BAYESIAN MODELING FRAMEWORK FOR DATA WITH GROUP STRUCTURE VIA VARIABLE FUSION

by

Yuko Kakikawa

Doctoral Thesis

Submitted to

Statistical Science Program

on March 2026

in Partial Fulfillment of the Requirements

for the degree of Doctor of Philosophy

Graduate Institute for Advanced Studies, SOKENDAI

Acknowledgements

I would like to express my heartfelt gratitude to Professor Yoshiyuki Ninomiya of the Institute of Statistical Mathematics for his devoted supervision and valuable discussions as my principal advisor throughout this research. In the course of my doctoral studies, his thoughtful guidance, continuous encouragement, and broad expertise were indispensable to my research, and his support greatly shaped my academic development.

I am also deeply grateful to Associate Professor Daisuke Murakami of the Institute of Statistical Mathematics for his insightful guidance and many fruitful discussions on the applications of this study and the latest research developments as my co-advisor. I would further like to sincerely thank Professor Hironori Fujisawa of the Institute of Statistical Mathematics and Professor Hidetoshi Matsui of Shiga University for kindly serving as examiners of this dissertation and for their valuable comments and careful review. I also wish to express my sincere appreciation to Professor Shuichi Kawano of Kyushu University for many helpful suggestions and valuable guidance throughout my research career.

In addition, I would like to extend my sincere thanks to all the members of the Institute of Statistical Mathematics for their kind support and cooperation during my research activities.

Finally, I wish to express my deepest gratitude to my family for their unwavering support, patience, and understanding throughout my academic journey.

Contents

Chapter 1

Introduction

Recent advances in computation and measurement have led to the accumulation of vast and diverse data, making their proper analysis increasingly important. Regression models provide interpretable links between explanatory variables and a response variable through estimated regression coefficients. In many modern applications, explanatory variables have an underlying structure such as spatial or temporal adjacency, and regression coefficients of neighboring variables are expected to show similar effects on the response. This adjacency-based dependence can be viewed as a form of group structure. Ignoring such group structure can lead to unstable estimates and reduced interpretability. To address this, variable fusion (Land and Friedman 1997) has been proposed to encourage adjacent or related coefficients to take similar values. Based on this idea, the fused lasso (Tibshirani et al., 2005) extends the lasso (Tibshirani 1996) which applies an L_1-penalty only to the regression coefficients to shrink them to zero, by introducing an additional one on differences between adjacent coefficients, thereby achieving both variable selection and variable fusion. As a result, it identifies important variables as well as groups of variables whose regression coefficients exhibit similar impact on the response. To handle a wider variety of data, the generalized fused lasso (Hoefling 2010) in which the penalty can be imposed on differences between arbitrary pairs of coefficients, was subsequently proposed. In addition, to alleviate the over-shrinkage of large coefficients caused by the L_1 penalty, concave penalties such as the smoothly clipped absolute deviation (SCAD; Fan and Li 2001) and the minimax concave penalty (MCP) (Zhang 2010) were developed for improved variable selection. Building on this idea, Jing et al. (2018) proposed the fused-MCP, which applies the MCP to fusion-type regularization. Similarly, Inoue et al. (2020) also applied the SCAD penalty to variable fusion and compared

its performance with that of the MCP in an apartment rent analysis.

Although variable fusion methods have mainly been discussed in the context of linear regression (Bondell and Reich 2008; Kim et al. 2009; She 2010), their application to binary data has attracted increasing attention in recent years. In fields such as genomics, chemometrics, and medical statistics, where binary outcomes (e.g., disease presence or class labels) are common, fused penalties have been utilized to obtain sparse yet smooth coefficient estimates in logistic regression models. For example, Yu et al. (2015) introduced a logistic regression model estimated using the fused lasso and demonstrated its usefulness in spectral data classification. For clinical data analysis, Lin et al. (2016) proposed a penalized logistic regression method that fuses adjacent categories, simultaneously achieving variable selection and category grouping in colon cancer screening data. Furthermore, Yan et al. (2022) proposed a heterogeneous logistic regression model that incorporates a concave fusion penalty across individual-specific coefficients to identify subgroups with similar effects in hypertension data, demonstrating its potential for individualized medicine.

Thus far, variable fusion methods have been discussed from a frequentist perspective. However, these approaches lack a principled way to quantify uncertainty and to treat different types of regularization within a unified probabilistic framework which enables consistent inference and model evaluation. Therefore, Bayesian formulations of variable fusion have also been proposed. Since the L_1 penalty corresponds to assuming a Laplace prior on the coefficients, the L_1-norm regularization methods can naturally be interpreted in a Bayesian framework. Park and Casella (2008) first introduced the Bayesian lasso by representing the Laplace prior as a scale mixture of normals, which enabled efficient Gibbs sampling instead of the inefficient Metropolis-Hastings updates required under the direct Laplace prior. Building on this framework, Kyung et al. (2010) extended it to variable fusion and proposed the Bayesian fused lasso, in which Laplace priors are placed on both the regression coefficients and their adjacent differences. For analyzing binary data, Genkin et al. (2007) and Betancourt et al. (2017) extended these Bayesian shrinkage methods to logistic regression models. However, existing Bayesian logistic regression models have not simultaneously achieved variable selection and variable fusion. In a related development, Masuda and Inoue (2022) applied the Bayesian generalized fused lasso to a Poisson regression model for spatial count data, where the fusion penalty was imposed across sample-specific

effects to identify spatial clusters. This approach, however, focuses on fusion among samples rather than among regression coefficients. At present, Bayesian extensions of the generalized fused lasso that perform fusion among regression coefficients corresponding to explanatory variables have received little attention beyond linear regression, even in logistic regression models.

While the above Bayesian approaches are based on the L_1 penalty, it should be noted that this penalty itself has a well-known limitation, as mentioned earlier: it tends to impose excessive shrinkage on large coefficients. To address this issue, nonconvex penalties such as SCAD and MCP have been developed in the frequentist framework. In contrast, within the Bayesian framework, global-local shrinkage priors (Polson and Scott 2010) have been proposed as alternatives to the Laplace prior. The horseshoe prior (Carvalho et al. 2010) is the widely used global-local shrinkage prior, which has both a sharper spike at zero and heavier tails than the Laplace prior. This property allows it to overcome not only the over-shrinkage of large coefficients but also the under-shrinkage of zero coefficients, both arising from the L_1 penalty. Makalic and Schmidt (2015) employed the horseshoe prior for variable selection and further constructed a Gibbs sampler based on its hierarchical representation. Subsequently, Banerjee 2022 applied the horseshoe prior to differences between adjacent regression coefficients in a one-dimensional fused lasso framework for signal approximation, showing its practical utility in signal denoising. While both approaches focus on either variable selection or variable fusion, Kakikawa and Kawano (2023) proposed a linear regression model that simultaneously incorporates both, assuming a Laplace prior on the regression coefficients and a horseshoe prior on the differences between adjacent coefficients. Furthermore, other global-local priors have also been applied to fusion penalties (Normal-exponential-gamma prior, Shimamura et al. 2019; Student-t prior, Song and Cheng 2020). For logistic regression, the use of global-local priors has been discussed in Makalic and Schmidt (2015) and Bhattacharyya et al. (2022); yet existing Bayesian logistic regression models still focus on variable selection, leaving variable fusion unaddressed. To date, attempts to incorporate global-local shrinkage priors into the fusion term have been made mostly for linear regression models, where variable selection is typically not considered.

As has become clear from the preceding discussion, the progress beyond linear regression has been limited both in the development of Bayesian generalized

fused lasso models and in the application of global-local shrinkage priors to fusion penalties. Motivated by these observations, this thesis first aims to provide a Bayesian logistic regression model based on the generalized fused lasso framework introducing global-local shrinkage priors. We first propose a logistic regression model that assumes a Laplace prior on both the regression coefficients and their adjacent differences. Then, owing to the horseshoe prior being theoretically well-understood and empirically effective, we focus on it among global-local shrinkage priors and propose an extended model by applying it to the differences to flexibly capture the group structure inherent in the regression coefficients. These formulations enable simultaneous variable selection and variable fusion within a hierarchical Bayesian framework. Furthermore, by employing the hierarchical representation of half-Cauchy distributions (Wand et al. 2011; Makalic and Schmidt 2015) and the data-augmentation technique using the Pólya–Gamma distribution (Polson et al. 2013), we derive efficient Gibbs sampling algorithms. We confirm the performance of the proposed methods through the Monte Carlo simulations and application to time-series sensor data from wafer manufacturing. The simulation study demonstrates that the proposed method achieves higher estimation accuracy and better predictive performance, while successfully capturing the group structure of the regression coefficients. In the real data analysis, the proposed methods achieved more simplified estimation by forming fewer variable groups than the existing method, which tended to produce many single-variable groups.

While such Bayesian shrinkage models can fit group-structured data and provide interpretable representations of the underlying structure, their performance crucially depends on the choice of hyperparameters in the prior distributions. In the Bayesian framework, priors can be placed on hyperparameters to enable inference without explicit tuning. However, the resulting inference inevitably depends on the choice of these priors, and thus the fully Bayesian approach is not a panacea. Moreover, the issue of how to select appropriate priors for the regression coefficients still remains. To address these challenges, it is essential to develop an objective measure for selecting hyperparameters and priors. Information criteria based on predictive performance serve as an effective tool for this purpose. They enable systematic and consistent evaluation of models not only when tuning hyperparameters within a fixed model structure, but also when comparing models with different prior specifications. Accordingly, the second aim of this thesis is

developing information criterion for the model with the Bayesian variable fusion.

Regarding information criteria in Bayesian modeling, the deviance information criterion (DIC; Spiegelhalter et al. 2002) was an early and pioneering attempt. However, the widely applicable information criterion (WAIC; Watanabe 2010b) has become the standard, as it more directly evaluates the Kullback-Leibler divergence between the true and predictive distributions and provides a theoretically justified, broadly applicable, and computationally convenient framework for model comparison. However, there is a concern about using WAIC for comparing models that employ different classes of prior distributions, because it lacks a penalty term that reflects the complexity of the prior. As a result, it inevitably selects the prior class with the highest complexity, potentially leading to overfitting in hyperparameter selection. Furthermore, WAIC is derived under the asymptotic setting in which the influence of the prior distribution becomes negligible; that is, the setting corresponds to assuming that Bayesian estimators are close to maximum likelihood estimators. However, in practice, the data size is finite, and WAIC might therefore fail to reflect the characteristics of the Bayesian estimator, which is in fact influenced by the prior distribution.

To overcome these issues, Ninomiya 2021 proposed the prior intensified information criterion (PIIC). To resolve the first problem, PIIC introduces a penalty term that accounts for the prior complexity. To resolve the second problem, it is derived under the setting where the logarithm of the prior distribution is of order $O(n)$. However, the PIIC has been developed only for the Bayesian lasso as an example of Bayesian regularization methods, and its practical usefulness in real data analysis has not yet been examined. Therefore, we develop the PIIC for the Bayesian generalized fused lasso. In particular, with an eye toward applications to spatial data, which are a representative example of data with group structures, we consider its use for model selection in the SVC (spatially varying coefficients) model. (Its theoretical framework can naturally be extended to generalized linear models, including the proposed logistic regression model with Bayesian generalized fused lasso.) SVC models have been developed as statistical frameworks for capturing spatial heterogeneity in regression analysis. Following the idea of geographically weighted regression (Brunsdon et al. 1996), Gelfand et al. (2003) formulated SVC models to model regression coefficients that vary smoothly according to the regions to which the data belong, under the assumption that they follow a Gaussian process. In this paper, however, we use the term SVC models

in a broader sense, referring not only to those based on Gaussian processes but also to probabilistic models that assume smooth spatial variation in regression coefficients. While SVC models are suitable for representing continuous spatial variation, real data often reveal that regression coefficients remain constant within certain areas but exhibit discontinuous changes at regional boundaries, such as those between administrative districts. To accommodate such patterns, extensions of the SVC framework have been proposed that group adjacent regions to share coefficients within a group and allow variation across groups (Lawson 2000; Huang and Yao 2012; Sugasawa and Murakami 2021). Furthermore, the methods based on the generalized fused lasso regularization (Hoefling 2010) have been developed, which maintain computational feasibility even when the number of groups increases (Zhao and Bondell 2020; Li and Sang 2019; Zhong et al. 2023). Following these developments, we focus on the SVC model with Bayesian generalized fused lasso regularization and develop a corresponding information criterion.

For the development of PIIC, it is necessary to derive the asymptotic properties of generalized fused lasso estimators for when the order of the logarithm of the prior distributions is $\mathrm{O}(n)$. Here, we modify the approach of Viallon et al. 2016, which derived the selection consistency of generalized fused lasso estimators for when regularization terms are of order $\mathrm{o}(\sqrt{n})$, to fit the $\mathrm{O}(n)$ case. As a result, although selection consistency does not hold, we can demonstrate fast convergence to zero for the estimators of the regression coefficients in the non-active set and the asymptotic normality for the estimators of the regression coefficients in the active set. Similar asymptotic properties for lasso estimators were obtained in Ninomiya and Kawano (2016). These properties were used for the derivation of AIC for lasso, and later, in Ninomiya 2021, for the derivation of PIIC. The approach in that prior study might also be applicable to generalized fused lasso estimators if one used parameter transformations, which involve overly cumbersome notation, but here we avoid parameter transformations by basing our work on the approach of Viallon et al. (2016). Thus, in this study, we derived a new version of PIIC without getting involved with AIC.

Through these two main developments, this dissertation aims to establish a comprehensive and practical Bayesian modeling framework for data with group structures by combining flexible variable fusion-based estimation with proper model selection method. The remainder of this paper is organized as follows.

Chapter 2 provides an overview of variable fusion methods and their Bayesian extensions, as well as the parameter sampling schemes employed in Bayesian estimation. Furthermore, several information criteria, including those developed for the lasso and their Bayesian counterparts, are also discussed. Chapter 3 focuses on developing hierarchical Bayesian logistic regression models capable of both variable selection and variable fusion. Specifically, we propose two types of Bayesian variable fusion models for binary data: one employs a Laplace prior for both variable selection and fusion, and the other combines a Laplace prior for variable selection with a horseshoe prior for fusion. Simulation studies and an application to time-series data are then presented. Chapter 4 constructs an information criterion for the Bayesian generalized fused lasso. Concretely, the asymptotic properties of our Bayesian generalized fused lasso estimator is shown and based on these properties, PIIC is adapted to SVC models with the Bayesian generalized fused lasso. In addition, the effectiveness of the proposed information criterion is demonstrated through simulation studies and real spatial data analyses. Chapter 5 concludes this thesis and briefly discusses possible directions for future work.

Chapter 2

Background

In this chapter, we describe L_1-regularization methods for variable fusion within the framework of generalized linear models, as well as their extensions using nonconvex penalties. We then introduce Bayesian approaches to variable fusion and major parameter sampling methods, with particular attention to global-local shrinkage priors that overcome the limitations of the Laplace prior. Finally, we review major information criteria for model selection, covering both frequentist and Bayesian approaches, as well as their applications to the lasso.

2.1 Sparse regularization and variable fusion

2.1.1 Generalized linear models

Let a random variable y_i $(i = 1, \ldots, n)$ follow a natural exponential family with a density

$$q(y_i \mid \phi_i) = \exp\{y_i\phi_i - R(\phi_i) + \Omega(y_i)\},$$

with respect to a σ-finite measure μ. Here, $\phi_i \in \Phi \subset \mathbb{R}$ is a natural parameter and we assume that ϕ_i satisfies $0 < \int \exp\{y\phi_i + \Omega(y)\}\mathrm{d}\mu(y) < \infty$, which implies that Φ is a natural parameter space. Under this setting, all the derivatives of $R(\phi_i)$ as well as all the moments of y_i exist in the interior Φ^{int} of Φ. In addition, the expectation and variance of y_i are expressed as

$$\mathrm{E}_{\phi_i}[y_i] = R'(\phi_i), \qquad \mathrm{Var}_{\phi_i}[y_i] = R''(\phi_i),$$

where $R'(\phi_i)$ and $R''(\phi_i)$ denote $\mathrm{d}R(\phi_i)/\mathrm{d}\phi_i$ and $\mathrm{d}^2R(\phi_i)/\mathrm{d}\phi_i^2$, respectively. We further assume that $\mathrm{Var}_{\phi_i}[y_i] > 0$. Under this assumption, the negative log-density $-\log q(y_i \mid \phi_i)$ is a convex function with respect to ϕ_i.

Let $(y_i, \boldsymbol{x}_i)$ $(i = 1, \ldots, n)$ denote the i-th observation, where y_i is a response variable and $\boldsymbol{x}_i$ is a p-dimensional vector of explanatory variables. We assume that the pairs $(y_i, \boldsymbol{x}_i)$ are independent across i. Based on the natural exponential family defined above, we consider generalized linear models (GLMs) with canonical link functions; that is, a class of density functions of y_i: $\{\, q(y_i \mid \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\theta}) : \boldsymbol{\theta} \in \boldsymbol{\Theta} \,\}$, where Θ is an open convex subset of $\mathbb{R}^p$.

2.1.2 Lasso

L_1-norm regularization is a fundamental approach that constrains the size of regression coefficients by penalizing their absolute values, which helps prevent overfitting and obtain simpler, more interpretable models. Among methods based on this idea, the lasso (Tibshirani 1996) is a representative example that simultaneously achieves parameter estimation and variable selection. The lasso is formulated as the following optimization problem:

$$\underset{\boldsymbol{\theta}\in\Theta}{\operatorname{argmin}} \left\{ -\sum_{i=1}^{n} \log q(y_i \mid \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\theta}) + \lambda \sum_{j=1}^{p} |\theta_j| \right\}, \tag{2.1}$$

where λ (> 0) is a regularization parameter that controls the strength of penalization. When λ takes a large value to some extent, some of the estimated coefficients shrink exactly to zero, and the corresponding variables are automatically excluded from the model. This property enables the lasso to select important variables while estimating regression coefficients.

2.1.3 Fused lasso

While the lasso imposes an L_1-penalty on each regression coefficient individually, it does not account for possible relationships among the coefficients. The fused lasso (Tibshirani et al. 2005) extends the lasso to incorporate structural information among adjacent variables. The fused lasso is formulated as the following optimization problem:

$$\underset{\boldsymbol{\theta}\in\Theta}{\operatorname{argmin}} \left\{ -\sum_{i=1}^{n} \log q(y_i \mid \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\theta}) + \lambda_1 \sum_{j=1}^{p} |\theta_j| + \lambda_2 \sum_{j=2}^{p} |\theta_j - \theta_{j-1}| \right\}, \tag{2.2}$$

where $\lambda_1 (> 0)$ and $\lambda_2 (> 0)$ are regularization parameters controlling the shrinkage of coefficients and the smoothness of their adjacent differences, respectively. Because of the second penalty term, the fused lasso sets not only some coefficients

but also some adjacent differences to exactly zero. Thus, it performs both variable selection and variable fusion. Specifically, it is suitable for situations where the regression coefficients corresponding to adjacent variables are expected to exert similar influences on the response, as in ordered variables, such as spatial or temporal data. For example, in comparative genomic hybridization (CGH) data, measurements are ordered along the genome, and therefore Tibshirani and Wang (2008) applied the fused lasso to estimate piecewise-constant copy number profiles and to identify genomic regions exhibiting amplifications or deletions associated with cancer.

Likewise, in spectral data measured over time, adjacent time points often exhibit similar patterns, and Yu et al. (2015) used logistic regression with the fused lasso for classification.

2.1.4 Generalized fused lasso

The fused lasso assumes a one-dimensional ordering among variables, penalizing only the differences between regression coefficients that are adjacent along this order. However, in many applications, relationships among variables are more complex than simple sequential adjacency. To address this, the generalized fused lasso (Hoefling 2010) extends the penalty to arbitrary pairs of regression coefficients. Concretely, the generalized fused lasso is defined as

$$\underset{\boldsymbol{\theta}\in\Theta}{\operatorname{argmin}} \left\{ -\sum_{i=1}^{n} \log q(y_i \mid \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\theta}) + \lambda_1 \sum_{j=1}^{p} |\theta_j| + \lambda_2 \sum_{(j,k)\in E_{gfl}} |\theta_j - \theta_k| \right\},$$

where E_{gfl} is a set of pairs (j,k) specifying which regression coefficients are adjacent. When $E_{gfl} = \{(1,2),\ldots,(p-1,p)\}$, this formulation reduces to the standard fused lasso. By appropriately defining E_{gfl}, the generalized fused lasso can represent diverse structures such as spatial grids, networks, clustered regions.

2.1.5 Nonconvex penalties

Although L_1-type regularization methods are effective in providing sparse estimations, the convexity of the penalty causes them to shrink large regression coefficients excessively toward zero, leading to biased estimates. To resolve this problem, several nonconvex penalties have been proposed, including the smoothly clipped absolute deviation (SCAD; Fan and Li 2001) and the minimax concave

penalty (MCP; Zhang 2010). These nonconvex penalties are designed so that their derivatives become zero for sufficiently large regression coefficients, which effectively removes shrinkage on large true regression coefficients and thereby prevents overshrinkage.

As a method that applies the MCP to variable fusion, the fused-MCP has been proposed (Jing et al. 2018; Inoue et al. 2020), and it is formulated as

$$\underset{\boldsymbol{\theta}\in\Theta}{\operatorname{argmin}} \left\{ -\sum_{i=1}^{n} \log q(y_i \mid \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\theta}) + \sum_{j=1}^{p} \rho_{\lambda_1,\gamma}(|\theta_j|) + \sum_{(j,k)\in E_{gfl}}^{p} \rho_{\lambda_2,\gamma}(|\theta_j - \theta_k|) \right\},$$

where $\rho_{\lambda,\gamma}(|\cdot|)$ is

$$\rho_{\lambda,\gamma}(|t|) = \begin{cases} \lambda|t| - \dfrac{t^2}{2\gamma}, & |t| \leq \gamma\lambda, \\ \dfrac{\gamma\lambda^2}{2}, & |t| > \gamma\lambda, \end{cases}$$

where $\lambda > 0$ and $\gamma > 1$, which controls the degree of concavity. In addition, when the SCAD penalty is adopted instead as the concave penalty, $\rho_{\lambda,\gamma}(\cdot)$ is given by

$$\rho_{\lambda,\gamma}(|t|) = \begin{cases} \lambda|t|, & |t| \leq \lambda, \\ \dfrac{-t^2 + 2\gamma\lambda|t| - \lambda^2}{2(\gamma-1)}, & \lambda < |t| \leq \gamma\lambda, \\ \dfrac{(\gamma+1)\lambda^2}{2}, & \gamma\lambda < |t|, \end{cases}$$

where $\lambda > 0$ and $\gamma > 2$, which controls the degree of concavity. Concave fusion penalties have been actively employed across various fields. For example, Jing et al. (2018) demonstrated that the fused-MCP enhances the quality of signal denoising and, when applied to two-dimensional images, yields sharper edges and stronger contrast.

2.2 Bayesian regularization framework for variable fusion

The former section reviewed frequentist variable fusion methods. In the Bayesian framework, regularization can be viewed as imposing priors on regression coefficients that shrink them toward zero. This section first outlines the concept of Bayesian estimation and parameter estimation methods, then introduces Bayesian penalized regression methods based on the Laplace prior, which correspond to regularization with the L_1-penalty, and finally presents global-local shrinkage priors as its extensions.

2.2.1 *Bayesian estimation*

In Bayesian estimation, model parameters are regarded as random variables that follow prior distributions reflecting prior beliefs. Given the observed data $(y_i, \boldsymbol{x}_i)$, we define $\boldsymbol{y} = (y_1, \ldots, y_n)^{\mathrm{T}}$ and $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)^{\mathrm{T}}$. The posterior distribution of the parameters $\boldsymbol{\theta}$ is given by Bayes' theorem as

$$\begin{aligned} \pi(\boldsymbol{\theta} \mid \boldsymbol{y}, \boldsymbol{X}) &= \frac{\left\{\prod_{i=1}^{n} q(y_i \mid \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\theta})\right\} \pi(\boldsymbol{\theta})}{\displaystyle\int_{\Theta} \left\{\prod_{i=1}^{n} q(y_i \mid \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\theta})\right\} \pi(\boldsymbol{\theta}) \, \mathrm{d}\boldsymbol{\theta}} \\ &\propto \left\{\prod_{i=1}^{n} q(y_i \mid \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\theta})\right\} \pi(\boldsymbol{\theta}), \end{aligned} \tag{2.3}$$

where $\pi(\boldsymbol{\theta})$ is the prior distribution. Bayesian estimation aims to infer $\boldsymbol{\theta}$ based on this posterior distribution, typically through posterior means, credible intervals, or other summary statistics. A simple and widely used point estimator is the maximum a posteriori (MAP) estimator, which is defined as a solution of the following optimization problem:

$$\operatorname*{argmax}_{\boldsymbol{\theta} \in \Theta} \pi(\boldsymbol{\theta} \mid \boldsymbol{y}, \boldsymbol{X}),$$

representing the parameter value that maximizes the posterior density. Furthermore, the Bayesian framework allows predictive inference for new observations using the posterior distribution. When the $\boldsymbol{x}$ are treated as fixed, the predictive distribution is given by

$$p(\tilde{y} \mid \boldsymbol{y}, \boldsymbol{X}) = \int_{\Theta} q(\tilde{y} \mid \tilde{\boldsymbol{x}}^{\mathrm{T}}\boldsymbol{\theta}) \; \pi(\boldsymbol{\theta} \mid \boldsymbol{y}, \boldsymbol{X}) \, \mathrm{d}\boldsymbol{\theta},$$

where $(\tilde{y}, \tilde{\boldsymbol{x}})$ is a new observation. This predictive distribution enables probabilistic prediction while naturally reflecting parameter uncertainty through the posterior distribution.

2.2.2 *Posterior sampling methods*

In general, the posterior distribution in (2.3) cannot be expressed in a closed form, especially when complex priors or hierarchical structures are involved. Therefore, direct sampling from $\pi(\boldsymbol{\theta} \mid \boldsymbol{y}, \boldsymbol{X})$ is difficult and simulation-based approaches are often employed to approximate the posterior distribution. Accordingly, Markov chain Monte Carlo (MCMC) methods provide a powerful framework for approximating such intractable posteriors by constructing a Markov chain whose stationary distribution equals the target posterior.

One of the most fundamental MCMC methods is the Metropolis-Hastings (MH) algorithm (Metropolis et al. 1953; Hastings 1970). Given a current state $\boldsymbol{\theta}^{[t]}$, a candidate $\boldsymbol{\zeta}$ is generated from a proposal distribution $w(\boldsymbol{\zeta} \mid \boldsymbol{\theta}^{[t]})$ and accepted with probability

$$\alpha = \min\left\{1, \frac{\pi(\boldsymbol{\zeta} \mid \boldsymbol{y}, \boldsymbol{X})\, w(\boldsymbol{\theta}^{[t]} \mid \boldsymbol{\zeta})}{\pi(\boldsymbol{\theta}^{[t]} \mid \boldsymbol{y}, \boldsymbol{X})\, w(\boldsymbol{\zeta} \mid \boldsymbol{\theta}^{[t]})}\right\}.$$

If the candidate is rejected, the current value is retained. By repeating this procedure, the sequence $\{\boldsymbol{\theta}^{[t]}\}$ converges to the target distribution $\pi(\boldsymbol{\theta} \mid \boldsymbol{y}, \boldsymbol{X})$ under mild regularity conditions.

In addition, updating all parameters simultaneously in high-dimensional problems can be inefficient because a jointly proposed candidate is likely to be rejected frequently. To improve efficiency, the parameter vector $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_{k_0})$ can be divided into several blocks, and each block can be updated conditionally on the others. This scheme is referred to as the block Metropolis-Hastings (block MH) algorithm. In the block MH algorithm, the j-th block $\boldsymbol{\theta}_j$ is updated while keeping $\boldsymbol{\theta}_{-j}$ fixed. The target distribution for this update is the full conditional distribution

$$\pi(\boldsymbol{\theta}_j \mid \boldsymbol{\theta}_{-j}, \boldsymbol{y}, \boldsymbol{X}) \propto \left\{\prod_{i=1}^{n} q(y_i \mid \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\theta})\right\} \pi(\boldsymbol{\theta}_j \mid \boldsymbol{\theta}_{-j}),$$

where $\boldsymbol{\theta}_{-j} = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_{j-1}, \boldsymbol{\theta}_{j+1}, \ldots, \boldsymbol{\theta}_{k_0})$ denotes all components of $\boldsymbol{\theta}$ except $\boldsymbol{\theta}_j$. A candidate $\boldsymbol{\zeta}$ is generated from a proposal distribution $w_j(\boldsymbol{\zeta} \mid \boldsymbol{\theta}_j^{[t]}, \boldsymbol{\theta}_{-j}^{[t]})$ and accepted with probability

$$\alpha_j = \min\left\{1, \frac{\pi(\boldsymbol{\zeta} \mid \boldsymbol{\theta}_{-j}^{[t]}, \boldsymbol{y}, \boldsymbol{X})\, w_j(\boldsymbol{\theta}_j^{[t]} \mid \boldsymbol{\zeta}, \boldsymbol{\theta}_{-j}^{[t]})}{\pi(\boldsymbol{\theta}_j^{[t]} \mid \boldsymbol{\theta}_{-j}^{[t]}, \boldsymbol{y}, \boldsymbol{X})\, w_j(\boldsymbol{\zeta} \mid \boldsymbol{\theta}_j^{[t]}, \boldsymbol{\theta}_{-j}^{[t]})}\right\}.$$

Although the MH method is broadly applicable, its performance strongly depends on the choice of the proposal distribution. For models with complex likelihood functions such as logistic regression, it is often difficult to find an appropriate proposal distribution, resulting in inefficient sampling or poor convergence.

To overcome this difficulty, a more efficient approach called Gibbs sampling (Geman and Geman 1984) can be employed. Gibbs sampling can be viewed as a special case of the block Metropolis-Hastings (block MH) algorithm, where the full conditional distributions are available in closed form. In Gibbs sampling, each block $\boldsymbol{\theta}_j$ is sequentially sampled from its full conditional distribution while

keeping the other components fixed:

$$\boldsymbol{\theta}_j^{[t+1]} \sim \pi(\boldsymbol{\theta}_j \mid \boldsymbol{\theta}_{-j}^{[t]}, \boldsymbol{y}, \boldsymbol{X}), \quad j = 1, \ldots, k_0.$$

In this case, the proposal distribution is set equal to the full conditional distribution, that is,

$$w_j(\boldsymbol{\zeta} \mid \boldsymbol{\theta}_j^{[t]}, \boldsymbol{\theta}_{-j}^{[t]}) = \pi(\boldsymbol{\zeta} \mid \boldsymbol{\theta}_{-j}^{[t]}, \boldsymbol{y}, \boldsymbol{X}),$$

so that the numerator and denominator in the acceptance probability α_j cancel out, yielding $\alpha_j = 1$. This property ensures that every generated sample is accepted, leading to high sampling efficiency when the full conditional distributions are tractable.

2.2.3 Bayesian lasso

The Bayesian lasso (Park and Casella 2008) provides a Bayesian formulation of the lasso by assuming Laplace priors on regression coefficients. Specifically, it assumes that each coefficient θ_j follows an independent Laplace prior,

$$\pi(\boldsymbol{\theta} \mid \lambda) = \prod_{j=1}^{p} \frac{\lambda}{2} \exp(-\lambda|\theta_j|), \tag{2.4}$$

where $\lambda > 0$ controls the overall shrinkage. Under this prior, the posterior mode of $\boldsymbol{\theta}$ coincides with the solution of the L_1-penalized optimization problem in the frequentist framework. Given λ, the posterior distribution of $\boldsymbol{\theta}$ is expressed as

$$\pi(\boldsymbol{\theta} \mid \boldsymbol{y}, \boldsymbol{X}) \propto \prod_{i=1}^{n} q(y_i \mid \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\theta}) \prod_{j=1}^{p} \frac{\lambda}{2} \exp\left(-\lambda|\theta_j|\right).$$

Then, the negative log posterior distribution is written as

$$-\log \pi(\boldsymbol{\theta} \mid \boldsymbol{y}, \boldsymbol{X}) \propto -\sum_{i=1}^{n} \log q(y_i \mid \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\theta}) + \lambda \sum_{j=1}^{p} |\theta_j|. \tag{2.5}$$

The MAP estimator of $\boldsymbol{\theta}$ under the Laplace prior, obtained by minimizing (2.5) with respect to $\boldsymbol{\theta} \in \Theta$, corresponds exactly to the lasso estimator defined in (2.1).

However, direct sampling from this posterior distribution is not straightforward because the Laplace prior is not conjugate to likelihoods in general settings. To address this issue, Park and Casella (2008) adopted a hierarchical formulation of the Laplace distribution by expressing it with a scale mixture of normals (Andrews and Mallows 1974), which is given by

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

Using this representation, the prior in (2.4) can be written hierarchically as

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

Equivalently, it can be expressed in a compact form as

$$\begin{aligned} \boldsymbol{\theta} \mid \tau_1^2,\ldots,\tau_p^2 &\sim \mathrm{N}_p(\mathbf{0}_p, \boldsymbol{\Sigma}_l), \\ \tau_j^2 &\sim \mathrm{EXP}\left(\frac{\lambda_1^2}{2}\right), \end{aligned} \tag{2.6}$$

where $\boldsymbol{\Sigma}_l = \mathrm{diag}\left(\tau_1^2, \tau_2^2, \ldots, \tau_p^2\right)$ and $\mathrm{EXP}(x \mid d)$ is an exponential distribution with a rate parameter d (> 0). From the perspective of posterior computation, this representation provides a convenient structure. Conditional on $\tau_1^2,\ldots,\tau_p^2$, the prior on $\boldsymbol{\theta}$ becomes Gaussian, which makes the posterior distribution more tractable and facilitates Gibbs sampling. In addition, Park and Casella (2008) assumed a Gamma distribution $\mathrm{Ga}(r_1, t_1)$, where r_1 (> 0) is a shape parameter and t_1 (> 0) is a rate parameter, on λ^2, enabling fully Bayesian approach.

2.2.4 Bayesian fused lasso

Kyung et al. (2010) extended the fused lasso into a Bayesian framework, proposing the Bayesian fused lasso. Specifically, Laplace priors are assumed not only on the regression coefficients but also on their adjacent differences, that is, the following prior distribution on $\boldsymbol{\theta}$ is considered:

$$\pi(\boldsymbol{\theta} \mid \lambda_1, \lambda_2) \propto \exp\left(-\lambda_1 \sum_{j=1}^{p} |\theta_j| - \lambda_2 \sum_{j=2}^{p} |\theta_j - \theta_{j-1}|\right), \tag{2.7}$$

where $\lambda_1, \lambda_2 > 0$ are regularization parameters controlling the degree of sparseness and fusion, respectively. The posterior mode of regression coefficients in the Bayesian fused lasso is equivalent to the fused lasso solution, defined in (2.2).

Following Andrews and Mallows (1974), the Laplace distributions in (2.7) can be expressed hierarchically using a scale mixture of normals as

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

Consequently, the priors on the parameters of the model with the Bayesian fused lasso can be written hierarchically as

$$\begin{aligned} \boldsymbol{\theta} \mid \tau_1^2, \ldots, \tau_p^2, \tilde{\tau}_2^2, \ldots, \tilde{\tau}_p^2 &\sim \mathrm{N}_p(\mathbf{0}_p, \boldsymbol{\Sigma}_{fl}), \\ \tau_j^2 &\sim \mathrm{EXP}\left(\frac{\lambda_1^2}{2}\right), \\ \tilde{\tau}_j^2 &\sim \mathrm{EXP}\left(\frac{\lambda_2^2}{2}\right), \end{aligned} \tag{2.8}$$

where the inverse matrix of $\boldsymbol{\Sigma}_{fl}$ is expressed as

$$\boldsymbol{\Sigma}_{fl}^{-1} = \begin{pmatrix} \frac{1}{\tau_1^2} + \frac{1}{\tilde{\tau}_2^2} & -\frac{1}{\tilde{\tau}_2^2} & 0 & \ldots & 0 & 0 \\ -\frac{1}{\tilde{\tau}_2^2} & \frac{1}{\tau_2^2} + \frac{1}{\tilde{\tau}_2^2} + \frac{1}{\tilde{\tau}_3^2} & -\frac{1}{\tilde{\tau}_3^2} & \ldots & 0 & 0 \\ 0 & -\frac{1}{\tilde{\tau}_3^2} & \frac{1}{\tau_3^2} + \frac{1}{\tilde{\tau}_3^2} + \frac{1}{\tilde{\tau}_4^2} & \ldots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \ldots & \frac{1}{\tau_{p-1}^2} + \frac{1}{\tilde{\tau}_{p-1}^2} + \frac{1}{\tilde{\tau}_p^2} & -\frac{1}{\tilde{\tau}_p^2} \\ 0 & 0 & 0 & \ldots & -\frac{1}{\tilde{\tau}_p^2} & \frac{1}{\tau_p^2} + \frac{1}{\tilde{\tau}_p^2} \end{pmatrix}.$$

As in the Bayesian lasso, Gamma distributions $\mathrm{Ga}(r_1, t_1)$ and $\mathrm{Ga}(r_2, t_2)$ can be assumed on λ_1^2 and λ_2^2, respectively, where r_1, r_2 (> 0) are shape parameters and t_1, t_2 (> 0) are rate parameters.

2.2.5 Bayesian generalized fused lasso

The generalized fused lasso can also be extended to the Bayesian method, which we call as the generalized fused lasso. In the Bayesian generalized fused lasso, the Laplace priors are imposed not only on the individual regression coefficients but also on the pairwise differences of coefficients specified by the adjacency set E_{gfl}, that is, the following prior distribution on $\boldsymbol{\theta}$ is assumed:

$$\pi(\boldsymbol{\theta} \mid \lambda_1, \lambda_2) \propto \exp\left(-\lambda_1 \sum_{j=1}^{p} |\theta_j| - \lambda_2 \sum_{(j,k) \in E_{gfl}} |\theta_j - \theta_k|\right), \tag{2.9}$$

where $\lambda_1, \lambda_2 > 0$ are regularization parameters controlling the degree of sparseness and the degree of fusion, respectively. For notational convenience, we introduce an ordered version of the adjacency set, $E_{gfl} = \{\boldsymbol{e}_1, \ldots, \boldsymbol{e}_\chi\}$ with $\boldsymbol{e}_i = (e_{i1}, e_{i2})$ corresponds to the i-th pair of adjacent coefficients, which will be used in the following.

