BAYESIAN METHODS FOR DATA INTEGRATION AND HIGH DIMENSIONAL LINEAR MODEL WITH NON-SPARSITY

GUANLIN ZHANG

A DISSERTATION SUBMITTED TO
THE FACULTY OF GRADUATE STUDIES
IN PARTIAL FULFILMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

GRADUATE PROGRAM IN MATHEMATICS AND STATISTICS
YORK UNIVERSITY
TORONTO, ONTARIO

March 2025

Abstract

We address the problem of data integration where correlated data are collected across multiple platforms. Within each platform, the relationship between responses and predictors is modelled linearly. To extend this framework, we incorporate random errors from a broader class of sub-Gaussian and sub-exponential distributions. The primary objective is to identify key predictors across platforms, even as the number of predictors and observations grows infinitely large.

Our approach combines marginal response densities from different platforms into a composite likelihood and introduces a model selection criterion based on Bayesian composite posterior probabilities. Under regularity conditions, we demonstrate the consistency of this criterion in selecting the true model, even with a diverging model size. Moreover, when true models differ across platforms, the proposed method successfully recovers the union support of predictors—those predictors that are relevant in at least one platform. A Monte Carlo Markov Chain (MCMC) algorithm is implemented to perform this model selection.

Simulation studies reveal that integrating data from multiple platforms enhances model selection accuracy compared to using a single data source. In an application to real financial data, our method combines information from three financial indices to identify a common set of important predictors, resulting in a predictive model with higher accuracy and lower mean squared error than models relying on individual data sources.

In high-dimensional linear regression models, the sparsity assumption on regression coefficients $\boldsymbol{\beta} \in \mathbb{R}^p$ is often standard. However, this assumption may break down when most or all regression coefficients are non-zero, leading to substantial bias when using methods tailored for sparse models. To address this, we propose a novel Bayesian Grouping-Gibbs Sampling (BGGS) method, which relaxes the sparsity assumption. The BGGS method partitions $\boldsymbol{\beta}$ into k distinct groups, enabling efficient sampling in high-dimensional spaces.

We explore the selection of k through simulations and recommend the use of an "elbow plot" for precise determination. Theoretical analysis guarantees model selection consistency, parameter estimation consistency, and bounded prediction error under regularity conditions. Numerical experiments on two finite examples demonstrate the competitive advantage of BGGS in terms of parameter estimation and prediction accuracy. Finally, the method is applied to a financial dataset, showcasing its practical effectiveness in identifying robust predictive models.

Keywords: Bayesian method, Data integration, Gibbs sampling, Model selection, Sub-Gaussian, Sub-exponential, Union support recovery, Bayesian Grouping-Gibbs sampling, Non-sparse, High-dimensional, Linear regression.

Acknowledgements

First and foremost, I would like to express my deepest gratitude to my supervisors, Professor Yuehua Wu and Professor Xin Gao. Their unwavering support, insightful guidance, and continuous encouragement have been the cornerstone of this work. Their expertise, patience, and dedication have been instrumental in the successful development and completion of this dissertation.

I am also sincerely grateful to Professor Yuejiao Fu and Professor Steven Wang, who served as members of my supervisory committee. Their valuable feedback and thoughtful suggestions have greatly enhanced the quality of my research. My heartfelt thanks also go to my classmates, Shanshan Qin and Yan Zhang, for their constructive input and valuable discussions, which have significantly contributed to this dissertation.

Finally, I extend my deepest appreciation to my parents and friends for their constant support and encouragement throughout this journey. A special thanks goes to my mother for her unwavering support, endless patience, and the guidance and strength she has provided throughout my life. This dissertation is dedicated to Professor Hongmei Zhu, whose support and encouragement have given me the opportunity to pursue this path and see it through to the end. I wish her happiness and fulfillment every day.

Contents

List of Tables

List of Figures

1 Introduction

Over the past decade, data volume has surged across various domains, presenting both opportunities and challenges. While large-scale data enables more precise analyses and informed decision-making in fields such as finance, biology, and healthcare, it also introduces computational and methodological hurdles. The increasing complexity of datasets necessitates the integration of multiple data sources and the consideration of model variations, leading to union support recovery. Additionally, handling high-dimensional variables requires imposing sparsity assumptions to address ill-posed problems while also accounting for non-sparse cases when these assumptions do not hold. This dissertation explores these two key issues—data integration and non-sparsity—in specific contexts and proposes innovative methods to address them.

For the first issue, data integration is the process of combining data generated from different platforms to provide a consolidated analysis across multiple sources. This process is crucial in many contexts. For instance, in financial markets, indices are often influenced by the fluctuations of numerous stock prices. Simplifying the forecasting of multiple market indices requires identifying key stocks that are important across these indices. By integrating data from various financial indices, a consolidated analysis helps pinpoint critical stocks (Gao and Carroll (2017), Li et al. (2022)). Similarly, in biomedical applications, classifying a disease condition often relies on a collection of genetic biomarkers. Experimental platforms like mRNA expression, protein expression,

and RNAseq expression provide complementary data on genes. The integration of these datasets aims to identify important genes influencing disease processes across these platforms (Wu et al. (2012), Tian et al. (2004)).

A common assumption in many existing data integration methods is that different tasks share the same set of true predictors. However, this assumption may not hold in practice, as many related tasks have their distinct sets of true predictors (Caruana (1997)). For example, in satellite imaging, tasks such as predicting satellite keypoints, direct pose regression, and binary segmentation of the satellite foreground are all related to detecting and estimating a target spacecraft. Each task has a unique model, yet a joint model can learn features relevant to the combination of these tasks (Park and D'Amico (2024)). This multitask learning approach effectively integrates information, leveraging commonalities while preserving task-specific distinctions. Let the underlying model of each platform i be represented by $S_i \subseteq \{1, 2, \ldots, p_n\}$, where S_i is the set of the indices of all true predictors to the ith platform.

The goal of the model selection approach is to recover the union of all individual models, denoted by $\mathcal{S} = \cup_{i=1}^{k} \mathcal{S}_i$, $i = 1, \ldots k$, where k is the total number of platforms. In the literature, Wainwright (2009) and Tang and Nehorai (2010) analyzed the performance of penalized variable selection methods for recovering the union support. These studies explored the theoretical and practical aspects of penalized approaches in the context of union support recovery. Further, Ramesh et al. (2019) investigated the performance of group LASSO for recovering the union support under the assumption of sub-Gaussian errors, demonstrating its utility in high-dimensional settings.

The question of model selection on one data source has been extensively explored. In linear models, the Bayesian variable selection approach has been studied by Clyde et al. (1998), Clyde and George (2000), and Wolfe et al. (2004) among others. The Bayesian models used in Bayesian

variable selection often rely on variations of the hierarchical Bayesian model (George and McCulloch (1993)), which facilitates the integration of marginal posterior probabilities of candidate models given the observed dataset. The model with the highest posterior probability is selected. For a fixed number of predictors p, the consistency of Bayesian variable selection has been established by Fernández et al. (2001), Moreno and Girón (2005), Liang et al. (2008) and Casella et al. (2009), etc. When p diverges with the sample size, Bayes factor consistency has been proved by Berger et al. (2003), and Moreno et al. (2010).However, as Bayes factor consistency applies to pairwise model comparisons, there is a need to establish stronger model selection consistency for the true model against all competing models. Shang and Clayton (2011) and Shang and Li (2014) investigated the asymptotic properties of Bayesian model selection in linear models, where the number of predictors p grows with but does not exceed the sample size n. Their results demonstrate that under this framework, the true model is chosen with the highest probability. For data integration across multiple platforms, Gao and Carroll (2017) proposed a pseudo-likelihood based Bayesian information criterion and proved its model selection consistency for divergent p. Their approach, however, used frequentist methods to perform variable selection via group penalization on the pseudo-likelihood. The data sets across different platforms can be different types including discrete and continuous responses. Maity et al. (2019) extended Bayesian variable selection to handle data integration problems involving both survival and binary data using a shrinkage prior. However, the model selection consistency of their Bayesian approach has not been formally established. Additionally, Dai et al. (2020) introduced a quantile-based method for selecting predictors that influence responses at any quantile level, estimating nonzero parameters in high-dimensional settings. Despite these advances, the model selection properties of Bayesian methods for data integration remain underdeveloped. There is a pressing need for Bayesian data integration methods

capable of selecting the true underlying model with a probability that tends to 1, addressing both theoretical consistency and practical challenges.

For the second issue, the rapid advancements in data collection and measurement techniques have led to datasets with an increasingly large number of predictors or variables p, often exceeding the number of samples n. This high-dimensional scenario ($p > n$) poses significant challenges for regression modeling, as the estimation problem becomes ill-posed without additional assumptions on the regression coefficients.

A commonly adopted and natural assumption to address this issue is sparsity: the idea that only a subset of predictors, associated with non-zero elements of $\boldsymbol{\beta}$, significantly influence the response. These predictors are referred to as active predictors. By assuming sparsity, the complexity of the high-dimensional problem can be reduced, enabling feasible parameter estimation and variable selection.

