P_{B} indicates the accuracy of identifying the true coefficient blocks. The higher the value, the more accurate variable selection is. We assessed the accuracy of prediction using the mean squared error (MSE) and prediction squared error (PSE) as follows:

$$\begin{aligned}
\mathrm{MSE} &= \frac{1}{200}\sum_{k=1}^{200} (\hat{\boldsymbol{\beta}}^{(k)} - \boldsymbol{\beta}^*)^T \Sigma (\hat{\boldsymbol{\beta}}^{(k)} - \boldsymbol{\beta}^*), \\
\mathrm{PSE} &= \frac{1}{200}\sum_{k=1}^{200} \left(\frac{1}{n} \|\hat{\boldsymbol{y}}^{(k)} - \widetilde{\boldsymbol{y}}^{(k)}\|_2^2 \right),
\end{aligned}$$

where $\widetilde{y}^{(k)} = X^{(k)}\boldsymbol{\beta}^* + \widetilde{\boldsymbol{\epsilon}}^{(k)}$, with $\widetilde{\boldsymbol{\epsilon}}^{(k)}$ being an observation independent of k-th error vector $\boldsymbol{\epsilon}^{(k)}$.

The simulation results are summarized in Table 2. First, the lasso showed low P_{B} in all cases, because it cannot handle regression coefficients as blocks, and hence blocks of zero coefficients exist. The fused lasso and Bayesian fused lasso outperformed the lasso because of accounting for the block structure. With respect to almost all criteria, the proposed method provided much better performance than the competitors. In particular, the true blocks were almost identified by the proposed method, which may be seen from the results that the values of P_{B} in Table 2 are close to 1. Moreover, the low values of MSE and PSE show that our method provides proper estimates for not only the true blocks but also their true regression coefficients.

Table 2: The results for Monte Carlo simulations. flasso indicates the fused lasso, Bflasso the Bayesian fused lasso, and NEG-flasso our proposed fused lasso-type modeling via the NEG prior distribution.

	Case 1 : $n = 50,\ p = 20$						
	MSE	(sd)	PSE	(sd)	P_{Z}	P_{NZ}	P_{B}
lasso	0.26	(0.13)	0.85	(0.21)	0.58	1.00	0.17
flasso	0.27	(0.20)	0.69	(0.15)	0.49	1.00	0.89
Bflasso	0.14	(0.10)	0.72	(0.12)	0.57	1.00	0.82
NEG-flasso	0.03	(0.05)	0.59	(0.12)	0.96	1.00	1.00

	Case 2 : $n = 50,\ p = 50$						
	MSE	(sd)	PSE	(sd)	P_{Z}	P_{NZ}	P_{B}
lasso	1.23	(0.71)	1.81	(0.80)	0.57	1.00	0.24
flasso	0.46	(0.24)	0.88	(0.20)	0.74	1.00	0.89
Bflasso	1.50	(1.27)	1.98	(1.35)	0.38	1.00	0.52
NEG-flasso	0.04	(0.03)	0.60	(0.12)	1.00	1.00	1.00

	Case 3 : $n = 30,\ p = 50$						
	MSE	(sd)	PSE	(sd)	P_{Z}	P_{NZ}	P_{B}
lasso	67.70	(23.77)	96.17	(33.17)	0.54	0.69	0.22
flasso	76.38	(36.55)	48.56	(12.40)	0.28	0.86	0.47
Bflasso	71.83	(32.11)	104.96	(42.83)	0.11	0.94	0.30
NEG-flasso	10.54	(8.92)	35.81	(10.56)	0.49	0.96	0.94

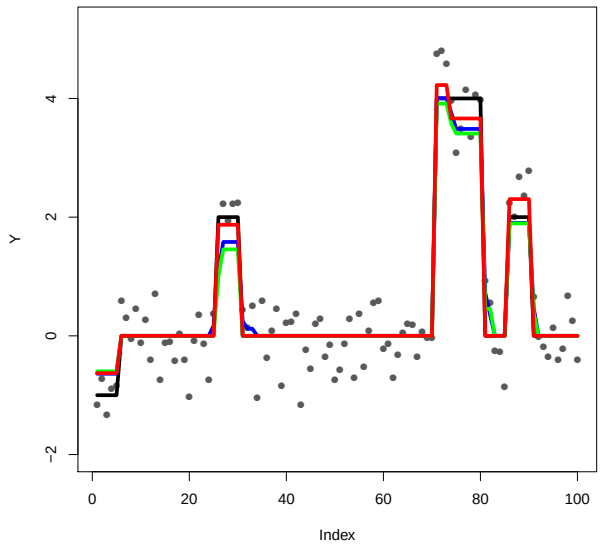


Figure 4: The result for the simulated data in Section 5.2. Black dots indicate the simulated data, the black line is the true model, the blue line is the estimate of fused lasso, the green line is the estimate of Bayesian fused lasso, and the red line is the estimate of the proposed method.