Based on Andrews and Mallows (1974), the Laplace distributions in (2.9) can be expressed hierarchically using a scale mixture of normals as

$$\pi(\boldsymbol{\theta} \mid \lambda_1, \lambda_2) \propto \prod_{j=1}^{p} \int \frac{1}{\sqrt{2\pi\tau_j^2}} \exp\left(-\frac{\theta_j^2}{2\tau_j^2}\right) \frac{\lambda_1^2}{2} \exp\left(-\frac{\lambda_1^2\tau_j^2}{2}\right) \mathrm{d}\tau_j^2 \\ \times \prod_{(j,k)\in E_{gfl}} \int \frac{1}{\sqrt{2\pi\tilde{\tau}_{j,k}^2}} \exp\left\{-\frac{(\theta_j - \theta_k)^2}{2\tilde{\tau}_{j,k}^2}\right\} \frac{\lambda_2^2}{2} \exp\left(-\frac{\lambda_2^2\tilde{\tau}_{j,k}^2}{2}\right) \mathrm{d}\tilde{\tau}_{j,k}^2. \tag{2.10}$$

Accordingly, the priors on parameters of the model with the Bayesian generalized fused lasso can be represented hierarchically as

$$\begin{aligned} \boldsymbol{\theta} \mid \tau_1^2, \ldots, \tau_p^2, \tilde{\tau}_{e_{11},e_{12}}^2, \ldots, \tilde{\tau}_{e_{\chi 1},e_{\chi 2}}^2 &\sim \mathrm{N}_p(\mathbf{0}_p, \boldsymbol{\Sigma}_{gfl}), \\ \tau_j^2 &\sim \mathrm{EXP}\left(\frac{\lambda_1^2}{2}\right), \\ \tilde{\tau}_{j,k}^2 &\sim \mathrm{EXP}\left(\frac{\lambda_2^2}{2}\right), \end{aligned} \tag{2.11}$$

where $\boldsymbol{\Sigma}_{gfl}^{-1}$ is defined as

$$\boldsymbol{\Sigma}_{gfl}^{-1} = \boldsymbol{T} + \boldsymbol{\Upsilon}^{\mathrm{T}} \tilde{\boldsymbol{T}} \boldsymbol{\Upsilon},$$

with $\boldsymbol{T} = \mathrm{diag}\left(1/\tau_1^2, \ldots, 1/\tau_p^2\right)$. Here, $\boldsymbol{\Upsilon}$ is a $\chi \times p$ matrix representing the adjacency in E_{gfl}. Each row of $\boldsymbol{\Upsilon}$ corresponds to an element $\boldsymbol{e}_i \in E_{gfl}$, where $\Upsilon_{i,e_{i1}} = 1$, $\Upsilon_{i,e_{i2}} = -1$, and all other elements are zero. In addition, $\tilde{\boldsymbol{T}}$ is defined as

$$\tilde{T}_{i,j} = \begin{cases} 1/\tilde{\tau}_{e_{i1},e_{i2}}^2, & \text{if } i = j, \\ 0, & \text{otherwise.} \end{cases}$$

As in the previous models, Gamma distributions can be assumed for λ_1^2 and λ_2^2.

2.2.6 Global-local shrinkage prior

In the previous sections, we have discussed Bayesian L_1-regularization methods. The Laplace prior, which places a high density around zero, shrinks regression coefficients toward zero and thereby induces sparsity in estimation. However, since the Laplace prior corresponds to the L_1 penalty in the frequentist framework, it inherits the same limitation: the convexity of the penalty causes over-shrinkage when the true coefficients are non-zero and under-shrinkage when the true coefficients are zero.

To mitigate this limitation and provide a more flexible and general framework for Bayesian regularization, a family of priors known as the global-local

(GL) shrinkage priors (Polson and Scott 2010) has been proposed. GL priors introduce a hierarchical structure consisting of a global scale parameter that controls the overall shrinkage intensity and local scale parameters that allow coefficient-specific adaptivity. The Laplace prior can be regarded as a special, non-heavy-tailed case within this framework, whereas heavier-tailed GL priors such as the horseshoe prior can strongly shrink small regression coefficients while leaving large ones relatively unshrunk.

To incorporate this flexibility into the Bayesian generalized fused lasso, we introduce GL priors for both variable selection and variable fusion components. Specifically, the priors on the parameters of the model with the Bayesian generalized fused lasso can be represented hierarchically as

$$\begin{aligned}
\boldsymbol{\theta} \mid \gamma_1^2, \ldots, \gamma_p^2, \tilde{\gamma}^2_{e_{i1},e_{i2}}, \ldots, \tilde{\gamma}^2_{e_{\chi 1},e_{\chi 2}}, \delta_1^2, \delta_2^2 &\sim \mathrm{N}_p(\mathbf{0}_p,\, \boldsymbol{\Sigma}_{gfl}), \\
\gamma_j^2 &\sim p_{\mathrm{local}}(\cdot), \\
\tilde{\gamma}^2_{j,k} &\sim \tilde{p}_{\mathrm{local}}(\cdot), \\
\delta_1 &\sim p_{\mathrm{global}}(\cdot), \\
\delta_2 &\sim \tilde{p}_{\mathrm{global}}(\cdot),
\end{aligned}$$

where the precision matrix $\boldsymbol{\Sigma}^{-1}_{gfl}$ is given by

$$\boldsymbol{\Sigma}^{-1}_{gfl} = \delta_1^{-2}\boldsymbol{T} + \delta_2^{-2}\boldsymbol{\Upsilon}^{\mathrm{T}}\tilde{\boldsymbol{T}}\boldsymbol{\Upsilon},$$

and $\boldsymbol{T}$ and $\tilde{\boldsymbol{T}}$ are defined as

$$\begin{aligned}
\boldsymbol{T} &= \mathrm{diag}(1/\gamma_1^2, \ldots, 1/\gamma_p^2), \\
\tilde{\boldsymbol{T}}_{ij} &= \begin{cases} 1/\tilde{\gamma}^2_{e_{i1},e_{i2}}, & \text{if } i = j, \\ 0, & \text{otherwise.} \end{cases}
\end{aligned}$$

The specific choices of p_{local}, $\tilde{p}_{\mathrm{local}}$, p_{global}, $\tilde{p}_{\mathrm{global}}$ determine the type of GL prior, such as the Laplace, normal-gamma, normal-exponential-gamma, horseshoe, or Dirichlet-Laplace priors. For example, the Laplace prior used in the Bayesian generalized fused lasso corresponds to the case where p_{local} and $\tilde{p}_{\mathrm{local}}$ are exponential distributions and the global scales are $\delta_1 = 1$ and $\delta_2 = 1$. If the priors of global and local shrinkage parameters admit hierarchical representations in terms of conditionally conjugate scale mixtures such as a gaussian, gamma, or inverse-gamma, the resulting model yields closed-form full conditional distributions for all parameters, enabling efficient posterior sampling via Gibbs updates.

A variety of Bayesian regularization models with a fusion penalty employing GL shrinkage priors have been proposed (e.g., Shimamura et al. 2019; Song and Cheng 2020; Banerjee 2022; Bhattacharyya et al. 2022; Kakikawa and Kawano 2023).

2.3 Information criteria

In the previous section, we introduced various Bayesian regularization models. However, it remains challenging to objectively compare these models across different prior formulations, and to select appropriate priors or values of hyperparameters. Information criteria provide an effective and principled approach to addressing these issues. In this section, we review a series of information criteria for model selection, beginning with the classical Akaike information criterion (AIC; Akaike 1973) and Bayesian information criterion (BIC; Schwarz 1978) and their lasso-based extensions, and then proceeding to Bayesian counterparts including the deviance information criterion (DIC; Spiegelhalter et al. 2002), the widely applicable information criterion (WAIC; Watanabe 2010a), and the prior intensified information criterion (PIIC; Ninomiya 2021).

2.3.1 Review of information criteria

Information criteria can be broadly divided into two theoretical frameworks. The first approach is based on minimizing the Kullback-Leibler (KL) divergence (Kullback and Leibler 1951) between the true distribution and the estimated one:

$$\tilde{\mathrm{E}}\Big\{\sum_{i=1}^{n} g_{\tilde{y}_i,\boldsymbol{x}_i}(\boldsymbol{\theta}^*)\Big\} - \tilde{\mathrm{E}}\Big\{\sum_{i=1}^{n} g_{\tilde{y}_i,\boldsymbol{x}_i}(\hat{\boldsymbol{\theta}})\Big\}, \tag{2.12}$$

where $g_{y_i,\boldsymbol{x}_i}(\boldsymbol{\theta}) = \log q(y_i \mid \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\theta})$, $\boldsymbol{\theta}^*$ denotes the true value of $\boldsymbol{\theta}$, and $\hat{\boldsymbol{\theta}}$ is its estimator. Here, $(\tilde{y}_1, \ldots, \tilde{y}_n)$ is a copy of $(y_1, \ldots, y_n)$ drawn from the same distribution, and $\tilde{\mathrm{E}}$ denotes the expectation with respect to $(\tilde{y}_1, \ldots, \tilde{y}_n)$ only. A model is selected when this quantity is small. From this perspective, AIC is the most well-known information criterion developed in the frequentist context. In the Bayesian setting, related criteria include DIC and WAIC, which provides a theoretical extension of AIC based on the KL divergence. In contrast, another branch of information criteria for Bayesian models is derived from the marginal

likelihood:

$$\int_{\Theta} \left\{ \prod_{i=1}^{n} q(y_i \mid \boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{\theta}) \right\} \pi(\boldsymbol{\theta}) \, \mathrm{d}\boldsymbol{\theta}$$

which is the term appearing in the denominator of (2.3) and is also known as the Bayesian evidence. It quantifies how likely the observed data are under the model by integrating over all possible parameter values. A larger value indicates that the model assigns higher likelihood to the observed data and hence receives a higher posterior probability. BIC arises as a large-sample approximation to this quantity, but its focus differs from AIC: BIC aims to identify the true model within the candidate set, while AIC seeks to minimize predictive loss even when the true model is not included among the candidates. In this study, we focus on information criteria derived in the KL-divergence framework, which evaluate model performance from a predictive viewpoint. Before proceeding to the proposed framework, however, we briefly review representative existing information criteria to establish the theoretical context for our approach.

2.3.2 Akaike information criterion

We begin by outlining the Akaike information criterion (AIC; Akaike 1973), which aims to minimize twice the KL divergence in (2.12). Since the first term of (2.12) is constant across candidate models, minimizing twice the KL divergence is equivalent to minimizing only twice the second term. Twice the second term can be simply estimated by $-2\sum_{i=1}^{n} g_{y_i,\boldsymbol{x}_i}(\hat{\boldsymbol{\theta}})$, but this tends to underestimate the target quantity because the same data are used for both estimation and evaluation. Therefore, AIC-type information criteria (Konishi and Kitagawa 2008) introduce a bias-corrected negative log-likelihood:

$$-2\sum_{i=1}^{n} g_{y_i,\boldsymbol{x}_i}(\hat{\boldsymbol{\theta}}) + 2\,\mathrm{E}\left[\sum_{i=1}^{n} g_{y_i,\boldsymbol{x}_i}(\hat{\boldsymbol{\theta}}) - \tilde{\mathrm{E}}\left\{\sum_{i=1}^{n} g_{\tilde{y}_i,\boldsymbol{x}_i}(\hat{\boldsymbol{\theta}})\right\}\right]. \tag{2.13}$$

By asymptotically evaluating the second term in (2.13), Akaike (1973) showed that it can be approximated by twice the number of parameters, $2p$ under the maximum likelihood estimation, and proposed the AIC:

$$\mathrm{AIC} = -2\sum_{i=1}^{n} g_{y_i,\boldsymbol{x}_i}(\hat{\boldsymbol{\theta}}) + 2p.$$

2.3.3 *AIC for lasso*

The derivation of AIC relies on the assumption that the estimator is a smooth function of the data. However, this assumption does not hold in regularized estimation methods, where the estimator is typically a non-smooth function of the data due to the non-differentiability of the penalty. To overcome this problem, extensions of AIC to the lasso have been developed. In applying AIC to the lasso, the bias-corrected negative log-likelihood in (2.13) using the lasso estimator $\hat{\boldsymbol{\theta}}^{\text{lasso}}$ is adopted. Note that $\hat{\boldsymbol{\theta}}^{\text{lasso}}$ is not the solution of (2.1), but the one of $\operatorname{argmin}_{\boldsymbol{\theta}\in\Theta}\left\{-\sum_{i=1}^{n}\log q(y_i|\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\theta})+n\lambda\sum_{j=1}^{p}|\theta_j|\right\}$. Let us consider the case of a Gaussian linear model with a known error variance, where $g_{y_i,\boldsymbol{x}_i}(\boldsymbol{\theta})$ is given by

$$-\frac{1}{2}(y_i-\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\theta})^2-\frac{1}{2}\log(2\pi), \tag{2.14}$$

and let $\hat{\mathcal{J}}^{(2)}=\{j:\hat{\theta}_j^{\text{lasso}}\neq 0\}$ denote the active set of non-zero coefficients in the lasso estimator $\hat{\boldsymbol{\theta}}^{\text{lasso}}$. Then, the number of non-zero coefficients $|\hat{\mathcal{J}}^{(2)}|$ provides an unbiased estimator of the second term in (2.13) (Efron et al. 2004; Zou et al. 2007) according to Stein's unbiased estimation theory (Stein 1981). Consequently, the following criterion serves as an AIC for the lasso:

$$\text{AIC}^{\text{lasso}}=-2\sum_{i=1}^{n}g_{y_i,\boldsymbol{x}_i}(\hat{\boldsymbol{\theta}}^{\text{lasso}})+2|\hat{\mathcal{J}}^{(2)}|. \tag{2.15}$$

This expression also can be regarded as the finite-sample corrected AIC (AICc; Sugiura 1978; Hurvich and Tsai 1989) for the lasso estimator in the Gaussian case.

However, AIC for the lasso in (2.15) cannot be utilized for more general likelihoods. Therefore, following the same approach as in the original derivation of the AIC, Ninomiya and Kawano (2016) derived an AIC for the lasso by asymptotically evaluating the second term in (2.13). Specifically, Ninomiya and Kawano (2016) considered that

$$\mathrm{E}\left[\sum_{i=1}^{n}g_{y_i,\boldsymbol{x}_i}(\hat{\boldsymbol{\theta}}^{\text{lasso}})-\tilde{\mathrm{E}}\left\{\sum_{i=1}^{n}g_{\tilde{y}_i,\boldsymbol{x}_i}(\hat{\boldsymbol{\theta}}^{\text{lasso}})\right\}\right]$$

can be rewritten as the expectation of

$$\sum_{i=1}^{n}\left\{g_{y_i,\boldsymbol{x}_i}(\hat{\boldsymbol{\theta}}^{\text{lasso}})-g_{y_i,\boldsymbol{x}_i}(\boldsymbol{\theta}^{**})\right\}-\sum_{i=1}^{n}\left\{g_{\tilde{y}_i,\boldsymbol{x}_i}(\hat{\boldsymbol{\theta}}^{\text{lasso}})-g_{\tilde{y}_i,\boldsymbol{x}_i}(\boldsymbol{\theta}^{**})\right\}, \tag{2.16}$$

where $\hat{\boldsymbol{\theta}}^{\text{lasso}}$ converges in probability to $\boldsymbol{\theta}^{**}$, and utilized the following expression as an information criterion:

$$-2\sum_{i=1}^{n}g_{y_i,\boldsymbol{x}_i}(\hat{\boldsymbol{\theta}}^{\text{lasso}})+2\,\mathrm{E}(z^{\text{limit}})$$

where z^{limit} is the limit to which (2.16) converges in distribution. More concretely, an asymptotic bias $\mathrm{E}(z^{\text{limit}})$ is given by

$$\operatorname{tr}\left\{\boldsymbol{J}^{(22)}(\boldsymbol{\theta}^*)\boldsymbol{J}^{(22)-1}(\boldsymbol{\theta}^{**})\right\}, \tag{2.17}$$

under standard regularity conditions, where a positive definite matrix $\boldsymbol{J}(\boldsymbol{\theta})$ is the asymptotic limit of $\boldsymbol{J}_n(\boldsymbol{\theta}) \equiv \sum_{i=1}^n \mathrm{E}\{\boldsymbol{x}_i^{\mathrm{T}} R''(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\theta})\boldsymbol{x}_i\}/n$ and $\boldsymbol{J}^{(22)}$ denotes $(\boldsymbol{J}_{ij})_{i\in\mathcal{J}^{(2)}, j\in\mathcal{J}^{(2)}}$ with $\mathcal{J}^{(2)} = \{j : \beta_j^{**} \neq 0\}$. In practice, the consistent estimator of (2.17):

$$\operatorname{tr}\left\{\hat{\boldsymbol{J}}_n^{*(22)}\hat{\boldsymbol{J}}_n^{**(22)-1}\right\}$$

is used for the place of an asymptotic bias, where $\hat{\boldsymbol{J}}_n^{*(22)} \equiv (\boldsymbol{J}_n(\hat{\boldsymbol{\theta}}_0)_{jk})_{j\in\hat{\mathcal{J}}^{(2)}, k\in\hat{\mathcal{J}}^{(2)}}$, $\hat{\boldsymbol{J}}_n^{**(22)} \equiv (\boldsymbol{J}_n(\hat{\boldsymbol{\theta}}^{\text{lasso}})_{jk})_{j\in\hat{\mathcal{J}}^{(2)}, k\in\hat{\mathcal{J}}^{(2)}}$, and $\hat{\boldsymbol{\theta}}_0$ is a consistent estimator of $\boldsymbol{\theta}^*$. Thus, a general form of the AIC for the lasso can be written as

$$\mathrm{AIC}^{\text{lasso}} = -2\sum_{i=1}^n g_{y_i,\boldsymbol{x}_i}(\hat{\boldsymbol{\theta}}^{\text{lasso}}) + 2\operatorname{tr}\left(\hat{\boldsymbol{J}}_n^{*(22)}\hat{\boldsymbol{J}}_n^{**(22)-1}\right). \tag{2.18}$$

When $g_{y_i,\boldsymbol{x}_i}(\boldsymbol{\theta})$ is given by (2.14), $\hat{\boldsymbol{J}}_n^{*(22)} = \hat{\boldsymbol{J}}_n^{**(22)}$ holds, and therefore (2.18) reduces to (2.15). Thus, (2.18) can be regarded as a generalization of the AICc for the Gaussian linear regression when the variance in known.

2.3.4 Bayesian information criterion

Following the development of AIC, which was motivated by the minimization of the Kullback-Leibler divergence, the Bayesian information criterion (BIC; Schwarz 1978) was proposed from a different theoretical standpoint. Whereas AIC is derived from a predictive perspective, BIC is based on the Bayesian principle of posterior model probability. Specifically, under a set of candidate models $\{M_1, \ldots, M_K\}$, the posterior probability of model M_j is expressed as

$$p(M_j \mid \boldsymbol{y}, \boldsymbol{X}) = \frac{p_j(\boldsymbol{y} \mid \boldsymbol{X})p(M_j)}{\sum_{k=1}^K p_k(\boldsymbol{y} \mid \boldsymbol{X})p(M_k)},$$

where $p(M_j)$ is the prior probability that model M_j is true, and

$$p_j(\boldsymbol{y} \mid \boldsymbol{X}) = \int_{\Theta_{(j)}} \left\{\prod_{i=1}^n q(y_i \mid \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\theta}_{(j)})\right\} \pi_j(\boldsymbol{\theta}_{(j)})\, \mathrm{d}\boldsymbol{\theta}_{(j)} \tag{2.19}$$

is the marginal likelihood, where $\boldsymbol{\theta}_{(j)} \in \Theta_{(j)}$ is the parameter of model M_j and $\pi_j(\boldsymbol{\theta}_{(j)})$ is a prior of parameters of model M_j. (2.19) is also referred to as the model evidence. When the prior probabilities of models $p(M_j)$ are equal for all

$j = 1, \ldots, K$, model comparison can be conducted solely based on the marginal likelihoods. To approximate the logarithm of (2.19), Schwarz (1978) applied a Laplace expansion around the posterior mode, which is approximately equal to the maximum likelihood estimator $\hat{\boldsymbol{\theta}}_{(k)}$ under a flat prior, yielding

$$\log p_j(\boldsymbol{y} \mid \boldsymbol{X}) \approx \sum_{i=1}^{n} \log q(y_i \mid \boldsymbol{x}_i^{\mathrm{T}} \hat{\boldsymbol{\theta}}_{(j)}) - \frac{e_j}{2} \log n,$$

where e_j is the number of parameters in model M_j. Multiplying by -2 leads to the familiar form of BIC:

$$\mathrm{BIC} = -2 \sum_{i=1}^{n} \log q(y_i \mid \boldsymbol{x}_i^{\mathrm{T}} \hat{\boldsymbol{\theta}}_{(j)}) + e_j \log n.$$

Under standard regularity conditions, BIC is consistent in the sense that the probability of selecting the true model converges to one as the sample size increases.

2.3.5 BIC for lasso

The traditional BIC described in the previous subsection is theoretically justified only when the number of parameters is fixed. To ensure consistent model selection when the number of parameters diverges, Wang et al. (2009) proposed a modified BIC and further proved that it also achieves model selection consistency for penalized estimators. The modified BIC is expressed as

$$\mathrm{BIC}_m = -2 \sum_{i=1}^{n} \log q(y_i \mid \boldsymbol{x}_i^{\mathrm{T}} \hat{\boldsymbol{\theta}}_{(j)}) + d_j C_n \log n,$$

where d_j is the number of non-zero components of $\hat{\boldsymbol{\theta}}_{(j)}$ and $C_n > 0$ is a constant. Unlike the classical BIC, however, the modified BIC is not derived from a Laplace approximation to the marginal likelihood; rather, it is designed to achieve selection consistency. Although consistency holds for any sequence C_n satisfying certain regularity conditions, there is no clear guideline for choosing an optimal value of C_n, which remains a practical limitation of this approach.

2.3.6 Deviance information criterion

Bayesian model selection had long relied on criteria based on the Bayes factor framework, such as the BIC and its modified versions introduced in the previous

subsections. These approaches compare models through their marginal likelihoods to identify the true model, assuming that it is included within the candidate set. However, this assumption is often unrealistic, since statistical models usually provide only an approximation to the underlying data-generating process. In practice, modern Bayesian modeling, such as hierarchical regression, mixture models, and regularization methods, tends to emphasize predictive performance rather than strict model identification.

In light of these limitations, the deviance information criterion (DIC; Spiegelhalter et al. 2002) brought a new perspective to Bayesian model assessment. Moving beyond the evidence-based perspective of BIC-type criteria, DIC represented the first general attempt to incorporate the KL divergence into Bayesian model comparison. It provides a practical approximation to the KL divergence between the true distribution and the estimated one, thereby introducing an information-theoretic viewpoint into Bayesian inference. In particular, DIC approximates the KL divergence using the posterior mean of the deviance, and introduces a bias correction term that accounts for the variability of the deviance under the posterior distribution. Let the deviance be defined as

$$D(\boldsymbol{\theta}) = -2 \sum_{i=1}^{n} \log q(y_i \mid \boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{\theta}).$$

Then, the bias in the deviance is given by the effective number of parameters:

$$p_D = \overline{D}(\boldsymbol{\theta}) - D(\hat{\boldsymbol{\theta}}),$$

where $\hat{\boldsymbol{\theta}}$ is a posterior mean of $\boldsymbol{\theta}$. Consequently, the DIC is expressed as

$$\mathrm{DIC} = \overline{D}(\boldsymbol{\theta}) + p_D = D(\hat{\boldsymbol{\theta}}) + 2p_D,$$

where

$$\overline{D}(\boldsymbol{\theta}) = \int_{\Theta} D(\boldsymbol{\theta}) \pi(\boldsymbol{\theta} \mid \boldsymbol{y}, \boldsymbol{X}) \, \mathrm{d}\boldsymbol{\theta}.$$

A notable advantage of DIC is its computational feasibility: both $\overline{D}(\boldsymbol{\theta})$ and p_D can be directly estimated from posterior samples generated by MCMC methods, without requiring closed-form likelihood functions or the existence of maximum likelihood estimates. This makes DIC applicable to complex Bayesian models, for which BIC-type criteria based on marginal likelihoods are often intractable.

2.3.7 Widely applicable information criterion

DIC can be regarded as bringing the idea of AIC into the Bayesian framework. However, because DIC is defined using the posterior mean as a plug-in point estimate, it is not fully coherent with the general Bayesian predictive framework, including non-regular models.

From a more general information-theoretic viewpoint, the widely applicable information criterion (WAIC; Watanabe 2010a) was later proposed as a fully Bayesian extension of AIC. WAIC provides a theoretically rigorous approximation of the KL divergence between the true distribution and the Bayesian predictive distribution. Conceptually, AIC, DIC, and WAIC share the common goal of approximating the KL divergence, but they differ in the distribution being compared to the true one: AIC and DIC use the likelihood evaluated at point estimates, whereas WAIC uses the Bayesian predictive distribution. Although the AIC for the lasso described in the previous subsection was derived under fixed covariates following Ninomiya and Kawano (2016), both AIC and WAIC can be formulated under either fixed or random covariates. In what follows, WAIC is described under the general setting where each pair $(y_i, \boldsymbol{x}_i)$ is treated as a random realization from the true joint distribution, for notational convenience and to maintain coherence with the later discussion of PIIC.

To present the formulation of WAIC in the Bayesian lasso, assume a probability density function $f(\cdot, \cdot \mid \boldsymbol{\theta})$ and a prior distribution $\pi_n(\cdot; \boldsymbol{\lambda})$, and consider independent random vectors satisfying

$$(y_1, \boldsymbol{x}_1), \ldots, (y_n, \boldsymbol{x}_n), (\tilde{y}_1, \tilde{\boldsymbol{x}}_1), \ldots, (\tilde{y}_n, \tilde{\boldsymbol{x}}_n) \sim f(\cdot, \cdot \mid \boldsymbol{\theta}), \qquad \boldsymbol{\theta} \sim \pi_n(\cdot; \boldsymbol{\lambda}).$$

Here, $(y_i, \boldsymbol{x}_i)$ is the i-th sample that gives data by realization, and $(\tilde{y}_i, \tilde{\boldsymbol{x}}_i)$ is a copy of $(y_i, \boldsymbol{x}_i)$ that appears as a notional random vector in deriving the information criterion. In addition, $\pi_n(\cdot; \boldsymbol{\lambda}) \propto \pi(\cdot; \lambda)$ with $\pi(\cdot; \lambda)$ in (2.4). Furthermore, let denote by $\mathrm{E}_{\boldsymbol{\theta}|\boldsymbol{y},\boldsymbol{X};\boldsymbol{\lambda}}[\cdot]$ the expectation based on the posterior distribution $f(\boldsymbol{y}, \boldsymbol{X} \mid \boldsymbol{\theta})\pi_n(\boldsymbol{\theta}; \boldsymbol{\lambda}) / \int_\Theta f(\boldsymbol{y}, \boldsymbol{X} \mid \boldsymbol{\theta})\pi_n(\boldsymbol{\theta}; \boldsymbol{\lambda}) \mathrm{d}\boldsymbol{\theta}$ of $\boldsymbol{\theta}$. Then the Bayesian predictive distribution $f(\cdot, \cdot \mid \boldsymbol{y}, \boldsymbol{X}; \boldsymbol{\lambda})$ can be written as $\mathrm{E}_{\boldsymbol{\theta}|\boldsymbol{y},\boldsymbol{X};\boldsymbol{\lambda}}[f(\cdot, \cdot \mid \boldsymbol{\theta})]$. Specifically, WAIC uses the KL divergence between the true distribution and the Bayesian predictive distribution minus a constant. This quantity is the so-called expected log-loss of the Bayesian predictive distribution, $-\sum_{i=1}^{n} \mathrm{E}_{\tilde{y}_i, \tilde{\boldsymbol{x}}_i} [\log f(\tilde{y}_i, \tilde{\boldsymbol{x}}_i \mid \boldsymbol{y}, \boldsymbol{X}; \boldsymbol{\lambda})]$ and serves as an evaluation index. However, in practice, this cannot be explicitly

calculated, and therefore $-\sum_{i=1}^{n} \log f(y_i, \boldsymbol{x}_i \mid \boldsymbol{y}, \boldsymbol{X}; \boldsymbol{\lambda})$ is considered as an initial estimator. However, this initial estimator uses $(y_i, \boldsymbol{x}_i)$ instead of $(\tilde{y}_i, \tilde{\boldsymbol{x}}_i)$, and hence it underestimates the target. Then, by regarding an asymptotic evaluation of the expectation

$$\begin{aligned}&\sum_{i=1}^{n} \log f(y_i, \boldsymbol{x}_i \mid \boldsymbol{y}, \boldsymbol{X}; \boldsymbol{\lambda}) - \sum_{i=1}^{n} \log f(y_i, \boldsymbol{x}_i \mid \boldsymbol{\theta}_{\boldsymbol{\lambda}}^{bl*}) \\ &- \sum_{i=1}^{n} \log f(\tilde{y}_i, \tilde{\boldsymbol{x}}_i \mid \boldsymbol{y}, \boldsymbol{X}; \boldsymbol{\lambda}) + \sum_{i=1}^{n} \log f(\tilde{y}_i, \tilde{\boldsymbol{x}}_i \mid \boldsymbol{\theta}_{\boldsymbol{\lambda}}^{bl*}),\end{aligned} \tag{2.20}$$

as asymptotic bias, an asymptotic bias correction of $-\sum_{i=1}^{n} \log f(y_i, \boldsymbol{x}_i \mid \boldsymbol{y}, \boldsymbol{X}; \boldsymbol{\lambda})$ can be achieved. Here, $\boldsymbol{\theta}_{\boldsymbol{\lambda}}^{bl*}$ denotes the probability limit of the Bayesian lasso estimator of $\boldsymbol{\theta}$. As a result, WAIC is obtained in the form of an initial estimator plus the following simple penalty term:

$$\sum_{i=1}^{n} \mathrm{E}_{\boldsymbol{\theta}|\boldsymbol{y},\boldsymbol{X};\boldsymbol{\lambda}}[\{\log f(y_i, \boldsymbol{x}_i \mid \boldsymbol{\theta})\}^2] - \sum_{i=1}^{n} \{\mathrm{E}_{\boldsymbol{\theta}|\boldsymbol{y},\boldsymbol{X};\boldsymbol{\lambda}}[\log f(y_i, \boldsymbol{x}_i \mid \boldsymbol{\theta})]\}^2.$$

Note that the expectation of WAIC asymptotically coincides with the expected log-loss of the Bayesian predictive distribution. Furthermore, it is asymptotically equivalent to AIC when the model is regular and the true distribution is contained within it.

2.3.8 Prior intensified information criterion

Despite WAIC being a powerful tool for the selection of Bayesian models, the following two problems should be considered when applying it to the selection of models using Bayesian regularization methods. First, WAIC adopts bias correction based on the second-order term of the asymptotic expansion without making the prior distribution depend on n. Then, there arises the problem that the influence of the prior distribution, which appears as higher-order terms, is not sufficiently reflected. That is, WAIC is targeted at Bayesian estimation that is closely similar to maximum likelihood estimation, and in particular, it might be poorly compatible with Bayesian estimation that induces sparsity. This is significant because emphasis of such estimation is placed on sparseness in reality, whereas WAIC is based on asymptotics close to that of maximum likelihood estimation that does not induce sparsity. The second problem is that WAIC always selects models with prior distributions having more hyperparameters when comparing models, which employ prior distributions from classes with different degrees of freedom.

PIIC was proposed to solve both these problems. The prior distribution was assumed to depend on n, thereby strengthening the influence of the prior distribution on the estimation, in order to solve the first problem. Described concretely in the Bayesian lasso, PIIC assumes $\pi_n(\boldsymbol{\theta};\boldsymbol{\lambda}) \propto \pi(\boldsymbol{\theta};\lambda)^{n/n_0}$. In addition, to solve the second problem, a penalty term for hyperparameters was introduced.

PIIC is the asymptotic bias-corrected statistic of $-\sum_{i=1}^{n}\log f(y_i,\boldsymbol{x}_i \mid \boldsymbol{y},\boldsymbol{X};\boldsymbol{\lambda})$, sharing the same concept as WAIC, and it asymptotically evaluates the expectation of (2.20). With defining the expectation of the weak limit of (2.20) as the asymptotic bias of $-\sum_{i=1}^{n}\log f(y_i,\boldsymbol{x}_i \mid \boldsymbol{y},\boldsymbol{X};\boldsymbol{\lambda})$, the bias is given by

$$\mathrm{tr}\{\boldsymbol{I}^{(2)}_{1,\boldsymbol{\lambda}}(\boldsymbol{\theta}^{bl*}_{\boldsymbol{\lambda}})^{-1}\boldsymbol{I}^{(2)}_{2,\boldsymbol{\lambda}}(\boldsymbol{\theta}^{bl*}_{\boldsymbol{\lambda}})\}, \tag{2.21}$$

where

$$\begin{aligned}
\boldsymbol{I}_{1,\boldsymbol{\lambda}}(\boldsymbol{\theta}) &\equiv \mathrm{E}_{y_i,\boldsymbol{x}_i}\Big\{-\frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^{\mathrm{T}}}\log g(y_i,\boldsymbol{x}_i,\boldsymbol{\theta};\boldsymbol{\lambda})\Big\},\\
\boldsymbol{I}_{2,\boldsymbol{\lambda}}(\boldsymbol{\theta}) &\equiv \mathrm{E}_{y_i,\boldsymbol{x}_i}\Big\{\frac{\partial}{\partial\boldsymbol{\theta}}\log g(y_i,\boldsymbol{x}_i,\boldsymbol{\theta};\boldsymbol{\lambda})\frac{\partial}{\partial\boldsymbol{\theta}^{\mathrm{T}}}\log g(y_i,\boldsymbol{x}_i,\boldsymbol{\theta};\boldsymbol{\lambda})\Big\}
\end{aligned}$$

with $\log g(y_i,\boldsymbol{x}_i,\boldsymbol{\theta};\boldsymbol{\lambda}) \equiv \log f(y_i,\boldsymbol{x}_i \mid \boldsymbol{\theta}) + \frac{1}{n_0}\log\pi(\boldsymbol{\theta};\boldsymbol{\lambda})$. Here, $\boldsymbol{\theta}^{(2)}$ denotes the subvector of $\boldsymbol{\theta}$ corresponding to the subvector of $\boldsymbol{\theta}^{bl*}_{\boldsymbol{\lambda}}$ consisting of its non-zero components, and $\boldsymbol{I}^{(2)}_{1,\boldsymbol{\lambda}}(\boldsymbol{\theta}^{bl*}_{\boldsymbol{\lambda}})$ and $\boldsymbol{I}^{(2)}_{2,\boldsymbol{\lambda}}(\boldsymbol{\theta}^{bl*}_{\boldsymbol{\lambda}})$ denote the sub-matrices of $\boldsymbol{I}_{1,\boldsymbol{\lambda}}(\boldsymbol{\theta}^{bl*}_{\boldsymbol{\lambda}})$ and $\boldsymbol{I}_{2,\boldsymbol{\lambda}}(\boldsymbol{\theta}^{bl*}_{\boldsymbol{\lambda}})$ corresponding to $\boldsymbol{\theta}^{(2)}$, respectively. Then, replacing this bias in (2.21) with its consistent estimator

$$\mathrm{tr}\{\hat{\boldsymbol{I}}^{(2)}_{1,\boldsymbol{\lambda}}(\hat{\boldsymbol{\theta}}^{bl}_{\boldsymbol{\lambda}})^{-1}\hat{\boldsymbol{I}}^{(2)}_{2,\boldsymbol{\lambda}}(\hat{\boldsymbol{\theta}}^{bl}_{\boldsymbol{\lambda}})\},$$

where

$$\begin{aligned}
\hat{\boldsymbol{I}}_{1,\boldsymbol{\lambda}}(\boldsymbol{\theta}) &\equiv -\frac{1}{n}\sum_{i=1}^{n}\frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^{\mathrm{T}}}\log g(y_i,\boldsymbol{x}_i,\boldsymbol{\theta};\boldsymbol{\lambda}),\\
\hat{\boldsymbol{I}}_{2,\boldsymbol{\lambda}}(\boldsymbol{\theta}) &\equiv \frac{1}{n}\sum_{i=1}^{n}\frac{\partial}{\partial\boldsymbol{\theta}}\log g(y_i,\boldsymbol{x}_i,\boldsymbol{\theta};\boldsymbol{\lambda})\frac{\partial}{\partial\boldsymbol{\theta}^{\mathrm{T}}}\log g(y_i,\boldsymbol{x}_i,\boldsymbol{\theta};\boldsymbol{\lambda})
\end{aligned}$$

and $\hat{\boldsymbol{I}}^{(2)}_{1,\boldsymbol{\lambda}}(\boldsymbol{\theta})$ and $\hat{\boldsymbol{I}}^{(2)}_{2,\boldsymbol{\lambda}}(\boldsymbol{\theta})$ are the sub-matrices of $\hat{\boldsymbol{I}}_{1,\boldsymbol{\lambda}}(\boldsymbol{\theta})$ and $\hat{\boldsymbol{I}}_{2,\boldsymbol{\lambda}}(\boldsymbol{\theta})$ corresponding to the subvector of the Bayesian lasso estimator $\hat{\boldsymbol{\theta}}^{bl}_{\boldsymbol{\lambda}}$ consisting of its non-zero components, the prior intensified information criterion is given by

$$\mathrm{PIIC1} = -\sum_{i=1}^{n}\log f(y_i,\boldsymbol{x}_i \mid \boldsymbol{y},\boldsymbol{X};\boldsymbol{\lambda}) + \mathrm{tr}\{\hat{\boldsymbol{I}}^{(2)}_{1,\boldsymbol{\lambda}}(\hat{\boldsymbol{\theta}}^{bl}_{\boldsymbol{\lambda}})^{-1}\hat{\boldsymbol{I}}^{(2)}_{2,\boldsymbol{\lambda}}(\hat{\boldsymbol{\theta}}^{bl}_{\boldsymbol{\lambda}})\}, \tag{2.22}$$

Unlike WAIC, PIIC1 incorporates the influence of a prior distribution that introduces sparsity to the Bayesian lasso estimator. However, as with WAIC,

PIIC1 always selects the model with the larger dimension of $\boldsymbol{\lambda}$ when comparing models that use prior distributions with hyperparameters $\boldsymbol{\lambda}$ of different dimensions. Therefore, by making a bias evaluation that also takes into account the fact that $\boldsymbol{\lambda}$ is selected from the data and adding a penalty term that affects the selection of the dimension of hyperparameters, it is made possible to perform appropriate model selection even when there are candidate prior distributions with different complexities.