Many frequentist methods apply well-known penalization techniques that shrink the regression coefficients towards zero. Some prominent methods include the least absolute shrinkage and selection operator (Lasso) (Tibshirani, 1996), the smoothly clipped absolute deviation (SCAD) (Fan and Li, 2001), the minimum concave penalty (MCP) (Zhang, 2010), and the adaptive LASSO (Zou, 2006). Many Bayesian methods have also been proposed for variable selection, including the stochastic search variable selection (George and McCulloch, 1993), empirical Bayes variable selection (George and Foster, 2000), spike and slab selection method (Ishwaran and Rao, 2005), penalized credible regions (Bondell and Reich, 2012), nonlocal prior method (Johnson and Rossell, 2012), among others. Shankar et al. (2015) evaluated a selection of both frequentist and Bayesian regression modeling ensembles based on Microbiome data, which are represented by variants of stability selection in conjunction with elastic net and spike-and-slab Bayesian model averaging. More

recent works have been proposed for variable selection under other model frameworks. Specifically, Narisetty et al. (2019) proposed a Skinny Gibbs sampling method for variable selection in logistic regression under the sparse assumption in the case of large p. Wang et al. (2021) proposed a one-step Bayesian variable selection method for partially linear models. Yi and Tang (2022) proposed a variational Bayesian approach for simultaneous variable selection and parameter estimation in high-dimensional linear mixed models. Lee and Goh (2023) introduced a Bayesian variable selection scheme employing the hybrid deterministic-deterministic variable selection (HD-DVS) algorithm, which guarantees rapid convergence to the global mode of the posterior model distribution.

Frequentist and Bayesian methods have been widely applied for simultaneous variable selection and parameter estimation in high-dimensional regression models under the assumption of sparsity. However, these methods may perform poorly when the majority of regression coefficients are non-zero, such as when the number of non-zero elements in $\boldsymbol{\beta}$ is comparable to or larger than the sample size n. This limitation arises because most existing approaches are fundamentally designed to exploit sparsity, which may not always hold in practical scenarios. Moreover, verifying the sparsity assumption can be challenging in real-world applications. Beyond sparsity, grouping structures are often present in high-dimensional regression problems. These structures reflect the relationships among predictors, where subsets of variables may exhibit shared effects. Methods for addressing grouping problems based on the penalization techniques have been well studied by shrinking the differences between regression coefficients towards zero, such as fused Lasso (Tibshirani et al., 2005), group lasso (Yuan and Lin, 2006), FGSG (Shen et al., 2012; Zhu et al., 2013), among others. However, the efficiency of model estimation via penalized methods may not be adequate due to the resulting nonconvex optimization problems and/or the mis-selection of regularization

parameters. More recently, sparse and grouping pursuit problems have also been studied in the Bayesian framework. Chen et al. (2015) proposed a Bayesian approach to tackle the sparse group selection problem in regression models by utilizing a Bayesian hierarchical formulation, where each candidate group is associated with a binary variable to indicate its activity status (active or not). Lee and Cao (2020) explored the use of a hierarchical group spike and slab prior for logistic regression models with group-structured covariates in high-dimensional settings.

In high-dimensional problems, linear regression models are often utilized under the assumption of sparsity. This framework typically focuses on prediction and the identification of significant predictors (variable selection). While sparsity is a common and effective assumption, it may not always reflect the underlying structure of the data. Departing from this assumption, our work addresses high-dimensional linear regression problems without relying on sparsity.

The challenge of non-sparsity has been explored in recent frequentist methods. For instance, Ding et al. (2021) investigated a scenario where $\boldsymbol{\beta}$ consists of two distinct groups: components in group I are large, while those in group II are small but not necessarily zero. Their work analyzed the asymptotic behavior of a ridge-regularized high-dimensional multivariate M-estimator for $\boldsymbol{\beta}$ in group II.

In this dissertation, regarding the first issue of data integration, we consider the problem in the Bayesian data integration framework and aim to develop a model selection method to recover the union support of the true predictors. Most existing data integration methods have focused on applications with Gaussian random errors. In real applications, light-tailed random errors such as uniform random errors or errors with bounded values may arise. Similarly, heavy-tailed random errors such as t distribution or χ^2 distribution can occur in data sets. To accommodate both light-tailed and heavy-tailed distributions, we extend the proposed method to sub-Gaussian and

sub-exponential family of error distributions so that the method is applicable to a broad range of data types. First, we consider random errors with a Sub-Gaussian distribution, which is a wide family with strong tail decay, that is the tail of sub-Gaussian distribution is dominated by the tail of normal distribution. A random variable X follows a sub-Gaussian distribution with a proxy variance σ^2, if for all $t \in \mathbb{R}$, $\mathbb{E}e^{tX} \leqslant e^{t^2\sigma^2/2}$. It has the property that $||X||_{\psi_2} \leqslant K$ for some $K > 0$, where $||X||_{\psi_2} = \sup_{p \geqslant 1} p^{-1/2}(\mathbb{E}X^p)^{1/p}$, which is another definition equivalent to the previous one. A random variable X follows $SubE(\sigma^2, b)$, if for all t such that $|t| < \frac{1}{b}$, $\mathbb{E}e^{tX} \leqslant e^{t^2\sigma^2/2}$. To further include more general cases of random errors with heavier tails than normal distribution, we consider the sub-exponential distribution which is a wider family of distributions including normal distribution, sub-Gaussian distributions and some distributions with heavier tails than normal distribution. Data from multiple platforms can be correlated. Instead of formulating the joint likelihood of all the data sets, we can formulate composite likelihood (Lindsay (1988); Cox and Reid (2004); Varin et al. (2011)) by compounding the marginal densities of the responses from different sources. Composite likelihood has also been adopted in a Bayesian framework (Cooley et al. (2007); Ribatet et al. (2012)). To solve the Bayesian variable selection problem on data integration, we can formulate the marginal density of the response variables from each data source, and combine the marginal densities to form a composite likelihood of the overall data. We assign priors to different competing models and calculate a composite posterior probability for each competing model. Then we formulate a model selection criterion based on composite posterior probabilities to select the union support of true predictors. To implement the Bayesian approach, a Gibbs sampling algorithm is used to explore the model space, compare different models and select the best model. Lastly, we apply our method to integrate the data from three different financial indices and the models learned from multiple platforms are shown to have better predictive performance

than other competing methods.

For the second issue addressing the challenge of releasing sparsity assumption in high-dimensional data, we continue to consider the general non-sparse case. Denote the group index sets by $\mathcal{G}_{\mathrm{I}} = \{j|\beta_j = O(1), j \in \{1, \ldots, p\}\}$ for Group I and $\mathcal{G}_{\mathrm{II}} = \{j|\beta_j = o(1), j \in \{1, \ldots, p\}\}$ for Group II. Then $\boldsymbol{\beta} = (\boldsymbol{\beta}_{\mathcal{G}_{\mathrm{I}}}^\top, \boldsymbol{\beta}_{\mathcal{G}_{\mathrm{II}}}^\top)^\top$, where $\boldsymbol{\beta}_{\mathcal{G}_{\mathrm{I}}} = \{\beta_j|j \in \mathcal{G}_{\mathrm{I}}\}$ and $\boldsymbol{\beta}_{\mathcal{G}_{\mathrm{II}}} = \{\beta_j|j \in \mathcal{G}_{\mathrm{II}}\}$. The components of $\boldsymbol{\beta}_{\mathcal{G}_{\mathrm{II}}}$ are relatively small but still significant. Specially, $\mathcal{G}_{\mathrm{I}}$ can be empty, implying that all the components of $\boldsymbol{\beta}$ are small. Ding et al. (2021) proposed to estimate the expected mean squared error of $\boldsymbol{\beta}_{\mathcal{G}_{\mathrm{II}}}$ using a two-step estimation procedure whereas $\boldsymbol{\beta}_{\mathcal{G}_{\mathrm{I}}}$ is estimated by some penalized methods including LASSO, SCAD and MCP. However, the two-step estimation may produce a large bias since these penalized methods are designed for sparse models while $\boldsymbol{\beta}$ is non-sparse in their study. In light of the grouping penalization methods, say, fused Lasso and FGSG, we adopt a flexible grouping technique to tackle this challenging estimation problem when $p > n$ with non-sparsity. It is noted that the $\boldsymbol{\beta}_{\mathcal{G}_{\mathrm{I}}}$ or $\boldsymbol{\beta}_{\mathcal{G}_{\mathrm{II}}}$ may also contain subgroups. Within a subgroup, the regression coefficients are similar or identical. Notably, the number of subgroups will not exceed the sample size n, while the number of nonzero coefficients can be larger than n. To the best of our knowledge, there are few Bayesian methods specifically designed to address non-sparse high-dimensional linear regression models using grouping techniques. One advantage of the Bayesian method for model estimation is that Markov Chain Monte Carlo (MCMC) techniques can be used to explore the posterior distributions, which often offer a more informative approach than the corresponding penalization method requiring a selection of tuning parameters and a highly nonconvex optimization problem (Narisetty et al., 2019). We devise a grouping-Gibbs sampler for high-dimensional posterior sampling, leveraging grouping information to reduce computational complexity and achieve

computational efficiency. Under regular conditions, theoretically, we demonstrate weak and strong model selection consistency, parameter estimation consistency and boundedness on mean squared error (MSE). In the context of linear regression models, parameter estimation and prediction are of prime concern. This dissertation focuses on the estimation of $\boldsymbol{\beta}$ and model's prediction accuracy rather than on identifying the groups of $\boldsymbol{\beta}$, although we assume $\boldsymbol{\beta}$ has a grouping structure. The competitive advantages of the proposed method are demonstrated by simulated examples and a financial dataset.

In this dissertation, we explore two key Bayesian approaches. In Chapter 2, we introduce model selection via composite likelihood, establishing model selection consistency, developing a Gibbs sampling algorithm for efficient model space exploration, and validating the approach through simulations and real data analysis. In Chapter 3, we present Bayesian grouping-Gibbs sampling (BGGS) for non-sparse linear models, demonstrating both model and parameter consistency in non-sparse settings, developing a group Gibbs sampling algorithm for coefficient estimation, and showing superior performance compared to existing methods in simulations and real data analysis. These techniques address traditional limitations while enhancing prediction accuracy and model selection consistency.