5.2 Demonstration with artificial data for FLSA model

We demonstrated our proposed method with artificial data generated from the FLSA model

$$\boldsymbol{y} = \boldsymbol{\beta}^* + \boldsymbol{\epsilon}, \tag{24}$$

where $\boldsymbol{\beta}^*$ is the p-dimensional true parameter and $\boldsymbol{\epsilon} \sim \mathrm{N}_p(\mathbf{0}_p, \sigma^2 I_p)$. We considered $\boldsymbol{\beta}^* = (-\mathbf{1}_5^T, \mathbf{0}_{20}^T, \mathbf{2}_5^T, \mathbf{0}_{40}^T, \mathbf{4}_{10}^T, \mathbf{0}_5^T, \mathbf{2}_5^T, \mathbf{0}_{10}^T)^T$ and $\sigma = 0.5$. This setting was inspired by Friedman et al. (2007). The hyper-parameters were tested for candidates given by (23), where $(m, \lambda_{\min}, \lambda_{\max}) = (200, 0.001, 30)$ for λ_1, $(m, \lambda_{\min}, \lambda_{\max}) = (200, 0.5, 30)$ for λ_2 , and $(m, \lambda_{\min}, \lambda_{\max}) = (5, 0.1, 2)$ for γ_2. We used the fused lasso and Bayesian fused lasso as competitors.

Fig. 4 gives estimates from the proposed method, fused lasso, and Bayesian fused lasso. In the fused lasso and Bayesian fused lasso, the blocks of estimated nonzero coefficients tended to shrunk toward zero. On the other hand, the proposed method could successfully estimate the true coefficients blocks. The proposed method gave no blocks consisting of single coefficient, while other methods had such seven blocks. In addition, the proposed method seems to capture the true structure better than other methods in whole.

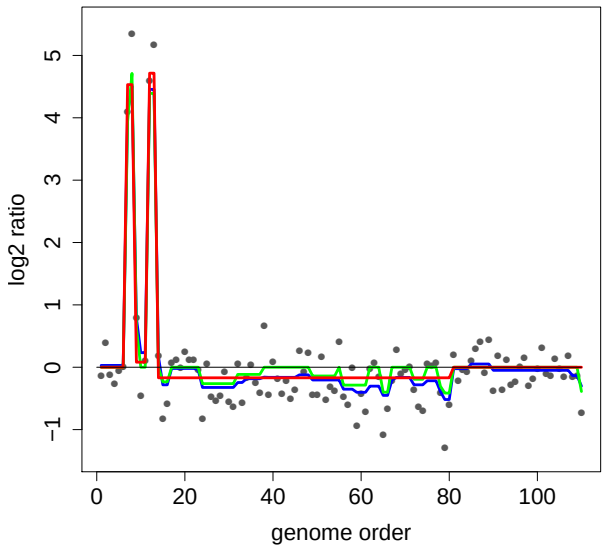


Figure 5: The result for the comparative genome hybridization (CGH) analysis. Black dots indicate data points, the blue line is the estimate of fused lasso, the green line is the estimate of Bayesian fused lasso, and the red line is the estimate of the proposed method.

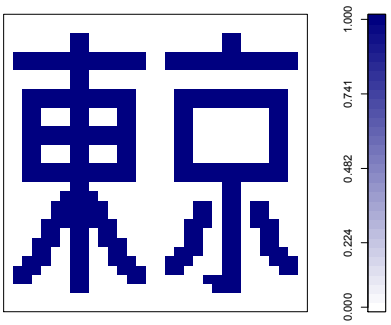
5.3 Comparative genomic hybridization analysis for FLSA model

We applied our proposed method to a real dataset: comparative genomic hybridization (CGH) data. The dataset is available from the `cghFLasso` package in the software `R`. We randomly extracted 110 samples from the dataset. We compared the proposed method to the FLSA procedure of Tibshirani and Wang (2008), which can be implemented in the `cghFLasso` package, and Bayesian FLSA procedure. The candidate values of the hyper-parameters λ_1, γ_2 were the same as those given in Section 5.2. For λ_2, we set $(m, \lambda_{\min}, \lambda_{\max}) = (200, 0.001, 60)$.