Let $\hat{\boldsymbol{\lambda}}$ be the minimizer with respect to $\boldsymbol{\lambda}$ of the information criterion in (2.22), and taking $-\sum_{i=1}^{n} \mathrm{E}_{\tilde{y}_i,\tilde{\boldsymbol{x}}_i}[\log f(\tilde{y}_i, \tilde{\boldsymbol{x}}_i \mid \boldsymbol{y}, \boldsymbol{X}; \hat{\boldsymbol{\lambda}})]$ as the target. Then, let substitute $\hat{\boldsymbol{\lambda}}$ for $\boldsymbol{\lambda}$ in the first and third terms of (2.20) and evaluate this quantity. Then, the asymptotic bias of $-\sum_{i=1}^{n} \log f(y_i, \boldsymbol{x}_i \mid \boldsymbol{y}, \boldsymbol{X}, \hat{\boldsymbol{\lambda}})$ is given by

$$\mathrm{tr}\{\boldsymbol{I}_{1,\boldsymbol{\lambda}^*}(\boldsymbol{\theta}^{bl*}_{\boldsymbol{\lambda}^*})^{-1}\boldsymbol{I}_{2,\boldsymbol{\lambda}^*}(\boldsymbol{\theta}^{bl*}_{\boldsymbol{\lambda}^*})\} + \mathrm{tr}\{\boldsymbol{J}_1(\boldsymbol{\lambda}^*)^{-1}\boldsymbol{J}_2(\boldsymbol{\lambda}^*)\},$$

where $\boldsymbol{\lambda}^*$ is the minimizer with respect to $\boldsymbol{\lambda}$ of $-\mathrm{E}[\log f(y, \boldsymbol{x} \mid \boldsymbol{\theta}^{bl*}_{\boldsymbol{\lambda}})]$ and

$$\begin{aligned}
\boldsymbol{J}_1(\boldsymbol{\lambda}) &\equiv \mathrm{E}\Big[-\frac{\partial^2}{\partial\boldsymbol{\lambda}\partial\boldsymbol{\lambda}^{\mathrm{T}}} \log f(y_i, \boldsymbol{x}_i \mid \boldsymbol{y}, \boldsymbol{X}, \boldsymbol{\lambda}^*)\Big], \\
\boldsymbol{J}_2(\boldsymbol{\lambda}) &\equiv \mathrm{E}\Big[\Big\{\frac{\partial}{\partial\boldsymbol{\lambda}} \log f(y_i, \boldsymbol{x}_i \mid \boldsymbol{y}, \boldsymbol{X}, \boldsymbol{\lambda}^*)\Big\}\Big\{\frac{\partial}{\partial\boldsymbol{\lambda}} \log f(y_i, \boldsymbol{x}_i \mid \boldsymbol{y}, \boldsymbol{X}, \boldsymbol{\lambda}^*)\Big\}^{\mathrm{T}}\Big].
\end{aligned}$$

Then, the prior intensified information criterion which also considers the complexity of the prior distribution is given by

$$\mathrm{PIIC2} \equiv -\sum_{i=1}^{n} \log f(y_i, \boldsymbol{x}_i \mid \boldsymbol{y}, \boldsymbol{X}; \hat{\boldsymbol{\lambda}}) + \mathrm{tr}\{\hat{\boldsymbol{I}}_{1,\hat{\boldsymbol{\lambda}}}(\hat{\boldsymbol{\theta}}^{bl}_{\hat{\boldsymbol{\lambda}}})^{-1}\hat{\boldsymbol{I}}_{2,\hat{\boldsymbol{\lambda}}}(\hat{\boldsymbol{\theta}}^{bl}_{\hat{\boldsymbol{\lambda}}}) + \hat{\boldsymbol{J}}_1(\hat{\boldsymbol{\lambda}})^{-1}\hat{\boldsymbol{J}}_2(\hat{\boldsymbol{\lambda}})\},$$

where

$$\begin{aligned}
\hat{\boldsymbol{J}}_1(\hat{\boldsymbol{\lambda}}) &\equiv -\frac{1}{n}\sum_{i=1}^{n} \frac{\partial^2}{\partial\boldsymbol{\lambda}\partial\boldsymbol{\lambda}^{\mathrm{T}}} \log f(y_i, \boldsymbol{x}_i \mid \boldsymbol{y}, \boldsymbol{X}; \hat{\boldsymbol{\lambda}}), \\
\hat{\boldsymbol{J}}_2(\hat{\boldsymbol{\lambda}}) &\equiv \frac{1}{n}\sum_{i=1}^{n} \frac{\partial}{\partial\boldsymbol{\lambda}} \log f(y_i, \boldsymbol{x}_i \mid \boldsymbol{y}, \boldsymbol{X}; \hat{\boldsymbol{\lambda}}) \frac{\partial}{\partial\boldsymbol{\lambda}^{\mathrm{T}}} \log f(y_i, \boldsymbol{x}_i \mid \boldsymbol{y}, \boldsymbol{X}; \hat{\boldsymbol{\lambda}}).
\end{aligned}$$

Chapter 3

Bayesian variable fusion method for binary data

In this chapter, we describe Bayesian methods for variable fusion in a logistic regression model. First, we review a Bayesian logistic regression model and a data-augmentation method using a Pólya-Gamma distribution. Then, we describe Bayesian logistic regression models based on the Bayesian lasso and also outline its extension employing horseshoe prior. We subsequently propose logistic regression models with the Bayesian fused lasso using Laplace and horseshoe priors and their extensions to the Bayesian generalized fused lasso. We also derive efficient Gibbs samplers for posterior inference. Finally, we present numerical experiments and a real data analysis, focusing only on the proposed models with the Bayesian fused lasso.

3.1 Data-augmentation method with Pólya-Gamma distribution

Suppose that we have a dataset $\{(y_i, \boldsymbol{x}_i); i = 1, \ldots, n\}$, where $y_i \in \{0, 1\}$ is a binary response variable and $\boldsymbol{x}_i$ is a p-dimensional vector of explanatory variables. A logistic regression model is formulated as

$$\Pr(y_i = 1 \mid \boldsymbol{x}_i) = \frac{1}{1 + e^{-(\theta_0 + \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\theta})}}, \quad i = 1, \ldots, n,$$

where θ_0 is an intercept and $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_p)^{\mathrm{T}}$ is a p-dimensional regression coefficient vector. The likelihood function is given by

$$\prod_{i=1}^{n} f(y_i \mid \boldsymbol{x}_i; \theta_0, \boldsymbol{\theta}),$$

where

$$f(y_i \mid \boldsymbol{x}_i; \theta_0, \boldsymbol{\theta}) = \frac{(e^{\theta_0 + \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\theta}})^{y_i}}{1 + e^{\theta_0 + \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\theta}}}. \tag{3.1}$$

Then, the log-likelihood function is given by

$$\ell_{LR}(\theta_0, \boldsymbol{\theta}) = \sum_{i=1}^{n} \left[y_i(\theta_0 + \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\theta}) - \log\left\{1 + \exp(\theta_0 + \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\theta})\right\} \right]. \tag{3.2}$$

In maximum likelihood estimation, the regression coefficient $\boldsymbol{\theta}$ is estimated by the maximization of the log-likelihood (3.2).

In the Bayesian framework, a logistic regression model is formulated as

$$y_i \mid \boldsymbol{x}_i, \boldsymbol{\theta}, \theta_0 \sim \mathrm{Binom}\left(1, \frac{1}{1 + e^{-(\theta_0 + \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\theta})}}\right), \tag{3.3}$$

where $\mathrm{Binom}(\cdot, \cdot)$ represents a binomial distribution.

To obtain samples from the posterior distribution under the model in (3.3) by Gibbs sampling, Polson et al. (2013) proposed a data-augmentation method with a Pólya-Gamma distribution. The probability density function of a Pólya-Gamma distribution is expressed by

$$\mathrm{PG}(x \mid a, b) = \frac{1}{2\pi^2} \sum_{k=1}^{\infty} \frac{g_k}{\left(k - \frac{1}{2}\right)^2 + \frac{b^2}{(4\pi^2)}}, \tag{3.4}$$

where a (> 0) and b are hyper-parameters and g_k is an independent random variable with a Gamma distribution $\mathrm{Ga}(a, 1)$. In addition, the Pólya-Gamma distribution (3.4) can be expressed hierarchically with $\mathrm{PG}(x \mid a, 0)$ as follows:

$$\mathrm{PG}(x \mid a, b) \propto \exp\left(-\frac{b^2 x}{2}\right) \mathrm{PG}(x \mid a, 0). \tag{3.5}$$

By using the hierarchical expression (3.5), the function (3.1) can be expressed hierarchically with the latent variables w_i:

$$\begin{aligned}
&f(y_i \mid \boldsymbol{x}_i, \boldsymbol{\theta}, \theta_0) \\
&= \frac{(e^{\theta_0 + \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\theta}})^{y_i}}{1 + e^{\theta_0 + \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\theta}}} \\
&= \frac{1}{2} \exp\{\kappa_i(\theta_0 + \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\theta})\} \int_0^{\infty} \exp\left\{-\frac{w_i(\theta_0 + \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\theta})^2}{2}\right\} \mathrm{PG}(w_i \mid 1, 0) dw_i,
\end{aligned}$$

where $\kappa_i = y_i - 0.5$. When the regression coefficient vector $\boldsymbol{\theta}$ has a Gaussian prior $\mathrm{N}_p(\mathbf{0}_p, \boldsymbol{B})$, the full conditional distributions can be obtained as follows:

$$\begin{aligned}
\boldsymbol{\theta} \mid \boldsymbol{y}, \boldsymbol{X}, \theta_0, \lambda_1, w_1, \ldots, w_n &\sim \mathrm{N}_p(\boldsymbol{A}^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{W}(\boldsymbol{z} - \theta_0 \mathbf{1}), \boldsymbol{A}^{-1}), \\
w_i \mid \boldsymbol{x}_i, \boldsymbol{\theta}, \theta_0 &\sim \mathrm{PG}(1, \theta_0 + \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\theta}),
\end{aligned} \tag{3.6}$$

where $\boldsymbol{X} = (\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n)^{\mathrm{T}}$, $\boldsymbol{A} = \boldsymbol{X}^{\mathrm{T}}\boldsymbol{W}\boldsymbol{X} + \boldsymbol{B}^{-1}$, $\boldsymbol{W} = \mathrm{diag}(w_1, \ldots, w_n)$, $\boldsymbol{z} = (\kappa_1/w_1, \kappa_2/w_2, \ldots, \kappa_n/w_n)^{\mathrm{T}}$, and $\mathbf{1}$ is an n-dimensional vector of which all components are one. From the full conditional distributions (3.6), the Gibbs sampling can be performed.

3.2 Bayesian logistic regression with lasso-type priors

3.2.1 Bayesian logistic regression model with Laplace prior

In Sec. 2.2, we reviewed the L_1-norm regularization methods. In the following, we discuss how these regularization techniques can be incorporated into the Bayesian logistic regression framework. We begin by revisiting the Bayesian lasso, for which several extensions have been proposed in the context of logistic regression models. Makalic and Schmidt (2016) proposed a logistic regression model with the priors (2.6) in the case of $\lambda_1^2 = 2$ and $\boldsymbol{\Sigma}_l = \mathrm{diag}\left(\psi^2\tau_1^2, \psi^2\tau_2^2, \ldots, \psi^2\tau_p^2\right)$, where ψ is a hyper-parameter which determines the overall shrinkage of the regression coefficients. Tian et al. (2019) also proposed a logistic regression model with the Bayesian lasso. In Tian et al. (2019), the priors (2.6) were assumed and the hyper-parameter λ was selected to make a predictive probability distribution be an approximately uniform distribution.

Among the models, we note the logistic regression model with the Bayesian lasso having priors (2.6) on regression coefficients. First, as in Makalic and Schmidt (2016), a uniform distribution $U(\iota, \iota)$ with a hyper-parameter ι is assumed on the intercept θ_0. In addition, a Gamma distribution $\mathrm{Ga}(r_1, t_1)$ is assumed on λ^2, following Park and Casella (2008). By combining the data-augmentation method in Sec. 3.1 and the expression for the prior on regression coefficients as a Gaussian scale-mixture prior (2.6), the full conditional distributions can be obtained as follows:

$$
\begin{aligned}
\boldsymbol{\theta} \mid \boldsymbol{y}, \boldsymbol{X}, \theta_0, \tau_1^2, \ldots, \tau_p^2, w_1, \ldots, w_n &\sim \mathrm{N}_p(\boldsymbol{A}^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{W}(\boldsymbol{z} - \theta_0\boldsymbol{1}), \boldsymbol{A}^{-1}), \\
w_i \mid \boldsymbol{x}_i, \boldsymbol{\theta}, \theta_0 &\sim \mathrm{PG}(1, \theta_0 + \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\theta}), \\
\frac{1}{\tau_j^2} \mid \theta_j, \lambda_1^2 &\sim \mathrm{IGauss}\left(\sqrt{\frac{\lambda_1^2}{\theta_j^2}}, \lambda_1^2\right), \\
\lambda_1^2 \mid \tau_1^2, \ldots, \tau_p^2 &\sim \mathrm{Ga}\left(p + r_1, \frac{1}{2}\sum_{j=1}^{p}\tau_j^2 + t_1\right), \\
\theta_0 \mid \boldsymbol{y}, \boldsymbol{X}, \tau_1^2, \ldots, \tau_p^2, w_1, \ldots, w_n &\sim \mathrm{N}\left(\frac{1}{S}\sum_{i=1}^{n} v_i, \frac{1}{S}\right),
\end{aligned}
$$

where $\boldsymbol{A} = \boldsymbol{X}^{\mathrm{T}}\boldsymbol{W}\boldsymbol{X} + \boldsymbol{\Sigma}_l^{-1}$, $\boldsymbol{v} = (\boldsymbol{z} - \boldsymbol{X}\boldsymbol{\theta}) \odot \boldsymbol{W}$, $S = \sum_{i=1}^{n} w_i$, and the symbol $\odot$ means the Hadamard product. $\mathrm{IGauss}\,(a, b)$ represents an inverse-Gaussian distribution, where a is a mean parameter and b is a shape parameter.

3.2.2 Bayesian logistic regression model with horseshoe prior

A Laplace prior tends to shrink regression coefficients too little when the true coefficients are zero and shrink the coefficients too much when the true coefficients are non-zero. This is why a Laplace prior has insufficient concentration at zero and an exponential tail. To avoid this problem, Carvalho et al. (2010) proposed a horseshoe prior given by

$$\theta_j \mid \lambda_j^2, \tau^2 \sim \mathrm{N}(0, \lambda_j^2 \tau^2), \ \lambda_j \sim \mathrm{C}^+(0,1), \ \tau \sim \mathrm{C}^+(0,1), \tag{3.7}$$

where $\mathrm{C}^+(\cdot,\cdot)$ represents a half-Cauchy distribution, λ_j adjusts the degree of the shrinkage of each regression coefficient θ_j, and τ adjusts the amount of the shrinkage of the overall regression coefficients $\boldsymbol{\theta}$. By having a hyper-parameter λ_j following a half-Cauchy distribution with a pole at zero and a polynomial tail, a horseshoe prior realizes an infinite spike at zero and a heavier tail than a Laplace prior. Thus, a horseshoe prior can strongly shrink the small regression coefficients towards zero and prevent over-shrinkage of large regression coefficients.

Makalic and Schmidt (2015) proposed a logistic regression model whose regression coefficients follow a horseshoe prior. They utilized the hierarchical expression of a half-Cauchy distribution (Wand et al., 2011). When x follows $\mathrm{C}^+(0,1)$, the relation

$$x^2 \mid y \sim \mathrm{IG}\left(\frac{1}{2}, \frac{1}{y}\right), \quad y \sim \mathrm{IG}\left(\frac{1}{2}, 1\right)$$

holds, where $\mathrm{IG}(a,b)$ represents an inverse-Gamma prior with a shape parameter a and a scale parameter b. Therefore, the priors (3.7) can be rewritten as follows:

$$\begin{aligned} \boldsymbol{\theta} \mid \lambda_1^2, \ldots, \lambda_p^2, \tau^2 &\sim \mathrm{N}_p(\mathbf{0}_p, \boldsymbol{D}), \\ \lambda_j^2 \mid \nu_j &\sim \mathrm{IG}\left(\frac{1}{2}, \frac{1}{\nu_j}\right), \\ \tau^2 \mid \xi &\sim \mathrm{IG}\left(\frac{1}{2}, \frac{1}{\xi}\right), \\ \nu_1, \ldots, \nu_p, \xi &\sim \mathrm{IG}\left(\frac{1}{2}, 1\right), \end{aligned} \tag{3.8}$$

where $\boldsymbol{D} = \mathrm{diag}\left(\tau^2\lambda_1^2, \tau^2\lambda_2^2 \ldots, \tau^2\lambda_p^2\right)$. By using the priors (3.8) and the data-augmentation method with a Pólya-Gamma distribution, the full conditional distributions can be obtained. The details are found in Makalic and Schmidt (2015).

3.3 Logistic regression model with Bayesian generalized fused lasso

As reviewed in Sec. 2.2.4, Kyung et al. (2010) extended the Bayesian lasso by proposing a linear regression model that employs the prior given in (2.7). Extending this framework to binary data, we propose a logistic regression model which assumes a Laplace prior on regression coefficients and differences between adjacent regression coefficients. Specifically, we assume the prior given in (2.7) on regression coefficients. The prior (2.7) induce shrinkage of both regression coefficients and differences of adjacent regression coefficients towards zero. In addition, we assume the uniform prior $\mathrm{U}(\iota, \iota)$ on an intercept θ_0 as in Makalic and Schmidt (2016). We also assume Gamma distributions $\mathrm{Ga}(r_1, t_1)$ and $\mathrm{Ga}(r_2, t_2)$ on λ_1^2 and λ_2^2, respectively. By combining a data-augmentation method with a Pólya-Gamma distribution and the hierarchical expression of the priors in (2.8), the full conditional distributions can be obtained as follows:

$$
\begin{aligned}
\boldsymbol{\theta} \mid \boldsymbol{y}, \boldsymbol{X}, \theta_0, \tau_1^2, \ldots, \tau_p^2, \tilde{\tau}_2^2, \ldots, \tilde{\tau}_p^2, w_1, \ldots, w_n &\sim \mathrm{N}_p(\boldsymbol{A}^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{W}(\boldsymbol{z} - \theta_0 \boldsymbol{1}), \boldsymbol{A}^{-1}), \\
w_i \mid \boldsymbol{x}_i, \boldsymbol{\theta}, \theta_0 &\sim \mathrm{PG}(1, \theta_0 + \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\theta}), \\
\frac{1}{\tau_j^2} \mid \theta_j, \lambda_1^2 &\sim \mathrm{IGauss}\left(\sqrt{\frac{\lambda_1^2}{\theta_j^2}}, \lambda_1^2\right), \\
\lambda_1^2 \mid \tau_1^2, \ldots, \tau_p^2 &\sim \mathrm{Ga}\left(p + r_1, \frac{1}{2}\sum_{j=1}^{p} \tau_j^2 + t_1\right), \\
\frac{1}{\tilde{\tau}_j^2} \mid \theta_j, \lambda_2^2 &\sim \mathrm{IGauss}\left(\sqrt{\frac{\lambda_2^2}{(\theta_j - \theta_{j-1})^2}}, \lambda_2^2\right), \\
\lambda_2^2 \mid \tilde{\tau}_2^2, \ldots, \tilde{\tau}_p^2 &\sim \mathrm{Ga}\left(p - 1 + r_2, \frac{1}{2}\sum_{j=2}^{p} \tilde{\tau}_j^2 + t_2\right), \\
\theta_0 \mid \boldsymbol{y}, \boldsymbol{X}, \boldsymbol{\theta}, \tau_1^2, \ldots, \tau_p^2, \tilde{\tau}_2^2, \ldots, \tilde{\tau}_p^2, w_1, \ldots, w_n &\sim \mathrm{N}\left(\frac{1}{S}\sum_{i=1}^{n} v_i, \frac{1}{S}\right),
\end{aligned}
$$

where $\boldsymbol{A} = \boldsymbol{X}^{\mathrm{T}}\boldsymbol{W}\boldsymbol{X} + \boldsymbol{\Sigma}_{fl}^{-1}$, $\boldsymbol{v} = (\boldsymbol{z} - \boldsymbol{X}\boldsymbol{\theta}) \odot \boldsymbol{W}$, and $S = \sum_{i=1}^{n} w_i$. From these full conditional distributions, Gibbs sampling can be performed.

Similar to the Bayesian fused lasso, the Bayesian generalized fused lasso can also be incorporated into a logistic regression model. Assuming the prior in (2.10) on regression coefficients, expressing it hierarchically as in (2.11), and combining the data-augmentation method, the full conditional distributions that

enable Gibbs sampling can be obtained as follows:

$$
\begin{aligned}
\boldsymbol{\theta} \mid \boldsymbol{y}, \boldsymbol{X}, \theta_0, \tau_1^2, \ldots, \tau_p^2, \tilde{\tau}^2_{e_{11},e_{12}}, \ldots, \tilde{\tau}^2_{e_{\chi 1},e_{\chi 2}}, w_1, \ldots, w_n &\sim \mathrm{N}_p(\boldsymbol{A}^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{W}(\boldsymbol{z}-\theta_0\boldsymbol{1}), \boldsymbol{A}^{-1}), \\
w_i \mid \boldsymbol{x}_i, \boldsymbol{\theta}, \theta_0 &\sim \mathrm{PG}(1, \theta_0 + \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\theta}), \\
\frac{1}{\tau_j^2} \mid \theta_j, \lambda_1^2 &\sim \mathrm{IGauss}\left(\sqrt{\frac{\lambda_1^2}{\theta_j^2}}, \lambda_1^2\right), \\
\lambda_1^2 \mid \tau_1^2, \ldots, \tau_p^2 &\sim \mathrm{Ga}\left(p + r_1, \frac{1}{2}\sum_{j=1}^{p}\tau_j^2 + t_1\right), \\
\frac{1}{\tilde{\tau}^2_{j,k}} \mid \theta_j, \theta_k, \lambda_2^2 &\sim \mathrm{IGauss}\left(\sqrt{\frac{\lambda_2^2}{(\theta_j-\theta_k)^2}}, \lambda_2^2\right), \\
\lambda_2^2 \mid \tilde{\tau}^2_{e_{11},e_{12}}, \ldots, \tilde{\tau}^2_{e_{\chi 1},e_{\chi 2}} &\sim \mathrm{Ga}\left(\chi + r_2 - 1, \frac{1}{2}\sum_{(j,k)\in E_{gfl}}\tilde{\tau}^2_{j,k} + t_2\right), \\
\theta_0 \mid \boldsymbol{y}, \boldsymbol{X}, \boldsymbol{\theta}, \tau_1^2, \ldots, \tau_p^2, \tilde{\tau}^2_{e_{11},e_{12}}, \ldots, \tilde{\tau}^2_{e_{\chi 1},e_{\chi 2}}, w_1, \ldots, w_n &\sim \mathrm{N}\left(\frac{1}{S}\sum_{i=1}^{n} v_i, \frac{1}{S}\right),
\end{aligned}
$$

where $\boldsymbol{A} = \boldsymbol{X}^{\mathrm{T}}\boldsymbol{W}\boldsymbol{X} + \boldsymbol{\Sigma}_{gfl}$.

3.4 Logistic regression model with Bayesian generalized fused lasso with horseshoe prior

We also propose an alternative Bayesian model in Sec. 3.3 by assuming a horseshoe prior on differences between adjacent regression coefficients. First, we introduce the priors given by

$$
\begin{aligned}
\pi(\boldsymbol{\theta}) \propto &\prod_{j=1}^{p} \mathrm{Laplace}\left(\theta_j \,\middle|\, 0, \frac{1}{\tilde{\lambda}_1}\right) \\
&\times \int \left[\prod_{j=2}^{p} \int \frac{1}{\sqrt{2\pi\lambda_j^2\tilde{\tau}^2}} \exp\left\{-\frac{(\theta_j-\theta_{j-1})^2}{2\lambda_j^2\tilde{\tau}^2}\right\} \frac{2}{\pi(1+\lambda_j^2)} \mathrm{d}\lambda_j^2\right] \frac{2}{\pi(1+\tilde{\tau}^2)} \mathrm{d}\tilde{\tau}^2.
\end{aligned} \tag{3.9}
$$

With the hierarchical expressions of a Laplace distribution and a half-Cauchy distribution, the priors (3.9) are expressed as

$$
\begin{aligned}
\boldsymbol{\theta} \mid \tau_1^2, \ldots, \tau_p^2, \lambda_2^2, \ldots, \lambda_p^2, \tilde{\tau}^2 &\sim \mathrm{N}_p(\boldsymbol{0}_p, \boldsymbol{F}), \\
\tau_j^2 &\sim \mathrm{EXP}\left(\frac{\tilde{\lambda}_1^2}{2}\right), \\
\tilde{\tau}^2 \mid \xi &\sim \mathrm{IG}\left(\frac{1}{2}, \frac{1}{\xi}\right), \\
\lambda_j^2 \mid \nu_j &\sim \mathrm{IG}\left(\frac{1}{2}, \frac{1}{\nu_j}\right), \\
\xi, \nu_j &\sim \mathrm{IG}\left(\frac{1}{2}, 1\right),
\end{aligned} \tag{3.10}
$$

where the inverse matrix of $\boldsymbol{F}$ is given by

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

By assuming a horseshoe prior on differences between adjacent regression coefficients, small differences can be shrunk more, while large differences are shrunk less, compared to a Laplace prior.

From the priors (3.10) and the data-augmentation method, we can get the full conditional distributions as follows:

$$\begin{aligned}
\boldsymbol{\theta} \mid \boldsymbol{y}, \boldsymbol{X}, \theta_0, \tau_1^2, \dots, \tau_p^2, \lambda_2^2, \dots, \lambda_p^2, \tilde{\tau}^2, w_1, \dots, w_n &\sim \mathrm{N}_p(\boldsymbol{A}^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{W}(\boldsymbol{z} - \theta_0\boldsymbol{1}), \boldsymbol{A}^{-1}), \\
w_i \mid \boldsymbol{x}_i, \boldsymbol{\theta}, \theta_0 &\sim \mathrm{PG}(1, \theta_0 + \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\theta}), \\
\frac{1}{\tau_j^2} \mid \theta_j, \tilde{\lambda}_1^2 &\sim \mathrm{IGauss}\left(\sqrt{\frac{\tilde{\lambda}_1^2}{\theta_j^2}}, \tilde{\lambda}_1^2\right), \\
\tilde{\lambda}_1^2 \mid \tau_1^2, \dots, \tau_p^2 &\sim \mathrm{Ga}\left(p + r_1, \frac{1}{2}\sum_{j=1}^{p}\tau_j^2 + t_1\right), \\
\tilde{\tau}^2 \mid \theta_1, \dots, \theta_p, \lambda_2^2, \dots, \lambda_p^2, \xi &\sim \mathrm{IG}\left(\frac{p}{2}, \frac{1}{2}\sum_{j=2}^{p}\frac{(\theta_j - \theta_{j-1})^2}{\lambda_j^2} + \frac{1}{\xi}\right), \\
\lambda_j^2 \mid \theta_j, \theta_{j-1}, \tilde{\tau}^2, \nu_j &\sim \mathrm{IG}\left(1, \frac{(\theta_j - \theta_{j-1})^2}{2\tilde{\tau}^2} + \frac{1}{\nu_j}\right), \\
\nu_j \mid \lambda_j^2 &\sim \mathrm{IG}\left(1, \frac{1}{\lambda_j^2} + 1\right), \\
\xi \mid \tilde{\tau}^2 &\sim \mathrm{IG}\left(1, \frac{1}{\tilde{\tau}^2} + 1\right), \\
\theta_0 \mid \boldsymbol{y}, \boldsymbol{X}, \tau_1^2, \dots, \tau_p^2, \lambda_2^2, \dots, \lambda_p^2, \tilde{\tau}^2, w_1, \dots, w_n &\sim \mathrm{N}\left(\frac{1}{S}\sum_{i=1}^{n} v_i, \frac{1}{S}\right),
\end{aligned}$$

where $\boldsymbol{A} = \boldsymbol{X}^{\mathrm{T}}\boldsymbol{W}\boldsymbol{X} + \boldsymbol{F}^{-1}$, $\boldsymbol{v} = (\boldsymbol{z} - \boldsymbol{X}\boldsymbol{\theta}) \odot \boldsymbol{W}$, and $S = \sum_{i=1}^{n} w_i$. Note that we did not assume a horseshoe prior on regression coefficients, because the MCMC chain did not converge.

This logistic regression model with the Bayesian fused lasso with a horseshoe prior can be extended to the Bayesian generalized fused lasso framework.

Specifically, we consider the priors on regression coefficients as follows:

$$\pi(\boldsymbol{\theta}) \propto \prod_{j=1}^{p} \text{Laplace}\left(\theta_j \,\middle|\, 0, \frac{1}{\tilde{\lambda}_1}\right) \times \int \left[\prod_{(j,k)\in E_{gfl}} \int \frac{1}{\sqrt{2\pi\lambda_{j,k}^2\tilde{\tau}^2}} \exp\left\{-\frac{(\theta_j-\theta_k)^2}{2\lambda_{j,k}^2\tilde{\tau}^2}\right\} \frac{2}{\pi(1+\lambda_{j,k}^2)} \mathrm{d}\lambda_{j,k}^2 \right] \frac{2}{\pi(1+\tilde{\tau}^2)} \mathrm{d}\tilde{\tau}^2. \tag{3.11}$$

Then, the hierarchical expressions of the priors in (3.11) are given by

$$\begin{aligned} \boldsymbol{\theta} \mid \tau_1^2, \ldots, \tau_p^2, \lambda_{e_{11},e_{12}}^2, \ldots, \lambda_{e_{\chi 1},e_{\chi 2}}^2, \tilde{\tau}^2 &\sim \mathrm{N}_p(\mathbf{0}_p, \boldsymbol{K}), \\ \tau_j^2 &\sim \text{EXP}\left(\frac{\tilde{\lambda}_1^2}{2}\right), \\ \tilde{\tau}^2 \mid \xi &\sim \text{IG}\left(\frac{1}{2}, \frac{1}{\xi}\right), \\ \lambda_{j,k}^2 \mid \nu_{j,k} &\sim \text{IG}\left(\frac{1}{2}, \frac{1}{\nu_{j,k}}\right) \\ \xi, \nu_{j,k} &\sim \text{IG}\left(\frac{1}{2}, 1\right), \end{aligned} \tag{3.12}$$

where the inverse matrix of $\boldsymbol{K}$ is given by

$$\boldsymbol{K}^{-1} = \boldsymbol{T} + \tilde{\tau}^{-2}\boldsymbol{\Upsilon}^{\mathrm{T}}\tilde{\boldsymbol{T}}\boldsymbol{\Upsilon},$$

and $\boldsymbol{T}$ and $\tilde{\boldsymbol{T}}$ are defined as

$$\boldsymbol{T} = \mathrm{diag}(1/\tau_1^2, \ldots, 1/\tau_p^2),$$

$$\tilde{\boldsymbol{T}}_{ij} = \begin{cases} 1/\lambda_{e_{i1},e_{i2}}^2, & \text{if } i = j, \\ 0, & \text{otherwise.} \end{cases}$$

From the hierarchical expressions of priors in (3.12) and the data-augmentation method, we can obtain the full conditional distributions that enable Gibbs sam-

pling as follows:

$$
\begin{aligned}
\boldsymbol{\theta} \mid \boldsymbol{y}, \boldsymbol{X}, \theta_0, \tau_1^2, \ldots, \tau_p^2, \lambda_{e_{11},e_{12}}^2, \ldots, \lambda_{e_{\chi 1},e_{\chi 2}}^2, \tilde{\tau}^2, w_1, \ldots, w_n &\sim \mathrm{N}_p(\boldsymbol{A}^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{W}(\boldsymbol{z}-\theta_0\boldsymbol{1}), \boldsymbol{A}^{-1}),\\
w_i \mid \boldsymbol{x}_i, \boldsymbol{\theta}, \theta_0 &\sim \mathrm{PG}(1, \theta_0 + \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\theta}),\\
\frac{1}{\tau_j^2} \mid \theta_j, \tilde{\lambda}_1^2 &\sim \mathrm{IGauss}\left(\sqrt{\frac{\tilde{\lambda}_1^2}{\theta_j^2}}, \tilde{\lambda}_1^2\right),\\
\tilde{\lambda}_1^2 \mid \tau_1^2, \ldots, \tau_p^2 &\sim \mathrm{Ga}\left(p + r_1, \frac{1}{2}\sum_{j=1}^{p}\tau_j^2 + t_1\right),\\
\tilde{\tau}^2 \mid \theta_1, \ldots, \theta_p, \lambda_{e_{11},e_{12}}^2, \ldots, \lambda_{e_{\chi 1},e_{\chi 2}}^2, \xi &\sim \mathrm{IG}\left(\frac{\chi}{2} - 1, \frac{1}{2}\sum_{(j,k)\in E_{gfl}}\frac{(\theta_j-\theta_k)^2}{\lambda_{j,k}^2} + \frac{1}{\xi}\right),\\
\lambda_{j,k}^2 \mid \theta_j, \theta_k, \tilde{\tau}^2, \nu_{j,k} &\sim \mathrm{IG}\left(1, \frac{(\theta_j-\theta_k)^2}{2\tilde{\tau}^2} + \frac{1}{\nu_{j,k}}\right),\\
\nu_{j,k} \mid \lambda_{j,k}^2 &\sim \mathrm{IG}\left(1, \frac{1}{\lambda_{j,k}^2} + 1\right),\\
\xi \mid \tilde{\tau}^2 &\sim \mathrm{IG}\left(1, \frac{1}{\tilde{\tau}^2} + 1\right),\\
\theta_0 \mid \boldsymbol{y}, \boldsymbol{X}, \tau_1^2, \ldots, \tau_p^2, \lambda_{e_{11},e_{12}}^2, \ldots, \lambda_{e_{\chi 1},e_{\chi 2}}^2, \tilde{\tau}^2, w_1, \ldots, w_n &\sim \mathrm{N}\left(\frac{1}{S}\sum_{i=1}^{n} v_i, \frac{1}{S}\right),
\end{aligned}
$$

where $\boldsymbol{A} = \boldsymbol{X}^{\mathrm{T}}\boldsymbol{W}\boldsymbol{X} + \boldsymbol{K}^{-1}$.

3.5 Monte Carlo simulations

We generated y_i $(i = 1, \ldots, n)$ according to the true model:

$$\Pr(y_i = 1 \mid \boldsymbol{x}_i) = \frac{1}{1 + e^{-\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\theta}^*}},$$

where $\boldsymbol{\theta}^* = (\theta_1^*, \theta_2^*, \ldots, \theta_p^*)^{\mathrm{T}}$ is a p-dimensional regression coefficient vector. The explanatory variable $\boldsymbol{x}_i$ $(i = 1, \ldots, n)$ followed the multivariate normal distribution $\mathrm{N}_p(\boldsymbol{0}_p, \Sigma)$. For $\boldsymbol{\theta}^*$ and Σ, we considered the following cases:

Case 1: $\boldsymbol{\theta}^* = \boldsymbol{\theta}_1^*$ or $\boldsymbol{\theta}_2^*$, $\Sigma_{ii} = 1$, $\Sigma_{ij} = \rho$, $\quad (i \neq j)$,

Case 2: $\boldsymbol{\theta}^* = \boldsymbol{\theta}_1^*$ or $\boldsymbol{\theta}_2^*$,

$$\Sigma_{ii} = 1,\ \Sigma_{ij} = \begin{cases} 0.5 & (\theta_i^* = \theta_j^* \text{ and } 1 \leq |i-j| \leq 4) \\ 0 & \text{otherwise} \end{cases} \quad (i \neq j),$$

Case 3: $\boldsymbol{\theta}^* = \boldsymbol{\theta}_1^*$ or $\boldsymbol{\theta}_2^*$,

$$\Sigma_{ii} = 1,\ \Sigma_{ij} = \begin{cases} 0.5^{|i-j|} & (\theta_i^* = \theta_j^* \text{ and } 1 \leq |i-j| \leq 4) \\ 0 & \text{otherwise} \end{cases} \quad (i \neq j),$$

Case 4: $\boldsymbol{\theta}^* = (\boldsymbol{1.0}_{20}^{\mathrm{T}}, \boldsymbol{-1.0}_{20}^{\mathrm{T}}, \boldsymbol{0.0}_{170}^{\mathrm{T}}, \boldsymbol{1.5}_{20}^{\mathrm{T}}, \boldsymbol{0.0}_{170}^{\mathrm{T}})^{\mathrm{T}}$, $\Sigma_{ii} = 1$, $\Sigma_{ij} = 0$, $\quad (i \neq j)$,

where $\boldsymbol{\theta}_1^* = (\boldsymbol{1.0}_5^{\mathrm{T}}, \boldsymbol{0.0}_5^{\mathrm{T}}, \boldsymbol{1.0}_5^{\mathrm{T}}, \boldsymbol{0.0}_5^{\mathrm{T}})^{\mathrm{T}}$, $\boldsymbol{\theta}_2^* = (\boldsymbol{-1.0}_5^{\mathrm{T}}, \boldsymbol{2.0}_5^{\mathrm{T}}, \boldsymbol{1.0}_5^{\mathrm{T}}, \boldsymbol{0.0}_5^{\mathrm{T}})^{\mathrm{T}}$, Σ_{ij} is

the (i, j)-th element of Σ, and $\rho = 0.0, 0.5$. We considered $n = 500, 1,000$ for Case 1 and $n = 500$ for Cases 2 and 3, which correspond to $n > p$ cases, while $n = 300$ for Case 4, which corresponds to an $n < p$ case. We simulated 100 datasets for each case. Note that Case 4 is based on the simulation in Bhattacharyya et al. (2022).