2 Bayesian Model Selection via Composite Likelihood for High-dimensional Data Integration

2.1 Methodology

Assume we have data collected from multiple experimental platforms, with responses that are related to the same research question. The primary objective is to identify important predictors that influence at least one of the responses across these platforms.

Let $\boldsymbol{y}_1, \boldsymbol{y}_2, ..., \boldsymbol{y}_k$ denote the response vectors from k different platforms, where $\boldsymbol{y}_i = (y_{i1}, \ldots, y_{in})^\top$, represents the responses observed for n subjects on the ith platform, for $i \in \{1, \ldots, k\}$. For each platform, we have an $n \times p_n$ design matrix X_i, where $X_{i[lm]}$ denotes the value of mth predictor for the ℓth subject, with $\ell \in \{1, \ldots, n\}$, and $m \in \{1, \ldots, p_n\}$. While the platforms share a common set of p_n predictors, the predictor values can vary from different platforms. To accommodate high-dimensional data, we allow the number of predictors p_n to increase with the sample size.

The response vectors $\boldsymbol{y}_i$'s may exhibit dependent relationships due to shared underlying factors or correlated measurement processes across platforms. Instead of modeling the complex joint distribution of the responses across platforms, we adopt a composite likelihood approach. Composite likelihood methods construct a pseudo-likelihood by combining the marginal likelihoods of the individual datasets (Lindsay (1988), Cox and Reid (2004)).

$$CL(\boldsymbol{y}_1, ..., \boldsymbol{y}_k) = \prod_{i=1}^{k} f(\boldsymbol{y}_i; \theta),$$

where θ represents the parameters in the marginal likelihoods. Although the composite likelihood is not a true likelihood, it retains many desirable properties similar to those of the true likelihood (Kent (1982), Varin et al. (2011), Chandler and Bate (2007)). For instance, the score function of the composite likelihood is a linear combination of the score functions derived from the true likelihoods. As a result, the composite score function has zero expectations under the true model, ensuring that the estimation procedure is unbiased in expectation.

Furthermore, the maximum composite likelihood estimator (MCLE) is consistent, meaning it converges to the true parameter values as the sample size increases. Additionally, the MCLE possesses an asymptotic normal distribution, which facilitates statistical inference, such as hypothesis testing and the construction of confidence intervals. These properties make the composite likelihood approach a robust and efficient tool for modeling and estimation, particularly in complex or high-dimensional settings where constructing a joint likelihood is infeasible or computationally expensive.

Consider the setting that $\boldsymbol{y}_i$ and X_i have linear relationship $\boldsymbol{y}_i = X_i\boldsymbol{\beta}_i + \boldsymbol{\epsilon}_i$, where $\boldsymbol{\beta}_i = (\beta_{i1}, ..., \beta_{ip_n})^\top$ are linear coefficients. We assume that the models are sparse and share identical sparsity patterns across the k models. In Section 2.3, this assumption will be relaxed, and the method will be extended to account for situations where the sparsity patterns differ among the models. Define the indicator vector $\boldsymbol{\gamma}_i = (\gamma_{i1}, ..., \gamma_{ip_n})^\top$, where $\gamma_{ij} = I(\beta_{ij} \neq 0)$, for $i \in \{1, \ldots, k\}$ and $j \in \{1, \ldots, p_n\}$. As for current section, the models across all platforms are assumed to be identical, we ignore the subscript i and use $\boldsymbol{\gamma} = (\gamma_{\cdot 1}, ..., \gamma_{\cdot p_n})^\top$ to denote the common model of all

platforms and $\boldsymbol{\gamma}_0$ to denote the values under the true model. This indicator vector uniquely defines a model in the model space, which includes all the models under consideration. Throughout the rest of this dissertation, the model under the indicator vector $\boldsymbol{\gamma}$ is written as "model $\boldsymbol{\gamma}$" with model size denoted as $|\boldsymbol{\gamma}|$. Let s_n denote the model size of the true model and let r_n denote the maximum size of the models in the model space. We have $|\boldsymbol{\gamma}| \leqslant r_n \leqslant p_n$.

We assume that the marginal distribution of $\boldsymbol{y}_i$ follows a hierarchical Bayesian model (Shang and Clayton (2011)), which is a variation of the model proposed by George and McCulloch (1993):

$$
\begin{aligned}
&\boldsymbol{y}_i|\boldsymbol{\beta}_i, \sigma_i^2 \sim X_i\boldsymbol{\beta}_i + \boldsymbol{\epsilon}_i,\\
&\beta_{ij}|\gamma_{\cdot j}, \sigma_i^2 \sim (1-\gamma_{\cdot j})\delta_0 + \gamma_{\cdot j}\mathcal{N}(0, c_{ij}\sigma_i^2),\\
&1/\sigma_i^2 \sim \chi^2_{v_i},\\
&p(\boldsymbol{\gamma}) = \begin{cases} \pi_{\boldsymbol{\gamma}}, & \text{if } |\boldsymbol{\gamma}| \leqslant r_n,\\ 0, & \text{otherwise.} \end{cases}
\end{aligned}
$$

In this model setup, the variable δ_0 is a point-mass measure concentrated at zero. The random errors are denoted by $\boldsymbol{\epsilon}_i = (\epsilon_{i1}, ..., \epsilon_{in})^\top$, where $\epsilon_{i\ell}$, $\ell \in \{1, \ldots, n\}$, are assumed to be independent and follow an identical sub-Gaussian or a sub-exponential distribution with the variance proxy σ_i^2. The inverse of the variance proxy σ_i^2 has a prior chi-square distribution with degrees of freedom v_i, where v_is are fixed hyper-parameters. Let $\Sigma_i = \text{diag}(\boldsymbol{c}_i)$ with $\boldsymbol{c}_i = (c_{ij})^\top_{1\leqslant j\leqslant p_n}$ being a vector of positive entries. Let $\Sigma_{i,\boldsymbol{\gamma}}$ be the $|\boldsymbol{\gamma}| \times |\boldsymbol{\gamma}|$ dimensional sub-diagonal matrix of Σ_i, $\boldsymbol{\beta}_{i,\boldsymbol{\gamma}}$ be the $|\boldsymbol{\gamma}| \times 1$ sub-vector of $\boldsymbol{\beta}_i$, and $X_{i,\boldsymbol{\gamma}}$ be the $n \times |\boldsymbol{\gamma}|$ sub-matrix comprising the columns of X_i corresponding to the nonzero entries of $\boldsymbol{\gamma}$. Let $Z_i = (\boldsymbol{y}_i, X_i)$ denote the full data set. When p_n is finite, we could use a uniform prior on the model space. However, in this dissertation, we address th scenario where p_n grows to infinity as the sample size increases. Consequently, it is necessary to enforce

sparsity in the model by imposing prior probabilities $p(\boldsymbol{\gamma})$ on the models, based on their model sizes. If $|\boldsymbol{\gamma}| > r_n$, the model has zero prior probability. Otherwise, the model is assigned a prior $\pi_{\boldsymbol{\gamma}} \propto q^{|\boldsymbol{\gamma}|}(1-q)^{r_n-|\boldsymbol{\gamma}|}$, where $\log q = -M \log p_n$, and M is a constant.

Within each platform, the errors from different subjects are assumed to be independent. However, across different platforms, the errors for the same subject $\epsilon_{1\ell}, \ldots, \epsilon_{k\ell}$ may be correlated. This leads to the dependent relationships among $\boldsymbol{y}_1, \boldsymbol{y}_2, ..., \boldsymbol{y}_k$. Instead of directly formulating the full likelihood, which is difficult to construct, the composite likelihood method is used to model multiple data sets together:

$$P_C(\sigma_1^2, ..., \sigma_k^2, \boldsymbol{\beta}_1, ..., \boldsymbol{\beta}_k, \boldsymbol{\gamma} | Z_1, ..., Z_k) \propto p(\boldsymbol{\gamma}) \prod_{i=1}^{k} \Big\{ f(\boldsymbol{y}_i | X_i, \boldsymbol{\beta}_i, \sigma_i) \\ \frac{1}{2^{\frac{v_i}{2}}\Gamma(\frac{v_i}{2})} \sigma_i^{-v_i+2} \exp[-\frac{1}{2\sigma_i^2}] \prod_{j=1}^{p_n} \Big\{ \frac{1}{\sqrt{2\pi c_{ij}}\sigma_i} \exp[-\frac{1}{2\sigma_i^2 c_{ij}} \beta_{ij}^2] \Big\}^{\gamma_j} \Big\}.$$

If $f(\boldsymbol{y}_i | X_i, \boldsymbol{\beta}_i, \sigma_i)$ is a normal density, we can integrate out $\boldsymbol{\beta}_1, ..., \boldsymbol{\beta}_k, \sigma_1, ..., \sigma_k$ and obtain the composite posterior probability (Shang and Clayton, 2011):

$$\begin{aligned} & P_C(\boldsymbol{\gamma} | Z_1, ..., Z_k) \\ =& c \cdot p(\boldsymbol{\gamma}) \prod_{i=1}^{k} \Big\{ |W_{i,\boldsymbol{\gamma}}|^{-\frac{1}{2}} (1 + \boldsymbol{y}_i^\top (I_n - X_{i,\boldsymbol{\gamma}} (X_{i,\boldsymbol{\gamma}}^\top X_{i,\boldsymbol{\gamma}} + \Sigma_{i,\boldsymbol{\gamma}}^{-1})^{-1} X_{i,\boldsymbol{\gamma}}^\top) \boldsymbol{y}_i)^{-\frac{n+v_i}{2}} \Big\}, \end{aligned} \tag{2.1}$$

where c is a normalizing constant, $U_{i,\boldsymbol{\gamma}} = \Sigma_{i,\boldsymbol{\gamma}}^{-1} + X_{i,\boldsymbol{\gamma}}^\top X_{i,\boldsymbol{\gamma}}$, and $W_{i,\boldsymbol{\gamma}} = \Sigma_{i,\boldsymbol{\gamma}}^{1/2} U_{i,\boldsymbol{\gamma}} \Sigma_{i,\boldsymbol{\gamma}}^{1/2}$. For the special case that $\boldsymbol{\gamma}_0 = \mathbf{0}$, this formula still works if we assign that $X_{i,\varnothing} = 0, \Sigma_{i,\varnothing} = U_{i,\varnothing} = W_{i,\varnothing} = 1$. Incorporating the normalizing constant, Equation (2.1) is a composite posterior probability distribution for all the competing models in the model space. If $f(\boldsymbol{y}_i | X_i, \boldsymbol{\beta}_i, \sigma_i)$ is a sub-Gaussian or sub-exponential density, analytically integrating out $\boldsymbol{\beta}_1, ..., \boldsymbol{\beta}_k$, and $\sigma_1, ..., \sigma_k$ becomes challenging. Consequently, Equation (2.1) does not represent a true composite posterior probability of the model given the data. However, it still defines a valid probability distribution over the model

space. Remarkably, it retains model selection consistency and assigns the highest probability to the true model, as demonstrated in Theorems 2.2.1 and 2.2.2 in Section 2.2. Leveraging this property, we adopt Equation (2.1) as a model selection criterion and select the model that maximizes this criterion.