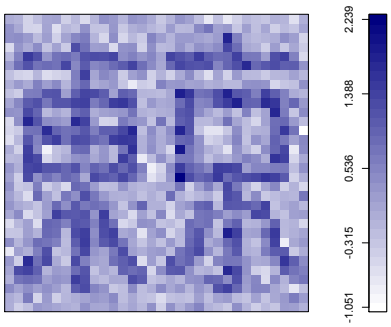
Fig. 5 gives the result for analyzing the CGH data. The FLSA procedure provided seemingly an over-fitted model. Bayesian FLSA procedure could avoid overfitting compared to the FLSA procedure, but it was unstable for a range from 20 to 80 genome orders. On the other hand, the proposed method seems to be stable for all ranges and gives more clear-cut estimates than other methods.

5.4 Demonstration with artificial data for 2d fused lasso model

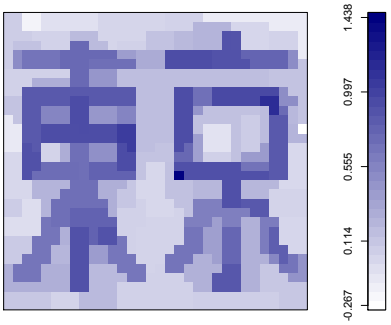
Next, we considered a numerical demonstration for the 2d fused lasso model applied to image reconstruction. A sample image was generated by simulation. The figure (a) in Fig. 6 shows the true image taking the values from 0 (blue) to 1 (white). The figure (b) in Fig. 6 shows a noisy image which has noises generated from normal distribution with mean 0 and standard deviation 0.35. These images are $32 \times 32 = 1024$ pixel in size. The hyper-



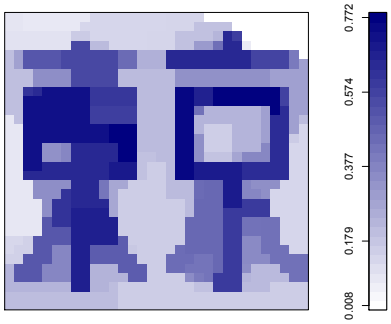
(a)



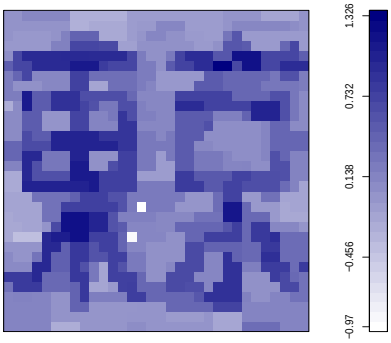
(b)



(c)



(d)



(e)

Figure 6: Results for artificial data generated from 2d fused lasso model. (a) is the true image, (b) the noisy image, (c) our proposed method, (d) the 2d fused lasso, and (e) the Bayesian 2d fused lasso.

parameters were tested for candidates given by (23), where $(m, \lambda_{\min}, \lambda_{\max}) = (200, 1, 200)$ for λ_2 and $(m, \lambda_{\min}, \lambda_{\max}) = (5, 0.1, 2)$ for γ_2. Note that we set $\lambda_1 = 0$ because this numerical study does not focus on estimating zeros of coefficients. We compared the proposed method to the 2d fused lasso by Friedman et al. (2007), implemented in the `genlasso` package in the software `R`, and Bayesian 2d fused lasso. The regularization parameters were chosen by the EBIC.

The figures (c), (d), (e) in Fig. 6 show the results of the proposed method, the 2d fused lasso, and the Bayesian 2d fused lasso, respectively. The 2d fused lasso and Bayesian 2d fused lasso often failed to differentiate between the blue and white areas in the true image. The proposed method more successfully recovered the true image. The result shows that the proposed method worked better than other methods. The squares error $\|\boldsymbol{\beta}^* - \hat{\boldsymbol{\beta}}\|_2^2$ by the proposed method was 50.38, while that by the 2d fused lasso and Bayesian 2d fused lasso was, respectively, 102.91 and 86.05. The results suggest that the proposed method may also be effective in image analysis.

6 Concluding remarks

We proposed the fused lasso-type estimation via NEG distribution for the penalty for differences between regression coefficients. Because the NEG distribution has a more extreme spike at zero and more tail flatness than the Laplace distribution, the proposed method enables us to estimate blocks of coefficients more clearly. In addition, we proposed the sparse fused algorithm to provide a solution which has exactly zero coefficients and produces exactly estimated blocks. Numerical examples showed that our proposed method provided a contrasted estimator, and worked better than existing methods.