We compared our proposed methods, which are the logistic regression model with the Bayesian fused lasso (LBFL) and that with the Bayesian fused lasso with horseshoe prior (LBFH), to the logistic regression model with the fused lasso (LFL), that with the MCP (LMCP), that with the SCAD (LSCAD), and that with no penalty (LN). For Case 4, the comparison was conducted except for LN, because it cannot be used for $n < p$ case. For LN and LFL, we used the package penalized of the software **R**, which is available from https://cran.r-project.org/web/packages/penalized/index.html. For LMCP and LSCAD, we used the package ncvreg, which can be obtained from https://cran.r-project.org/web/packages/ncvreg/index.html. The values of the hyper-parameters λ_1 and λ_2 for LFL were selected by the Bayesian information criterion (BIC). As the term of the degrees of freedom in BIC, we used the number of groups which consists of non-zero fused estimated regression coefficients (Tibshirani et al., 2005; Tibshirani and Taylor, 2011). For LMCP and LSCAD, the tuning parameter was set as 3.0 for LMCP and 3.7 for LSCAD. The regularization parameter was selected by 10-fold cross-validation. For LBFL and LBFH, the Gibbs sampling was run with 10,000 iterations, and then the first 6,000 iterations were discarded as burn-in.

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

$$\mathrm{MSE} = \frac{1}{100}\sum_{k=1}^{100}\left[(\hat{\theta}_0^{(k)})^2 + \left\{\hat{\boldsymbol{\theta}}^{(k)} - \boldsymbol{\theta}^*\right\}^{\mathrm{T}}\left\{\hat{\boldsymbol{\theta}}^{(k)} - \boldsymbol{\theta}^*\right\}\right],$$

where $\hat{\theta}_0^{(k)}$ and $\hat{\boldsymbol{\theta}}^{(k)} = (\hat{\theta}_1^{(k)}, \ldots, \hat{\theta}_p^{(k)})^{\mathrm{T}}$ are an intercept and a vector of regression coefficients estimated from the k-th dataset, respectively. To evaluate the prediction accuracy, we used the negative expected log-likelihood. We generated 1,000 test data, and then computed the mean of the empirical negative expected log-likelihood:

$$\mathrm{EL} = \frac{1}{100}\sum_{k=1}^{100}\left\{-\frac{1}{1000}\sum_{j=1}^{1000}\left[y_j^{\dagger(k)}(\boldsymbol{x}^{\dagger}{}_j^{(k)\mathrm{T}}\hat{\boldsymbol{\theta}}^{(k)} + \hat{\theta}_0^{(k)}) - \log\{1 + \exp(\hat{\boldsymbol{\theta}}^{(k)} + \hat{\theta}_0^{(k)})\}\right]\right\},$$

where $y_j^{\dagger(k)}$ and $\boldsymbol{x}^{\dagger}{}_j^{(k)}$ are the j-th data in the test data for the k-th dataset.

To assess the performance of variable selection and variable fusion for the Bayesian methods, we computed the following measures:

$$\mathrm{PV} = \frac{1}{100}\sum_{k=1}^{100}\frac{|\{j \mid (\theta_j^* \neq 0) \wedge (0 \notin \mathrm{CIV}_j(k))\}|}{|\{j \mid \theta_j^* \neq 0\}|} \quad (1 \leq j \leq p),$$

$$\mathrm{PZV} = \frac{1}{100}\sum_{k=1}^{100}\frac{|\{j \mid (\theta_j^* = 0) \wedge (0 \in \mathrm{CIV}_j(k))\}|}{|\{j \mid \theta_j^* = 0\}|} \quad (1 \leq j \leq p),$$

$$\mathrm{AV} = \frac{1}{100}\sum_{k=1}^{100}\frac{|\{j \mid (\theta_j^* \neq 0) \wedge (0 \notin \mathrm{CIV}_j(k))\}|+|\{j \mid (\theta_j^* = 0) \wedge (0 \in \mathrm{CIV}_j(k))\}|}{p} \quad (1 \leq j \leq p),$$

$$\mathrm{PF} = \frac{1}{100}\sum_{k=1}^{100}\frac{|\{j \mid (\theta_j^* - \theta_{j-1}^* \neq 0) \wedge (0 \notin \mathrm{CIF}_j(k))\}|}{|\{j \mid \theta_j^* - \theta_{j-1}^* \neq 0\}|} \quad (2 \leq j \leq p),$$

$$\mathrm{PNF} = \frac{1}{100}\sum_{k=1}^{100}\frac{|\{j \mid (\theta_j^* - \theta_{j-1}^* = 0) \wedge (0 \in \mathrm{CIF}_j(k))\}|}{|\{j \mid \theta_j^* - \theta_{j-1}^* = 0\}|} \quad (2 \leq j \leq p),$$

$$\mathrm{AF} = \frac{1}{100}\sum_{k=1}^{100}\left\{\frac{|\{j \mid (\theta_j^* - \theta_{j-1}^* \neq 0) \wedge (0 \notin \mathrm{CIF}_j(k))\}|}{p-1} + \frac{|\{j \mid (\theta_j^* - \theta_{j-1}^* = 0) \wedge (0 \in \mathrm{CIF}_j(k))\}|}{p-1}\right\} \quad (2 \leq j \leq p),$$

where $\mathrm{CIV}_j(k)$ denotes the 95% credible interval for $\hat{\theta}_j^{(k)}$, while $\mathrm{CIF}_j(k)$ denotes the 50% credible interval for the difference between $\hat{\theta}_j^{(k)}$ and $\hat{\theta}_{j-1}^{(k)}$. PV and PZV measure the accuracy of estimating regression coefficients when the corresponding true regression coefficients are non-zero and zero, respectively. PF and PNF measure the accuracy of the differences between estimated adjacent regression coefficients when the differences between the true adjacent regression coefficients are non-zero and zero, respectively. AV and AF measure the accuracy of variable selection and variable fusion, respectively. For the frequentist methods, the measures were computed by substituting $\hat{\theta}_j^{(k)} \neq 0$ for $0 \notin \mathrm{CIV}_j(k)$, $\hat{\theta}_j^{(k)} = 0$ for $0 \in \mathrm{CIV}_j(k)$, $\hat{\theta}_j^{(k)} - \hat{\theta}_{j-1}^{(k)} \neq 0$ for $0 \notin \mathrm{CIF}_j(k)$, and $\hat{\theta}_j^{(k)} - \hat{\theta}_{j-1}^{(k)} = 0$ for $0 \in \mathrm{CIF}_j(k)$.

Note that we utilized credible intervals of MCMC samples in evaluating the performance of variable selection and variable fusion. The reason why LBFL and LBFH do not estimate regression coefficients and their differences as exactly zero is that the posteriors of them are continuous. For variable selection, we computed 95% credible intervals as in Bhattacharyya et al. (2022), and then judged that the regression coefficient is estimated as zero when the credible interval of its MCMC samples includes zero. For variable fusion, we computed 50% credible intervals of the differences between adjacent regression coefficients as in Banerjee (2022), and

then judged that the difference is estimated as zero, similar to variable selection.

Table 3.1: MSE (standard deviation), EL, PV, PZV, AV, PF, PNF, and AF for Case 1 and $\rho = 0$. Bold font indicates the smallest value of MSE and EL and the largest value of PV, PZV, AV, PF, PNF, and AF before rounding among LN, LMCP, LSCAD, LFL, LBFL, and LBFH.

			MSE	EL	PV	PZV	AV	PF	PNF	AF
			(sd)	(sd)	(sd)	(sd)	(sd)	(sd)	(sd)	(sd)
$n = 500$	$\boldsymbol{\theta}_1^*$	LN	0.686	186.876	**1.000**	0.000	0.500	**1.000**	0.000	0.158
			(0.478)	(6.344)	**(0.000)**	(0.000)	(0.000)	**(0.000)**	(0.000)	(0.000)
		LMCP	0.359	180.272	**1.000**	0.914	0.957	**1.000**	0.415	0.507
			(0.278)	(4.286)	**(0.000)**	(0.103)	(0.052)	**(0.000)**	(0.097)	(0.082)
		LSCAD	0.361	180.229	**1.000**	0.838	0.919	**1.000**	0.349	0.452
			(0.294)	(4.395)	**(0.000)**	(0.147)	(0.073)	**(0.000)**	(0.127)	(0.107)
		LFL	0.619	180.551	**1.000**	0.501	0.751	**1.000**	0.766	0.803
			(0.287)	(3.032)	**(0.000)**	(0.288)	(0.144)	**(0.000)**	(0.113)	(0.095)
		LBFL	0.490	183.965	**1.000**	0.911	0.956	**1.000**	0.541	0.613
			(0.303)	(4.997)	**(0.000)**	(0.094)	(0.047)	**(0.000)**	(0.143)	(0.120)
		LBFH	**0.252**	**178.379**	**1.000**	**0.961**	**0.981**	**1.000**	**0.850**	**0.874**
			(0.167)	**(3.135)**	**(0.000)**	**(0.076)**	**(0.038)**	**(0.000)**	**(0.105)**	**(0.089)**
	$\boldsymbol{\theta}_2^*$	LN	1.930	127.062	**1.000**	0.000	0.750	**1.000**	0.000	0.158
			(1.953)	(5.513)	**(0.000)**	(0.000)	(0.000)	**(0.000)**	(0.000)	(0.000)
		LMCP	1.337	123.225	**1.000**	0.828	0.957	**1.000**	0.177	0.307
			(1.367)	(4.989)	**(0.000)**	(0.243)	(0.061)	**(0.000)**	(0.085)	(0.072)
		LSCAD	1.325	122.928	**1.000**	0.730	0.933	**1.000**	0.131	0.268
			(1.383)	(5.050)	**(0.000)**	(0.200)	(0.050)	**(0.000)**	(0.075)	(0.063)
		LFL	3.722	123.790	**1.000**	0.480	0.870	**1.000**	0.769	0.805
			(0.898)	(3.398)	**(0.000)**	(0.417)	(0.104)	**(0.000)**	(0.101)	(0.085)
		LBFL	0.905	122.723	**1.000**	0.944	0.986	**1.000**	0.541	0.613
			(0.494)	(3.846)	**(0.000)**	(0.107)	(0.027)	**(0.000)**	(0.135)	(0.114)
		LBFH	**0.809**	**117.895**	**1.000**	**0.972**	**0.993**	**1.000**	**0.868**	**0.888**
			(0.522)	**(3.084)**	**(0.000)**	**(0.103)**	**(0.026)**	**(0.000)**	**(0.083)**	**(0.069)**
$n = 1000$	$\boldsymbol{\theta}_1^*$	LN	0.268	359.573	**1.000**	0.000	0.500	**1.000**	0.000	0.158
			(0.112)	(4.005)	**(0.000)**	(0.000)	(0.000)	**(0.000)**	(0.000)	(0.000)
		LMCP	0.154	353.921	**1.000**	0.963	0.982	**1.000**	0.464	0.548
			(0.083)	(2.905)	**(0.000)**	(0.071)	(0.035)	**(0.000)**	(0.068)	(0.057)
		LSCAD	0.151	353.780	**1.000**	0.943	0.972	**1.000**	0.449	0.536
			(0.080)	(2.772)	**(0.000)**	(0.098)	(0.049)	**(0.000)**	(0.083)	(0.070)
		LFL	0.270	354.261	**1.000**	0.378	0.689	**1.000**	0.649	0.705
			(0.127)	(2.770)	**(0.000)**	(0.232)	(0.116)	**(0.000)**	(0.111)	(0.093)
		LBFL	0.224	357.885	**1.000**	0.926	0.963	**1.000**	0.508	0.586
			(0.087)	(3.572)	**(0.000)**	(0.072)	(0.036)	**(0.000)**	(0.139)	(0.117)
		LBFH	**0.121**	**352.405**	**1.000**	**0.968**	**0.984**	**1.000**	**0.843**	**0.868**
			(0.058)	**(2.455)**	**(0.000)**	**(0.066)**	**(0.033)**	**(0.000)**	**(0.121)**	**(0.102)**
	$\boldsymbol{\theta}_2^*$	LN	0.634	237.349	**1.000**	0.000	0.750	**1.000**	0.000	0.158
			(0.488)	(4.668)	**(0.000)**	(0.000)	(0.000)	**(0.000)**	(0.000)	(0.000)
		LMCP	0.473	233.374	**1.000**	0.900	0.975	**1.000**	0.208	0.333
			(0.424)	(3.813)	**(0.000)**	(0.200)	(0.050)	**(0.000)**	(0.075)	(0.064)
		LSCAD	0.469	233.262	**1.000**	0.840	0.960	**1.000**	0.181	0.311
			(0.416)	(3.810)	**(0.000)**	(0.216)	(0.054)	**(0.000)**	(0.081)	(0.068)
		LFL	1.642	234.649	**1.000**	0.424	0.856	**1.000**	0.716	0.761
			(0.600)	(3.378)	**(0.000)**	(0.358)	(0.089)	**(0.000)**	(0.113)	(0.095)
		LBFL	0.429	235.016	**1.000**	0.920	0.980	**1.000**	0.526	0.601
			(0.265)	(3.875)	**(0.000)**	(0.130)	(0.033)	**(0.000)**	(0.122)	(0.103)
		LBFH	**0.311**	**229.710**	**1.000**	**0.974**	**0.994**	**1.000**	**0.859**	**0.882**
			(0.226)	**(2.799)**	**(0.000)**	**(0.112)**	**(0.028)**	**(0.000)**	**(0.090)**	**(0.076)**

Table 3.2: MSE (standard deviation), EL, PV, PZV, AV, PF, PNF, and AF for Case 1 and $\rho = 0.5$. Bold font indicates the smallest value of MSE and EL and the largest value of PV, PZV, AV, PF, PNF, and AF before rounding among LN, LMCP, LSCAD, LFL, LBFL, and LBFH.

			MSE	EL	PV	PZV	AV	PF	PNF	AF
			(sd)	(sd)	(sd)	(sd)	(sd)	(sd)	(sd)	(sd)
$n = 500$	$\boldsymbol{\theta}_1^*$	LN	3.749	104.036	**1.000**	0.000	0.500	**1.000**	0.000	0.158
			(2.638)	(9.752)	**(0.000)**	(0.000)	(0.000)	**(0.000)**	(0.000)	(0.000)
		LMCP	2.013	97.558	0.985	0.763	0.874	0.990	0.303	0.411
			(1.244)	(5.991)	(0.041)	(0.196)	(0.100)	(0.057)	(0.136)	(0.114)
		LSCAD	2.069	97.968	0.993	0.641	0.817	0.997	0.220	0.343
			(1.234)	(6.202)	(0.029)	(0.217)	(0.109)	(0.033)	(0.129)	(0.108)
		LFL	1.735	94.449	**1.000**	0.236	0.618	0.940	**0.901**	**0.907**
			(0.634)	(4.134)	**(0.000)**	(0.261)	(0.130)	(0.137)	**(0.066)**	**(0.065)**
		LBFL	1.716	96.289	0.979	**0.920**	0.950	0.967	0.581	0.642
			(0.870)	(5.362)	(0.048)	**(0.090)**	(0.051)	(0.101)	(0.145)	(0.126)
		LBFH	**1.062**	**92.329**	0.995	0.905	**0.950**	0.970	0.871	0.887
			(0.480)	**(3.813)**	(0.022)	(0.134)	**(0.067)**	(0.096)	(0.089)	(0.078)
	$\boldsymbol{\theta}_2^*$	LN	5.357	96.020	**1.000**	0.000	0.750	**1.000**	0.000	0.158
			(4.293)	(7.091)	**(0.000)**	(0.000)	(0.000)	**(0.000)**	(0.000)	(0.000)
		LMCP	4.183	94.090	0.984	0.612	0.891	**1.000**	0.112	0.252
			(3.034)	(7.304)	(0.040)	(0.328)	(0.078)	**(0.000)**	(0.097)	(0.081)
		LSCAD	4.203	94.057	0.993	0.518	0.874	**1.000**	0.077	0.223
			(3.056)	(7.040)	(0.025)	(0.286)	(0.070)	**(0.000)**	(0.077)	(0.065)
		LFL	6.180	91.536	**1.000**	0.482	0.871	0.980	0.847	0.868
			(1.205)	(3.603)	**(0.000)**	(0.447)	(0.112)	(0.080)	(0.097)	(0.083)
		LBFL	1.742	88.395	0.977	0.930	0.966	0.977	0.598	0.657
			(0.791)	(3.967)	(0.037)	(0.104)	(0.039)	(0.085)	(0.125)	(0.107)
		LBFH	**1.520**	**84.530**	0.997	**0.966**	**0.989**	0.970	**0.881**	**0.895**
			(0.914)	**(3.086)**	(0.015)	**(0.110)**	**(0.029)**	(0.107)	**(0.091)**	**(0.079)**
$n = 1000$	$\boldsymbol{\theta}_1^*$	LN	1.134	184.764	**1.000**	0.000	0.500	**1.000**	0.000	0.158
			(0.511)	(5.685)	**(0.000)**	(0.000)	(0.000)	**(0.000)**	(0.000)	(0.000)
		LMCP	0.707	179.290	**0.999**	0.821	0.910	**1.000**	0.349	0.452
			(0.400)	(4.278)	**(0.010)**	(0.175)	(0.087)	**(0.000)**	(0.134)	(0.113)
		LSCAD	0.700	179.165	**1.000**	0.707	0.854	**1.000**	0.257	0.374
			(0.407)	(4.358)	**(0.000)**	(0.169)	(0.084)	**(0.000)**	(0.119)	(0.101)
		LFL	0.736	178.005	**1.000**	0.275	0.638	**1.000**	0.844	0.869
			(0.295)	(3.728)	**(0.000)**	(0.252)	(0.126)	**(0.000)**	(0.080)	(0.067)
		LBFL	0.779	181.135	0.999	0.930	0.965	**1.000**	0.559	0.628
			(0.323)	(4.651)	(0.010)	(0.081)	(0.040)	**(0.000)**	(0.124)	(0.104)
		LBFH	**0.459**	**176.787**	**1.000**	**0.958**	**0.979**	**1.000**	**0.859**	**0.881**
			(0.245)	**(3.995)**	**(0.000)**	**(0.077)**	**(0.038)**	**(0.000)**	**(0.085)**	**(0.072)**
	$\boldsymbol{\theta}_2^*$	LN	1.676	172.321	**1.000**	0.000	0.750	**1.000**	0.000	0.158
			(1.229)	(5.502)	**(0.000)**	(0.000)	(0.000)	**(0.000)**	(0.000)	(0.000)
		LMCP	1.372	168.946	0.999	0.784	0.945	**1.000**	0.162	0.294
			(1.095)	(5.437)	(0.009)	(0.258)	(0.064)	**(0.000)**	(0.086)	(0.073)
		LSCAD	1.378	168.971	0.999	0.668	0.917	**1.000**	0.122	0.261
			(1.114)	(5.469)	(0.007)	(0.262)	(0.065)	**(0.000)**	(0.081)	(0.068)
		LFL	2.943	169.180	**1.000**	0.442	0.861	0.997	0.808	0.838
			(0.834)	(3.830)	**(0.000)**	(0.421)	(0.105)	(0.033)	(0.085)	(0.072)
		LBFL	0.993	168.401	0.999	0.930	0.982	0.997	0.551	0.621
			(0.475)	(4.111)	(0.007)	(0.118)	(0.030)	(0.033)	(0.122)	(0.103)
		LBFH	**0.745**	**163.983**	**1.000**	**0.974**	**0.994**	0.997	**0.843**	**0.867**
			(0.461)	**(3.280)**	**(0.000)**	**(0.088)**	**(0.022)**	(0.033)	**(0.094)**	**(0.080)**

Table 3.3: MSE (standard deviation), EL, PV, PZV, AV, PF, PNF, and AF for Case 2. Bold font indicates the smallest value of MSE and EL and the largest value of PV, PZV, AV, PF, PNF, and AF before rounding among LN, LMCP, LSCAD, LFL, LBFL, and LBFH.

		MSE	EL	PV	PZV	AV	PF	PNF	AF
		(sd)	(sd)	(sd)	(sd)	(sd)	(sd)	(sd)	(sd)
$\boldsymbol{\theta}_1^*$	LN	1.719	126.161	**1.000**	0.000	0.500	**1.000**	0.000	0.158
		(0.913)	(5.458)	**(0.000)**	(0.000)	(0.000)	**(0.000)**	(0.000)	(0.000)
	LMCP	1.143	121.746	0.999	0.798	0.899	0.997	0.319	0.426
		(0.724)	(5.060)	(0.010)	(0.158)	(0.080)	(0.033)	(0.127)	(0.107)
	LSCAD	1.116	121.528	**1.000**	0.716	0.858	**1.000**	0.256	0.373
		(0.687)	(4.926)	**(0.000)**	(0.151)	(0.075)	**(0.000)**	(0.119)	(0.100)
	LFL	0.718	118.113	**1.000**	0.481	0.741	**1.000**	**0.898**	**0.914**
		(0.297)	(2.373)	**(0.000)**	(0.306)	(0.153)	**(0.000)**	**(0.083)**	**(0.070)**
	LBFL	0.949	121.116	0.999	0.937	0.968	0.993	0.581	0.646
		(0.435)	(3.396)	(0.010)	(0.087)	(0.045)	(0.047)	(0.148)	(0.125)
	LBFH	**0.381**	**116.649**	**1.000**	**0.985**	**0.993**	**1.000**	0.888	0.905
		(0.194)	**(2.220)**	**(0.000)**	**(0.046)**	**(0.023)**	**(0.000)**	(0.094)	(0.079)
$\boldsymbol{\theta}_2^*$	LN	8.141	90.274	**1.000**	0.000	0.750	**1.000**	0.000	0.158
		(10.931)	(12.539)	**(0.000)**	(0.000)	(0.000)	**(0.000)**	(0.000)	(0.000)
	LMCP	5.611	87.099	0.967	0.686	0.897	0.990	0.134	0.269
		(6.057)	(9.831)	(0.046)	(0.292)	(0.067)	(0.057)	(0.085)	(0.071)
	LSCAD	5.424	87.142	0.977	0.656	0.897	0.993	0.114	0.253
		(5.943)	(9.862)	(0.039)	(0.248)	(0.061)	(0.047)	(0.072)	(0.060)
	LFL	5.350	80.432	**1.000**	0.364	0.841	**1.000**	0.846	0.871
		(0.790)	(2.902)	**(0.000)**	(0.424)	(0.106)	**(0.000)**	(0.078)	(0.066)
	LBFL	1.860	79.103	0.982	0.942	0.972	0.980	0.561	0.627
		(0.894)	(4.442)	(0.031)	(0.118)	(0.040)	(0.080)	(0.138)	(0.117)
	LBFH	**1.375**	**74.886**	0.994	**0.990**	**0.993**	0.987	**0.871**	**0.889**
		(0.730)	**(3.396)**	(0.019)	**(0.044)**	**(0.019)**	(0.066)	**(0.094)**	**(0.082)**

Table 3.4: MSE (standard deviation), EL, PV, PZV, AV, PF, PNF, and AF for Case 3. Bold font indicates the smallest value of MSE and EL and the largest value of PV, PZV, AV, PF, PNF, and AF before rounding among LN, LMCP, LSCAD, LFL, LBFL, and LBFH.

		MSE (sd)	EL (sd)	PV (sd)	PZV (sd)	AV (sd)	PF (sd)	PNF (sd)	AF (sd)
$\boldsymbol{\theta}_1^*$	LN	1.507	142.051	**1.000**	0.000	0.500	**1.000**	0.000	0.158
		(0.928)	(6.508)	**(0.000)**	(0.000)	(0.000)	**(0.000)**	(0.000)	(0.000)
	LMCP	0.960	136.540	0.996	0.785	0.891	**1.000**	0.319	0.427
		(0.764)	(5.918)	(0.024)	(0.215)	(0.106)	**(0.000)**	(0.142)	(0.120)
	LSCAD	0.929	136.312	0.998	0.685	0.842	**1.000**	0.242	0.362
		(0.707)	(5.620)	(0.014)	(0.214)	(0.106)	**(0.000)**	(0.136)	(0.115)
	LFL	0.690	134.037	**1.000**	0.465	0.733	**1.000**	0.846	0.870
		(0.243)	(2.457)	**(0.000)**	(0.254)	(0.127)	**(0.000)**	(0.089)	(0.075)
	LBFL	0.815	137.327	0.999	0.925	0.962	**1.000**	0.564	0.633
		(0.458)	(4.360)	(0.010)	(0.093)	(0.046)	**(0.000)**	(0.147)	(0.123)
	LBFH	**0.297**	**132.313**	**1.000**	**0.982**	**0.991**	**1.000**	**0.883**	**0.902**
		(0.167)	**(2.760)**	**(0.000)**	**(0.046)**	**(0.023)**	**(0.000)**	**(0.086)**	**(0.072)**
$\boldsymbol{\theta}_2^*$	LN	6.184	97.942	**1.000**	0.000	0.750	**1.000**	0.000	0.158
		(6.682)	(9.432)	**(0.000)**	(0.000)	(0.000)	**(0.000)**	(0.000)	(0.000)
	LMCP	4.645	94.429	0.985	0.660	0.904	0.993	0.126	0.263
		(4.403)	(7.976)	(0.039)	(0.321)	(0.075)	(0.047)	(0.095)	(0.080)
	LSCAD	4.577	94.589	0.992	0.574	0.888	0.997	0.096	0.238
		(4.454)	(8.478)	(0.030)	(0.287)	(0.068)	(0.033)	(0.081)	(0.068)
	LFL	4.681	89.716	**1.000**	0.440	0.860	**1.000**	0.807	0.837
		(0.973)	(3.554)	**(0.000)**	(0.405)	(0.101)	**(0.000)**	(0.094)	(0.080)
	LBFL	1.531	88.550	0.996	0.940	0.982	0.993	0.626	0.684
		(0.923)	(4.763)	(0.019)	(0.104)	(0.029)	(0.047)	(0.149)	(0.127)
	LBFH	**1.030**	**84.256**	0.999	**0.980**	**0.994**	**1.000**	**0.904**	**0.919**
		(0.657)	**(3.792)**	(0.009)	**(0.060)**	**(0.016)**	**(0.000)**	**(0.077)**	**(0.065)**

Table 3.5: MSE (standard deviation), EL, PV, PZV, AV, PF, PNF, and AF for Case 4. Bold font indicates the smallest value of MSE and EL and the largest value of PV, PZV, AV, PF, PNF, and AF before rounding among LMCP, LSCAD, LFL, LBFL, and LBFH.

	MSE (sd)	EL (sd)	PV (sd)	PZV (sd)	AV (sd)	PF (sd)	PNF (sd)	AF (sd)
LMCP	75.855	184.816	0.363	0.973	0.882	0.475	0.873	0.869
	(3.689)	(10.424)	(0.105)	(0.012)	(0.014)	(0.260)	(0.036)	(0.035)
LSCAD	74.618	177.210	0.536	0.927	0.868	0.668	0.770	0.769
	(2.871)	(9.031)	(0.074)	(0.017)	(0.014)	(0.216)	(0.033)	(0.032)
LFL	66.735	124.512	**0.998**	0.947	0.954	**0.890**	0.989	0.988
	(0.984)	(3.965)	**(0.008)**	(0.061)	(0.051)	**(0.144)**	(0.005)	(0.005)
LBFL	**42.765**	115.630	0.578	0.999	0.936	0.840	0.924	0.923
	(1.639)	(11.928)	(0.068)	(0.001)	(0.010)	(0.157)	(0.013)	(0.013)
LBFH	60.976	**109.058**	0.857	**1.000**	**0.978**	0.563	**0.999**	**0.994**
	(1.855)	**(10.577)**	(0.151)	**(0.0004)**	**(0.023)**	(0.231)	**(0.002)**	**(0.004)**

The results are summarized in Tables 3.1, 3.2, 3.3, 3.4, and 3.5. LN always gives the largest MSEs and ELs. LBFH gives the smallest ELs in all cases. LBFH also gives the smallest MSEs in almost all cases except for Case 4. These results show that LBFH outperformed the other methods in terms of the estimation and prediction accuracy. In addition, LBFH achieves the largest AVs and AFs in most cases. This shows that LBFH provides superior performance of variable selection and variable fusion compared to the other methods. The main cause of these performance of LBFH is assuming a horseshoe prior, which introduces less bias than the Laplace prior, on the differences between regression coefficients. By comparing LBFL with LFL, LBFL gives the smaller MSEs in many cases. LFL often gives larger PVs than LBFL, but LBFL gives larger PZVs and AVs in almost all cases. These results show that LBFL performs better than LFL in terms of the accuracy of estimation and variable selection. For the performance of variable fusion, PFs, PNFs, and AFs of LFL are competitive or larger than those of LBFL. By comparing LBFL with LMCP and LSCAD, LBFL gives smaller MSEs in most cases and smaller ELs in more than half of the cases. For PVs and PFs, the performance of the three methods are comparable. LBFL usually gives larger PZVs, PNFs, AVs, and AFs than LMCP and LSCAD. LBFL provides the better performance than LMCP and LSCAD with respect to various measures. As opposed to LMCP and LSCAD, LBFL has a variable fusion term and can

give more accurate estimation for regression coefficients with a group structure.

In addition, we also measured the computational time required for Gibbs sampling in the proposed methods LBFL and LBFH, based on CPU time. The results are summarized in Tables 3.6, 3.7, and 3.8. The values of the tables are the average time of 100 runs. From Table 3.6, we can see that the computational time increases substantially as the sample size grows. For Cases 1–3, the computational times of LBFL and LBFH are comparable across different correlation structures of explanatory variables and values of true regression coefficients. In Case 4, which corresponds to a high-dimensional setting with $n < p$, LBFH required approximately 30 seconds less computational time than LBFL.

Table 3.6: Computational times (sec.) for estimation of regression coefficients for Case 1. Figures in parentheses give the estimated standard deviation.

			$\rho = 0$	$\rho = 0.5$
			Time	Time
n=500	$\boldsymbol{\theta}_1^*$	LBFL	49.620	51.310
			(6.305)	(3.432)
		LBFH	47.690	48.246
			(5.117)	(4.674)
	$\boldsymbol{\theta}_2^*$	LBFL	40.136	40.361
			(1.968)	(1.695)
		LBFH	48.877	48.491
			(3.758)	(4.657)
n=1000	$\boldsymbol{\theta}_1^*$	LBFL	159.112	159.017
			(13.311)	(12.755)
		LBFH	157.801	158.540
			(14.286)	(13.698)
	$\boldsymbol{\theta}_2^*$	LBFL	158.884	158.646
			(13.484)	(13.131)
		LBFH	157.571	157.021
			(13.696)	(16.178)

Table 3.7: Computational times (sec.) for estimation of regression coefficients for Case 2 and 3. Figures in parentheses give the estimated standard deviation.

		Case 2	Case 3
		Time	Time
$\boldsymbol{\theta}_1^*$	LBFL	51.499	51.319
		(2.177)	(2.556)
	LBFH	48.276	48.524
		(4.370)	(4.170)
$\boldsymbol{\theta}_2^*$	LBFL	40.279	40.215
		(2.050)	(2.340)
	LBFH	48.100	48.034
		(5.032)	(5.359)

Table 3.8: Computational times (sec.) for estimation of regression coefficients for Case 4. Figures in parentheses give the estimated standard deviation.

	Time
LBFL	184.579
	(10.431)
LBFH	152.794
	(0.976)

3.6 Application

We applied our proposed methods LBFL and LBFH to the Wafer dataset, which was formatted in Olszewski (2001) and analyzed in Deng et al. (2014). The dataset can be obtained from the UCR Time Series Classification Archive (https://www.cs.ucr.edu/%7Eeamonn/time_series_data_2018). By utilizing the dataset, the normal and abnormal etching processes of a wafer in semiconductor microelectronics manufacturing were classified based on time series data from six sensors (which monitor radio frequency forward power, radio frequency reflected power, chamber pressure, 405 nanometer emission, 520 nanometer emission, and direct current bias, respectively). Each time series data contains the value from one of the six sensors for one wafer and has length $p = 152$. We labeled abnormal data as one and normal data as zero. The dataset contains

$n = 1{,}000$ training data and 6,164 test data. The abnormal data constitute 9.7% of the training data and 10.8% of the test data, meaning that the dataset has a large class imbalance.

We compared LBFL and LBFH to LFL. As in Deng et al. (2014), we selected the hyper-parameters λ_1 and λ_2 for LFL from four candidates, 0.05, 0.1, 0.3, and 0.5, by BIC. For LBFL and LBFH, the Gibbs sampler was run with 10,000 iterations, and then the first 6,000 samples were discarded as burn-in.

We evaluated the performance of LBFL, LBFH, and LFL by Area Under the ROC Curve (AUC). AUCs tend to be large in situations where the true positive rate is large when the false positive rate is small. The values of AUC are summarized in Table 3.9. From Table 3.9, LBFL gives the largest AUC. Meanwhile, AUC is not enough to evaluate the performance of the model when there is a large class imbalance in the dataset. In the case of such a large class imbalance, Area Under the Precision-Recall Curve (PR-AUC) is a more suitable indicator. PR-AUCs tend to be large in situations where the true positive rate is large when the precision is large. For the details of PR-AUC, we refer the reader to Davis and Goadrich (2006). The values of PR-AUC are also summarized in Table 3.9, which shows that LBFH gives the largest PR-AUC.

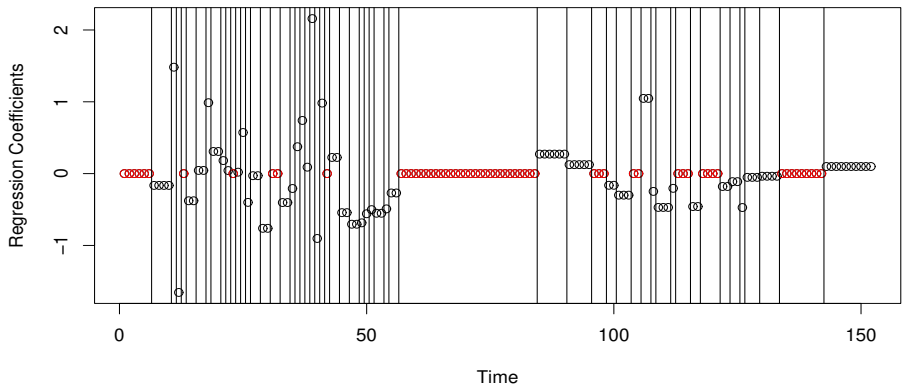
As with the Monte Carlo simulations in Sec. 3.5, we judged that the regression coefficient is estimated as zero if the corresponding 95% credible interval includes zero for LBFL and LBFH. We also judged that the difference between adjacent regression coefficients is regarded as non-zero when the 50% credible interval for the difference does not include zero. The estimated regression coefficients are shown in Figure 3.1. From Figure 3.1, we observe that LFL estimated 60 regression coefficients as zero, whereas LBFL estimated the most as zero, at 133, and LBFH estimated the second most, at 127. Thus, the number of points which are considered to be unnecessary for the model is larger for LBFL and LBFH than for LFL, meaning that the former two make clearer which points are necessary for the prediction.

From Figure 3.1(c), we see that LBFH split the regression coefficients into four groups. The first group contains the 1st to 27th coefficients, the second contains the 28th to 35th coefficients, the third one the 36th to 44th, and the fourth one the 45th to 152nd. On the other hand, from Figure 3.1(b), we see that LBFL split the regression coefficients into 21 groups, whereas from Figure 3.1(a), we see that LFL split the coefficients into 57 groups. Thus, LBFH gave the

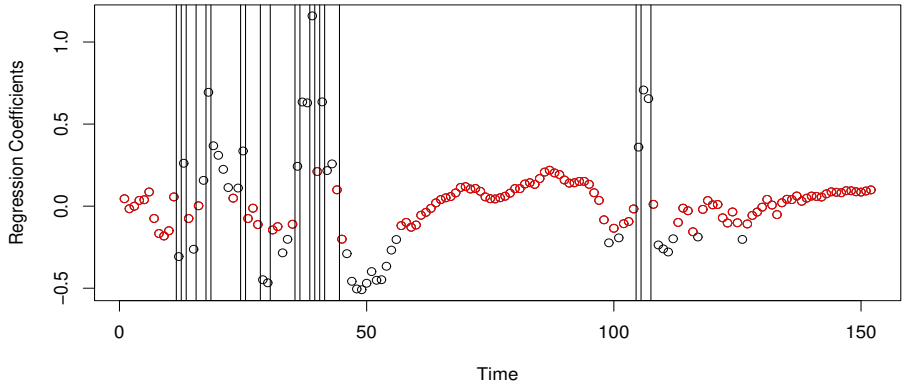
smallest number of groups of variables, whereas LBFL gave the second smallest. Based on the number of detected groups, LBFH seems to provide smoother and more simplified estimation than the other methods. Many of the groups detected by LFL contain only one variable, meaning that LFL has weak performance regarding grouping multiple variables.

Table 3.9: AUC and PR-AUC for the Wafer dataset. Bold font indicates the largest value among LFL, LBFL, and LBFH.

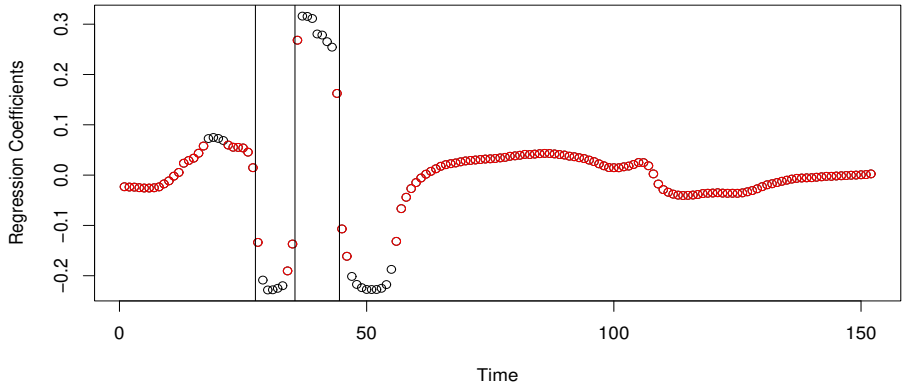
	AUC	PR-AUC
LFL	0.880	0.615
LBFL	**0.886**	0.592
LBFH	0.864	**0.626**



(a) LFL



(b) LBFL



(c) LBFH

Figure 3.1: Estimated regression coefficients for LFL, LBFL, and LBFH. Dots indicate the values of the estimated regression coefficients. For LFL, a red dot indicates a zero element of the estimated regression coefficients and a black vertical line indicates where a non-zero difference between adjacent estimated regression coefficients exists. For LBFL and LBFH, a red dot indicates an estimated regression coefficient whose MCMC samples give a 95% credible interval including zero and a black vertical line indicates where a difference between adjacent regression coefficients whose MCMC samples give a 50% credible interval not including zero exists.