2.2 Model selection consistency

In this section, we establish the model selection consistency property for the criterion introduced in Section 2.1. Specifically, we show that the proposed method selects the true model with a probability tending to one as the sample size increases. This result holds for a broad class of error distributions, including sub-Gaussian and sub-exponential distributions.

Let $\boldsymbol{\gamma}\backslash\boldsymbol{\gamma}'$ denote a vector with the jth element $(\boldsymbol{\gamma}\backslash\boldsymbol{\gamma}')_j = I(\gamma_{\cdot j} = 1, \gamma'_{\cdot j} = 0), j = 1, ..., p_n$. If $\boldsymbol{\gamma}\backslash\boldsymbol{\gamma}' = \mathbf{0}$, then $\boldsymbol{\gamma}$ is nested in model $\boldsymbol{\gamma}'$. Let $\boldsymbol{\gamma}_0$ denote the true model and $\beta^{\mathbf{0}}$ denote the true regression coefficients. Define the collection of overfitting models $S_1 = \{\boldsymbol{\gamma} \mid \boldsymbol{\gamma}_0\backslash\boldsymbol{\gamma} = \mathbf{0}, \boldsymbol{\gamma_0} \neq \boldsymbol{\gamma}\}$ and the collection of wrong models $S_2 = \{\boldsymbol{\gamma} \mid \boldsymbol{\gamma}_0\backslash\boldsymbol{\gamma} \neq \mathbf{0}\}$. The collection of all models is denoted by $S_1 \cup S_2 \cup \{\boldsymbol{\gamma}_0\}$.

2.2.1 Model selection with sub-Gaussian random error

First, we focus on the setting where the random errors follow sub-Gaussian distributions. When $\boldsymbol{\gamma}_0 \neq \mathbf{0}$, we define

$$\varphi_{\min}(n) = \min_{i\in\{1,...,k\}} \min_{\boldsymbol{\gamma}\in S_2} \lambda_-(\frac{1}{n} X^\top_{i,\boldsymbol{\gamma}_0\backslash\boldsymbol{\gamma}}(I_n - P_{i,\boldsymbol{\gamma}})X_{i,\boldsymbol{\gamma}_0\backslash\boldsymbol{\gamma}})$$

and

$$\varphi_{\max}(n) = \max_{i\in\{1,...,k\}} \max_{\boldsymbol{\gamma}\in S_2} \lambda_+(\frac{1}{n} X^\top_{i,\boldsymbol{\gamma}_0\backslash\boldsymbol{\gamma}}X_{i,\boldsymbol{\gamma}_0\backslash\boldsymbol{\gamma}}),$$

where $P_{i,\boldsymbol{\gamma}} = X_{i,\boldsymbol{\gamma}}(X_{i,\boldsymbol{\gamma}}^\top X_{i,\boldsymbol{\gamma}})^{-1}X_{i,\boldsymbol{\gamma}}^\top$ is the projection matrix, and $\lambda_-(A)$ and $\lambda_+(A)$ denote the minimal and maximal eigenvalues of a square matrix A, respectively. Under the following assumptions, we prove the consistency of the model selection criterion.

Assumption 2.2.1 *There exist positive constants C_1 and C_2 such that*

$$\liminf_n \varphi_{\min}(n) \geqslant C_1 \text{ and } \limsup_n \varphi_{\max}(n) \leqslant C_2.$$

There exists a constant $\delta \in (0,1)$ and a constant $C_3 > 0$ such that when n is large enough, for $i \in \{1,\ldots,k\}$,

$$\inf_{\boldsymbol{\gamma}\in S_1} \lambda_-(\frac{1}{n}X_{i,\boldsymbol{\gamma}\setminus\boldsymbol{\gamma}_0}^\top(I_n - P_{i,\boldsymbol{\gamma}_0})X_{i,\boldsymbol{\gamma}\setminus\boldsymbol{\gamma}_0}) \geqslant C_3 n^{-\delta}.$$

There exists a positive constant C_4 such that every entry of X_i is uniformly bounded by C_4. There exists C_5, C_6 greater than zero such that

$$C_5 = \liminf_n \min_{i\in\{1,\ldots,k\}} \min_{\boldsymbol{\gamma}} \frac{1}{n}\lambda_-(X_{i,\boldsymbol{\gamma}}^\top X_{i,\boldsymbol{\gamma}}),$$

$$C_6 = \limsup_n \max_{i\in\{1,\ldots,k\}} \max_{\boldsymbol{\gamma}} \frac{1}{n}\lambda_+(X_{i,\boldsymbol{\gamma}}^\top X_{i,\boldsymbol{\gamma}}).$$

Remark 2.2.1 *Assumption 2.2.1 is commonly employed in high-dimensional settings to ensure that the eigenvalues of the design matrix remain bounded as the dimensions of the design matrix grow to infinity.*

Assumption 2.2.2 *It is assumed that there exist the following constants and sequences:*

When $\boldsymbol{\gamma}_0 \neq \mathbf{0}$, there exists a positive sequence ψ_n such that

$$\min_{i\in\{1,\ldots,k\}} \min_{j\in\boldsymbol{\gamma}_0} |\beta_{ij}^0| \geqslant \psi_n \text{ and } n\psi_n^2 \to \infty.$$

There exists a positive sequence $k_n = \max_{i\in\{1,\ldots,k\}} \max_{j\in\boldsymbol{\gamma}^0} |\beta_{ij}^0|$, such that

$$k_n^2 s_n = O(\underline{\phi}_n),$$

where $\underline{\phi}_n$ *is a positive sequence such that*

$$\min_{1 \leqslant j \leqslant p_n, 1 \leqslant i \leqslant k} c_{ij} \geqslant \underline{\phi}_n.$$

There exists a positive sequence $\overline{\phi}_n = O(n^{\delta_0})$ *for some* $\delta_0 > 0$ *such that the proportions of variance in different normal distributions are bounded by*

$$\max_{i \in \{1,\ldots,k\}} \max_{1 \leqslant j \leqslant p_n} c_{ij} \leqslant \overline{\phi}_n.$$

Remark 2.2.2 *Assumption 2.2.2 requires the nonzero parameters to be bounded away from zero at a specific rate. Additionally, it imposes both lower and upper bounds on* c_{ij}*, the variance parameters of the prior distribution for the nonzero parameters, which ensures that the true values of the regression coefficients have a high probability of being selected while keeping the variance of the prior within the necessary range.*

Assumption 2.2.3 *It is assumed that* $s_n \leqslant r_n < n$*, and*

$$r_n \max\{\log p_n, \log n\} = o(n \log(1 + \min\{1, \psi_n^2\})),$$

where ψ_n *is defined in Assumption 2.2.2.*

Remark 2.2.3 *Assumption 2.2.3 specifies the rate of the maximum dimension* r_n *of candidate models relative to the sample size* n*. When* $\psi_n > 0$*, this assumption can be simplified as* $r_n \max\{\log p_n, \log n\} = o(n)$*. It clarifies that the maximum model size considered cannot exceed the size of the dataset and is often much smaller.*

Assumption 2.2.4 *The prior* $p(\boldsymbol{\gamma})$ *satisfies*

$$p(\boldsymbol{\gamma}) = q^{|\boldsymbol{\gamma}|}(1-q)^{p_n - |\boldsymbol{\gamma}|},$$

where $q = p_n^{-M}$ *for some constant* $M > 0$*.*

Remark 2.2.4 *The specified prior favors models that contain only a few relevant predictors, allowing us to efficiently initialize from a small model.*

Theorem 2.2.1 *Under Assumptions 2.2.1, 2.2.2, 2.2.3, 2.2.4, if* $p_n^{\alpha_0-2M/k} = o(n^{1-\delta}\underline{\phi}_n)$ *for some* $\alpha_0 > 1$, *then*

$$\sup_{c_{i,1},...,c_{i,p_n}\in[\underline{\phi}_n,\overline{\phi}_n],i\in\{1,...,k\}} \max_{\boldsymbol{\gamma}\neq\boldsymbol{\gamma}_0} P_C(\boldsymbol{\gamma}|Z_1,Z_2,...,Z_k)/P_C(\boldsymbol{\gamma}_0|Z_1,Z_2,...,Z_k) \xrightarrow{p} 0.$$