An extension of the proposed method to other types of the generalized fused lasso method is important. For example, we may also replace the Laplace prior for the regression coefficients by the NEG prior. This extension would be useful for the situation in which estimating regression coefficients to be zeros is important in addition to merging regression coefficients. However, as described in Section 3.3 this additional extension increases computational cost, and hence, we need to balance between computational feasibility and estimation accuracy. For example, in Case 1 of Section 5.1, the computational time is about 11.4 hours at each dataset even if the NEG is applied only to the fusion penalty.

It is also interesting to develop information criteria such as the generalized Bayesian information criterion (GBIC; Konishi et al. 2004) for evaluating these methods. We leave these interesting topics as future work.

Acknowledgments

The authors would like to thank the anonymous reviewers for their constructive and helpful comments. M. Ueki was supported by Grant-in-Aid for Young Scientist (B) (25870074) and Grants-in-Aid for Scientific Research (C) (25330049 and 25460403). S. Kawano was supported by Grant-in-Aid for Young Scientist (B) (15K15947) and Grants-in-Aid for Scientific Research on Innovative Areas (16H06429, 16K21723, and 16H06430). The computational resource was provided by the Super Computer System, Human Genome Center, Institute of Medical Science, The University of Tokyo.

References

Akaike H (1973) Information theory and an extension of the maximum likelihood principle. In: (Petrov, B.N. and Csaki, F., eds.), Akademiai Kiado, Budapest, pp. 267–281. (Reproduced in *Breakthroughs in Statistics*, Volume 1, S. Kotz and N. L. Johnson, eds., Springer Verlag, New York, (1992))

Andrews DF, Mallows CL (1974) Scale mixtures of normal distributions. Journal of the Royal Statistical Society, Ser B, 36:99–102

Chen J, Chen Z (2008) Extended bayesian information criterion for model selection with large model space. Biometrika 94:759–771

Efron B, Hastie T, Johnstone I, Tibshirani R (2004) Least angle regression (with discussion). Annals of Statistics 32:407–499

Fan J, Li R (2001) Variable selection via nonconcave penalized likelihood and its oracle properties. Journal of the American Statistical Association 96:1348–1360

Friedman J, Hastie T, Hofling H, Tibshirani R (2007) Pathwise coordinate optimization. Annals of Applied Statistics 1:302–332

Griffin J, Brown P (2005) Alternative prior distributions for variable selection with very many more variables than observations. Tech. Rep. Technical Report, University of Warwick, Coventry, UK.

Griffin J, Brown P (2011) Bayesian hyper-lassos with non-convex penalization. Australian & New Zealand Journal of Statistics 53:423–442

Hoerl AE, Kennard RW (1970) Ridge regression: biased estimation for nonorthogonal problem. Technometrics 12:55–67

Hoggart CJ, Whittaker JC, Iroio MD, Balding DJ (2008) Simultaneous analysis of all snps in genome-wide and re-sequencing association studies. PLoS Genetics 4:e1000,130

Jang W, Lim J, Lazar M, Loh J, Yu D (2013) Regression shrinkage and grouping of highly correlated predictors with horses. Tech. Rep. arXiv:1302.0256

Konishi S, Kitagawa G (2008) Information Criteria and Statistical Modeling. Springer, New York

Konishi S, Ando T, Imoto S (2004) Bayesian information criteria and smoothing parameter selection in radial basis function networks. Biometrika 91:27–43

Kyung M, Gill J, Ghosh M, Casalla G (2010) Penalized regression, standard error, and bayesian lasso. Bayesian Analysis 5:369–412

Park T, Casella G (2008) The bayesian lasso. Journal of the American Statistical Association 103:681–686

Rockova V, Lesaffre E (2014) Incorporating grouping information in bayesian variable selection with applications genomics. Bayesian Analysis 9:221–258

Schwarz G (1978) Estimating the dimension of a model. Annals of Statistics 6:461–464

Shen X, Huang HC (2010) Grouping pursuit through a regularization solution surface. Journal of the American Statistical Association 105(490):727–739

Tibshirani R (1996) Regression shrinkage and selection via lasso. Journal of the Royal Statistical Society Series B 58:267–288

Tibshirani R, Wang P (2008) Spatial smoothing and hot spot detection for cgh data using the fused lasso. Biostatistics 9:18–29

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 67:91–108

Zhang CH (2010) Nearly unbiased variable selection under minimax concave penalty. The Annals of Statistics 38:894–942