Chapter 4

Information criterion for Bayesian generalized fused lasso

In this chapter, we develop an information criterion for the Bayesian generalized fused lasso. For clarity, the notation used in this chapter is newly defined. We first specify the model structure, beginning with a spatially varying coefficients (SVC) model, and then formulate the generalized fused lasso in the model. Next, we derive asymptotic properties of the generalized fused lasso estimators and construct an information criterion based on the Bayesian predictive distribution. Finally, the performance of the proposed method is examined through numerical experiments and a real data analysis.

4.1 Model specification and derivation of estimator properties

In this section, we consider as SVC models, in a broad sense, those probability models in which regression coefficients vary smoothly across regions to which the data belong, and present them as representative models used in spatial data analysis. Then, as its estimation method, we take up the generalized fused lasso regularization method by Hoefling (2010). With the construction of a model selection criterion in mind, we derive asymptotic properties of estimators under a setting different from that of Viallon et al. (2016). In addition, we also provide a Bayesian interpretation of the generalized fused lasso in SVC models.

4.1.1 Spatially varying coefficients model

For the i $(\in 1, \ldots, n)$-th sample, suppose that the response variable y_i $(\in \mathbb{R})$, the explanatory variable vector $\tilde{\boldsymbol{x}}_i = (\tilde{x}_{i,1}, \ldots, \tilde{x}_{i,\tilde{p}})^{\mathrm{T}}$ $(\in \mathbb{R}^{\tilde{p}})$, and the indicator variable ψ_i $(\in 1, \ldots, M)$ representing which region the sample is associated with

are observed, and consider the following SVC model:

$$y_i = \sum_{m=1}^{M} I(\psi_i = m)\tilde{\boldsymbol{x}}_i^{\mathrm{T}}\boldsymbol{\theta}_m + \varepsilon_i, \tag{4.1}$$

where $\boldsymbol{\theta}_m = (\theta_{m,1}, \ldots, \theta_{m,\tilde{p}})^{\mathrm{T}}$ $(\in \mathbb{R}^{\tilde{p}})$ is the regression coefficient vector for the m-th region. For simplicity, suppose that $\{(\psi_i, \tilde{\boldsymbol{x}}_i, \varepsilon_i) \mid i \in \{1, \ldots, n\}\}$ are independent and identically distributed, and that the distribution of $(\psi_i, \tilde{\boldsymbol{x}}_i)$ is known. If the explanatory variables need to be non-random, it suffices to impose conditions on them as in ordinary regression analysis (for example, in the case of the framework for constructing an information criterion for sparse estimation, see Ninomiya and Kawano 2016). Let the error term ε_i follow $\mathrm{N}(0, \sigma^2)$ independently of $(\psi_i, \tilde{\boldsymbol{x}}_i)$. The error variance σ^2 is assumed to be unknown, but since deriving the model selection criterion, which is the main theme of this paper, under the assumption that σ^2 is unknown causes unnecessary complications, we derive the criterion assuming the error variance is known and then later substitute in an estimator. Of course, if the variance structure is also made spatio-temporal, the situation becomes different, but we do not deal with that in this paper. In this setting, letting $\boldsymbol{e}_m$ be the M-dimensional unit vector whose m-th component is 1, and defining $\boldsymbol{x}_i$ as $\boldsymbol{e}_{\psi_i} \otimes \tilde{\boldsymbol{x}}_i$ and $\boldsymbol{\theta}$ as $(\boldsymbol{\theta}_1^{\mathrm{T}}, \ldots, \boldsymbol{\theta}_M^{\mathrm{T}})^{\mathrm{T}}$, we express (4.1) simply as $y_i = \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\theta} + \varepsilon_i$. Here, $\{\boldsymbol{x}_i \mid i \in \{1, \ldots, n\}\}$ are independent and identically distributed.

In this model, we write the joint probability density function of $(y_i, \boldsymbol{x}_i)$, $(2\pi\sigma^2)^{-1/2}\exp\{-(y_i - \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\theta})^2/(2\sigma^2)\}$, as $f(y_i, \boldsymbol{x}_i \mid \boldsymbol{\theta})$. In addition, as conditions for constructing the asymptotic theory of sparse estimation for this model, we assume the following.

(C1) The parameter space Θ of $\boldsymbol{\theta}$ is compact, and $\boldsymbol{x}_i$ has moments up to the fourth order. In particular, $\boldsymbol{J} \equiv \mathrm{E}[\boldsymbol{x}_i\boldsymbol{x}_i^{\mathrm{T}}]$ is a positive definite matrix.

4.1.2 Generalized fused lasso

In the SVC model, regression coefficients corresponding to neighboring regions are set to similar values. The generalized fused lasso (Hoefling 2010) is a method that efficiently performs estimation in which neighboring regression coefficients tend to have the same value. Here, when region $m^{\dagger}$ and $m^{\ddagger}$ are adjacent, we designate $(m^{\dagger}, m^{\ddagger})$ as an edge, and denote the set of all edges by $\mathcal{E}$. When

defining the graph $(\mathcal{V}, \mathcal{E})$ with vertex set $\mathcal{V} = \{1, \ldots, M\}$, the generalized fused lasso is expressed as the following optimization problem:

$$\begin{aligned}\hat{\boldsymbol{\theta}} &= (\hat{\theta}_{1,1}, \ldots, \hat{\theta}_{1,\tilde{p}}, \hat{\theta}_{2,1}, \ldots, \hat{\theta}_{2,\tilde{p}}, \ldots, \hat{\theta}_{M,1}, \ldots, \hat{\theta}_{M,\tilde{p}}) \\ &\equiv \underset{\boldsymbol{\theta} \in \Theta}{\operatorname{argmin}} \left\{ -\sum_{i=1}^{n} \log f(y_i, \boldsymbol{x}_i \mid \boldsymbol{\theta}) + n \sum_{j=1}^{\tilde{p}} \lambda_{1,j} \sum_{m \in \mathcal{V}} |\theta_{m,j}| \right. \\ &\qquad \left. + n \sum_{j=1}^{\tilde{p}} \lambda_{2,j} \sum_{(m^\dagger, m^\ddagger) \in \mathcal{E}} |\theta_{m^\dagger,j} - \theta_{m^\ddagger,j}| \right\}, \end{aligned} \tag{4.2}$$

where $\lambda_{1,j}$ ($j \in \{1, \ldots, \tilde{p}\}$) and $\lambda_{2,j}$ ($j \in \{1, \ldots, \tilde{p}\}$) are regularization parameters. This optimization problem includes a penalty term $\lambda_{2,j}|\theta_{m^\dagger,j} - \theta_{m^\ddagger,j}|$ such that the difference between the regression coefficients $\theta_{m^\dagger,j}$ and $\theta_{m^\ddagger,j}$ corresponding to neighboring regions $(m^\dagger, m^\ddagger) \in \mathcal{E}$ is likely to be estimated as 0. It also includes a penalty term $\lambda_{1,j}|\theta_{m,j}|$ such that all regression coefficients $\theta_{m,j}$ are likely to be estimated as 0, and variable selection is performed simultaneously.

Usually, the regularization term of the generalized fused lasso is given without an n. However, for example, in constructing an asymptotic theory for providing an information criterion, it is unreasonable to assume an asymptotic setting in which the influence of the regularization term is small so that selection consistency is derived (see Ninomiya and Kawano 2016). Therefore, a regularization term with order $\mathrm{O}(n)$ is considered. As a preparatory result for the later lemma and theorems, the following are derived from (C1).

(R1) For each $\boldsymbol{\theta}$, there exists a convex and differentiable function $h(\boldsymbol{\theta})$ such that $-\sum_{i=1}^{n} \log f(y_i, \boldsymbol{x}_i \mid \boldsymbol{\theta})/n \xrightarrow{\mathrm{p}} h(\boldsymbol{\theta})$.

(R2) For $h(\boldsymbol{\theta})$ given in (R1), $\sum_{i=1}^{n} \mathrm{E}[-\partial \log f(y_i, \boldsymbol{x}_i \mid \boldsymbol{\theta})/\partial\boldsymbol{\theta}]/n - \partial h(\boldsymbol{\theta})/\partial\boldsymbol{\theta} = \mathbf{0}$.

(R3) $\sum_{i=1}^{n} (\partial \log f(y_i, \boldsymbol{x}_i \mid \boldsymbol{\theta})/\partial\boldsymbol{\theta} - \mathrm{E}[\partial \log f(y_i, \boldsymbol{x}_i \mid \boldsymbol{\theta})/\partial\boldsymbol{\theta}])/\sqrt{n} \xrightarrow{\mathrm{d}} \mathrm{N}(\mathbf{0}, \boldsymbol{J})$.

The proofs are easy; they can be obtained, for example, by simplifying the proofs in Ninomiya and Kawano (2016).

Next, we investigate the asymptotic properties of the generalized fused lasso estimator $\hat{\boldsymbol{\theta}}$. For simplicity of the following discussion, we set $\lambda_{1,1} = \ldots = \lambda_{1,\tilde{p}} = \lambda_1$ and $\lambda_{2,1} = \ldots = \lambda_{2,\tilde{p}} = \lambda_2$. To avoid notational complexity, we newly denote $(\theta_{1,1}, \ldots, \theta_{1,\tilde{p}}, \theta_{2,1}, \ldots, \theta_{2,\tilde{p}}, \ldots, \theta_{M,1}, \ldots, \theta_{M,\tilde{p}})$ as $(\xi_1, \ldots, \xi_p)$, and accordingly change the notation of the parameter space from Θ to Ξ. In addition, when $j^\dagger > j^\ddagger$ and $\xi_{j^\dagger}$ and $\xi_{j^\ddagger}$ are coefficient parameters corresponding to the same

variable and the corresponding regions are adjacent, that is, when $|\xi_{j^\dagger} - \xi_{j^\ddagger}|$ appears in the penalty term in (4.2), the pair of indices $(j^\dagger, j^\ddagger)$ is regarded as an edge, and we denote the set of all edges by E. When defining graph (V, E) with vertex set $V = \{1, \ldots, p\}$, (4.2) is rewritten as

$$\begin{aligned}\hat{\boldsymbol{\xi}} &= (\hat{\xi}_1, \ldots, \hat{\xi}_p) \\ &\equiv \operatorname*{argmin}_{\boldsymbol{\xi} \in \Xi} \left\{ -\sum_{i=1}^{n} \log f(y_i, \boldsymbol{x}_i \mid \boldsymbol{\xi}) + n\lambda_1 \sum_{j \in V} |\xi_j| + n\lambda_2 \sum_{(j^\dagger, j^\ddagger) \in E} |\xi_{j^\dagger} - \xi_{j^\ddagger}| \right\}.\end{aligned} \tag{4.3}$$

Since this estimator depends on $\boldsymbol{\lambda} = (\lambda_1, \lambda_2)$, we could write it as $\hat{\boldsymbol{\xi}}_{\boldsymbol{\lambda}}$, but for simplicity, we will write $\hat{\boldsymbol{\xi}}$ until it becomes essential to emphasize the dependence on $\boldsymbol{\lambda}$. Regarding the first-order asymptotics, the following lemma is easily obtained, where $\boldsymbol{\xi}^* = (\xi_1^*, \ldots, \xi_p^*)$ is the minimizer of $-\mathrm{E}[\log f(y, \boldsymbol{x} \mid \boldsymbol{\xi})] + \lambda_1 \sum_{j \in V} |\xi_j| + n\lambda_2 \sum_{(j^\dagger, j^\ddagger) \in E} |\xi_{j^\dagger} - \xi_{j^\ddagger}|$ with respect to $\boldsymbol{\xi}$.

Lemma 1. Under condition (C1), the generalized fused lasso estimator $\hat{\boldsymbol{\xi}}$ converges in probability to $\boldsymbol{\xi}^*$.

Proof. To develop the first-order asymptotics for the generalized fused lasso estimator $\hat{\boldsymbol{\xi}}$, we define the following random function:

$$u_n(\boldsymbol{\xi}) \equiv -\frac{1}{n} \sum_{i=1}^{n} \log f(y_i, \boldsymbol{x}_i \mid \boldsymbol{\xi}) + \lambda_1 \sum_{j \in V} |\xi_j| + \lambda_2 \sum_{(j,k) \in E} |\xi_j - \xi_k|.$$

It is obvious that $\hat{\boldsymbol{\xi}} = \operatorname{argmin}_{\boldsymbol{\xi} \in \Xi} u_n(\boldsymbol{\xi})$, and $u_n(\boldsymbol{\xi})$ converges in probability for each $\boldsymbol{\xi}$ to $h(\boldsymbol{\xi}) + \lambda_1 \sum_{j \in V} |\xi_j| + \lambda_2 \sum_{(j^\dagger, j^\ddagger) \in E} |\xi_{j^\dagger} - \xi_{j^\ddagger}|$ from (R1). Since $u_n(\boldsymbol{\xi})$ is convex with respect to $\boldsymbol{\xi}$, the convexity lemma of Andersen and Gill (1982) or Pollard (1991) can be applied, and the conclusion is obtained. □

Next, we consider the second-order asymptotics. Let $\mathcal{J}^{(1)} = \{j : \xi_j^* = 0\}$ and $\mathcal{J}^{(2)} = \{j : \xi_j^* \neq 0\}$, and consider the set of connected components in the graph (V^*, E^*) consisting of the vertex set $V^* \equiv \mathcal{J}^{(2)}$ and the edge set $E^* \equiv \{(j^\dagger, j^\ddagger) \in E : \xi_{j^\dagger}^* = \xi_{j^\ddagger}^*\}$. An isolated point not connected to any edge is regarded as a one-connected component. Then, from each connected component whose vertex values are non-zero, one representative vertex is arbitrarily chosen, and the set of all representatives is denoted as $\mathcal{J}^{(3)}$ $(\subset \mathcal{J}^{(2)})$. That is, if $j \in \mathcal{J}^{(2)}$, then $\exists j^\dagger \in \mathcal{J}^{(3)}$; $\xi_j^* = \xi_{j^\dagger}^*$, and furthermore, if $j^\ddagger \neq j^\dagger$, then $\xi_j^* \neq \xi_{j^\ddagger}^*$. For any vector $\boldsymbol{c} = (c_j)_{j \in \{1, \ldots, p\}}$, let $\boldsymbol{c}^{(\ell)}$ denote the vector $(c_j)_{j \in \mathcal{J}^{(\ell)}}$ $(\ell \in \{1, 2, 3\})$,

and for any matrix $\boldsymbol{C} = (C_{j^\dagger j^\ddagger})_{j^\dagger \in \{1,\ldots,p\}, j^\ddagger \in \{1,\ldots,p\}}$, let $\boldsymbol{C}^{(\ell^\dagger \ell^\ddagger)}$ denote the matrix $(C_{j^\dagger j^\ddagger})_{j^\dagger \in \mathcal{J}^{(\ell^\dagger)}, j^\ddagger \in \mathcal{J}^{(\ell^\ddagger)}}$ $(\ell^\dagger, \ell^\ddagger \in \{1,2,3\})$. In addition, we define

$$\boldsymbol{A} \equiv (I(\xi^*_{j^\dagger} = \xi^*_{j^\ddagger}))_{j^\dagger \in \{1,\ldots,p\}, j^\ddagger \in \{1,\ldots,p\}}, \tag{4.4}$$

which is a matrix whose components are all either 0 or 1, and we will use $\boldsymbol{A}^{(32)}$ and $\boldsymbol{A}^{(23)}$ later. Here, in addition to (C1), we assume the following condition.

(C2) $\boldsymbol{\xi}^*$ exists in the interior of Ξ. In addition, when $\boldsymbol{\xi}^*$ gives a sparse solution, that is, when ξ^*_j becomes 0 or $\xi^*_{j^\dagger} - \xi^*_{j^\ddagger}$ becomes 0, then $\boldsymbol{\lambda} = (\lambda_1, \lambda_2)$ is defined so that $\boldsymbol{\xi}^*$ is not the solution of the minimization problem with the absolute value operator removed, such as $\mathrm{argmin}_{\boldsymbol{\xi} \in \Xi}\{h(\boldsymbol{\xi}) + \lambda_1 \sum_{j \in V}(\pm \xi_j) + \lambda_2 \sum_{(j^\dagger, j^\ddagger) \in E} |\xi_{j^\dagger} - \xi_{j^\ddagger}|\}$ or
$\mathrm{argmin}_{\boldsymbol{\xi} \in \Xi}[h(\boldsymbol{\xi}) + \lambda_1 \sum_{j \in V} |\xi_j| + \lambda_2 \sum_{(j^\dagger, j^\ddagger) \in E}\{\pm(\xi_{j^\dagger} - \xi_{j^\ddagger})\}]$.

The objective function minimized by $\boldsymbol{\xi}^*$ is convex with respect to $\boldsymbol{\xi}$, and therefore if Ξ is sufficiently large, the first sentence of condition (C2) can be expected to hold. Furthermore, since the second sentence of condition (C2) is satisfied for almost all $\boldsymbol{\lambda} = (\lambda_1, \lambda_2)$, it is only a technical condition, and we do not consider it when actually applying the method to data. We now are prepared to prove the following asymptotic property.

Theorem 1. Under conditions (C1) and (C2), for the generalized fused lasso estimator $\hat{\boldsymbol{\xi}}$, it holds that

$$\begin{aligned}
&n(\hat{\boldsymbol{\xi}}^{(1)} - \boldsymbol{\xi}^{*(1)}) \xrightarrow{\mathrm{p}} \boldsymbol{0}, \\
&j^\dagger \in \mathcal{J}^{(2)} \quad \Rightarrow \quad \exists j^\ddagger \in \mathcal{J}^{(3)};\ \sqrt{n}(\hat{\xi}_{j^\dagger} - \hat{\xi}_{j^\ddagger}) \xrightarrow{\mathrm{p}} 0, \\
&\sqrt{n}(\hat{\boldsymbol{\xi}}^{(3)} - \boldsymbol{\xi}^{*(3)}) \xrightarrow{\mathrm{d}} \mathrm{N}(\boldsymbol{0}, (\boldsymbol{A}^{(32)} \boldsymbol{J}^{(22)} \boldsymbol{A}^{(23)})^{-1}).
\end{aligned}$$

Proof. The approach of the proof is close to that of Viallon et al. (2016), but since the order of the penalty term of the generalized fused lasso is $\mathrm{O}(n)$ and the limit of the estimator is not even the true value, we also incorporate the approach of Ninomiya and Kawano (2016). Let $\mathcal{K}_1 = \{(j^\dagger, j^\ddagger) : (j^\dagger, j^\ddagger) \in E,\ \xi^*_{j^\dagger} = \xi^*_{j^\ddagger} = 0\}$, $\mathcal{K}_2 = \{(j^\dagger, j^\ddagger) : (j^\dagger, j^\ddagger) \in E,\ \xi^*_{j^\dagger} = 0 \neq \xi^*_{j^\ddagger}\}$, $\mathcal{K}_3 = \{(j^\dagger, j^\ddagger) : (j^\dagger, j^\ddagger) \in E,\ \xi^*_{j^\dagger} = \xi^*_{j^\ddagger} \neq 0\}$, $\mathcal{K}_4 = \{(j^\dagger, j^\ddagger) : (j^\dagger, j^\ddagger) \in E,\ 0 \neq \xi^*_{j^\dagger} \neq \xi^*_{j^\ddagger} \neq 0\}$, $\mathcal{K}_5 = \{(j^\dagger, j^\ddagger) : (j^\dagger, j^\ddagger) \in E,\ \xi^*_{j^\dagger} \neq 0 = \xi^*_{j^\ddagger}\}$. Also, let $\mathcal{K}_{1,j} = \{k : (j,k) \in \mathcal{K}_1\} \cup \{k : (k,j) \in \mathcal{K}_1\}$, $\mathcal{K}_{2,j} = \{k : (j,k) \in \mathcal{K}_2\} \cup \{k : (k,j) \in \mathcal{K}_5\}$, $\mathcal{K}_{3,j} = \{k : (j,k) \in \mathcal{K}_3\} \cup \{k : (k,j) \in$

$\mathcal{K}_3\}$, $\mathcal{K}_{4,j} = \{k : (j,k) \in \mathcal{K}_4\} \cup \{k : (k,j) \in \mathcal{K}_4\}$, $\mathcal{K}_{5,j} = \{k : (j,k) \in \mathcal{K}_5\} \cup \{k : (k,j) \in \mathcal{K}_2\}$. Since $h(\boldsymbol{\theta})$ is a convex and differentiable function, from the Karush-Kuhn-Tucker (KKT) conditions and (C2), we obtain

$$\xi_j^* = 0 \quad \Rightarrow \quad \frac{\partial h}{\partial \xi_j}(\boldsymbol{\xi}^*) = -\lambda_1 a_j - \lambda_2 \left(\sum_{k \in \mathcal{K}_{1,j}} \mathrm{sgn}(k-j) b_{j,k} - \sum_{k \in \mathcal{K}_{2,j}} \mathrm{sgn}(\xi_k^*) \right), \tag{4.5}$$

$$\begin{aligned} \xi_j^* \neq 0 \quad \Rightarrow \quad \frac{\partial h}{\partial \xi_j}(\boldsymbol{\xi}^*) = & -\lambda_1 \mathrm{sgn}(\xi_j^*) \\ & - \lambda_2 \left(\sum_{k \in \mathcal{K}_{3,j}} \mathrm{sgn}(k-j) b_{j,k} + \sum_{k \in \mathcal{K}_{4,j} \cup \mathcal{K}_{5,j}} \mathrm{sgn}(\xi_j^* - \xi_k^*) \right). \end{aligned} \tag{4.6}$$

In addition, $-1 < a_j < 1$ and $-1 < b_{j,k} < 1$. Furthermore, letting $\mathcal{J}_1 = \{j : \exists k;\ (j,k) \in \mathcal{K}_1 \ \vee \ (k,j) \in \mathcal{K}_1\}$ and $\mathcal{J}_3 = \{j : \exists k;\ (j,k) \in \mathcal{K}_3 \ \vee \ (k,j) \in \mathcal{K}_3\}$, it holds that

$$b_{j,k} = b_{k,j} \quad (j \in \mathcal{J}_1,\ k \in \mathcal{K}_{1,j}), \tag{4.7}$$

$$b_{j,k} = b_{k,j} \quad (j \in \mathcal{J}_3,\ k \in \mathcal{K}_{3,j}). \tag{4.8}$$

Since $j \in \mathcal{J}_3$ and $k \in \mathcal{K}_{3,j}$ mean that $k \in \mathcal{J}_3$ and $j \in \mathcal{K}_{3,k}$, (4.8) implies that

$$\sum_{j \in \mathcal{J}_3} \sum_{k \in \mathcal{K}_{3,j}} \mathrm{sgn}(k-j) b_{j,k} = 0. \tag{4.9}$$

Now, recalling that for a vector $\boldsymbol{u} = (u_1, \ldots, u_p)^{\mathrm{T}}$, $\boldsymbol{u}^{(\ell)}$ represents $(u_j)_{j \in \mathcal{J}^{(\ell)}}$, we define the following random function:

$$\begin{aligned} & v_n(\boldsymbol{u}^{(1)}, \boldsymbol{u}^{(2)}) \\ & \equiv \sum_{i=1}^n \left\{ \log f(y_i, \boldsymbol{x}_i \mid \boldsymbol{\xi}^*) - \log f\left(y_i, \boldsymbol{x}_i \;\middle|\; \frac{\boldsymbol{u}^{(1)}}{n}, \frac{\boldsymbol{u}^{(2)}}{\sqrt{n}} + \boldsymbol{\xi}^{*(2)} \right) \right\} \\ & \quad + n\lambda_1 \left\| \frac{\boldsymbol{u}^{(1)}}{n} \right\|_1 + n\lambda_1 \left\| \frac{\boldsymbol{u}^{(2)}}{\sqrt{n}} + \boldsymbol{\xi}^{*(2)} \right\|_1 - n\lambda_1 \|\boldsymbol{\xi}^*\|_1 \\ & \quad + n\lambda_2 \sum_{(j^\dagger, j^\ddagger) \in \mathcal{K}_1} \left| \frac{u_{j^\dagger} - u_{j^\ddagger}}{n} \right| + n\lambda_2 \sum_{(j^\dagger, j^\ddagger) \in \mathcal{K}_2} \left| \frac{u_{j^\dagger}}{n} - \frac{u_{j^\ddagger}}{\sqrt{n}} + \xi_{j^\dagger}^* - \xi_{j^\ddagger}^* \right| \\ & \quad + n\lambda_2 \sum_{(j^\dagger, j^\ddagger) \in \mathcal{K}_3} \left| \frac{u_{j^\dagger} - u_{j^\ddagger}}{\sqrt{n}} + \xi_{j^\dagger}^* - \xi_{j^\ddagger}^* \right| + n\lambda_2 \sum_{(j^\dagger, j^\ddagger) \in \mathcal{K}_4} \left| \frac{u_{j^\dagger} - u_{j^\ddagger}}{\sqrt{n}} + \xi_{j^\dagger}^* - \xi_{j^\ddagger}^* \right| \\ & \quad + n\lambda_2 \sum_{(j^\dagger, j^\ddagger) \in \mathcal{K}_5} \left| \frac{u_{j^\dagger}}{\sqrt{n}} - \frac{u_{j^\ddagger}}{n} + \xi_{j^\dagger}^* - \xi_{j^\ddagger}^* \right| - n\lambda_2 \sum_{(j^\dagger, j^\ddagger) \in E} |\xi_{j^\dagger}^* - \xi_{j^\ddagger}^*|. \end{aligned} \tag{4.10}$$

Note that $\mathrm{argmin}_{(\boldsymbol{u}^{(1)}, \boldsymbol{u}^{(2)})} v_n(\boldsymbol{u}^{(1)}, \boldsymbol{u}^{(2)}) = (n\hat{\boldsymbol{\xi}}^{(1)}, \sqrt{n}(\hat{\boldsymbol{\xi}}^{(2)} - \boldsymbol{\xi}^{*(2)}))$. In the following, we perform a Taylor expansion of $v_n(\boldsymbol{u}^{(1)}, \boldsymbol{u}^{(2)})$ around $(\boldsymbol{u}^{(1)}, \boldsymbol{u}^{(2)}) = (\boldsymbol{0}, \boldsymbol{0})$; we consider the expansion in two cases.

(i) The case of $u_{j^\dagger} = u_{j^\ddagger}$ for any $(j^\dagger, j^\ddagger) \in E$ such that $\xi^*_{j^\dagger} = \xi^*_{j^\ddagger} \neq 0$.

Let $\partial \log f(y_i, \boldsymbol{x}_i \mid \boldsymbol{\xi})/\partial \boldsymbol{\xi}$ be denoted by $\boldsymbol{g}'_{y_i,\boldsymbol{x}_i}(\boldsymbol{\xi})$, and $\partial^2 \log f(y_i, \boldsymbol{x}_i \mid \boldsymbol{\xi})/\partial \boldsymbol{\xi} \partial \boldsymbol{\xi}^{\mathrm{T}}$ be denoted by $\boldsymbol{G}''_{y_i,\boldsymbol{x}_i}(\boldsymbol{\xi})$, and recalling also that $\boldsymbol{G}''^{(\ell^\dagger \ell^\ddagger)}_{y_i,\boldsymbol{x}_i}(\boldsymbol{\xi})$ represents $(\boldsymbol{G}''_{y_i,\boldsymbol{x}_i}(\boldsymbol{\xi})_{j^\dagger j^\ddagger})_{j^\dagger \in \mathcal{J}^{(\ell^\dagger)}, j^\ddagger \in \mathcal{J}^{(\ell^\ddagger)}}$, (4.10) can be rewritten as

$$
\begin{aligned}
& v_n(\boldsymbol{u}^{(1)}, \boldsymbol{u}^{(2)}) \\
&= -\sum_{i=1}^{n} \left\{ \boldsymbol{g}'^{(1)}_{y_i,\boldsymbol{x}_i}(\boldsymbol{\xi}^*)^{\mathrm{T}} \frac{\boldsymbol{u}^{(1)}}{n} + \boldsymbol{g}'^{(2)}_{y_i,\boldsymbol{x}_i}(\boldsymbol{\xi}^*)^{\mathrm{T}} \frac{\boldsymbol{u}^{(2)}}{\sqrt{n}} \right\} + \lambda_1 \sum_{j \in \mathcal{J}^{(1)}} |u_j| + \sqrt{n}\lambda_1 \boldsymbol{u}^{(2)\mathrm{T}} \mathrm{sgn}(\boldsymbol{\xi}^{*(2)}) \\
&\quad + \lambda_2 \sum_{(j^\dagger, j^\ddagger) \in \mathcal{K}_1} |u_{j^\dagger} - u_{j^\ddagger}| + \lambda_2 \sum_{(j^\dagger, j^\ddagger) \in \mathcal{K}_2} \mathrm{sgn}(\xi^*_{j^\dagger} - \xi^*_{j^\ddagger}) u_{j^\dagger} \\
&\quad - \sqrt{n}\lambda_2 \sum_{(j^\dagger, j^\ddagger) \in \mathcal{K}_2} \mathrm{sgn}(\xi^*_{j^\dagger} - \xi^*_{j^\ddagger}) u_{j^\ddagger} \\
&\quad + \sqrt{n}\lambda_2 \sum_{(j^\dagger, j^\ddagger) \in \mathcal{K}_4} \mathrm{sgn}(\xi^*_{j^\dagger} - \xi^*_{j^\ddagger})(u_{j^\dagger} - u_{j^\ddagger}) \\
&\quad + \sqrt{n}\lambda_2 \sum_{(j^\dagger, j^\ddagger) \in \mathcal{K}_5} \mathrm{sgn}(\xi^*_{j^\dagger} - \xi^*_{j^\ddagger}) u_{j^\dagger} - \lambda_2 \sum_{(j^\dagger, j^\ddagger) \in \mathcal{K}_5} \mathrm{sgn}(\xi^*_{j^\dagger} - \xi^*_{j^\ddagger}) u_{j^\ddagger} \\
&\quad - \sum_{i=1}^{n} \left[\frac{\boldsymbol{u}^{(1)\mathrm{T}}}{n} \left\{ \boldsymbol{G}''^{(11)}_{y_i,\boldsymbol{x}_i}(\boldsymbol{\xi}^*) \frac{\boldsymbol{u}^{(1)}}{2n} + \boldsymbol{G}''^{(12)}_{y_i,\boldsymbol{x}_i}(\boldsymbol{\xi}^*) \frac{\boldsymbol{u}^{(2)}}{\sqrt{n}} \right\} + \frac{1}{2} \frac{\boldsymbol{u}^{(2)\mathrm{T}}}{\sqrt{n}} \boldsymbol{G}''^{(22)}_{y_i,\boldsymbol{x}_i}(\boldsymbol{\xi}^*) \frac{\boldsymbol{u}^{(2)}}{\sqrt{n}} \right] \\
&\quad + \mathrm{o_P}(1),
\end{aligned}
\tag{4.11}
$$

where $\mathrm{sgn}(\boldsymbol{\xi}^{(2)})$ is a vector whose elements are $\mathrm{sgn}(\xi^*_j)$ $(j \in \mathcal{J}^{(2)})$. Now, defining $\mathcal{J}_2 = \{j : \exists k;\ (j,k) \in \mathcal{K}_2 \vee (k,j) \in \mathcal{K}_5\}$, $\mathcal{J}_4 = \{j : \exists k;\ (j,k) \in \mathcal{K}_4 \vee (k,j) \in \mathcal{K}_4\}$, and $\mathcal{J}_5 = \{j : \exists k;\ (k,j) \in \mathcal{K}_2 \ \vee \ (j,k) \in \mathcal{K}_5\}$, the fifth and ninth terms on the right-hand side can be combined as

$$
\lambda_2 \sum_{(j^\dagger, j^\ddagger) \in \mathcal{K}_2} \mathrm{sgn}(-\xi^*_{j^\ddagger}) u_{j^\dagger} - \lambda_2 \sum_{(j^\dagger, j^\ddagger) \in \mathcal{K}_5} \mathrm{sgn}(\xi^*_{j^\dagger}) u_{j^\ddagger} = -\lambda_2 \sum_{j \in \mathcal{J}_2} \sum_{k \in \mathcal{K}_{2,j}} \mathrm{sgn}(\xi^*_k) u_j,
$$

the sixth and eighth terms can be combined as

$$
\begin{aligned}
& -\sqrt{n}\lambda_2 \sum_{(j^\dagger, j^\ddagger) \in \mathcal{K}_2} \mathrm{sgn}(-\xi^*_{j^\ddagger}) u_{j^\ddagger} + \sqrt{n}\lambda_2 \sum_{(j^\dagger, j^\ddagger) \in \mathcal{K}_5} \mathrm{sgn}(\xi^*_{j^\dagger}) u_{j^\dagger} \\
&= \sqrt{n}\lambda_2 \sum_{j \in \mathcal{J}_5} \sum_{k \in \mathcal{K}_{5,j}} \mathrm{sgn}(\xi^*_j) u_j,
\end{aligned}
$$

and the seventh term can be rewritten as $\sqrt{n}\lambda_2 \sum_{j \in \mathcal{J}_4} \sum_{k \in \mathcal{K}_{4,j}} \mathrm{sgn}(\xi^*_j - \xi^*_k) u_j$. By also using the fact that the quadratic term involving $\boldsymbol{u}^{(1)}$ is $\mathrm{o_P}(1)$, after

rearrangement, (4.11) becomes

$$
\begin{aligned}
&v_n(\boldsymbol{u}^{(1)}, \boldsymbol{u}^{(2)}) \\
&= -\sum_{i=1}^{n} \left\{ \boldsymbol{g}'^{(1)}_{y_i,\boldsymbol{x}_i}(\boldsymbol{\xi}^*)^{\mathrm{T}} \frac{\boldsymbol{u}^{(1)}}{n} + \boldsymbol{g}'^{(2)}_{y_i,\boldsymbol{x}_i}(\boldsymbol{\xi}^*)^{\mathrm{T}} \frac{\boldsymbol{u}^{(2)}}{\sqrt{n}} \right\} + \lambda_1 \sum_{j \in \mathcal{J}^{(1)}} |u_j| + \sqrt{n}\lambda_1 \boldsymbol{u}^{(2)\mathrm{T}} \mathrm{sgn}(\boldsymbol{\xi}^{*(2)}) \\
&\quad + \lambda_2 \sum_{(j^\dagger, j^\ddagger) \in \mathcal{K}_1} |u_{j^\dagger} - u_{j^\ddagger}| - \lambda_2 \sum_{j \in \mathcal{J}_2} \sum_{k \in \mathcal{K}_{2,j}} \mathrm{sgn}(\xi^*_k) u_j + \sqrt{n}\lambda_2 \sum_{j \in \mathcal{J}_4} \sum_{k \in \mathcal{K}_{4,j}} \mathrm{sgn}(\xi^*_j - \xi^*_k) u_j \\
&\quad + \sqrt{n}\lambda_2 \sum_{j \in \mathcal{J}_5} \sum_{k \in \mathcal{K}_{5,j}} \mathrm{sgn}(\xi^*_j) u_j - \frac{1}{2} \sum_{i=1}^{n} \frac{\boldsymbol{u}^{(2)\mathrm{T}}}{\sqrt{n}} \boldsymbol{G}''^{(22)}_{y_i,\boldsymbol{x}_i}(\boldsymbol{\xi}^*) \frac{\boldsymbol{u}^{(2)}}{\sqrt{n}} + \mathrm{o_P}(1).
\end{aligned}
\tag{4.12}
$$

Next, from (R2), we have that $-\sum_{i=1}^{n} \boldsymbol{g}'^{(1)}_{y_i,\boldsymbol{x}_i}(\boldsymbol{\xi}^*)/n$ converges in probability to $(\partial h/\partial \boldsymbol{\xi}^{(1)})(\boldsymbol{\xi}^*)$. In addition, from (R2), (R3), and (4.6), it follows that there exists a $|\mathcal{J}^{(2)}|$-dimensional random vector $\boldsymbol{s}^{(2)}$ following $\mathrm{N}(\boldsymbol{0}, \boldsymbol{J}^{(22)})$, and that $\sum_{i=1}^{n} \boldsymbol{g}'^{(2)}_{y_i,\boldsymbol{x}_i}(\boldsymbol{\xi}^*)/\sqrt{n} - \sqrt{n}\{\lambda_1 \mathrm{sgn}(\boldsymbol{\xi}^{*(2)})$
$+\lambda_2(\sum_{k \in \mathcal{K}_{3,j}} \mathrm{sgn}(k-j) b_{j,k})_{j \in \mathcal{J}^{(2)}} + \lambda_2(\sum_{k \in \mathcal{K}_{4,j} \cup \mathcal{K}_{5,j}} \mathrm{sgn}(\xi^*_j - \xi^*_k))_{j \in \mathcal{J}^{(2)}}\}$ converges in distribution to $\boldsymbol{s}^{(2)}$. Furthermore, from (C1), we also have that $-\sum_{i=1}^{n} (\boldsymbol{u}^{(2)}/\sqrt{n})^{\mathrm{T}} \boldsymbol{G}''^{(22)}_{y_i,\boldsymbol{x}_i}(\boldsymbol{\xi}^*)(\boldsymbol{u}^{(2)}/\sqrt{n})$ converges to $\boldsymbol{u}^{(2)\mathrm{T}} \boldsymbol{J}^{(22)} \boldsymbol{u}^{(2)}$. Therefore, for each $(\boldsymbol{u}^{(1)}, \boldsymbol{u}^{(2)})$, (4.12) converges in distribution to

$$
\begin{aligned}
&v(\boldsymbol{u}^{(1)}, \boldsymbol{u}^{(2)}) \\
&= \sum_{j \in \mathcal{J}^{(1)}} \left\{ \frac{\partial h}{\partial \xi_j}(\boldsymbol{\xi}^*) u_j + \lambda_1 |u_j| \right\} + \lambda_2 \sum_{(j^\dagger, j^\ddagger) \in \mathcal{K}_1} |u_{j^\dagger} - u_{j^\ddagger}| - \lambda_2 \sum_{j \in \mathcal{J}_2} \sum_{k \in \mathcal{K}_{2,j}} \mathrm{sgn}(\xi^*_k) u_j \\
&\quad - \boldsymbol{u}^{(2)\mathrm{T}} \boldsymbol{s}^{(2)} - \sqrt{n}\lambda_2 \sum_{j \in \mathcal{J}^{(2)}} \sum_{k \in \mathcal{K}_{3,j}} \mathrm{sgn}(k-j) b_{j,k} u_j + \frac{1}{2} \boldsymbol{u}^{(2)\mathrm{T}} \boldsymbol{J}^{(22)} \boldsymbol{u}^{(2)}.
\end{aligned}
$$

By transforming the sum of the first and third terms on the right-hand side using (4.5), we obtain

$$
\begin{aligned}
&v(\boldsymbol{u}^{(1)}, \boldsymbol{u}^{(2)}) \\
&= -\lambda_1 \sum_{j \in \mathcal{J}^{(1)}} a_j u_j - \lambda_2 \sum_{j \in \mathcal{J}^{(1)}} \sum_{k \in \mathcal{K}_{1,j}} \mathrm{sgn}(k-j) b_{j,k} u_j + \lambda_1 \sum_{j \in \mathcal{J}^{(1)}} |u_j| \\
&\quad + \lambda_2 \sum_{(j^\dagger, j^\ddagger) \in \mathcal{K}_1} |u_{j^\dagger} - u_{j^\ddagger}| \\
&\quad - \boldsymbol{u}^{(2)\mathrm{T}} \boldsymbol{s}^{(2)} - \sqrt{n}\lambda_2 \sum_{j \in \mathcal{J}^{(2)}} \sum_{k \in \mathcal{K}_{3,j}} \mathrm{sgn}(k-j) b_{j,k} u_j \\
&\quad + \frac{1}{2} \boldsymbol{u}^{(2)\mathrm{T}} \boldsymbol{J}^{(22)} \boldsymbol{u}^{(2)}.
\end{aligned}
\tag{4.13}
$$

From (4.7), the second term on the right-hand side can be rewritten as $-\lambda_2 \sum_{(j^\dagger, j^\ddagger) \in \mathcal{K}_1} b_{j^\dagger, j^\ddagger}(u_{j^\dagger} - u_{j^\ddagger})$. Regarding the sixth term, $\sum_{j \in \mathcal{J}^{(2)}}$ can be replaced with $\sum_{j \in \mathcal{J}_3}$, and in the present case, all u_j $(j \in \mathcal{J}_3)$ take the same value;

hence, by (4.9), the sum equals 0. By also applying this property of u_j to the fifth and seventh terms, (4.13) can be rewritten using the matrix $\boldsymbol{A}$ defined in (4.4) as

$$\begin{aligned} & v(\boldsymbol{u}^{(1)}, \boldsymbol{u}^{(2)}) \\ & = \lambda_1 \sum_{j \in \mathcal{J}^{(1)}} (|u_j| - a_j u_j) + \lambda_2 \sum_{(j^\dagger, j^\ddagger) \in \mathcal{K}_1} \{|u_{j^\dagger} - u_{j^\ddagger}| - b_{j^\dagger, j^\ddagger}(u_{j^\dagger} - u_{j^\ddagger})\} \\ & \quad - \boldsymbol{u}^{(3)\mathrm{T}} \boldsymbol{A}^{(32)} \boldsymbol{s}^{(2)} + \frac{1}{2} \boldsymbol{u}^{(3)\mathrm{T}} \boldsymbol{A}^{(32)} \boldsymbol{J}^{(22)} \boldsymbol{A}^{(23)} \boldsymbol{u}^{(3)}. \end{aligned} \tag{4.14}$$

Noting that $-1 < a_j < 1$ and $-1 < b_{j^\dagger, j^\ddagger} < 1$, it follows that $v(\boldsymbol{u}^{(1)}, \boldsymbol{u}^{(2)})$ has a unique minimum at $\boldsymbol{u}^{(1)} = \boldsymbol{0}$ and $\boldsymbol{u}^{(3)} = (\boldsymbol{A}^{(32)} \boldsymbol{J}^{(22)} \boldsymbol{A}^{(23)})^{-1} \boldsymbol{A}^{(32)} \boldsymbol{s}^{(2)}$.