If $p_n^{\alpha_0-2(M-1)/k} = o(n^{1-\delta}\underline{\phi}_n)$, *then*

$$\sup_{c_{i,1},...,c_{i,p_n}\in[\underline{\phi}_n,\overline{\phi}_n],i\in\{1,...,k\}} \sum_{\boldsymbol{\gamma}\neq\boldsymbol{\gamma}_0} P_C(\boldsymbol{\gamma}|Z_1,Z_2,...,Z_k) \xrightarrow{p} 0.$$

Without loss of generality, we use C to represent different constants throughout the proof. For notational simplicity, we assume that all σ_i^2's are identical and represent them as σ^2. We begin by decomposing the composite posterior probability into five terms. For each term, we derive the asymptotic bound in the corresponding technical lemmas. We have

$$\begin{aligned}
&-\log(P_C(\boldsymbol{\gamma}|Z_1,...,Z_k)/P_C(\boldsymbol{\gamma}_0|Z_1,...,Z_k)) = -\log(\frac{p(\boldsymbol{\gamma})}{p(\boldsymbol{\gamma}_0)})+\\
&\qquad\prod_{i=1}^{k}\Big\{\frac{1}{2}\log(\frac{|W_{i,\boldsymbol{\gamma}}|}{|W_{i,\boldsymbol{\gamma}_0}|}) + \frac{n+v_i}{2}\log(\frac{1+\boldsymbol{y}_i^\top(I_n - X_{i,\boldsymbol{\gamma}}U_{i,\boldsymbol{\gamma}}^{-1}X_{i,\boldsymbol{\gamma}}^\top)\boldsymbol{y}_i}{1+\boldsymbol{y}_i^\top(I_n - X_{i,\boldsymbol{\gamma}_0}U_{i,\boldsymbol{\gamma}_0}^{-1}X_{i,\boldsymbol{\gamma}_0}^\top)\boldsymbol{y}_i})\Big\}\\
&= -\log(\frac{p(\boldsymbol{\gamma})}{p(\boldsymbol{\gamma}_0)}) + \prod_{i=1}^{k}\Big\{\frac{1}{2}\log(\frac{|W_{i,\boldsymbol{\gamma}}|}{|W_{i,\boldsymbol{\gamma}_0}|})\\
&+ \frac{n+v_i}{2}\log(\frac{1+\boldsymbol{y}_i^\top(I_n - X_{i,\boldsymbol{\gamma}}U_{i,\boldsymbol{\gamma}}^{-1}X_{i,\boldsymbol{\gamma}}^\top)\boldsymbol{y}_i}{1+\boldsymbol{y}_i^\top(I_n - P_{i,\boldsymbol{\gamma}})\boldsymbol{y}_i})\\
&- \frac{n+v_i}{2}\log(\frac{1+\boldsymbol{y}_i^\top(I_n - X_{i,\boldsymbol{\gamma}_0}U_{i,\boldsymbol{\gamma}_0}^{-1}X_{i,\boldsymbol{\gamma}_0}^\top)\boldsymbol{y}_i}{1+\boldsymbol{y}_i^\top(I_n - P_{i,\boldsymbol{\gamma}_0})\boldsymbol{y}_i})\\
&+ \frac{n+v_i}{2}\log(\frac{1+\boldsymbol{y}_i^\top(I_n - P_{i,\boldsymbol{\gamma}})\boldsymbol{y}_i}{1+\boldsymbol{y}_i^\top(I_n - P_{i,\boldsymbol{\gamma}_0})\boldsymbol{y}_i})\Big\} = -T_1 + \prod_{i=1}^{k}\{T_2^{(i)} + T_3^{(i)} - T_4^{(i)} + T_5^{(i)}\}.
\end{aligned} \tag{2.2}$$

Denote the above summands by $T_1, T_2^{(i)}, T_3^{(i)}, T_4^{(i)}, T_5^{(i)}$.

Lemma 2.2.1 *If $\boldsymbol{\gamma} \in S_1$, then uniformly for $c_{i1}, ..., c_{ip} \in [\underline{\phi}_n, \overline{\phi}_n]$,*

$$T_2^{(i)} \geqslant 1/2(|\boldsymbol{\gamma}| - s_n)\log(1 + C_3 n^{1-\delta}\underline{\phi}_n).$$

If $\boldsymbol{\gamma}_0 \neq \mathbf{0}$, when $\boldsymbol{\gamma} \in S_2$,

$$T_2^{(i)} \geqslant -1/2 s_n \log(1 + C_2 n\overline{\phi}_n).$$

If $\boldsymbol{\gamma}_0 = \mathbf{0}$, $S_2 = \varnothing$. In this case, $T_2^{(i)}$ is not considered.

Proof of Lemma 2.2.1. First we consider the scenario when $\boldsymbol{\gamma} \in S_1$.

$$\begin{aligned}
|U_{i,\boldsymbol{\gamma}}| =& |U_{i,\boldsymbol{\gamma}_0}||\Sigma^{-1}_{i,\boldsymbol{\gamma}\setminus\boldsymbol{\gamma}_0} + X^\top_{i,\boldsymbol{\gamma}\setminus\boldsymbol{\gamma}_0}(I_n - X_{i,\boldsymbol{\gamma}_0}U^{-1}_{i,\boldsymbol{\gamma}_0}X^\top_{i,\boldsymbol{\gamma}_0})X_{i,\boldsymbol{\gamma}\setminus\boldsymbol{\gamma}_0}| \\
\geqslant& |U_{i,\boldsymbol{\gamma}_0}||\Sigma^{-1}_{i,\boldsymbol{\gamma}\setminus\boldsymbol{\gamma}_0} + X^\top_{i,\boldsymbol{\gamma}\setminus\boldsymbol{\gamma}_0}(I_n - P_{i,\boldsymbol{\gamma}_0})X_{i,\boldsymbol{\gamma}\setminus\boldsymbol{\gamma}_0}| \\
\geqslant& |U_{i,\boldsymbol{\gamma}_0}||\Sigma^{-1}_{i,\boldsymbol{\gamma}\setminus\boldsymbol{\gamma}_0} + C_3 n^{1-\delta} I_{\boldsymbol{\gamma}\setminus\boldsymbol{\gamma}_0}|.
\end{aligned}$$

Thus

$$\frac{|W_{i,\boldsymbol{\gamma}}|}{|W_{i,\boldsymbol{\gamma}_0}|} \geqslant \det((1 + C_3 n^{1-\delta}\underline{\phi}_n)I_{|\boldsymbol{\gamma}\setminus\boldsymbol{\gamma}_0|}) = (1 + C_3 n^{1-\delta}\underline{\phi}_n)^{|\boldsymbol{\gamma}|-s_n}.$$

So

$$T_2^{(i)} \geqslant 1/2(|\boldsymbol{\gamma}| - s_n)\log(1 + C_3 n^{1-\delta}\underline{\phi}_n).$$

If $\boldsymbol{\gamma} \in S_2$, noting that $|W_{i,\boldsymbol{\gamma}}| \geqslant 1$, by Assumption 2.2.1,2.2.2,

$$\begin{aligned}
T_2^{(i)} &= 1/2\log(\frac{|W_{i,\boldsymbol{\gamma}}|}{|W_{i,\boldsymbol{\gamma}_0}|}) \geqslant -1/2\log|W_{i,\boldsymbol{\gamma}_0}| = -1/2\log(|U_{i,\boldsymbol{\gamma}_0}||\Sigma_{i,\boldsymbol{\gamma}_0}|) \\
&= -1/2\log(|I_{s_n} + X^\top_{\boldsymbol{i},\boldsymbol{\gamma}_0}X_{i,\boldsymbol{\gamma}_0}\Sigma_{i,\boldsymbol{\gamma}_0}|) \geqslant -1/2 s_n\log(1 + C_2 n\overline{\phi}_n).
\end{aligned}$$

□

Lemma 2.2.2 *Suppose ϵ_ℓ follows sub-Gaussian distribution, $\mathrm{var}(\boldsymbol{\epsilon}) \leqslant \sigma^2$, for $\ell \in \{1, \ldots, n\}$.*

(a). Let $\boldsymbol{v}_{\boldsymbol{\gamma}} = (I_n - P_{\boldsymbol{\gamma}})X_{\boldsymbol{\gamma}_0\setminus\boldsymbol{\gamma}}\boldsymbol{\beta}^0_{\boldsymbol{\gamma}_0\setminus\boldsymbol{\gamma}}$. If S_2 is not empty, then

$$\max_{\boldsymbol{\gamma}\in S_2} |\boldsymbol{v}^\top_{\boldsymbol{\gamma}}\boldsymbol{\epsilon}|/||\boldsymbol{v}_{\boldsymbol{\gamma}}||_2 = O_p(\sqrt{r_n \log p_n}),$$

where we adopt the convention that $|\boldsymbol{v}_{\boldsymbol{\gamma}}^{\top}\boldsymbol{\epsilon}|/||\boldsymbol{v}_{\boldsymbol{\gamma}}||_2 = 0$ *when* $\boldsymbol{v}_{\boldsymbol{\gamma}} = \mathbf{0}$.

(b). If S_1 *is not empty, then there exists a constant* $C > 0$ *so that for any* $\alpha > 1$*, with probability approaching 1,*

$$\max_{\boldsymbol{\gamma} \in S_1} \boldsymbol{\epsilon}^{\top}(P_{\boldsymbol{\gamma}} - P_{\boldsymbol{\gamma}_0})\boldsymbol{\epsilon} \leqslant C\alpha(|\boldsymbol{\gamma}| - s_n)\log p_n.$$

(c). If S_2 *is not empty, and we adopt the convention that* $\boldsymbol{\epsilon}^{\top}P_{\boldsymbol{\gamma}}\boldsymbol{\epsilon}/|\boldsymbol{\gamma}| = 0$ *when* $\boldsymbol{\gamma}$ *is null, then there exists a constant* $C > 0$*, so that for any* $\alpha > 1$*, with probability approaching 1,*

$$\max_{\boldsymbol{\gamma} \in S_2} \boldsymbol{\epsilon}^{\top}P_{\boldsymbol{\gamma}}\boldsymbol{\epsilon} \leqslant C\alpha|\boldsymbol{\gamma}|\log p_n.$$

The proof of this lemma follows similar arguments as in the proof of Lemma 2.2.4 using Proposition 1.1 in Götze et al. (2021).