(ii) The case other than (i)

Equation (4.10) can be rewritten as (4.11) with the addition of $\sqrt{n}\lambda_2 \sum_{(j^\dagger, j^\ddagger) \in \mathcal{K}_3} |u_{j^\dagger} - u_{j^\ddagger}|$, which, according to (4.13), converges in distribution to

$$\begin{aligned} & v(\boldsymbol{u}^{(1)}, \boldsymbol{u}^{(2)}) \\ & = -\lambda_1 \sum_{j \in \mathcal{J}^{(1)}} a_j u_j - \lambda_2 \sum_{j \in \mathcal{J}^{(1)}} \sum_{k \in \mathcal{K}_{1,j}} \mathrm{sgn}(k - j) b_{j,k} u_j + \lambda_1 \sum_{j \in \mathcal{J}^{(1)}} |u_j| \\ & + \lambda_2 \sum_{\mathrm{sgn}(j^\dagger, j^\ddagger) \in \mathcal{K}_1} |u_{j^\dagger} - u_{j^\ddagger}| + \sqrt{n}\lambda_2 \sum_{\mathrm{sgn}(j^\dagger, j^\ddagger) \in \mathcal{K}_3} |u_{j^\dagger} - u_{j^\ddagger}| \\ & - \boldsymbol{u}^{(2)\mathrm{T}} \boldsymbol{s}^{(2)} - \sqrt{n}\lambda_2 \sum_{j \in \mathcal{J}^{(2)}} \sum_{k \in \mathcal{K}_{3,j}} \mathrm{sgn}(k - j) b_{j,k} u_j + \frac{1}{2} \boldsymbol{u}^{(2)\mathrm{T}} \boldsymbol{J}^{(22)} \boldsymbol{u}^{(2)}. \end{aligned} \tag{4.15}$$

The seventh term on the right-hand side can also be expressed as $-\sqrt{n}\lambda_2 \sum_{(j^\dagger, j^\ddagger) \in \mathcal{K}_3} b_{j^\dagger, j^\ddagger}(u_{j^\dagger} - u_{j^\ddagger})$ according to (4.8). Therefore, just as in deriving (4.14), (4.15) can be rewritten as

$$\begin{aligned} & v(\boldsymbol{u}^{(1)}, \boldsymbol{u}^{(2)}) \\ & = \lambda_1 \sum_{j \in \mathcal{J}^{(1)}} (|u_j| - a_j u_j) + \lambda_2 \sum_{(j^\dagger, j^\ddagger) \in \mathcal{K}_1} \{|u_{j^\dagger} - u_{j^\ddagger}| - b_{j^\dagger, j^\ddagger}(u_{j^\dagger} - u_{j^\ddagger})\} \\ & \quad + \sqrt{n}\lambda_2 \sum_{(j^\dagger, j^\ddagger) \in \mathcal{K}_3} \{|u_{j^\dagger} - u_{j^\ddagger}| - b_{j^\dagger, j^\ddagger}(u_{j^\dagger} - u_{j^\ddagger})\} \\ & \quad - \boldsymbol{u}^{(3)\mathrm{T}} \boldsymbol{A}^{(32)} \boldsymbol{s}^{(2)} + \frac{1}{2} \boldsymbol{u}^{(3)\mathrm{T}} \boldsymbol{A}^{(32)} \boldsymbol{J}^{(22)} \boldsymbol{A}^{(23)} \boldsymbol{u}^{(3)}. \end{aligned} \tag{4.16}$$

In the present case, there exists $(j^\dagger, j^\ddagger) \in \mathcal{K}_3$ such that $u_{j^\dagger} \neq u_{j^\ddagger}$. Therefore, the third term on the right-hand side is positive and $\mathrm{O_P}(\sqrt{n})$. On the other hand, all other terms are $\mathrm{O_P}(1)$; hence, it follows that (4.16) diverges to positive infinity.

From (i) and (ii), it follows that $v(\boldsymbol{u}^{(1)}, \boldsymbol{u}^{(2)})$ has the unique minimizer given by

$$\begin{aligned} &\boldsymbol{u}^{(1)} = \boldsymbol{0}, \\ &j^\dagger \in \mathcal{J}^{(2)} \quad \Rightarrow \quad \exists j^\ddagger \in \mathcal{J}^{(3)};\ u_{j^\dagger} = u_{j^\ddagger}, \\ &\boldsymbol{u}^{(3)} = (\boldsymbol{A}^{(32)}\boldsymbol{J}^{(22)}\boldsymbol{A}^{(23)})^{-1}\boldsymbol{A}^{(32)}\boldsymbol{s}^{(2)}. \end{aligned} \tag{4.17}$$

Since $v_n(\boldsymbol{u}^{(1)}, \boldsymbol{u}^{(2)})$ is convex, it follows from the convexity lemma of Hjort and Pollard (1993) or Geyer (1996) that

$$\operatorname*{argmin}_{\boldsymbol{u}^{(1)},\boldsymbol{u}^{(2)}} v_n(\boldsymbol{u}^{(1)}, \boldsymbol{u}^{(2)}) \xrightarrow{\mathrm{d}} \operatorname*{argmin}_{\boldsymbol{u}^{(1)},\boldsymbol{u}^{(2)}} v(\boldsymbol{u}^{(1)}, \boldsymbol{u}^{(2)}),$$

and hence the conclusion follows. □

4.1.3 Bayesian generalized fused lasso

Park and Casella (2008) regarded the lasso regularization method as a Bayesian method with the Laplace distribution as the prior distribution, and the generalized fused lasso regularization method can be regarded in the same way. Specifically, we assume the Laplace distribution with the location parameter 0 and the scale parameter $1/\lambda_{1,j}$ or $1/\lambda_{2,j}$ for the regression coefficients and the differences between regression coefficients as

$$\pi_{\mathrm{gfl}}(\boldsymbol{\theta}; \boldsymbol{\lambda}) \propto \exp\left(-n\sum_{j=1}^{\tilde{p}} \lambda_{1,j} \sum_{m\in\mathcal{V}} |\theta_{m,j}| - n\sum_{j=1}^{\tilde{p}} \lambda_{2,j} \sum_{(m^\dagger,m^\ddagger)\in\mathcal{E}} |\theta_{m^\dagger,j} - \theta_{m^\ddagger,j}| \right). \tag{4.18}$$

As in Section 4.1.2, when we assume the setting $\lambda_{1,1} = \ldots = \lambda_{1,\tilde{p}} = \lambda_1$ and $\lambda_{2,1} = \ldots = \lambda_{2,\tilde{p}} = \lambda_2$, and if we newly express $\boldsymbol{\theta} = (\theta_{1,1}, \ldots, \theta_{1,\tilde{p}}, \theta_{2,1}, \ldots, \theta_{2,\tilde{p}}, \ldots, \theta_{M,1}, \ldots, \theta_{M,\tilde{p}})$ as $\boldsymbol{\xi} = (\xi_1, \ldots, \xi_p)$, then this prior distribution can be written as

$$\pi_{\mathrm{gfl}}(\boldsymbol{\xi}; \boldsymbol{\lambda}) \propto \exp\left(-n\lambda_1 \sum_{j\in V} |\xi_j| - n\lambda_2 \sum_{(j^\dagger,j^\ddagger)\in E} |\xi_{j^\dagger} - \xi_{j^\ddagger}| \right). \tag{4.19}$$

The maximum a posteriori estimator for the regression coefficients assuming the prior distribution (4.18) or (4.19) is equivalent to the generalized fused lasso estimator expressed by (4.2) or (4.3), respectively. What were treated as regularization parameters in the generalized fused lasso become hyperparameters of the Laplace distribution in the Bayesian generalized fused lasso.

In the SVC model of this paper, the complexity of the model changes depending on whether the values of $\lambda_{2,j}$ in (4.18) are varied for each j and how many

different values are used. If we are interested in considering two Bayesian models with extremely different complexities, then the following two models (Models 1 and 2) are suitable, so only those two models are treated in the numerical experiments and real data analysis of this paper for simplicity. In the first model, Model 1, is as specified above; that is, the prior distribution is as in (4.18) with settings $\lambda_{1,1} = \cdots = \lambda_{1,\tilde{p}} = \lambda_1$ and $\lambda_{2,1} = \cdots = \lambda_{2,\tilde{p}} = \lambda_2$, which corresponds to the prior distribution (4.19). This setting assumes that regression coefficients for all the explanatory variables follow a common Laplace distribution, and that the differences between adjacent coefficients also do so. In the second model, Model 2, the prior distribution of (4.18) is used as is. This means that the regression coefficients and the differences between adjacent regression coefficients are assumed to follow different Laplace distributions for all explanatory variables. Obviously, the latter model has greater complexity. For data in which how all variables affect the outcome is assumed to have the same spatial structure, Model 1 is suitable, whereas Model 2 is more appropriate for data in which how the effects differ depends on the variable type. This means that which model is suitable depends on the data. To provide an analytical method that improves performance by changing the model according to the data, the construction of an appropriate model selection method is a key point.

4.2 Construction of information criterion

In this section, we review the definition of WAIC used in this chapter and discuss PIIC, which attempts to resolve some concerns of WAIC, and adapt it to the SVC model with the Bayesian generalized fused lasso.

The definition of WAIC used in this chapter corresponds to that in Sec. 2.3.7, with $\boldsymbol{\theta}$ replaced by $\boldsymbol{\xi}$. Specifically, assume a probability density function $f(\cdot, \cdot \mid \boldsymbol{\xi})$ and a prior distribution $\pi_n(\cdot; \boldsymbol{\lambda})$, and consider independent random vectors satisfying

$$(y_1, \boldsymbol{x}_1), \ldots, (y_n, \boldsymbol{x}_n), (\tilde{y}_1, \tilde{\boldsymbol{x}}_1), \ldots, (\tilde{y}_n, \tilde{\boldsymbol{x}}_n) \sim f(\cdot, \cdot \mid \boldsymbol{\xi}), \qquad \boldsymbol{\xi} \sim \pi_n(\cdot; \boldsymbol{\lambda}).$$

Then, by evaluating the expectation

$$\begin{aligned} &\sum_{i=1}^{n} \log f(y_i, \boldsymbol{x}_i \mid \boldsymbol{y}, \boldsymbol{X}; \boldsymbol{\lambda}) - \sum_{i=1}^{n} \log f(y_i, \boldsymbol{x}_i \mid \boldsymbol{\xi}^*) \\ &- \sum_{i=1}^{n} \log f(\tilde{y}_i, \tilde{\boldsymbol{x}}_i \mid \boldsymbol{y}, \boldsymbol{X}; \boldsymbol{\lambda}) + \sum_{i=1}^{n} \log f(\tilde{y}_i, \tilde{\boldsymbol{x}}_i \mid \boldsymbol{\xi}^*), \end{aligned} \tag{4.20}$$

as asymptotic bias, an asymptotic bias correction of $-\sum_{i=1}^{n} \log f(y_i, \boldsymbol{x}_i \mid \boldsymbol{y}, \boldsymbol{X}; \boldsymbol{\lambda})$ can be obtained. Subsequently, the WAIC is given by

$$\begin{aligned}\mathrm{WAIC} = -\sum_{i=1}^{n} \log f(y_i, \boldsymbol{x}_i \mid \boldsymbol{y}, \boldsymbol{X}; \boldsymbol{\lambda}) + \sum_{i=1}^{n} \mathrm{E}_{\boldsymbol{\xi}\mid\boldsymbol{y},\boldsymbol{X};\boldsymbol{\lambda}}[\log f(y_i, \boldsymbol{x}_i \mid \boldsymbol{\xi})^2] \\ - \sum_{i=1}^{n} \left\{\mathrm{E}_{\boldsymbol{\xi}\mid\boldsymbol{y},\boldsymbol{X};\boldsymbol{\lambda}}[\log f(y_i, \boldsymbol{x}_i \mid \boldsymbol{\xi})]\right\}^2.\end{aligned}$$

As reviewed in Sec. 2.3.8, while WAIC provides a general framework for Bayesian model comparison, it faces two issues in the context of SVC models with Bayesian regularization. It insufficiently reflects the effect of the prior distribution in sparse settings, and it always selects models with prior distributions containing more hyperparameters when comparing models, such as those discussed in Section 4.1.3, whose prior distributions belong to classes with different degrees of freedom.

To solve the first problem, the prior distribution in PIIC was assumed to depend on n. Described concretely using the $\pi_{\mathrm{gfl}}(\boldsymbol{\xi}; \boldsymbol{\lambda})$ term defined in (4.19), PIIC assumes $\pi_n(\boldsymbol{\xi}; \boldsymbol{\lambda}) \propto \pi_{\mathrm{gfl}}(\boldsymbol{\xi}; \boldsymbol{\lambda})$, whereas WAIC assumes $\pi_n(\boldsymbol{\xi}; \boldsymbol{\lambda}) \propto \pi_{\mathrm{gfl}}(\boldsymbol{\xi}; \boldsymbol{\lambda})^{1/n}$. In addition, to solve the second problem, a penalty term for hyperparameters was added. However, since Ninomiya (2021) did not consider the Bayesian generalized fused lasso, we developed PIIC for the SVC model with the Bayesian generalized fused lasso, as will be presented in this section.

PIIC is the asymptotic bias-corrected statistic of $-\sum_{i=1}^{n} \log f(y_i, \boldsymbol{x}_i \mid \boldsymbol{y}, \boldsymbol{X}; \boldsymbol{\lambda})$, sharing the same concept as WAIC, and it asymptotically evaluates the expectation of (4.20). Here, we avoid a discussion on moment convergence, and define the expectation of the weak limit of (4.20) as the asymptotic bias. With that in mind, we henceforth asymptotically evaluate (4.20). To proceed simultaneously with the expansion for $(y_i, \boldsymbol{x}_i)$ and the expansion for $(\tilde{y}_i, \tilde{\boldsymbol{x}}_i)$, we adopt the notation $(\breve{y}_i, \breve{\boldsymbol{x}}_i)$, which denotes either $(y_i, \boldsymbol{x}_i)$ or $(\tilde{y}_i, \tilde{\boldsymbol{x}}_i)$. The predictive distribution $\log f(y_i, \boldsymbol{x}_i \mid \boldsymbol{y}, \boldsymbol{X}; \boldsymbol{\lambda})$ or $\log f(\tilde{y}_i, \tilde{\boldsymbol{x}}_i \mid \boldsymbol{y}, \boldsymbol{X}; \boldsymbol{\lambda})$ in (4.20) is expressed as

$$\log \int f(\breve{y}_i, \breve{\boldsymbol{x}}_i \mid \boldsymbol{\xi}) f(\boldsymbol{y}, \boldsymbol{X} \mid \boldsymbol{\xi}) \pi_{\mathrm{gfl}}(\boldsymbol{\xi}; \boldsymbol{\lambda}) \mathrm{d}\boldsymbol{\xi} - \log \int f(\boldsymbol{y}, \boldsymbol{X} \mid \boldsymbol{\xi}) \pi_{\mathrm{gfl}}(\boldsymbol{\xi}; \boldsymbol{\lambda}) \mathrm{d}\boldsymbol{\xi}.$$

For these two integrals, by applying higher-order Laplace approximation (Tierney and Kadane 1986), we can express them as follows using the function $h(y, \boldsymbol{x} \mid \boldsymbol{\xi})$:

$$\begin{aligned}&\log f(\breve{y}_i, \breve{\boldsymbol{x}}_i \mid \boldsymbol{y}, \boldsymbol{X}; \boldsymbol{\lambda}) \\ &= \log f(\breve{y}_i, \breve{\boldsymbol{x}}_i \mid \hat{\boldsymbol{\xi}}) + \frac{1}{n} h(\breve{y}_i, \breve{\boldsymbol{x}}_i \mid \hat{\boldsymbol{\xi}}) + \mathrm{o_P}\left(\frac{1}{n}\right) = \log f(\breve{y}_i, \breve{\boldsymbol{x}}_i \mid \hat{\boldsymbol{\xi}}) + \mathrm{O_P}\left(\frac{1}{n}\right).\end{aligned} \tag{4.21}$$

As a result, (4.20) can be rewritten as

$$\sum_{i=1}^{n} \log f(y_i, \boldsymbol{x}_i \mid \hat{\boldsymbol{\xi}}) + \frac{1}{n} \sum_{i=1}^{n} h(y_i, \boldsymbol{x}_i \mid \hat{\boldsymbol{\xi}}) - \sum_{i=1}^{n} \log f(y_i, \boldsymbol{x}_i \mid \boldsymbol{\xi}^*)$$
$$- \sum_{i=1}^{n} \log f(\tilde{y}_i, \tilde{\boldsymbol{x}}_i \mid \hat{\boldsymbol{\xi}}) - \frac{1}{n} \sum_{i=1}^{n} h(\tilde{y}_i, \tilde{\boldsymbol{x}}_i \mid \hat{\boldsymbol{\xi}}) + \sum_{i=1}^{n} \log f(\tilde{y}_i, \tilde{\boldsymbol{x}}_i \mid \boldsymbol{\xi}^*) + \mathrm{o_P}(1). \tag{4.22}$$

Since the expectation of $h(y_i, \boldsymbol{x}_i \mid \hat{\boldsymbol{\xi}})$ and the expectation of $h(\tilde{y}_i, \tilde{\boldsymbol{x}}_i \mid \hat{\boldsymbol{\xi}})$ are asymptotically equal, the sum of the second and fifth terms of (4.22) is $\mathrm{o_P}(1)$. In addition, when we perform a Taylor expansion on the sum of the first, third, fourth, and sixth terms, the terms involving $\hat{\boldsymbol{\xi}}^{(1)}$ become $\mathrm{o_P}(1)$ from the first equation in Theorem 1, $n\hat{\boldsymbol{\xi}}^{(1)} = \mathrm{o_P}(1)$, and therefore the sum can be written as

$$\sqrt{n}(\hat{\boldsymbol{\xi}}^{(2)} - \boldsymbol{\xi}^{*(2)})^{\mathrm{T}} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left\{ \frac{\partial}{\partial \boldsymbol{\xi}^{(2)}} \log f(y_i, \boldsymbol{x}_i \mid \boldsymbol{\xi}^*) - \frac{\partial}{\partial \boldsymbol{\xi}^{(2)}} \log f(\tilde{y}_i, \tilde{\boldsymbol{x}}_i \mid \boldsymbol{\xi}^*) \right\} + \mathrm{o_P}(1). \tag{4.23}$$

Here, note that the second-order terms of the Taylor expansion cancel each other out, between those from $f(y_i, \boldsymbol{x}_i \mid \boldsymbol{\xi}^*)$ and those from $f(\tilde{y}_i, \tilde{\boldsymbol{x}}_i \mid \boldsymbol{\xi}^*)$. Removing $\mathrm{o_P}(1)$, we can rewrite this as

$$\sum_{j \in \mathcal{J}^{(2)}} \sqrt{n}(\hat{\xi}_{j^\dagger} - \xi^*_{j^\dagger}) \left[\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left\{ \frac{\partial}{\partial \xi_j} \log f(y_i, \boldsymbol{x}_i \mid \boldsymbol{\xi}^*) - \frac{\partial}{\partial \xi_j} \log f(\tilde{y}_i, \tilde{\boldsymbol{x}}_i \mid \boldsymbol{\xi}^*) \right\} \right]$$
$$= \sum_{j^\dagger \in \mathcal{J}^{(3)}} \sum_{j^\ddagger : \xi^*_{j^\ddagger} = \xi^*_{j^\dagger}} \sqrt{n}(\hat{\xi}_{j^\ddagger} - \xi^*_{j^\ddagger}) \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left\{ \frac{\partial}{\partial \xi_{j^\ddagger}} \log f(y_i, \boldsymbol{x}_i \mid \boldsymbol{\xi}^*) - \frac{\partial}{\partial \xi_{j^\ddagger}} \log f(\tilde{y}_i, \tilde{\boldsymbol{x}}_i \mid \boldsymbol{\xi}^*) \right\}.$$

Then, applying the second equation of Theorem 1, we ultimately find that (4.23) becomes

$$\sum_{j^\dagger \in \mathcal{J}^{(3)}} \sqrt{n}(\hat{\xi}_{j^\dagger} - \xi^*_{j^\dagger}) \sum_{j^\ddagger : \xi^*_{j^\ddagger} = \xi^*_{j^\dagger}} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left\{ \frac{\partial}{\partial \xi_{j^\ddagger}} \log f(y_i, \boldsymbol{x}_i \mid \boldsymbol{\xi}^*) - \frac{\partial}{\partial \xi_{j^\ddagger}} \log f(\tilde{y}_i, \tilde{\boldsymbol{x}}_i \mid \boldsymbol{\xi}^*) \right\}$$
$$+ \mathrm{o_P}(1).$$

Let us now consider the weak limit of the above expression. First, from (R3), there exist $\boldsymbol{s}^{(2)}$ and its copy $\tilde{\boldsymbol{s}}^{(2)}$, which are $|\mathcal{J}^{(2)}|$-dimensional sub-vectors of a random vector $\boldsymbol{s}$ following $\mathrm{N}(\boldsymbol{0}, \boldsymbol{J})$ and themselves follow $\mathrm{N}(\boldsymbol{0}, \boldsymbol{J}^{(22)})$. Then, it holds that

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left\{ \frac{\partial}{\partial \boldsymbol{\xi}^{(2)}} \log f(y_i, \boldsymbol{x}_i \mid \boldsymbol{\xi}^*) - \frac{\partial}{\partial \boldsymbol{\xi}^{(2)}} \log f(\tilde{y}_i, \tilde{\boldsymbol{x}}_i \mid \boldsymbol{\xi}^*) \right\} \xrightarrow{\mathrm{d}} \boldsymbol{s}^{(2)} + \tilde{\boldsymbol{s}}^{(2)}.$$

Here, we are using the fact that the expectations of the first and second terms inside the sum are equal. Therefore, using the matrix $\boldsymbol{A}$ defined in (4.4), we

obtain

$$\left\{ \sum_{j^\ddagger:\xi^*_{j^\ddagger}=\xi^*_{j^\dagger}} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left\{ \frac{\partial}{\partial \xi_{j^\ddagger}} \log f(y_i, \boldsymbol{x}_i \mid \boldsymbol{\xi}^*) - \frac{\partial}{\partial \xi_{j^\ddagger}} \log f(\tilde{y}_i, \tilde{\boldsymbol{x}}_i \mid \boldsymbol{\xi}^*) \right\} \right\}_{j^\dagger \in \mathcal{J}^{(3)}}$$
$$\xrightarrow{\mathrm{d}} \boldsymbol{A}^{(32)}(\boldsymbol{s}^{(2)} + \tilde{\boldsymbol{s}}^{(2)}).$$

In addition, from (4.17) in Sec. 4.1.2, which is the basis of the third equation in Theorem 1, it is known that $\sqrt{n}(\hat{\xi}_{j^\dagger} - \xi^*_{j^\dagger})_{j^\dagger \in \mathcal{J}^{(3)}}$ converges in distribution to $(\boldsymbol{A}^{(32)}\boldsymbol{J}^{(22)}\boldsymbol{A}^{(23)})^{-1}\boldsymbol{A}^{(32)}\boldsymbol{s}^{(2)}$. As can be imagined from these, (4.23) converges in distribution to

$$\boldsymbol{s}^{(2)\mathrm{T}}\boldsymbol{A}^{(23)}(\boldsymbol{A}^{(32)}\boldsymbol{J}^{(22)}\boldsymbol{A}^{(23)})^{-1}\boldsymbol{A}^{(32)}(\boldsymbol{s}^{(2)} + \tilde{\boldsymbol{s}}^{(2)}). \tag{4.24}$$

The expectation of this is

$$\begin{aligned}
&\mathrm{E}[\mathrm{tr}\{(\boldsymbol{A}^{(32)}\boldsymbol{J}^{(22)}\boldsymbol{A}^{(23)})^{-1}\boldsymbol{A}^{(32)}(\boldsymbol{s}^{(2)} + \tilde{\boldsymbol{s}}^{(2)})\boldsymbol{s}^{(2)\mathrm{T}}\boldsymbol{A}^{(23)}\}] \\
&= \mathrm{tr}\{(\boldsymbol{A}^{(32)}\boldsymbol{J}^{(22)}\boldsymbol{A}^{(23)})^{-1}\boldsymbol{A}^{(32)}\mathrm{E}[\boldsymbol{s}^{(2)}\boldsymbol{s}^{(2)\mathrm{T}}]\boldsymbol{A}^{(23)}\} = \mathrm{tr}(\boldsymbol{I}_{|\mathcal{J}^{(3)}|}) = |\mathcal{J}^{(3)}|,
\end{aligned}$$

and therefore the following theorem holds.

Theorem 2. Under conditions (C1) and (C2), the asymptotic bias of $-\sum_{i=1}^{n} \log f(y_i, \boldsymbol{x}_i \mid \boldsymbol{y}, \boldsymbol{X}; \boldsymbol{\lambda})$ with respect to $-\sum_{i=1}^{n} \mathrm{E}_{\tilde{y}_i, \tilde{\boldsymbol{x}}_i}[\log f(\tilde{y}_i, \tilde{\boldsymbol{x}}_i \mid \boldsymbol{y}, \boldsymbol{X}; \boldsymbol{\lambda})]$ is $|\mathcal{J}^{(3)}|$.

Based on Theorem 2, we propose the following information criterion:

$$\mathrm{PIIC1} = -\sum_{i=1}^{n} \log f(y_i, \boldsymbol{x}_i \mid \boldsymbol{y}, \boldsymbol{X}; \boldsymbol{\lambda}) + |\hat{\mathcal{J}}^{(3)}|.$$

Here, for the graph consisting of the vertex set $\{j \in V : \hat{\xi}_j \neq 0\}$ and the edge set $\{(j^\dagger, j^\ddagger) \in E : \hat{\xi}_{j^\dagger} = \hat{\xi}_{j^\ddagger}\}$, $\hat{\mathcal{J}}^{(3)}$ is the collection obtained by arbitrarily choosing one vertex as a representative from each connected component. It follows trivially that $|\hat{\mathcal{J}}^{(3)}|$ is a consistent estimator of $|\mathcal{J}^{(3)}|$. For the selection of $\boldsymbol{\lambda}$, it suffices to use a $\hat{\boldsymbol{\lambda}}$ that minimizes PIIC1.

Unlike WAIC, PIIC1 incorporates the influence of a prior distribution that introduces sparsity to the Bayesian generalized fused lasso estimator. However, as with WAIC, PIIC1 always selects the model with the larger dimension of $\boldsymbol{\lambda}$ when comparing models that use prior distributions with hyperparameters $\boldsymbol{\lambda}$ of different dimensions. Therefore, by making a bias evaluation that also takes into account the fact that $\boldsymbol{\lambda}$ is selected from the data and adding a penalty term that affects the selection of the dimension of hyperparameters, it is made

possible to perform appropriate model selection even when there are candidate prior distributions with different complexities. Since the selection of $\boldsymbol{\lambda}$ becomes the topic, in the following part of this section, we attach $\boldsymbol{\lambda}$ to estimators and their limits, denoting them by $\hat{\boldsymbol{\xi}}_{\boldsymbol{\lambda}}$ and $\boldsymbol{\xi}^*_{\boldsymbol{\lambda}}$, respectively. In addition, as regularity conditions, we add the following, with $\boldsymbol{\lambda}^*$ being the minimizer with respect to $\boldsymbol{\lambda}$ of $-\mathrm{E}[\log f(y, \boldsymbol{x} \mid \boldsymbol{\xi}^*_{\boldsymbol{\lambda}})]$.

(C3) $\boldsymbol{\lambda}^*$ exists in the interior of the compact set Λ, which is the parameter space of $\boldsymbol{\lambda}$, and $\log f(y, \boldsymbol{x} \mid \boldsymbol{\xi}^*_{\boldsymbol{\lambda}})$ is of class C^2 in a neighborhood of $\boldsymbol{\lambda} = \boldsymbol{\lambda}^*$.

In view of the fact that PIIC1 can be written as

$$
\begin{aligned}
&-\sum_{i=1}^{n} \log f(y_i, \boldsymbol{x}_i \mid \boldsymbol{y}, \boldsymbol{X}; \boldsymbol{\lambda})\{1 + \mathrm{o_P}(1)\} \\
&= -\sum_{i=1}^{n} \log f(y_i, \boldsymbol{x}_i \mid \hat{\boldsymbol{\xi}}_{\boldsymbol{\lambda}})\{1 + \mathrm{o_P}(1)\} = -\sum_{i=1}^{n} \log f(y_i, \boldsymbol{x}_i \mid \boldsymbol{\xi}^*_{\boldsymbol{\lambda}})\{1 + \mathrm{o_P}(1)\}
\end{aligned}
$$

from (4.21), letting $\tilde{\boldsymbol{\lambda}}$ be the minimizer with respect to $\boldsymbol{\lambda}$ of $-\sum_{i=1}^{n} \log f(y_i, \boldsymbol{x}_i \mid \boldsymbol{\xi}^*_{\boldsymbol{\lambda}})$ and taking $-\sum_{i=1}^{n} \mathrm{E}_{\tilde{y}_i, \tilde{\boldsymbol{x}}_i}[\log f(\tilde{y}_i, \tilde{\boldsymbol{x}}_i \mid \boldsymbol{y}, \boldsymbol{X}; \tilde{\boldsymbol{\lambda}})]$ as the target, we substitute $\tilde{\boldsymbol{\lambda}}$ for $\boldsymbol{\lambda}$ in the first and third terms of (4.20) as in

$$
\begin{aligned}
&\sum_{i=1}^{n} \log f(y_i, \boldsymbol{x}_i \mid \boldsymbol{y}, \boldsymbol{X}; \tilde{\boldsymbol{\lambda}}) - \sum_{i=1}^{n} \log f(y_i, \boldsymbol{x}_i \mid \boldsymbol{\xi}^*_{\boldsymbol{\lambda}^*}) \\
&- \sum_{i=1}^{n} \log f(\tilde{y}_i, \tilde{\boldsymbol{x}}_i \mid \boldsymbol{y}, \boldsymbol{X}; \tilde{\boldsymbol{\lambda}}) - \sum_{i=1}^{n} \log f(\tilde{y}_i, \tilde{\boldsymbol{x}}_i \mid \boldsymbol{\xi}^*_{\boldsymbol{\lambda}^*}),
\end{aligned} \tag{4.25}
$$

and asymptotically evaluate this to derive the information criterion. First, since $\tilde{\boldsymbol{\lambda}} - \boldsymbol{\lambda}^* = \mathrm{o_P}(1)$ from the uniform law of large numbers, we perform a Taylor expansion of $\mathbf{0} = \sum_{i=1}^{n} \partial \log f(y_i, \boldsymbol{x}_i \mid \boldsymbol{\xi}^*_{\tilde{\boldsymbol{\lambda}}})/\partial \boldsymbol{\lambda}$ around $\tilde{\boldsymbol{\lambda}} = \boldsymbol{\lambda}^*$. Then, letting

$$
\boldsymbol{J}_1(\boldsymbol{\lambda}) \equiv \mathrm{E}\bigg[-\frac{\partial^2}{\partial \boldsymbol{\lambda} \partial \boldsymbol{\lambda}^{\mathrm{T}}} \log f(y_i, \boldsymbol{x}_i \mid \boldsymbol{\xi}^*_{\boldsymbol{\lambda}}) \bigg],
$$

and

$$
\boldsymbol{J}_2(\boldsymbol{\lambda}) \equiv \mathrm{E}\bigg[\bigg\{ \frac{\partial}{\partial \boldsymbol{\lambda}} \log f(y_i, \boldsymbol{x}_i \mid \boldsymbol{\xi}^*_{\boldsymbol{\lambda}}) \bigg\} \bigg\{ \frac{\partial}{\partial \boldsymbol{\lambda}} \log f(y_i, \boldsymbol{x}_i \mid \boldsymbol{\xi}^*_{\boldsymbol{\lambda}}) \bigg\}^{\mathrm{T}} \bigg],
$$

we obtain

$$
\tilde{\boldsymbol{\lambda}} - \boldsymbol{\lambda}^* = \boldsymbol{J}_1(\boldsymbol{\lambda}^*)^{-1} \frac{1}{n} \sum_{i=1}^{n} \frac{\partial}{\partial \boldsymbol{\lambda}} \log f(y_i, \boldsymbol{x}_i \mid \boldsymbol{\xi}^*_{\boldsymbol{\lambda}^*})\{1 + \mathrm{o_P}(1)\}, \tag{4.26}
$$

where $\boldsymbol{J}_2(\boldsymbol{\lambda})$ has been defined for later use. When (4.25) is asymptotically evaluated using (4.21) and (4.26), what is obtained is an expression in which $\boldsymbol{\lambda}^*$ is

substituted for $\boldsymbol{\lambda}$ in (4.20), with the addition of

$$\left\{\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{\partial}{\partial\boldsymbol{\lambda}^{\mathrm{T}}}\log f(y_i,\boldsymbol{x}_i\mid\boldsymbol{\xi}^*_{\boldsymbol{\lambda}^*})\right\}\boldsymbol{J}_1(\boldsymbol{\lambda}^*)^{-1}\left\{\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{\partial}{\partial\boldsymbol{\lambda}}\log f(y_i,\boldsymbol{x}_i\mid\boldsymbol{\xi}^*_{\boldsymbol{\lambda}^*})\right\}\{1+\mathrm{o_P}(1)\}$$
$$-\left\{\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{\partial}{\partial\boldsymbol{\lambda}^{\mathrm{T}}}\log f(\tilde{y}_i,\tilde{\boldsymbol{x}}_i\mid\boldsymbol{\xi}^*_{\boldsymbol{\lambda}^*})\right\}\boldsymbol{J}_1(\boldsymbol{\lambda}^*)^{-1}\left\{\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{\partial}{\partial\boldsymbol{\lambda}}\log f(y_i,\boldsymbol{x}_i\mid\boldsymbol{\xi}^*_{\boldsymbol{\lambda}^*})\right\}\{1+\mathrm{o_P}(1)\}.$$

Therefore, letting $\boldsymbol{s}_1$ and $\boldsymbol{s}_2$ be random vectors independently following $\mathrm{N}(\boldsymbol{0},\boldsymbol{J}_2(\boldsymbol{\lambda}^*))$, we find that the weak limit of (4.25) is that of (4.24) with the addition of

$$\boldsymbol{s}_1^{\mathrm{T}}\boldsymbol{J}_1(\boldsymbol{\lambda}^*)^{-1}\boldsymbol{s}_1-\boldsymbol{s}_2^{\mathrm{T}}\boldsymbol{J}_1(\boldsymbol{\lambda}^*)^{-1}\boldsymbol{s}_1.$$

The expectation of the weak limit is

$$|\mathcal{J}^{(3)}|+\mathrm{tr}\{\boldsymbol{J}_1(\boldsymbol{\lambda}^*)^{-1}\mathrm{E}[\boldsymbol{s}_1\boldsymbol{s}_1^{\mathrm{T}}]\}+\mathrm{E}[\boldsymbol{s}_2^{\mathrm{T}}]\boldsymbol{J}_1(\boldsymbol{\lambda}^*)^{-1}\mathrm{E}[\boldsymbol{s}_1]=|\mathcal{J}^{(3)}|+\mathrm{tr}\{\boldsymbol{J}_1(\boldsymbol{\lambda}^*)^{-1}\boldsymbol{J}_2(\boldsymbol{\lambda}^*)\},$$

and therefore the following theorem holds.

Theorem 3. Under conditions (C1), (C2), and (C3), the asymptotic bias of $-\sum_{i=1}^{n}\log f(y_i,\boldsymbol{x}_i\mid\boldsymbol{y},\boldsymbol{X},\tilde{\boldsymbol{\lambda}})$ with respect to $-\sum_{i=1}^{n}\mathrm{E}_{\tilde{y}_i,\tilde{\boldsymbol{x}}_i}[\log f(\tilde{y}_i,\tilde{\boldsymbol{x}}_i\mid\boldsymbol{y},\boldsymbol{X};\tilde{\boldsymbol{\lambda}})]$ is $|\mathcal{J}^{(3)}|+\mathrm{tr}\{\boldsymbol{J}_1(\boldsymbol{\lambda}^*)^{-1}\boldsymbol{J}_2(\boldsymbol{\lambda}^*)\}$.

In evaluating $\boldsymbol{J}_1(\boldsymbol{\lambda}^*)$ and $\boldsymbol{J}_2(\boldsymbol{\lambda}^*)$ of Theorem 3, based on (4.21), we use

$$\hat{\boldsymbol{J}}_1(\hat{\boldsymbol{\lambda}})\equiv-\frac{1}{n}\sum_{i=1}^{n}\frac{\partial^2}{\partial\boldsymbol{\lambda}\partial\boldsymbol{\lambda}^{\mathrm{T}}}\log f(y_i,\boldsymbol{x}_i\mid\boldsymbol{y},\boldsymbol{X};\hat{\boldsymbol{\lambda}})$$

and

$$\hat{\boldsymbol{J}}_2(\hat{\boldsymbol{\lambda}})\equiv\frac{1}{n}\sum_{i=1}^{n}\frac{\partial}{\partial\boldsymbol{\lambda}}\log f(y_i,\boldsymbol{x}_i\mid\boldsymbol{y},\boldsymbol{X};\hat{\boldsymbol{\lambda}})\frac{\partial}{\partial\boldsymbol{\lambda}^{\mathrm{T}}}\log f(y_i,\boldsymbol{x}_i\mid\boldsymbol{y},\boldsymbol{X};\hat{\boldsymbol{\lambda}}).$$

As a result, we propose the following information criterion with an added penalty term for hyperparameters:

$$\mathrm{PIIC2}\equiv-\sum_{i=1}^{n}\log f(y_i,\boldsymbol{x}_i\mid\boldsymbol{y},\boldsymbol{X};\hat{\boldsymbol{\lambda}})+|\hat{\mathcal{J}}^{(3)}|+\mathrm{tr}\{\hat{\boldsymbol{J}}_1(\hat{\boldsymbol{\lambda}})^{-1}\hat{\boldsymbol{J}}_2(\hat{\boldsymbol{\lambda}})\}.$$

4.3 Numerical experiments

In order to evaluate how the developed method performs compared with the existing method, we conducted numerical experiments as follows. In the SVC model (4.1), we set $\tilde{p}=3$, consider $\boldsymbol{\theta}_{[j]}=(\theta_{1,j},\ldots,\theta_{M,j})$ $(j=1,2,3)$, and appropriately specify $\boldsymbol{\theta}$ with them, independently generate $\psi_i\sim\mathrm{Multi}(1,\boldsymbol{p}_\psi)$, $\tilde{x}_{i,j}\sim\mathrm{N}(0,5^2)$ and $\varepsilon_i\sim\mathrm{N}(0,\sigma^2)$, and then construct y_i.

Here, σ^2 is taken to be 1.0, 1.5, or 2.0. In this numerical experiment, the data are generated without including an intercept term, and the estimation of the intercept is not performed. Moreover, in order to focus on evaluating the performance of information criteria for the regularization to fuse adjacent regression coefficients, we do not impose the regularization for variable selection that shrinks regression coefficients to zero. As the graph $(\mathcal{V}, \mathcal{E})$ representing the regions and their adjacency in the SVC model, we consider Graphs 1–8 in Figure 4.1, and the experiments using these are referred to as Cases 1–8, respectively. Two regions connected by an edge are regarded as adjacent. For each of Cases 1–8, we consider two settings for the sample size and the true values of the parameters, which are denoted by Setting 1 and Setting 2. Specifically, the detailed setups for Cases 1–4 are summarized in Table 4.1, for Cases 5 and 6 in Table 4.2, and for Cases 7 and 8 in Table 4.3.

Table 4.1: Simulation setups for Cases 1–4.

	Case 1		Case 2	
	Setting 1	Setting 2	Setting 1	Setting 2
n	20	35	45	50
M	3		5	
$\boldsymbol{p}_\psi$	$(1/3, 1/3, 1/3)$	$(1/3, 1/6, 1/2)$	$(1/5, \ldots, 1/5)$	$(1/10, 3/10, 2/10, 2/10, 2/10)$
$\boldsymbol{\theta}_{[1]}$	$(1.0, 1.0, 1.0)^{\mathrm{T}}$		$(1.0, 1.0, 1.0, 1.0, 1.0)^{\mathrm{T}}$	
$\boldsymbol{\theta}_{[2]}$	$(2.0, -2.0, -3.5)^{\mathrm{T}}$		$(2.0, 2.0, 1.5, 2.5, 1.5)^{\mathrm{T}}$	
$\boldsymbol{\theta}_{[3]}$	$(3.0, -3.0, 1.5)^{\mathrm{T}}$		$(3.0, -2.5, -3.0, 0.5, -0.5)^{\mathrm{T}}$	
	Case 3		Case 4	
	Setting 1	Setting 2	Setting 1	Setting 2
n	45	50	45	50
M	5		5	
$\boldsymbol{p}_\psi$	$(1/5, \ldots, 1/5)$	$(1/10, 3/10, 2/10, 2/10, 2/10)$	$(1/5, \ldots, 1/5)$	$(1/10, 2/10, 1/10, 2/10, 4/10)$
$\boldsymbol{\theta}_{[1]}$	$(1.0, 1.0, 1.0, -1.5, -1.5)^{\mathrm{T}}$		$(1.0, 1.0, -1.5, -1.5, 5.0)^{\mathrm{T}}$	
$\boldsymbol{\theta}_{[2]}$	$(2.0, -2.0, -3.5, 0.5, -0.5)^{\mathrm{T}}$		$(2.0, -2.0, -2.5, -0.5, -3.5)^{\mathrm{T}}$	
$\boldsymbol{\theta}_{[3]}$	$(3.0, -3.0, 1.5, -1, -2.5)^{\mathrm{T}}$		$(3.0, -3.0, -1.0, 0.5, -5.0)^{\mathrm{T}}$	

Table 4.2: Simulation setups for Cases 5 and 6. $\boldsymbol{a}_m$ denotes an m-dimensional vector with all components a.

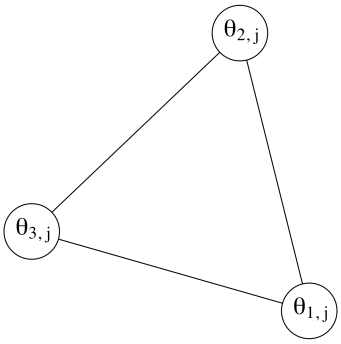
	Case 5	
	Setting 1	Setting 2
n	400	
M	50	
$\boldsymbol{p}_\psi$	$(1/50, \ldots, 1/50)$	
$\boldsymbol{\theta}_{[1]}$	$\mathbf{1.0}_{50}^{\mathrm{T}}$	$\boldsymbol{\theta}_{[1]} = A_{\beta_1} z_{\beta_1},\ A_{\beta_1} = \mathbf{1}_{50},\ z_{\beta_1} \sim \mathrm{N}(0, 3)$
$\boldsymbol{\theta}_{[2]}$	$(\mathbf{-2.0}_{10}^{\mathrm{T}}, \mathbf{-1.5}_{10}^{\mathrm{T}}, \mathbf{-1.0}_{10}^{\mathrm{T}}, \mathbf{0.5}_{10}^{\mathrm{T}}, \mathbf{1.5}_{10}^{\mathrm{T}})^{\mathrm{T}}$	$\boldsymbol{\theta}_{[2]} = A_{\beta_2} z_{\beta_2},\ A_{\beta_2} = \mathbf{1}_{50},\ z_{\beta_2} \sim \mathrm{N}(0, 3)$
$\boldsymbol{\theta}_{[3]}$	$(\mathbf{-3.0}_5^{\mathrm{T}}, \mathbf{-2.5}_5^{\mathrm{T}}, \mathbf{-2.0}_5^{\mathrm{T}}, \mathbf{-1.5}_5^{\mathrm{T}}, \mathbf{-1.0}_5^{\mathrm{T}}, \mathbf{0.5}_5^{\mathrm{T}}, \mathbf{1.5}_5^{\mathrm{T}}, \mathbf{2.0}_5^{\mathrm{T}}, \mathbf{2.5}_5^{\mathrm{T}}, \mathbf{3.0}_5^{\mathrm{T}})^{\mathrm{T}}$	$\boldsymbol{\theta}_{[3]} = z_{\beta_3},\ z_{\beta_3} \sim \mathrm{N}(\mathbf{0}, \Sigma_{\beta_3}),\ \Sigma_{\beta_3} = \underbrace{\begin{pmatrix} 5 & 1.5 & \cdots & 1.5 \\ 1.5 & 5 & \cdots & 1.5 \\ \vdots & \vdots & \ddots & \vdots \\ 1.5 & 1.5 & \cdots & 5 \end{pmatrix}}_{50 \times 50}$

	Case 6	
	Setting 1	Setting 2
n	400	
M	36	
$\boldsymbol{p}_\psi$	$(1/36, \ldots, 1/36)$	
$\boldsymbol{\theta}_{[1]}$	$\mathbf{2.0}_{36}^{\mathrm{T}}$	$\boldsymbol{\theta}_{[1]} = A_{\beta_1} z_{\beta_1},\ A_{\beta_1} = \mathbf{1}_{50},\ z_{\beta_1} \sim \mathrm{N}(0, 3)$
$\boldsymbol{\theta}_{[2]}$	$(\mathbf{-2.0}_2^{\mathrm{T}}, \mathbf{-1.5}_2^{\mathrm{T}}, \mathbf{-1.0}_2^{\mathrm{T}}, \mathbf{-2.0}_2^{\mathrm{T}}, \mathbf{-1.5}_2^{\mathrm{T}}, \mathbf{-1.0}_2^{\mathrm{T}}, \mathbf{-2.0}_2^{\mathrm{T}}, \mathbf{-1.5}_2^{\mathrm{T}}, \mathbf{-1.0}_2^{\mathrm{T}}, \mathbf{0.5}_2^{\mathrm{T}}, \mathbf{1.5}_2^{\mathrm{T}}, \mathbf{2.0}_2^{\mathrm{T}}, \mathbf{0.5}_2^{\mathrm{T}}, \mathbf{1.5}_2^{\mathrm{T}}, \mathbf{2.0}_2^{\mathrm{T}}, \mathbf{0.5}_2^{\mathrm{T}}, \mathbf{1.5}_2^{\mathrm{T}}, \mathbf{2.0}_2^{\mathrm{T}})^{\mathrm{T}}$	$\boldsymbol{\theta}_{[2]} = A_{\beta_2} z_{\beta_2},\ A_{\beta_2} = \mathbf{1}_{50},\ z_{\beta_2} \sim \mathrm{N}(0, 3)$
$\boldsymbol{\theta}_{[3]}$	$(\mathbf{-3.0}_2^{\mathrm{T}}, \mathbf{-2.5}_2^{\mathrm{T}}, \mathbf{-2.0}_2^{\mathrm{T}}, \mathbf{-3.0}_2^{\mathrm{T}}, \mathbf{-2.5}_2^{\mathrm{T}}, \mathbf{-2.0}_2^{\mathrm{T}}, \mathbf{-1.5}_2^{\mathrm{T}}, \mathbf{-1.0}_2^{\mathrm{T}}, \mathbf{0.5}_2^{\mathrm{T}}, \mathbf{-1.5}_2^{\mathrm{T}}, \mathbf{-1.0}_2^{\mathrm{T}}, \mathbf{0.5}_2^{\mathrm{T}}, \mathbf{1.0}_2^{\mathrm{T}}, \mathbf{2.0}_2^{\mathrm{T}}, \mathbf{2.5}_2^{\mathrm{T}}, \mathbf{1.0}_2^{\mathrm{T}}, \mathbf{2.0}_2^{\mathrm{T}}, \mathbf{2.5}_2^{\mathrm{T}})^{\mathrm{T}}$	$\boldsymbol{\theta}_{[3]} = z_{\beta_3},\ z_{\beta_3} \sim \mathrm{N}(\mathbf{0}, \Sigma_{\beta_3}),\ \Sigma_{\beta_3} = \underbrace{\begin{pmatrix} 5 & 1.5 & \cdots & 1.5 \\ 1.5 & 5 & \cdots & 1.5 \\ \vdots & \vdots & \ddots & \vdots \\ 1.5 & 1.5 & \cdots & 5 \end{pmatrix}}_{36 \times 36}$

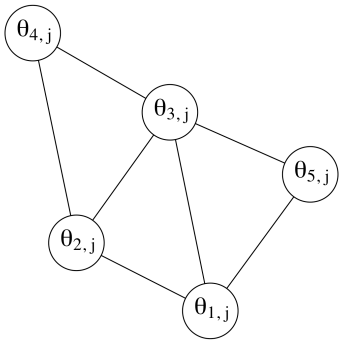
Table 4.3: Simulation setups for Cases 7 and 8. $\boldsymbol{a}_m$ denotes an m-dimensional vector with all components a.

	Case 7	
	Setting 1	Setting 2
n		400
M		36
$\boldsymbol{p}_\psi$		$(1/36, \ldots, 1/36)$
$\boldsymbol{\theta}_{[1]}$	$(\boldsymbol{-2.0}_{18}^{\mathrm{T}}, \boldsymbol{2.0}_{18}^{\mathrm{T}})^{\mathrm{T}}$	$\boldsymbol{\theta}_{[1]} = A_{\beta_1} z_{\beta_1},\ A_{\beta_1} = \begin{pmatrix} \mathbf{1}_{18} & \mathbf{0}_{18} \\ \mathbf{0}_{18} & \mathbf{1}_{18} \end{pmatrix},\ z_{\beta_1} \sim \mathrm{N}(0, 3I_2)$
$\boldsymbol{\theta}_{[2]}$	$(\boldsymbol{-2.0}_2^{\mathrm{T}}, \boldsymbol{-1.5}_2^{\mathrm{T}}, \boldsymbol{-1.0}_2^{\mathrm{T}}, \boldsymbol{-2.0}_2^{\mathrm{T}}, \boldsymbol{-1.5}_2^{\mathrm{T}}, \boldsymbol{-1.0}_2^{\mathrm{T}}, \boldsymbol{-2.0}_2^{\mathrm{T}}, \boldsymbol{-1.5}_2^{\mathrm{T}}, \boldsymbol{-1.0}_2^{\mathrm{T}}, \boldsymbol{1.0}_2^{\mathrm{T}}, \boldsymbol{1.5}_2^{\mathrm{T}}, \boldsymbol{2.0}_2^{\mathrm{T}}, \boldsymbol{1.0}_2^{\mathrm{T}}, \boldsymbol{1.5}_2^{\mathrm{T}}, \boldsymbol{2.0}_2^{\mathrm{T}}, \boldsymbol{1.0}_2^{\mathrm{T}}, \boldsymbol{1.5}_2^{\mathrm{T}}, \boldsymbol{2.0}_2^{\mathrm{T}})^{\mathrm{T}}$	$\boldsymbol{\theta}_{[2]} = A_{\beta_2} z_{\beta_2},\ A_{\beta_2} = \begin{pmatrix} \mathbf{1}_{18} & \mathbf{0}_{18} \\ \mathbf{0}_{18} & \mathbf{1}_{18} \end{pmatrix},\ z_{\beta_2} \sim \mathrm{N}(0, 3I_2)$
$\boldsymbol{\theta}_{[3]}$	$(\boldsymbol{-3.0}_3^{\mathrm{T}}, \boldsymbol{-2.5}_3^{\mathrm{T}}, \boldsymbol{-3.0}_3^{\mathrm{T}}, \boldsymbol{-2.5}_3^{\mathrm{T}}, \boldsymbol{-2.0}_3^{\mathrm{T}}, \boldsymbol{-1.5}_3^{\mathrm{T}}, \boldsymbol{0.5}_3^{\mathrm{T}}, \boldsymbol{1.0}_3^{\mathrm{T}}, \boldsymbol{0.5}_3^{\mathrm{T}}, \boldsymbol{1.0}_3^{\mathrm{T}}, \boldsymbol{1.5}_3^{\mathrm{T}}, \boldsymbol{2.0}_3^{\mathrm{T}})^{\mathrm{T}}$	$\boldsymbol{\theta}_{[3]} = z_{\beta_3},\ z_{\beta_3} \sim \mathrm{N}(\mathbf{0}, \Sigma_{\beta_3}),\ \Sigma_{\beta_3} = \begin{pmatrix} \Sigma_1 & 0 \\ 0 & \Sigma_2 \end{pmatrix},$ $\Sigma_1 = \underbrace{\begin{pmatrix} 5 & 0.5 & \cdots & 0.5 \\ 0.5 & 5 & \cdots & 0.5 \\ \vdots & \vdots & \ddots & \vdots \\ 0.5 & 0.5 & \cdots & 5 \end{pmatrix}}_{18\times 18}, \Sigma_2 = \underbrace{\begin{pmatrix} 3 & 0.9 & \cdots & 0.9 \\ 0.9 & 3 & \cdots & 0.9 \\ \vdots & \vdots & \ddots & \vdots \\ 0.9 & 0.9 & \cdots & 3 \end{pmatrix}}_{18\times 18}$

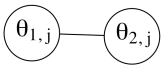
	Case 8	
	Setting 1	Setting 2
n		400
M		36
$\boldsymbol{p}_\psi$		$(1/36, \ldots, 1/36)$
$\boldsymbol{\theta}_{[1]}$	$(\boldsymbol{-2.0}_3^{\mathrm{T}}, \boldsymbol{-1.0}_3^{\mathrm{T}}, \boldsymbol{-2.0}_3^{\mathrm{T}}, \boldsymbol{-1.0}_3^{\mathrm{T}}, \boldsymbol{-2.0}_3^{\mathrm{T}}, \boldsymbol{-1.0}_3^{\mathrm{T}}, \boldsymbol{1.0}_3^{\mathrm{T}}, \boldsymbol{2.0}_3^{\mathrm{T}}, \boldsymbol{1.0}_3^{\mathrm{T}}, \boldsymbol{2.0}_3^{\mathrm{T}})^{\mathrm{T}}$	$\boldsymbol{\theta}_{[1]} = A_{\beta_1} z_{\beta_1},\ A_{\beta_1} = \begin{pmatrix} \mathbf{1}_9 & \mathbf{0}_9 & \mathbf{0}_9 & \mathbf{0}_9 \\ \mathbf{0}_9 & \mathbf{1}_9 & \mathbf{0}_9 & \mathbf{0}_9 \\ \mathbf{0}_9 & \mathbf{0}_9 & \mathbf{1}_9 & \mathbf{0}_9 \\ \mathbf{0}_9 & \mathbf{0}_9 & \mathbf{0}_9 & \mathbf{1}_9 \end{pmatrix},\ z_{\beta_1} \sim \mathrm{N}(0, 0.5I_4)$
$\boldsymbol{\theta}_{[2]}$	$(\boldsymbol{-2.0}_2^{\mathrm{T}}, -1.5, \boldsymbol{-1.0}_2^{\mathrm{T}}, -0.5, \boldsymbol{-2.0}_2^{\mathrm{T}}, -1.5, \boldsymbol{-1.0}_2^{\mathrm{T}}, -0.5, \boldsymbol{-2.0}_2^{\mathrm{T}}, -1.5, \boldsymbol{-1.0}_2^{\mathrm{T}}, -0.5, \boldsymbol{0.5}_2^{\mathrm{T}}, 1.0, \boldsymbol{1.5}_2^{\mathrm{T}}, 2.0, \boldsymbol{0.5}_2^{\mathrm{T}}, 1.0, \boldsymbol{1.5}_2^{\mathrm{T}}, 2.0, \boldsymbol{0.5}_2^{\mathrm{T}}, 1.0, \boldsymbol{1.5}_2^{\mathrm{T}}, 2.0)^{\mathrm{T}}$	$\boldsymbol{\theta}_{[2]} = A_{\beta_2} z_{\beta_2},\ A_{\beta_2} = \begin{pmatrix} \mathbf{1}_9 & \mathbf{0}_9 & \mathbf{0}_9 & \mathbf{0}_9 \\ \mathbf{0}_9 & \mathbf{1}_9 & \mathbf{0}_9 & \mathbf{0}_9 \\ \mathbf{0}_9 & \mathbf{0}_9 & \mathbf{1}_9 & \mathbf{0}_9 \\ \mathbf{0}_9 & \mathbf{0}_9 & \mathbf{0}_9 & \mathbf{1}_9 \end{pmatrix},\ z_{\beta_2} \sim \mathrm{N}(0, 0.5I_4)$
$\boldsymbol{\theta}_{[3]}$	$(\boldsymbol{-3.0}_3^{\mathrm{T}}, \boldsymbol{-1.5}_3^{\mathrm{T}}, \boldsymbol{-2.5}_3^{\mathrm{T}}, \boldsymbol{-1.0}_3^{\mathrm{T}}, \boldsymbol{-2.0}_3^{\mathrm{T}}, \boldsymbol{-0.5}_3^{\mathrm{T}}, \boldsymbol{0.5}_3^{\mathrm{T}}, \boldsymbol{2.0}_3^{\mathrm{T}}, \boldsymbol{1.0}_3^{\mathrm{T}}, \boldsymbol{2.5}_3^{\mathrm{T}}, \boldsymbol{1.5}_3^{\mathrm{T}}, \boldsymbol{3.0}_3^{\mathrm{T}})^{\mathrm{T}}$	$\boldsymbol{\theta}_{[3]} = z_{\beta_3},\ \ z_{\beta_3} \sim \mathrm{N}(\mathbf{0}, \Sigma_{\beta_3}),$ $\Sigma_{\beta_3} = \begin{pmatrix} \Sigma & 0 & 0 & 0 \\ 0 & \Sigma & 0 & 0 \\ 0 & 0 & \Sigma & 0 \\ 0 & 0 & 0 & \Sigma \end{pmatrix},\ \Sigma = \underbrace{\begin{pmatrix} 5 & 0.5 & \cdots & 0.5 \\ 0.5 & 5 & \cdots & 0.5 \\ \vdots & \vdots & \ddots & \vdots \\ 0.5 & 0.5 & \cdots & 5 \end{pmatrix}}_{9\times 9}$

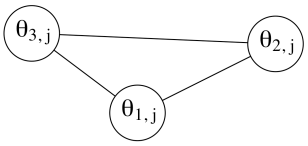


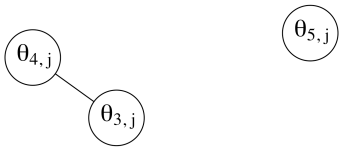
(a) Graph 1



(b) Graph 2

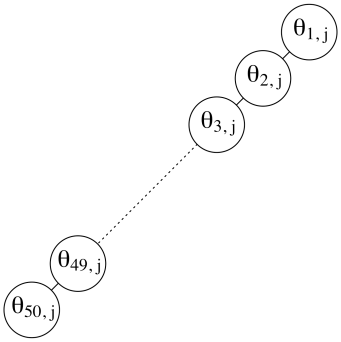




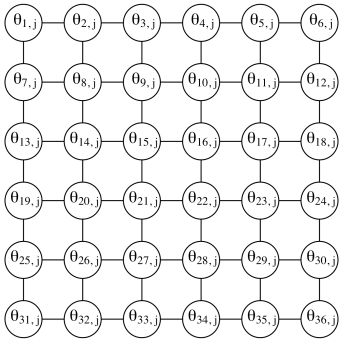


(c) Graph 3

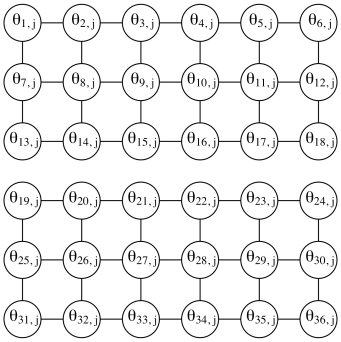
(d) Graph 4



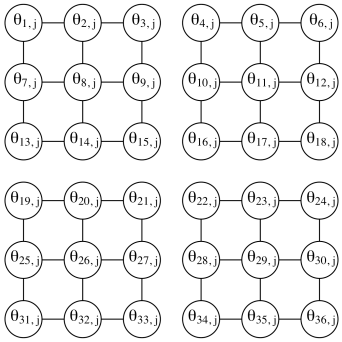
(e) Graph 5



(f) Graph 6



(g) Graph 7



(h) Graph 8

Figure 4.1: Graphs representing regional adjacency relationships in SVC model in (4.1).

For each experimental setting, we generate 100 datasets and compute four information criteria: WAIC (WAIC1) and PIIC1 for Model 1 explained in Section 4.1.3, and WAIC (WAIC2) and PIIC2 for the combination of Model 1 and Model 2. Then, when using WAIC1 and PIIC1, we select the hyperparameter of Model 1 and construct the Bayesian predictive distribution, whereas when using WAIC2 and PIIC2, we select not only the hyperparameters but also between Model 1 and Model 2 and construct the Bayesian predictive distribution, to compare them. It should be noted, however, that whereas PIIC2 attempts to select the more appropriate of Model 1 and Model 2, WAIC2 eventually always selects Model 2.

Model 1 is the model in which the prior distribution of the differences of adjacent regression coefficients is given by a Laplace distribution with a common hyperparameter for all explanatory variables, so the number of hyperparameters is one in that case. In contrast, Model 2 is the model in which the prior distribution is given by Laplace distributions with different hyperparameters for each explanatory variable. Specifically, the number of hyperparameters is 3, giving it a larger complexity than Model 1. As the concrete procedure of hyperparameter selection, in Model 1, we simply generate 20 candidate values and search for the optimal one. In Model 2, we sequentially focus on one hyperparameter at a time, generating 20 candidate values for it and then searching for the optimal value, and repeat this until a predetermined level of convergence is achieved.

As an evaluation measure for the information criteria, we use as risk the Kullback-Leibler divergence between the predictive distribution obtained and the true distribution, that is, $\mathrm{E}_{\tilde{y}_i,\tilde{\boldsymbol{x}}_i}[-\log f(\tilde{y}_i,\tilde{\boldsymbol{x}}_i \mid \boldsymbol{y},\boldsymbol{X};\boldsymbol{\lambda})]$ or $\mathrm{E}_{\tilde{y}_i,\tilde{\boldsymbol{x}}_i}[-\log f(\tilde{y}_i,\tilde{\boldsymbol{x}}_i \mid \boldsymbol{y},\boldsymbol{X};\hat{\boldsymbol{\lambda}})]$. More precisely, the expectation is evaluated by its empirical version; that is, the expectation operator is replaced with $n^{-1}\sum_{i=1}^{n}$. As a reference measure, we also report the rates in which WAIC yields a smaller or larger risk than PIIC.

Table 4.4 summarizes the results for Cases 1–4. In Cases 1–4, for almost all experimental settings, the risk given by PIIC1 is smaller than that given by WAIC1. This suggests that the asymptotic setting on which PIIC1 is based is more appropriate; nevertheless, the difference between them is small. This agrees with the experimental results of Ninomiya (2021), which show that although the superiority of PIIC1 over WAIC1 is clear for non-sparse estimation, PIIC1 has only a slight superiority for sparse estimation. On the other hand, for all experimental settings, the risk given by PIIC2 is clearly smaller than that given

by WAIC2. The reason for the clear difference is that in Cases 1–4, the number of vertices in the graph is small, and simpler models tend to perform better, but WAIC2 always selects the more complex model. In fact, when comparing PIIC1 and PIIC2, PIIC2 always yields smaller risk; however, the difference is not large, indicating that little benefit is gained from allowing the selection of the more complex model.

Table 4.5 summarizes the results for Cases 5–8. Compared with Cases 1–4, the number of vertices in the graph is greater, meaning these are settings where more complex models might sometimes perform better. In fact, when comparing WAIC2 and PIIC2, the differences are not as large as in Cases 1–4, and there are six settings in which, although the differences are small, WAIC2 yields smaller risk. Here, it should be noted that when σ^2 is larger, there is a tendency for the more complex model to be more appropriate. Furthermore, when comparing PIIC1 and PIIC2, the effect of allowing the number of hyperparameters to be selected becomes evident, and it can be confirmed that PIIC2 outperforms PIIC1 for all experimental settings by a larger margin than in Cases 1–4.

Let us summarize what we wish to emphasize in Table 4.5. For these experimental settings, sometimes the simpler Model 1 is more appropriate, and sometimes the more complex Model 2 is more appropriate. PIIC1 only considers Model 1, and WAIC2 always selects Model 2, but PIIC2 generally yields smaller risk than either of them. More precisely, settings are considered in which always selecting Model 2 becomes preferable, but even then PIIC2 provides results comparable to those of WAIC2. In other words, PIIC2 appropriately selects the number of hyperparameters according to the data and provides predictive distributions with good performance.

Table 4.4: Comparison of information criteria (Cases 1–4). Each value for WAIC1, PIIC1, WAIC2, and PIIC2 is the average of evaluations of Kullback-Leibler divergence between the true and Bayesian predictive distributions. For pairs WAIC1 and PIIC1, and WAIC2 and PIIC2, the smaller of the two values is shown in bold. Each vector for Rate 1 (Rate 2) indicates the rates at which the evaluation for WAIC1 (WAIC2) is smaller than, equal to, and larger than the evaluation for PIIC1 (PIIC2).

Setting	σ^2	Case	WAIC1	PIIC1	Rate1	WAIC2	PIIC2	Rate2
1	1.0	1	1.314	**1.311**	(3,91,6)	1.515	**1.293**	(28,0,72)
		2	0.875	**0.868**	(0,94,6)	1.033	**0.864**	(16,0,84)
		3	0.992	**0.991**	(1,95,4)	1.136	**0.982**	(24,2,74)
		4	1.097	**1.091**	(0,94,6)	1.228	**1.088**	(27,0,73)
	1.5	1	1.473	**1.468**	(1,95,4)	1.639	**1.458**	(26,0,74)
		2	1.123	**1.117**	(0,97,3)	1.272	**1.111**	(22,0,78)
		3	1.280	**1.277**	(0,98,2)	1.382	**1.258**	(27,1,72)
		4	1.315	**1.314**	(2,95,3)	1.493	**1.305**	(30,0,70)
	2.0	1	1.723	**1.721**	(0,97,3)	1.905	**1.715**	(34,0,66)
		2	1.371	**1.370**	(0,97,3)	1.458	**1.369**	(30,1,69)
		3	1.556	**1.554**	(1,96,3)	1.598	**1.539**	(33,2,65)
		4	1.595	**1.593**	(0,97,3)	1.797	**1.592**	(31,0,69)
2	1.0	1	0.990	**0.984**	(1,91,8)	1.077	**0.978**	(22,0,78)
		2	0.815	**0.814**	(2,94,4)	0.918	**0.813**	(24,0,76)
		3	0.892	**0.891**	(2,93,5)	0.979	**0.888**	(28,0,72)
		4	0.972	**0.965**	(0,94,6)	1.122	**0.962**	(19,0,81)
	1.5	1	1.283	**1.281**	(1,92,7)	1.327	**1.276**	(30,0,70)
		2	1.060	**1.059**	(0,98,2)	1.132	**1.054**	(31,0,69)
		3	1.179	**1.179**	(0,98,2)	1.242	**1.175**	(28,0,72)
		4	1.252	**1.251**	(0,97,3)	1.386	**1.246**	(31,0,69)
	2.0	1	1.552	**1.550**	(1,96,3)	1.581	**1.549**	(37,0,63)
		2	1.365	**1.359**	(0,96,4)	1.420	**1.346**	(33,1,66)
		3	1.484	**1.483**	(0,98,2)	1.540	**1.474**	(37,0,63)
		4	1.505	**1.504**	(1,97,2)	1.603	**1.493**	(37,0,63)

Table 4.5: Comparison of information criteria (Cases 5–8). Each value for WAIC1, PIIC1, WAIC2, and PIIC2 is the average of evaluations of Kullback-Leibler divergence between the true and Bayesian predictive distributions. For pairs WAIC1 and PIIC1, and WAIC2 and PIIC2, the smaller of the two values is shown in bold. Each vector for Rate 1 (Rate 2) indicates the rates at which the evaluation for WAIC1 (WAIC2) is smaller than, equal to, and larger than the evaluation for PIIC1 (PIIC2).

Setting	σ^2	Case	WAIC1	PIIC1	Rate1	WAIC2	PIIC2	Rate2
1	1.0	5	1.024	1.024	(0,100,0)	1.044	**0.958**	(35,5,60)
		6	0.754	0.754	(0,100,0)	0.745	**0.727**	(38,0,62)
		7	**0.703**	0.703	(4,96,0)	0.709	**0.693**	(29,1,70)
		8	**0.686**	0.686	(8,92,0)	0.698	**0.680**	(22,1,77)
	1.5	5	**1.474**	1.474	(1,99,0)	1.467	**1.397**	(34,4,62)
		6	1.095	1.095	(0,100,0)	**1.049**	1.053	(49,4,47)
		7	**1.034**	1.034	(1,99,0)	1.023	**1.007**	(42,2,56)
		8	**1.013**	1.013	(1,99,0)	1.014	**0.981**	(25,5,70)
	2.0	5	1.959	**1.959**	(0,99,1)	1.935	**1.877**	(46,3,51)
		6	1.459	1.459	(0,100,0)	**1.403**	1.420	(51,7,42)
		7	1.386	1.386	(0,100,0)	1.354	**1.342**	(39,7,54)
		8	1.359	1.359	(0,100,0)	1.352	**1.315**	(32,8,60)
2	1.0	5	**0.939**	0.939	(2,98,0)	0.962	**0.913**	(37,0,63)
		6	0.771	0.771	(0,100,0)	0.755	**0.754**	(47,0,53)
		7	0.769	0.769	(0,100,0)	0.768	**0.758**	(45,0,55)
		8	0.767	0.767	(0,100,0)	0.769	**0.759**	(42,0,58)
	1.5	5	1.360	1.360	(0,100,0)	1.401	**1.317**	(42,0,58)
		6	**1.088**	1.089	(1,99,0)	1.066	**1.056**	(48,3,49)
		7	1.116	1.116	(0,100,0)	**1.066**	1.082	(50,2,48)
		8	1.096	1.096	(0,100,0)	1.086	**1.062**	(41,2,57)
	2.0	5	1.807	1.807	(0,100,0)	1.857	**1.760**	(40,1,59)
		6	**1.477**	1.478	(2,98,0)	**1.414**	1.418	(45,6,49)
		7	1.462	1.462	(0,100,0)	**1.416**	1.431	(55,5,40)
		8	1.467	1.467	(0,100,0)	**1.418**	1.421	(56,5,39)

We also consider scenarios in which the explanatory variables $\tilde{\boldsymbol{x}}$ exhibit spatial correlation. Specifically, we consider the same scenario as Setting 1 of Cases 1–4, except that the correlation between $\tilde{x}_{i_1,j}$ and $\tilde{x}_{i_2,j}$ is set to cor_1 if they belong to the same region, to $\mathrm{cor}_2^{\mathrm{edges}(i_1,i_2)}$ if they belong to different but connected regions, and to zero otherwise, where $\mathrm{edges}(i_1, i_2)$ denotes the smallest number of edges between the regions to which $\tilde{\boldsymbol{x}}_{i_1}$ and $\tilde{\boldsymbol{x}}_{i_2}$ belong. Similarly, we consider the same scenario as Setting 1 of Cases 5–8, except that the correlation between $\tilde{x}_{i_1,j}$ and $\tilde{x}_{i_2,j}$ is set to cor_1 if they belong to the same region, to $\mathrm{cor}_2^{\mathrm{dist}(i_1,i_2)}$ if they belong to different but connected regions, and to zero otherwise, where $\mathrm{dist}(i_1, i_2)$ represents the Euclidean distance between the corresponding regions. We consider two parameter settings: $(\mathrm{cor}_1, \mathrm{cor}_2) = (0.9, 0.8)$ and $(0.99, 0.97)$. For $j = 1, \ldots, \tilde{p}$, $(\tilde{x}_{1,j}, \ldots, \tilde{x}_{n,j})^{\mathrm{T}}$ is generated independently from a multivariate normal distribution with mean zero and covariance matrix determined by the spatial correlations described above, with marginal variance 5^2.

The results are summarized in Table 4.6 and 4.7. Overall, the tendencies observed in the absence of spatial correlation remain largely unchanged when spatial correlation is introduced. However, in Case 6, across different values of σ^2, PIIC2 yields a smaller mean KL divergence than WAIC2 in more settings under spatial correlation than in its absence. This suggests that, although more complex models are often appropriate in Case 6, the ability of PIIC2 to select appropriate model complexity becomes more evident when spatial correlation is present. By contrast, Case 5 under the stronger spatial correlation setting in Table 4.7 exhibits a different behavior. WAIC2 shows substantial overfitting in this case. This suggests that, the data do not support the more complex model and the tendency of WAIC2 to always select the complex model becomes particularly problematic.