Lemma 2.2.3 *As* $n \to \infty$*, uniformly for* $\boldsymbol{\gamma} \in S_1$*,*

$$T_5^{(i)} \geqslant -1/2(|\boldsymbol{\gamma}| - s_n)\alpha_0 \log p_n$$

with probability tending to 1, where α_0 *is a constant greater than 1. Similarly, when* $\boldsymbol{\gamma}_0 \neq \mathbf{0}$*, uniformly for* $\boldsymbol{\gamma} \in S_2$*,*

$$T_5^{(i)} \geqslant -1/2(n + v_i)\log(1 + C'\psi_n^2)$$

with probability tending to 1, where C' *is a constant. When* $\boldsymbol{\gamma}_0 = \mathbf{0}$*,* S_2 *is empty, thus* $T_5^{(i)}$ *can be ignored.*

Proof of Lemma 2.2.3. We consider the two cases $\boldsymbol{\gamma} \in S_1$ and $\boldsymbol{\gamma} \in S_2$ separately. Assumption 2.2.3 implies that $r_n \log p_n = o(n\log\{1 + \min\{1, \psi_n^2\}\})$. This entails $r_n \log p_n = o(n\psi_n^2)$. Let $\boldsymbol{v}_{i,\boldsymbol{\gamma}} = (I_n - P_{i,\boldsymbol{\gamma}})X_{i,\boldsymbol{\gamma}_0\setminus\boldsymbol{\gamma}}\boldsymbol{\beta}^0_{i,\boldsymbol{\gamma}_0\setminus\boldsymbol{\gamma}}$. From Lemma 2.2.2, there exists $C > 0$ such that for any $\boldsymbol{\gamma} \in S_2$ and with

probability tending to 1,

$$\begin{aligned}
\boldsymbol{y_i}^\top(I_n - P_{i,n})\boldsymbol{y_i} &= ||\boldsymbol{v}_{i,\boldsymbol{\gamma}}||_2^2 + 2\boldsymbol{v}_{i,\boldsymbol{\gamma}}^\top\boldsymbol{\epsilon}_i + \boldsymbol{\epsilon}_i^\top(I_n - P_{i,\boldsymbol{\gamma}})\boldsymbol{\epsilon}_i \\
&\geqslant ||\boldsymbol{v}_{i,\boldsymbol{\gamma}}||_2^2 - 2C\sqrt{r_n \log p_n}||\boldsymbol{v}_{i,\boldsymbol{\gamma}}||_2 + \boldsymbol{\epsilon}_i^\top\boldsymbol{\epsilon}_i - C|\boldsymbol{\gamma}|\log p_n \\
&\geqslant ||\boldsymbol{v}_{i,\boldsymbol{\gamma}}||_2^2\Big(1 - 2C\frac{\sqrt{r_n \log p_n}}{||\boldsymbol{v}_{i,\boldsymbol{\gamma}}||_2} - C\frac{r_n \log p_n}{n\varphi_{\min}(n)\psi_n^2}\Big) + \boldsymbol{\epsilon}_i^\top\boldsymbol{\epsilon}_i \\
&= ||\boldsymbol{v}_{i,\boldsymbol{\gamma}}||_2^2(1 + o(1)) + \boldsymbol{\epsilon}_i^\top\boldsymbol{\epsilon}_i \\
&\geqslant n\varphi_{\min}(n)||\boldsymbol{\beta}^0_{\boldsymbol{\gamma}_0\setminus\boldsymbol{\gamma}}||_2^2(1 + o(1)) + \boldsymbol{\epsilon}_i^\top\boldsymbol{\epsilon}_i \\
&\geqslant n\varphi_{\min}(n)\psi_n^2(1 + o(1)) + \boldsymbol{\epsilon}_i^\top\boldsymbol{\epsilon}_i.
\end{aligned}$$

Assumption 2.2.3 implies that $s_n = o(n)$, and therefore,

$$\boldsymbol{\epsilon}_i^\top(I_n - P_{\boldsymbol{\gamma}_0})\boldsymbol{\epsilon}_i = (n - s_n)\mathrm{var}(\boldsymbol{\epsilon}_i)(1 + o_p(1)).$$

Thus, there exists a C' such that with large probability approaching 1,

$$T_5^{(i)} \geqslant \frac{n + v_i}{2}\log\Big(\frac{1 + n\varphi_{\min}(n)\psi_n^2(1 + o(1)) + \boldsymbol{\epsilon}_i^\top\boldsymbol{\epsilon}_i}{1 + \boldsymbol{\epsilon}_i^\top(I_n - P_{i,\boldsymbol{\gamma}_0})\boldsymbol{\epsilon}_i}\Big) \geqslant \frac{n + v_i}{2}\log(1 + C'\psi_n^2)$$

uniformly for $\boldsymbol{\gamma} \in S_2$. On the other hand, by the properties of projection matrices and part (b) of Lemma 2.2.2, with probability approaching 1, we have

$$\begin{aligned}
&\frac{1 + \boldsymbol{y}_i^\top(I_n - P_{i,\boldsymbol{\gamma}})\boldsymbol{y}_i}{1 + \boldsymbol{y}_i^\top(I_n - P_{i,\boldsymbol{\gamma}_0})\boldsymbol{y}_i} = 1 - \frac{\boldsymbol{y}_i^\top(P_{i,\boldsymbol{\gamma}} - P_{i,\boldsymbol{\gamma}_0})\boldsymbol{y}_i}{1 + \boldsymbol{y}_i^\top(I_n - P_{i,\boldsymbol{\gamma}_0})\boldsymbol{y}_i} \\
= \; & 1 - \frac{\boldsymbol{\beta}^0_{\boldsymbol{\gamma}_0}X^\top_{\boldsymbol{\gamma}_0}(P_{i,\boldsymbol{\gamma}} - P_{i,\boldsymbol{\gamma}_0})X_{\boldsymbol{\gamma}_0}\boldsymbol{\beta}_{\boldsymbol{\gamma}_0} + 2\boldsymbol{\beta}^{0\top}_{\boldsymbol{\gamma}_0}X^\top_{\boldsymbol{\gamma}_0}(P_{i,\boldsymbol{\gamma}} - P_{i,\boldsymbol{\gamma}_0})\boldsymbol{\epsilon}_i + \boldsymbol{\epsilon}_i^\top(P_{i,\boldsymbol{\gamma}} - P_{i,\boldsymbol{\gamma}_0})\boldsymbol{\epsilon}_i}{1 + \boldsymbol{y}_i^\top(I_n - P_{i,\boldsymbol{\gamma}_0})\boldsymbol{y}_i} \\
= \; & 1 - \frac{\boldsymbol{\epsilon}_i^\top(P_{i,\boldsymbol{\gamma}} - P_{i,\boldsymbol{\gamma}_0})\boldsymbol{\epsilon}_i}{1 + \boldsymbol{\epsilon}_i^\top(I_n - P_{i,\boldsymbol{\gamma}_0})\boldsymbol{\epsilon}_i} \geqslant 1 - \frac{\alpha(|\boldsymbol{\gamma}| - s_n)\log p_n}{n},
\end{aligned}$$

uniformly for $\boldsymbol{\gamma} \in S_1$, where α is a constant such that $1 < \alpha$. By the inequality that $\log(1-x) \geqslant -x$ for $x \in (0, 0.5)$, and by Assumption 2.2.3, it follows that uniformly for $\boldsymbol{\gamma} \in S_1$,

$$T_5^{(i)} \geqslant \frac{n + v_i}{2}\log\Big(1 - \frac{\alpha(|\boldsymbol{\gamma}| - s_n)\log p_n}{n}\Big) \geqslant -2^{-1}(|\boldsymbol{\gamma}| - s_n)\alpha_0 \log p_n,$$

for some $\alpha_0 > (n + v_i)\alpha/n$ with probability tending to 1. □

Proof of Theorem 2.2.1. First, we consider $\boldsymbol{\gamma}_0 \neq \mathbf{0}$. By Assumption 2.2.1, we have that

$$T_1 = \log(\frac{p(\boldsymbol{\gamma})}{p(\boldsymbol{\gamma}_0)}) = (|\boldsymbol{\gamma}| - s_n) \log \frac{q}{1-q}.$$

Since for $i \in \{1, \ldots, k\}$, $U_{i,\boldsymbol{\gamma}} = X_{i,\boldsymbol{\gamma}}^\top X_{i,\boldsymbol{\gamma}} + \Sigma_{i,\boldsymbol{\gamma}}^{-1} \geqslant X_{i,\boldsymbol{\gamma}}^\top X_{i,\boldsymbol{\gamma}}$, we have $T_3^{(i)} \geqslant 0$. To approximate $T_4^{(i)}$, let

$$\Delta_i = \boldsymbol{y}_i^\top X_{i,\boldsymbol{\gamma}_0} (X_{i,\boldsymbol{\gamma}_0}^\top X_{i,\boldsymbol{\gamma}_0})^{-1} (\Sigma_{\boldsymbol{\gamma}_0} + (X_{i,\boldsymbol{\gamma}_0}^\top X_{i,\boldsymbol{\gamma}_0})^{-1})^{-1} (X_{i,\boldsymbol{\gamma}_0}^\top X_{i,\boldsymbol{\gamma}_0})^{-1} X_{i,\boldsymbol{\gamma}_0}^\top \boldsymbol{y}_i.$$