Table 4.6: Comparison of information criteria (Setting 1 in Cases 1–8 under the spatial correlation structure with $(\mathrm{cor}_1, \mathrm{cor}_2) = (0.9, 0.8)$). Each value for WAIC1, PIIC1, WAIC2, and PIIC2 is the average of evaluations of Kullback-Leibler divergence between the true and Bayesian predictive distributions. For pairs WAIC1 and PIIC1, and WAIC2 and PIIC2, the smaller of the two values is shown in bold. Each vector for Rate 1 (Rate 2) indicates the rates at which the evaluation for WAIC1 (WAIC2) is smaller than, equal to, and larger than the evaluation for PIIC1 (PIIC2).

σ^2	Case	WAIC1	PIIC1	Rate1	WAIC2	PIIC2	Rate2
1.0	1	2.146	**2.118**	(2,92,6)	2.438	**2.108**	(37,0,63)
	2	1.004	**1.000**	(5,90,5)	1.225	**1.000**	(22,0,78)
	3	2.099	**2.063**	(0,95,5)	2.481	**2.057**	(17,0,83)
	4	2.669	**2.647**	(2,89,9)	3.410	**2.647**	(9,0,91)
	5	3.126	3.126	(0,100,0)	3.497	**3.012**	(35,6,59)
	6	1.758	**1.758**	(0,99,1)	1.704	**1.652**	(43,3,54)
	7	7.052	7.052	(0,100,0)	7.704	**6.869**	(17,3,80)
	8	**1.873**	1.873	(1,98,1)	2.045	**1.763**	(22,2,76)
1.5	1	2.108	**2.091**	(0,95,5)	2.274	**2.094**	(42,0,58)
	2	1.278	**1.272**	(0,95,5)	1.474	**1.269**	(18,0,82)
	3	2.229	**2.216**	(1,95,4)	2.577	**2.188**	(26,0,74)
	4	2.632	**2.614**	(0,94,6)	3.227	**2.619**	(23,0,77)
	5	4.260	4.260	(0,100,0)	5.049	**4.158**	(26,2,72)
	6	2.114	2.114	(0,100,0)	2.005	**1.955**	(44,4,52)
	7	5.901	5.901	(0,100,0)	6.174	**5.679**	(23,6,71)
	8	2.299	2.299	(0,100,0)	2.363	**2.148**	(28,7,65)
2.0	1	2.178	**2.167**	(0,96,4)	2.364	**2.158**	(29,1,70)
	2	1.554	**1.554**	(0,99,1)	1.690	**1.545**	(33,1,66)
	3	2.475	**2.471**	(0,97,3)	2.807	**2.426**	(26,0,74)
	4	2.724	**2.718**	(0,96,4)	3.286	**2.679**	(26,1,73)
	5	5.429	**5.425**	(0,96,4)	6.743	**5.368**	(33,2,65)
	6	2.559	2.559	(0,100,0)	**2.405**	2.416	(51,1,48)
	7	2.067	2.067	(0,100,0)	2.201	**2.026**	(31,1,68)
	8	3.148	**3.142**	(2,95,3)	3.770	**3.098**	(18,0,82)

Table 4.7: Comparison of information criteria (Setting 1 in Cases 1–8 under the spatial correlation structure with $(\mathrm{cor}_1, \mathrm{cor}_2) = (0.99, 0.97)$). Each value for WAIC1, PIIC1, WAIC2, and PIIC2 is the average of evaluations of Kullback-Leibler divergence between the true and Bayesian predictive distributions. For pairs WAIC1 and PIIC1, and WAIC2 and PIIC2, the smaller of the two values is shown in bold. Each vector for Rate 1 (Rate 2) indicates the rates at which the evaluation for WAIC1 (WAIC2) is smaller than, equal to, and larger than the evaluation for PIIC1 (PIIC2).

σ^2	Case	WAIC1	PIIC1	Rate1	WAIC2	PIIC2	Rate2
1.0	1	2.472	**2.419**	(1,90,9)	2.769	**2.433**	(30,0,70)
	2	1.104	**1.101**	(0,94,6)	1.329	**1.100**	(15,0,85)
	3	2.416	**2.411**	(1,93,6)	4.648	**2.391**	(13,0,87)
	4	3.892	**3.587**	(1,78,21)	6.534	**3.529**	(9,0,91)
	5	7.078	7.078	(0,100,0)	16.797	**7.066**	(9,5,86)
	6	2.395	2.395	(0,100,0)	2.351	**2.216**	(37,7,56)
	7	**4.680**	4.681	(2,98,0)	6.716	**4.676**	(15,11,74)
	8	4.356	4.356	(0,100,0)	7.807	**4.468**	(11,5,84)
1.5	1	2.389	**2.381**	(0,94,6)	2.618	**2.387**	(33,0,67)
	2	1.358	**1.356**	(0,97,3)	1.557	**1.345**	(23,0,77)
	3	2.808	**2.805**	(1,96,3)	4.881	**2.803**	(19,1,80)
	4	3.830	**3.766**	(1,91,8)	6.893	**3.742**	(8,2,90)
	5	10.080	**10.063**	(0,99,1)	34.824	**10.012**	(10,0,90)
	6	3.043	3.403	(0,100,0)	2.892	**2.848**	(41,2,57)
	7	5.974	**5.969**	(0,98,2)	8.481	**5.925**	(23,8,69)
	8	6.301	6.301	(0,100,0)	11.710	**6.549**	(9,8,83)
2.0	1	2.576	**2.545**	(0,93,7)	2.755	**2.527**	(38,0,62)
	2	1.664	**1.664**	(1,98,1)	1.808	**1.660**	(39,0,61)
	3	3.364	**3.353**	(0,97,3)	5.545	**3.325**	(20,3,77)
	4	4.228	**4.224**	(1,95,4)	7.949	**4.208**	(10,0,90)
	5	24.456	**13.966**	(0,97,3)	70.285	**13.881**	(13,1,86)
	6	3.614	**3.614**	(0,99,1)	3.588	**3.290**	(38,6,56)
	7	8.165	**8.152**	(0,98,2)	12.230	**8.029**	(20,4,68)
	8	9.125	9.125	(0,100,0)	15.551	**9.644**	(15,4,81)

Furthermore, we evaluate the computational time of the information criteria for higher dimensionality of explanatory variables and larger sample sizes. Specifically, we consider four scenarios similar to Setting 1 of Case 6 with $\sigma^2 = 1.0$. In Scenario 1, we set $(n, \tilde{p}) = (700, 10)$ and assign the hyperparameters $\lambda_{2,1}$, $\lambda_{2,2}$, and $\lambda_{2,3}$ associated with the fused penalty in (4.2) for $j = 1, \ldots, 3$, $j = 4, \ldots, 7$, and $j = 8, \ldots, 10$, respectively. Thus, the number of types of hyperparameters associated with the fused penalty, denoted by q_{fuse}, is 3. In Scenario 2, we set $(n, \tilde{p}) = (500, 5)$ and assign hyperparameters $\lambda_{2,j}$ for the fused penalty for $j = 1, \ldots, \tilde{p}$, yielding $q_{\text{fuse}} = 5$. In Scenario 3 and 4, we set $n = 500$ and $\tilde{p} = 3$, $n = 700$ and $\tilde{p} = 3$, and assign hyperparameters $\lambda_{2,j}$ for the fused penalty for $j = 1, \ldots, \tilde{p}$. In both scenarios, q_{fuse} is 3. When calculating information criteria, the predictive distribution is evaluated using two approaches: Monte Carlo integration and the Laplace approximation. For WAIC1, PIIC1, WAIC2, and PIIC2, we measured the total computational time required to calculate the criteria using Monte Carlo integration, including the entire hyperparameter selection procedure. In addition, the total computational time required to compute PIIC1 and PIIC2 using the Laplace approximation was measured in the same manner. However, for WAIC1 and WAIC2, the Laplace approximation would require approximating the bias correction term in addition to the predictive distribution; therefore, only results based on Monte Carlo integration are used. The computational time is measured based on CPU time.

Table 4.8: Computational times (sec.) for model evaluation based on information criteria using Monte Carlo integration. Figures in parentheses give the estimated standard deviation.

Scenario	n	$\tilde{p}$	q_{fuse}	WAIC1	PIIC1	WAIC2	PIIC2
1	700	10	3	145.731	127.089	1064.665	1060.251
				(1.480)	(1.363)	(12.503)	(5.744)
2	500	5	5	99.283	66.287	1999.210	1637.317
				(0.277)	(0.315)	(12.238)	(33.049)
3	500	3	3	94.358	57.954	600.942	549.819
				(0.174)	(0.363)	(1.625)	(1.478)
4	700	3	3	133.007	81.750	812.972	779.145
				(0.615)	(0.613)	(4.036)	(7.431)

Table 4.9: Computational times (sec.) for model evaluation based on information criteria using Laplace approximation. Figures in parentheses give the estimated standard deviation.

Scenario	n	$\tilde{p}$	q_{fuse}	PIIC1	PIIC2
1	700	10	3	13.274 (0.197)	97.511 (2.000)
2	500	5	5	4.076 (0.028)	40.879 (0.114)
3	500	3	3	2.602 (0.048)	15.575 (0.075)
4	700	3	3	6.133 (0.041)	37.297 (0.946)

The results are summarized in Table 4.8 and 4.9. The values reported in the table represent the average over eight runs. For the results based on Monte Carlo integration, the comparison between Scenarios 3 and 4 shows that the computational time increases approximately linearly with the sample size n.

Furthermore, across the results for WAIC2 and PIIC2, a comparison between Scenarios 2 and 3 shows that when both $\tilde{p}$ and q_{fuse} increase, the computational time increases more than linearly. In contrast, a comparison between Scenarios 1 and 4 indicates that increasing $\tilde{p}$ alone leads to a milder increase in computational time. These results suggest that the computational cost for WAIC2 and PIIC2 is strongly affected by q_{fuse}, through the increased burden of the sequential hyperparameter selection procedure rather than by $\tilde{p}$ alone.

In addition, across all scenarios and information criteria, the calculations based on the Laplace approximation are substantially faster than those based on Monte Carlo integration, achieving speedups of several tens of times depending on the criterion. Although a theoretical guarantee for applying the Laplace approximation to the predictive distribution is not provided, our numerical experiments consistently demonstrate that replacing Monte Carlo integration with the Laplace approximation leads to a significant reduction in computational time.

4.4 Real data analysis

In order to check how much difference the proposed PIIC makes compared with the existing WAIC in an actual analysis, we conducted an analysis using dust data collected from houses in the United States. This dataset was gathered in the Wild Life of Our Homes project, and multiple analyses have been conducted, starting with Barberán et al. (2015). Among those studies, the one having the greatest influence on the present paper is Zhao and Bondell (2020), in which the generalized fused lasso was applied, although no choice was made about how the regularization parameters were assigned.

For the analysis present here, we use samples from 1,070 houses, excluding samples containing missing values and data from Alaska and Hawaii. The response variable is fungal diversity (the proportion of the number of fungal species in each sample to the total number of species, 763), and we consider two explanatory variables. Specifically, for Setting 1, Variables 1 and 2 are mean annual temperature and elevation, and for Setting 2, they are mean annual precipitation and net primary productivity (NPP), respectively. Applying the k-means method, we divide the data into 85 clusters according to the latitude and longitude of the houses, and regard these clusters as regions. Then, based on the centroids of each cluster, we perform a Voronoi tessellation (see Figure 4.2), which determines the adjacency relationships of the regions. As in Section 4.4, we focus on the regularization for fusing adjacent regression coefficients, and do not perform variable selection. As the models to be fitted to the data, we consider Model 1 and Model 2 explained in Section 4.1.3. In addition, for WAIC1, we compute it in Model 1 with 20 candidate values of λ_2, and for WAIC2 and PIIC2, we compute them in Model 2 by repeatedly generating 20 candidate values of $\lambda_{2,j}$ as j varies. Note that we consider computing PIIC1 to not be necessary, based on the results of the numerical experiments.

Figures 4.3, 4.4, and 4.5 show the group structures of the intercept, the regression coefficients of Variable 1, and the regression coefficients of Variable 2 in the models selected by each information criterion, respectively. In the figures, regions displayed as belonging to the same group represent those whose regression coefficients have been fused. First, regarding WAIC1, in both Setting 1 and Setting 2, more than 60 groups are selected for the intercept, Variable 1, and Variable 2. As a result, despite the application of sparse regularization, the estimation

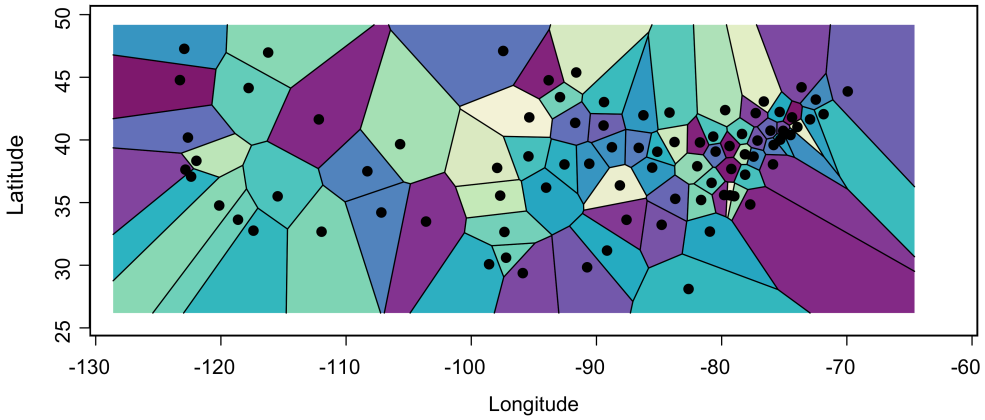
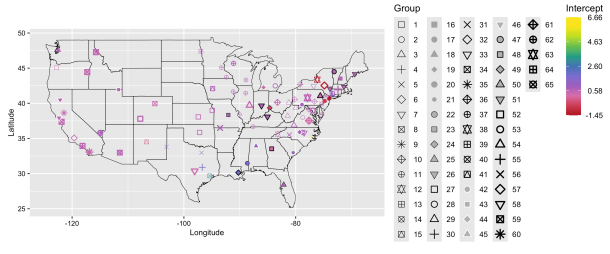


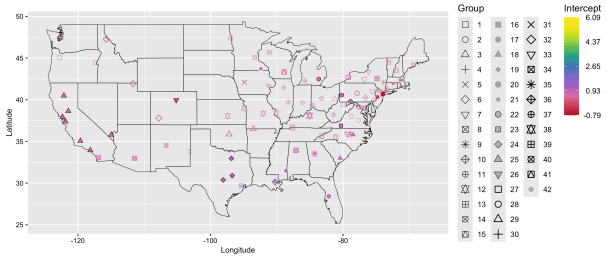
Figure 4.2: Adjacency structure of regions determined by Voronoi tessellation.

produces results with almost no sparsity. Next, regarding WAIC2, for Setting 1, more than 40 groups are selected for the intercept and Variable 2, whereas for Setting 2, all regression coefficients are fused, so the number of groups becomes 1. In the case of Setting 2, perhaps due to excessive regularization, the estimation produces results with no spatial heterogeneity, despite using the SVC model for capturing spatial heterogeneity. PIIC2 gives moderate results: for Setting 1, it selects around 20 groups, and for Setting 2, it selects 2 or 3 groups. When the number of groups is 2 or 3, it is easier to make interpretations regarding spatial heterogeneity. In fact, the groups are roughly divided into the western United States and all other regions, and this is considered to relate to fungal diversity, since the West has low precipitation. Moreover, because the West has high net primary productivity (NPP) due to soil characteristics and other reasons, the vegetation types are limited, which is also considered to relate to fungal diversity.

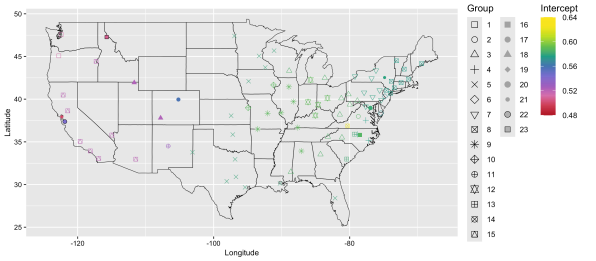
For the real data analysis, since the true structure is unknown, it is difficult to discuss which—the proposed method and the existing method—is superior. What we would like to emphasize in this section is that the results differ considerably between WAIC and PIIC. Although not in the sense of being inappropriate, WAIC provides models in which the selected number of groups is either too large or too small, making it difficult to interpret spatial heterogeneity, whereas PIIC gives models with moderate numbers of groups.



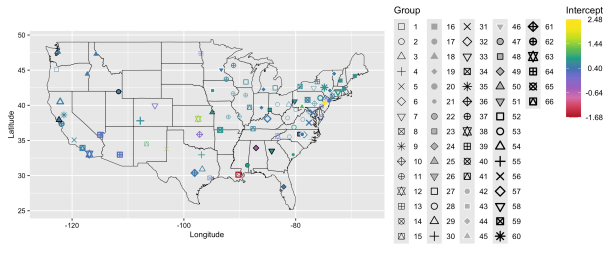
(a) WAIC1 for Setting 1



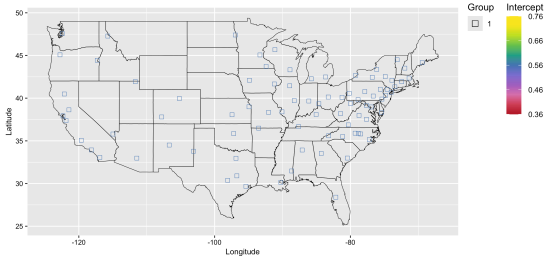
(b) WAIC2 for Setting 1



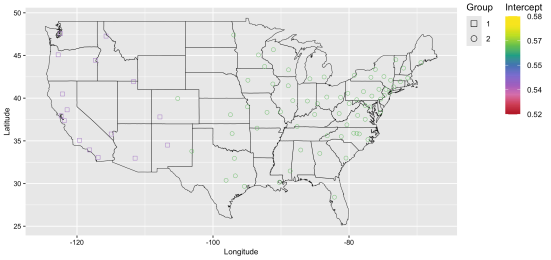
(c) PIIC2 for Setting 1



(d) WAIC1 for Setting 2

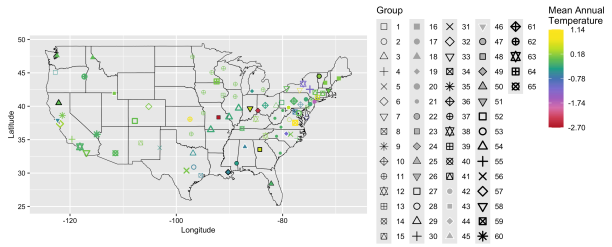


(e) WAIC2 for Setting 2

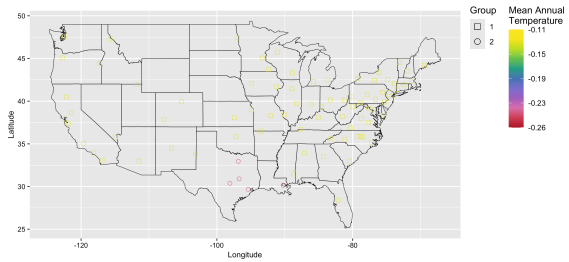


(f) PIIC2 for Setting 2

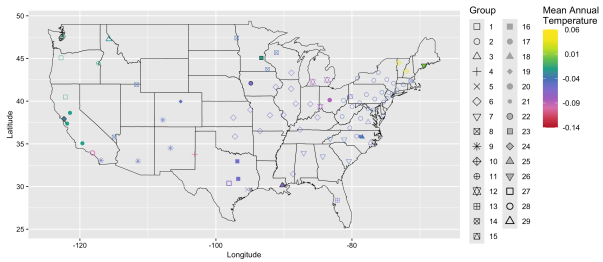
Figure 4.3: Group structure in intercept of model selected by each information criterion.



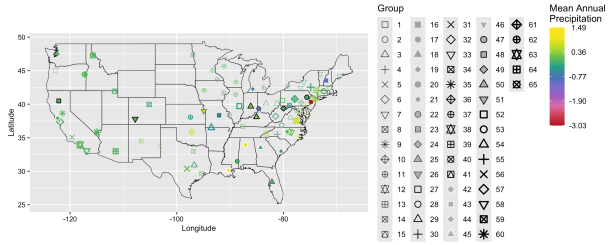
(a) WAIC1 for Setting 1



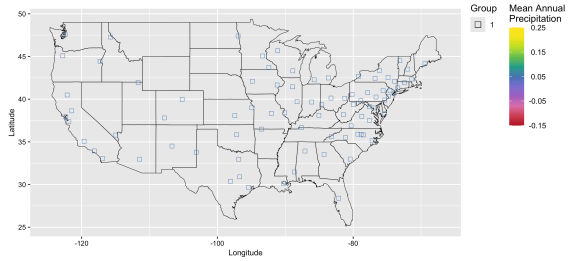
(b) WAIC2 for Setting 1



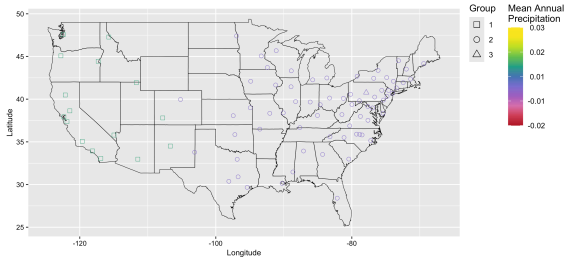
(c) PIIC2 for Setting 1



(d) WAIC1 for Setting 2

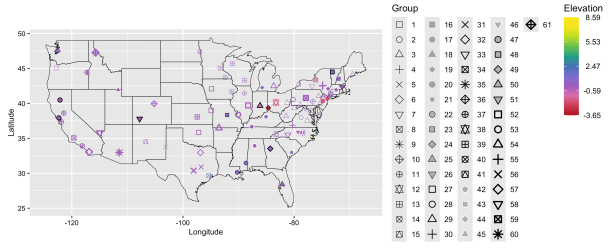


(e) WAIC2 for Setting 2

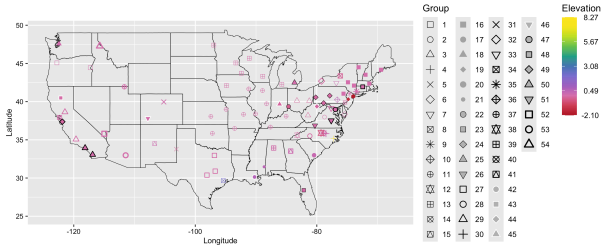


(f) PIIC2 for Setting 2

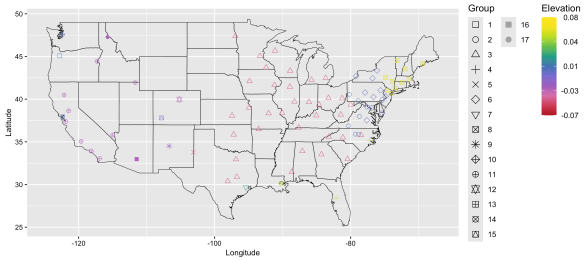
Figure 4.4: Group structure in regression coefficient for Variable 1 of model selected by each information criterion.



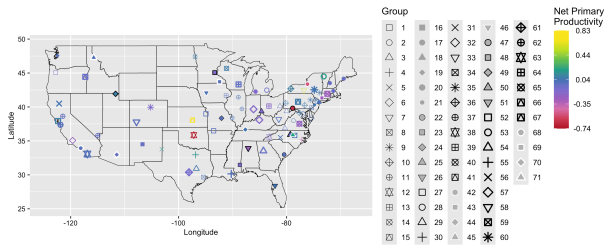
(a) WAIC1 for Setting 1



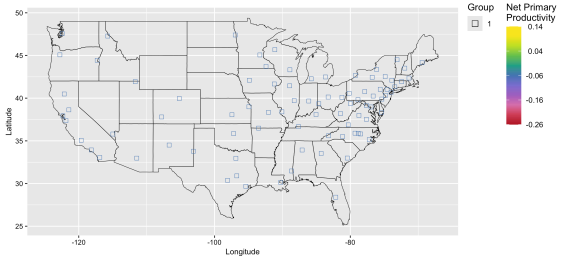
(b) WAIC2 for Setting 1



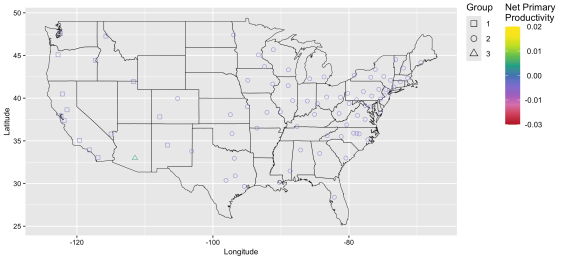
(c) PIIC2 for Setting 1



(d) WAIC1 for Setting 2



(e) WAIC2 for Setting 2



(f) PIIC2 for Setting 2

Figure 4.5: Group structure in regression coefficient for Variable 2 of model selected by each information criterion.

Chapter 5

Conclusion

This dissertation has established a comprehensive Bayesian modeling framework for data with group structures by developing (i) hierarchical Bayesian logistic regression models capable of simultaneous variable selection and variable fusion, and (ii) an information criterion suitable for model comparison within such Bayesian regularization frameworks. The proposed methods address two major challenges that have been insufficiently explored in previous studies: extending the Bayesian generalized fused lasso with global-local shrinkage priors beyond linear regression, and developing an information criterion that properly evaluates models with such complex priors.

In the first part, we proposed hierarchical Bayesian logistic regression models that achieve both variable selection and variable fusion. Specifically, we formulated a logistic regression model assuming Laplace priors on both regression coefficients and their adjacent differences, and an extended model combining a Laplace prior for regression coefficients with a horseshoe prior for their adjacent differences. By exploiting the hierarchical representations of these priors and the data-augmentation method with the Pólya-Gamma distribution, we derived efficient Gibbs sampling algorithms. Through simulation studies and real time-series data analyses, the proposed models demonstrated higher estimation accuracy and predictive performance than existing methods, while achieving automatic and interpretable grouping of explanatory variables. These results confirmed the practical advantage of incorporating global-local shrinkage priors into fusion-type regularization for flexible modeling of group structures. Nevertheless, computational efficiency remains an important issue, especially in high-dimensional settings, suggesting that future work should explore more scalable sampling schemes or variational approximations.

In the second part, motivated by the need for objective model selection in Bayesian regularization, we developed the prior intensified information criterion (PIIC) for the Bayesian generalized fused lasso, with a particular focus on applications to spatially varying coefficients (SVC) models. Unlike existing criteria such as DIC or WAIC, PIIC explicitly accounts for the complexity of prior distributions and is derived for when the order of the logarithm of the prior distributions is $O(n)$. We established asymptotic properties of the generalized fused lasso estimator under this setting and adapted PIIC for SVC models with the Bayesian generalized fused lasso. Numerical experiments demonstrated that PIIC provides more appropriate model selection than WAIC, especially in settings where the class of prior distributions is also being selected. In the U.S. house-dust dataset, PIIC yielded spatial groupings that were both more parsimonious and interpretable than those selected by WAIC, illustrating its practical usefulness.

Overall, this dissertation contributes to the field of Bayesian regularization in two complementary ways. First, it extends the scope of Bayesian variable fusion to binary and group-structured data by incorporating global-local shrinkage priors and employing a data augmentation scheme with a Pólya-Gamma distribution, thereby enabling efficient Gibbs sampling while achieving both sparsity and flexible variable fusion. Second, it establishes an information criterion that enables theoretically grounded model comparison for models with the Bayesian generalized fused lasso. These developments together form a unified Bayesian modeling framework for estimation and model evaluation of data with group structures.

Future research directions include: (i) extending the proposed Bayesian variable fusion models to high-dimensional settings or to more general non-Gaussian data through scalable inference techniques; (ii) theoretical investigation of credible intervals for variable selection and fusion decisions; and (iii) the extension of the information criterion to models with other hierarchical priors such as the normal-exponential-gamma or Dirichlet-Laplace priors, as well as to group level and bi-level model selection. Such extensions will further enhance the interpretability and applicability of Bayesian variable fusion methods in modern statistical modeling.

References

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. in B. N. Petrov and F. Csaki (Eds.) 2nd International Symposium on Information Theory (pp. 267–281). Budapest: Akademiai Kiado.

Andersen, P. and Gill, R. (1982). Cox's regression model for counting processes: a large sample study. *The Annals of Statistics*, 10(4):1100–1120.

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

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

Barberán, A., Ladau, J., Leff, J. W., Pollard, K. S., Menninger, H. L., Dunn, R. R., and Fierer, N. (2015). Continental-scale distributions of dust-associated bacteria and fungi. *Proceedings of the National Academy of Sciences*, 112(18):5756–5761.

Betancourt, B., Rodríguez, A., and Boyd, N. (2017). Bayesian fused lasso regression for dynamic binary networks. *Journal of Computational and Graphical Statistics*, 26(4):840–850.

Bhattacharyya, A., Pal, S., Mitra, R., and Rai, S. (2022). Applications of Bayesian shrinkage prior models in clinical research with categorical responses. *BMC Medical Research Methodology*, 22(1):1–19.

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

Brunsdon, C., Fotheringham, A. S., and Charlton, M. E. (1996). Geographically weighted regression: a method for exploring spatial nonstationarity. *Geographical Analysis*, 28(4):281–298.

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

Davis, J. and Goadrich, M. (2006). The relationship between Precision-Recall and ROC curves. In *Proceedings of the 23rd international conference on Machine learning*, pages 233–240.

Deng, H., Baydogan, M. G., and Runger, G. (2014). SMT: Sparse multivariate tree. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, 7(1):53–69.

Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. (2004). Least angle regression. *The Annals of Statistics*, 32(2):407 – 499.

Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American statistical Association*, 96(456):1348–1360.

Gelfand, A. E., Kim, H.-J., Sirmans, C., and Banerjee, S. (2003). Spatial modeling with spatially varying coefficient processes. *Journal of the American Statistical Association*, 98(462):387–396.

Geman, S. and Geman, D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-6(6):721–741.

Genkin, A., Lewis, D. D., and Madigan, D. (2007). Large-scale Bayesian logistic regression for text categorization. *Technometrics*, 49(3):291–304.

Geyer, C. J. (1996). On the asymptotics of convex stochastic optimization. *Unpublished manuscript*.

Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57(1):97–109.

Hjort, N. L. and Pollard, D. (1993). Asymptotics for minimisers of convex processes. *Unpublished manuscript*.

Hoefling, H. (2010). A path algorithm for the fused lasso signal approximator. *Journal of Computational and Graphical Statistics*, 19(4):984–1006.

Huang, M. and Yao, W. (2012). Mixture of regression models with varying mixing proportions: a semiparametric approach. *Journal of the American Statistical Association*, 107(498):711–724.

Hurvich, C. M. and Tsai, C.-L. (1989). Regression and time series model selection in small samples. *Biometrika*, 76(2):297–307.

Inoue, R., Ishiyama, R., and Sugiura, A. (2020). Identifying local differences with fused-mcp: an apartment rental market case study on geographical segmentation detection. *Japanese Journal of Statistics and Data Science*, 3:183–214.

Jing, B., Yang, G., Yu, X., and Zhang, C. (2018). Fused-mcp with application to signal processing. *Journal of Computational and Graphical Statistics*, 27(4):872–886.

Kakikawa, Y. and Kawano, S. (2023). Bayesian fused lasso modeling via horseshoe prior. *Japanese Journal of Statistics and Data Science*, 6(2):705–727.

Kim, S.-J., Koh, K., Boyd, S., and Gorinevsky, D. (2009). l_1 trend filtering. *SIAM review*, 51(2):339–360.

Konishi, S. and Kitagawa, G. (2008). *Information criteria and statistical modeling*. Springer, New York.

Kullback, S. and Leibler, R. A. (1951). On Information and Sufficiency. *The Annals of Mathematical Statistics*, 22(1):79 – 86.

Kyung, M., Gill, J., Ghosh, M., and Casella, G. (2010). Penalized regression, standard errors, and Bayesian lassos. *Bayesian Analysis*, 5(2):369–411.

Land, S. R. and Friedman, J. H. (1997). Variable fusion: A new adaptive signal regression method. *Dept. Statistics, Carnegie Mellon Univ. Pittsburgh, Pittsburgh, PA, USA, Rep*, 656.

Lawson, A. B. (2000). Cluster modelling of disease incidence via rjmcmc methods: a comparative evaluation. *Statistics in Medicine*, 19(17-18):2361–2375.

Li, F. and Sang, H. (2019). Spatial homogeneity pursuit of regression coefficients for large datasets. *Journal of the American Statistical Association*, 114(527):1050–1062.

Lin, Y., Yu, M., Wang, S., Chappell, R., and Imperiale, T. F. (2016). Advanced colorectal neoplasia risk stratification by penalized logistic regression. *Statistical Methods in Medical Research*, 25(4):1677–1691.

Makalic, E. and Schmidt, D. F. (2015). A simple sampler for the horseshoe estimator. *IEEE Signal Processing Letters*, 23(1):179–182.

Makalic, E. and Schmidt, D. F. (2016). High-dimensional Bayesian regularised regression with the BayesReg package. *arXiv preprint arXiv:1611.06649*.

Masuda, R. and Inoue, R. (2022). Point event cluster detection via the bayesian generalized fused lasso. *ISPRS International Journal of Geo-Information*, 11(3).

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., and Teller, E. (1953). Equation of state calculations by fast computing machines. *The Journal of Chemical Physics*, 21(6):1087–1092.

Ninomiya, Y. (2021). Prior intensified information criterion. *arXiv preprint arXiv:2110.12145*.

Ninomiya, Y. and Kawano, S. (2016). AIC for the lasso in generalized linear models. *Electronic Journal of Statistics*, 10(2):2537–2560.

Olszewski, R. T. (2001). *Generalized feature extraction for structural pattern recognition in time-series data*. PhD thesis, Carnegie Mellon University.

Park, T. and Casella, G. (2008). The bayesian lasso. *Journal of the American Statistical Association*, 103(482):681–686.

Pollard, D. (1991). Asymptotics for least absolute deviation regression estimators. *Econometric Theory*, 7(2):186–199.

Polson, N. and Scott, J. (2010). Shrink globally, act locally: sparse Bayesian regularization and prediction. *Bayesian Statistics*, 9:501–538.

Polson, N. G., Scott, J. G., and Windle, J. (2013). Bayesian inference for logistic models using pólya–gamma latent variables. *Journal of the American Statistical Association*, 108(504):1339–1349.

Schwarz, G. (1978). Estimating the dimension of a model. *The annals of statistics*, 6(2):461–464.

She, Y. (2010). Sparse regression with exact clustering. *Electronic Journal of Statistics*, 4(none):1055 – 1096.

Shimamura, K., Ueki, M., Kawano, S., and Konishi, S. (2019). Bayesian generalized fused lasso modeling via neg distribution. *Communications in Statistics-Theory and Methods*, 48(16):4132–4153.

Song, Q. and Cheng, G. (2020). Bayesian fusion estimation via t shrinkage. *Sankhya A*, 82(2):353–385.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., and Van Der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B*, 64(4):583–639.

Stein, C. M. (1981). Estimation of the mean of a multivariate normal distribution. *The Annals of Statistics*, 9(6):1135–1151.

Sugasawa, S. and Murakami, D. (2021). Spatially clustered regression. *Spatial Statistics*, 44(100525).

Sugiura, N. (1978). Further analysis of the data by akaike's information criterion and the finite corrections. *Communications in Statistics - Theory and Methods*, 7(1):13–26.

Tian, Y., Bondell, H. D., and Wilson, A. (2019). Bayesian variable selection for logistic regression. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, 12(5):378–393.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288.

Tibshirani, R., Saunders, M., Rosset, S., Zhu, J., and Knight, K. (2005). Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(1):91–108.

Tibshirani, R. and Wang, P. (2008). Spatial smoothing and hot spot detection for cgh data using the fused lasso. *Biostatistics*, 9(1):18–29.

Tibshirani, R. J. and Taylor, J. (2011). The solution path of the generalized lasso. *The Annals of Statistics*, 39(3):1335–1371.

Tierney, L. and Kadane, J. B. (1986). Accurate approximations for posterior moments and marginal densities. *Journal of the American Statistical Association*, 81(393):82–86.

Viallon, V., Lambert-Lacroix, S., Hoefling, H., and Picard, F. (2016). On the robustness of the generalized fused lasso to prior specifications. *Statistics and Computing*, 26(1):285–301.

Wand, M. P., Ormerod, J. T., Padoan, S. A., and Frühwirth, R. (2011). Mean field variational Bayes for elaborate distributions. *Bayesian Analysis*, 6(4):847–900.

Wang, H., Li, B., and Leng, C. (2009). Shrinkage tuning parameter selection with a diverging number of parameters. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 71(3):671–683.

Watanabe, S. (2010a). Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory. *Journal of Machine Learning Research*, 11(12).

Watanabe, S. (2010b). Equations of states in singular statistical estimation. *Neural Networks*, 23(1):20–34.

Yan, X., Wang, H., Zhou, Y., Yan, J., Wang, Y., Wang, W., Xie, J., Yang, S., Zeng, Z., and Chen, X. (2022). Heterogeneous logistic regression for estimation of subgroup effects on hypertension. *Journal of Biopharmaceutical Statistics*, 32(6):969–985.

Yu, D., Lee, S. J., Lee, W. J., Kim, S. C., Lim, J., and Kwon, S. W. (2015). Classification of spectral data using fused lasso logistic regression. *Chemometrics and Intelligent Laboratory Systems*, 142:70–77.

Zhang, C.-H. (2010). Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics*, 38(2):894–942.

Zhao, Y. and Bondell, H. (2020). Solution paths for the generalized lasso with applications to spatially varying coefficients regression. *Computational Statistics & Data Analysis*, 142(106821).

Zhong, Y., Sang, H., Cook, S. J., and Kellstedt, P. M. (2023). Sparse spatially clustered coefficient model via adaptive regularization. *Computational Statistics & Data Analysis*, 177(107581).

Zou, H., Hastie, T., and Tibshirani, R. (2007). On the "degrees of freedom" of the lasso. *The Annals of Statistics*, 35(5):2173 – 2192.