By the Sherman-Morrison-Woodbury matrix identity,

$$\begin{aligned} U_{i,\boldsymbol{\gamma}_0}^{-1} =& (X_{i,\boldsymbol{\gamma}_0}^\top X_{i,\boldsymbol{\gamma}_0} + \Sigma_{i,\boldsymbol{\gamma}_0}^{-1})^{-1} = (X_{i,\boldsymbol{\gamma}_0}^\top X_{i,\boldsymbol{\gamma}_0})^{-1} \\ & - (X_{i,\boldsymbol{\gamma}_0}^\top X_{i,\boldsymbol{\gamma}_0})^{-1} (\Sigma_{\boldsymbol{\gamma}_0} + (X_{i,\boldsymbol{\gamma}_0}^\top X_{i,\boldsymbol{\gamma}_0})^{-1})^{-1} (X_{i,\boldsymbol{\gamma}_0}^\top X_{i,\boldsymbol{\gamma}_0})^{-1}. \end{aligned}$$

This implies that

$$U_{i,\boldsymbol{\gamma}_0}^{-1} - (X_{i,\boldsymbol{\gamma}_0}^\top X_{i,\boldsymbol{\gamma}_0})^{-1} = - (X_{i,\boldsymbol{\gamma}_0}^\top X_{i,\boldsymbol{\gamma}_0})^{-1} (\Sigma_{\boldsymbol{\gamma}_0} + (X_{i,\boldsymbol{\gamma}_0}^\top X_{i,\boldsymbol{\gamma}_0})^{-1})^{-1} (X_{i,\boldsymbol{\gamma}_0}^\top X_{i,\boldsymbol{\gamma}_0})^{-1}.$$

Thus,

$$\begin{aligned} & \frac{1 + \boldsymbol{y}_i^\top (I_n - X_{i,\boldsymbol{\gamma}_0} U_{i,\boldsymbol{\gamma}_0}^{-1} X_{i,\boldsymbol{\gamma}_0}^\top) \boldsymbol{y}_i}{1 + \boldsymbol{y}_i^\top (I_n - P_{i,\boldsymbol{\gamma}_0}) \boldsymbol{y}_i} = 1 + \frac{\boldsymbol{y}_i^\top (P_{i,\boldsymbol{\gamma}_0} - X_{i,\boldsymbol{\gamma}_0} U_{i,\boldsymbol{\gamma}_0}^{-1} X_{i,\boldsymbol{\gamma}_0}^\top) \boldsymbol{y}_i}{1 + \boldsymbol{y}_i^\top (I_n - P_{i,\boldsymbol{\gamma}_0}) \boldsymbol{y}_i} \\ =& 1 + \frac{\boldsymbol{y}_i^\top X_{i,\boldsymbol{\gamma}_0} ((X_{i,\boldsymbol{\gamma}_0}^\top X_{i,\boldsymbol{\gamma}_0})^{-1} - U_{i,\boldsymbol{\gamma}_0}^{-1}) X_{i,\boldsymbol{\gamma}_0}^\top \boldsymbol{y}_i}{1 + \boldsymbol{y}_i^\top (I_n - P_{i,\boldsymbol{\gamma}_0}) \boldsymbol{y}_i} = 1 + \frac{\Delta_i}{1 + \boldsymbol{y}_i^\top (I_n - P_{i,\boldsymbol{\gamma}_0}) \boldsymbol{y}_i}. \end{aligned}$$

Using the fact that $(a+b)^2 \leqslant 2(a^2 + b^2)$, and

$$\boldsymbol{y}_i = X_{i,\boldsymbol{\gamma}_0} \boldsymbol{\beta}_{i,\boldsymbol{\gamma}_0} + \boldsymbol{\epsilon}_i, (\Sigma_{i,\boldsymbol{\gamma}_0} + (X_{i,\boldsymbol{\gamma}_0}^\top X_{i,\boldsymbol{\gamma}_0})^{-1})^{-1} \leqslant \Sigma_{i,\boldsymbol{\gamma}_0}^{-1},$$

we have

$$\begin{aligned} & \frac{\Delta_i}{1 + \boldsymbol{y}_i^\top (I_n - P_{i,\boldsymbol{\gamma}_0}) \boldsymbol{y}_i} \\ \leqslant & 2 \frac{\boldsymbol{\beta}_{i,\boldsymbol{\gamma}_0}^{0\top} \Sigma_{i,\boldsymbol{\gamma}_0}^{-1} \boldsymbol{\beta}_{i,\boldsymbol{\gamma}_0}^0 + \boldsymbol{\epsilon}_i^\top X_{i,\boldsymbol{\gamma}_0}^\top (X_{i,\boldsymbol{\gamma}_0}^\top X_{i,\boldsymbol{\gamma}_0})^{-1} \Sigma_{i,\boldsymbol{\gamma}_0}^{-1} (X_{i,\boldsymbol{\gamma}_0}^\top X_{i,\boldsymbol{\gamma}_0})^{-1} X_{i,\boldsymbol{\gamma}_0} \boldsymbol{\epsilon}_i}{1 + \boldsymbol{y}_i^\top (I_n - P_{i,\boldsymbol{\gamma}_0}) \boldsymbol{y}_i}. \end{aligned}$$

By Assumption 2.2.2, we obtain that $\underline{\phi}_n I_n \leqslant \Sigma_{i,\boldsymbol{\gamma}_0}$ and

$$1+\frac{\Delta_i}{1+\boldsymbol{y}_i^\top(I_n-P_{i,\boldsymbol{\gamma}_0})\boldsymbol{y}_i} \leqslant 1+2\underline{\phi}_n^{-1}\frac{\|\boldsymbol{\beta}_{i,\boldsymbol{\gamma}_0}\|_2^2+\boldsymbol{\epsilon}_i^\top X_{i,\boldsymbol{\gamma}_0}(X_{i,\boldsymbol{\gamma}_0}^\top X_{i,\boldsymbol{\gamma}_0})^{-2}X_{i,\boldsymbol{\gamma}_0}^\top\boldsymbol{\epsilon}_i}{1+\boldsymbol{y}_i^\top(I_n-P_{i,\boldsymbol{\gamma}_0})\boldsymbol{y}_i}.$$

For the denominator, we have that $\boldsymbol{y}_i^\top(I_n-P_{i,\boldsymbol{\gamma}_0})\boldsymbol{y}_i/n=\boldsymbol{\epsilon}_i^\top(I_n-P_{i,\boldsymbol{\gamma}_0})\boldsymbol{\epsilon}_i/n\to\sigma_i^2$ in probability. As $\varnothing\in S_2$,

$$\varphi_{\min}(n)=\min_{i=1,,,k}\min_{\boldsymbol{\gamma}\in S_2}\lambda_-(\frac{1}{n}X_{i,\boldsymbol{\gamma}_0\setminus\boldsymbol{\gamma}}(I_n-P_{i,\boldsymbol{\gamma}})X_{i,\boldsymbol{\gamma}_0\setminus\boldsymbol{\gamma}})\leqslant\lambda_-(\frac{1}{n}X_{i,\boldsymbol{\gamma}_0}^\top X_{i,\boldsymbol{\gamma}_0}).$$

The eigenvalues of $(X_{i,\boldsymbol{\gamma}_0}^\top X_{i,\boldsymbol{\gamma}_0})^{-1}$ are the same as these of $X_{i,\boldsymbol{\gamma}_0}(X_{i,\boldsymbol{\gamma}_0}^\top X_{i,\boldsymbol{\gamma}_0})^{-2}X_{i,\boldsymbol{\gamma}_0}^\top$. By eigenvalue decomposition, there exists orthogonal matrix U with dimension $s_n\times s_n$ and eigenvalues $\lambda_1,...,\lambda_{s_n}$ such that $X_{i,\boldsymbol{\gamma}_0}(X_{i,\boldsymbol{\gamma}_0}^\top X_{i,\boldsymbol{\gamma}_0})^{-2}X_{i,\boldsymbol{\gamma}_0}^\top=U^\top\Lambda U$, where Λ is a diagonal matrix whose diagonal elements are zeros and $\lambda_1,...,\lambda_{s_n}$. We have

$$\begin{aligned}&\boldsymbol{\epsilon}_i^\top X_{i,\boldsymbol{\gamma}_0}(X_{i,\boldsymbol{\gamma}_0}^\top X_{i,\boldsymbol{\gamma}_0})^{-2}X_{i,\boldsymbol{\gamma}_0}^\top\boldsymbol{\epsilon}_i=\boldsymbol{\epsilon}_i^\top U^\top\Lambda U\boldsymbol{\epsilon}_i\leqslant\boldsymbol{\epsilon}_i^\top U^\top\Lambda U\boldsymbol{\epsilon}_i\\&\leqslant\boldsymbol{\epsilon}_i^\top B\boldsymbol{\epsilon}_i\lambda_-(X_{i,\boldsymbol{\gamma}_0}^\top X_{i,\boldsymbol{\gamma}_0})^{-1}\leqslant\boldsymbol{\epsilon}_i^\top B\boldsymbol{\epsilon}_i(n\varphi_{\min}(n))^{-1},\end{aligned}$$

where B is a diagonal matrix whose first s_n diagonal elements are 1s and the rest of them are 0s. Furthermore,

$$\mathbb{E}(\boldsymbol{\epsilon}_i^\top X_{i,\boldsymbol{\gamma}_0}(X_{i,\boldsymbol{\gamma}_0}^\top X_{i,\boldsymbol{\gamma}_0})^{-2}X_{i,\boldsymbol{\gamma}_0}^\top\boldsymbol{\epsilon}_i)\leqslant\mathrm{var}(\boldsymbol{\epsilon}_i)s_n(n\varphi_{\min}(n))^{-1}\to 0.$$

Next, we prove that $\boldsymbol{\epsilon}_i^\top X_{i,\boldsymbol{\gamma}_0}(X_{i,\boldsymbol{\gamma}_0}^\top X_{i,\boldsymbol{\gamma}_0})^{-2}X_{i,\boldsymbol{\gamma}_0}^\top\boldsymbol{\epsilon}_i=O_p(1)$. The errors ϵ_{ij} are from sub-Gaussian distribution with $\max_{j=1,...,n}||\epsilon_{ij}||_{\psi_2}\leqslant K$ for some $K>0$, where $||X||_{\psi_2}=\sup_{p\geqslant 1}p^{-1/2}(\mathbb{E}X^p)^{1/p}$. By Hanson-Wright inequality, there exists a constant $C>0$ such that for any $n\times n$ matrix A and for all $t>0$,

$$P(\boldsymbol{\epsilon}_i^\top A\boldsymbol{\epsilon}_i-\mathbb{E}(\boldsymbol{\epsilon}_i^\top A\boldsymbol{\epsilon}_i)>t)\leqslant\exp\left\{-C\min\left(\frac{t^2}{K^4||A||_F^2},\frac{t}{K^2||A||_2}\right)\right\}.$$

When t is sufficiently large, we get

$$P(\boldsymbol{\epsilon}_i^\top A\boldsymbol{\epsilon}_i-\mathbb{E}(\boldsymbol{\epsilon}_i^\top A\boldsymbol{\epsilon}_i)>t)\leqslant\exp\left(\frac{-Ct}{K^2||A||_2}\right).$$

Let $A = X_{i,\boldsymbol{\gamma}_0}(X_{i,\boldsymbol{\gamma}_0}^\top X_{i,\boldsymbol{\gamma}_0})^{-2}X_{i,\boldsymbol{\gamma}_0}^\top$. Then $||A||_2 = (n\psi_{\min}(n))^{-1}$ and

$$||A||_F = \sqrt{\mathrm{tr}(X_{i,\boldsymbol{\gamma}_0}(X_{i,\boldsymbol{\gamma}_0}^\top X_{i,\boldsymbol{\gamma}_0})^{-3}X_{i,\boldsymbol{\gamma}_0}^\top)} = \sqrt{\mathrm{tr}((X_{i,\boldsymbol{\gamma}_0}^\top X_{i,\boldsymbol{\gamma}_0})^{-2})} \leqslant \sqrt{(n\psi_{\min}(n))^{-2}s_n}.$$

Thus,

$$P(\boldsymbol{\epsilon}_i^\top A\boldsymbol{\epsilon}_i - \mathbb{E}(\boldsymbol{\epsilon}_i^\top A\boldsymbol{\epsilon}_i) > t) \leqslant \exp(-n\psi_{\min}(n)Ct/K^2) \to 0,$$

as $n \to \infty$. So, for every $\delta > 0$, there exist t, and $n_\delta > 0$ such that $\mathbb{E}(\boldsymbol{\epsilon}_i^\top A\boldsymbol{\epsilon}_i) \leqslant 1$ and $\exp(-n\psi_{\min}(n)Ct)) \leqslant \delta$, for all $n > n_\delta$ and

$$P(\boldsymbol{\epsilon}_i^\top A\boldsymbol{\epsilon}_i > t+1) \leqslant P(\boldsymbol{\epsilon}_i^\top A\boldsymbol{\epsilon}_i - \mathbb{E}(\boldsymbol{\epsilon}_i^\top A\boldsymbol{\epsilon}_i) > t) \leqslant \exp(-n\psi_{\min}(n)Ct/K^2)) \leqslant \delta.$$

This implies $\boldsymbol{\epsilon}_i^\top X_{i,\boldsymbol{\gamma}_0}(X_{i,\boldsymbol{\gamma}_0}^\top X_{i,\boldsymbol{\gamma}_0})^{-2}X_{i,\boldsymbol{\gamma}_0}^\top\boldsymbol{\epsilon}_i = O_p(1)$. Continue from Equation (2.3), combining all the results above, there exists a constant $\widetilde{C} > 0$

$$\begin{aligned}
\frac{1+\boldsymbol{y}_i^\top(I_n - X_{i,\boldsymbol{\gamma}_0}U_{i,\boldsymbol{\gamma}_0}^{-1}X_{i,\boldsymbol{\gamma}_0}^\top)\boldsymbol{y}_i}{1+\boldsymbol{y}_i^\top(I_n - P_{i,\boldsymbol{\gamma}_0})\boldsymbol{y}_i} &\leqslant 1 + 2\underline{\phi}_n^{-1}\frac{k_n^2 s_n + O_p(1)}{1+n\mathrm{var}(\boldsymbol{\epsilon}_i)} \\
&\leqslant 1 + \frac{2k_n^2 s_n}{n\underline{\phi}_n\mathrm{var}(\boldsymbol{\epsilon}_i)}(1+O_p(\frac{1}{s_n k_n^2})) \\
&\leqslant 1 + \frac{2k_n^2 s_n}{n\underline{\phi}_n\mathrm{var}(\boldsymbol{\epsilon}_i)}(1+o_p(1)) \leqslant 1 + \frac{2\widetilde{C}}{n}(1+o_p(1)).
\end{aligned}$$

By Assumption 2.2.3. $-T_4^{(i)} \geqslant -\widetilde{C}, i \in \{1,\ldots,k\}$. By Lemma 2.2.1, 2.2.2, Assumption 2.2.3, 2.2.4 and the fact that $p_n^{\alpha_0 - \frac{2M}{k}} = o(\rho_n)$ with $\rho_n = n^{1-\delta}\underline{\phi}_n$,

$$\begin{aligned}
& P_C(\boldsymbol{\gamma}|Z_1,...,Z_k)/P_C(\boldsymbol{\gamma}_0|Z_1,...,Z_k) \\
\leqslant\ & \widetilde{C}\Big(\exp(-2^{-1}(|\boldsymbol{\gamma}| - s_n)\log((1+C_3\rho_n)/p_n^{\alpha_0}))\Big)^k\Big(\frac{q}{1-q}\Big)^{(|\boldsymbol{\gamma}|-s_n)} \\
\leqslant\ & \widetilde{C}\Big(\frac{(2p_n)^{\alpha_0-\frac{2M}{k}}}{1+C_3\rho_n}\Big)^{k(|\gamma|-s_n)/2} \xrightarrow{p} 0,
\end{aligned}$$

with probability tending to 1. By Assumptions 2.2.2 and 2.2.3, it can be verified that $s_n\log(1+$

$C_2 n\overline{\phi}_n) \ll \frac{n+v_i}{2}\log(1+C'\psi_n^2)$. So,

$$\begin{aligned}
&P_C(\boldsymbol{\gamma}|Z_1,...,Z_k)/P_C(\boldsymbol{\gamma}_0|Z_1,...,Z_k)\\
&\leqslant \widetilde{C}\exp[k(2^{-1}s_n\log(1+C_2n\overline{\phi}_n)-\frac{n+v_i}{2}\log(1+C'\psi_n^2))]\\
&\leqslant \widetilde{C}(1+C'\psi_n^2)^{-k\frac{n+v_i}{4}} \xrightarrow{p} 0.
\end{aligned}$$

Next we establish the strong model selection consistency results under two scenarios. First for overfitting models,

$$\begin{aligned}
&\sum_{\boldsymbol{\gamma}\in S_1} P_C(\boldsymbol{\gamma}|Z_1,...,Z_k)/P_C(\boldsymbol{\gamma}_0|Z_1,...,Z_k) \leqslant \widetilde{C}^k \sum_{\boldsymbol{\gamma}\in S_1}\Big(\frac{1+C_3\rho_n}{p_n^{\alpha_0-2M/k}}\Big)^{-2^{-1}k(|\boldsymbol{\gamma}|-s_n)}\\
=\ & \widetilde{C}^k\sum_{r=1}^{r_n-s_n}\binom{p_n-s_n}{r}\Big(\frac{1+C_3\rho_n}{p_n^{\alpha_0-2M/k}}\Big)^{-rk/2} \leqslant \widetilde{C}^k\sum_{r=1}^{r_n-s_n}\frac{p_n^r}{r!}\Big(\frac{p_n^{\alpha_0-2M/k}}{1+C_3\rho_n}\Big)^{rk/2}\\
=\ & \widetilde{C}^k\sum_{r=1}^{r_n-s_n}\frac{1}{r!}\Big(\frac{p_n^{\alpha_0-(2M-2)/k}}{1+C_3\rho_n}\Big)^{rk/2} \leqslant \widetilde{C}^k\sum_{r=1}^{r_n-s_n}\Big(\frac{p_n^{\alpha_0-(2M-2)/k}}{1+C_3\rho_n}\Big)^{kr/2}\\
\leqslant\ & \widetilde{C}^k\frac{\Big(\frac{p_n^{\alpha_0-(2M-2)/k}}{1+C_3\rho_n}\Big)^{k/2}}{1-\Big(\frac{p_n^{\alpha_0-(2M-2)/k}}{1+C_3\rho_n}\Big)^{k/2}} \xrightarrow{p} 0,
\end{aligned}$$

where the last limit result follows from the assumption that $p_n^{\alpha_0-2(M-1-L)/k} = o(\rho_n)$. For wrong models, by $r_n\log p = o(n\log(1+\psi_n^2))$, we have

$$\sum_{\boldsymbol{\gamma}\in S_2} P_C(\boldsymbol{\gamma}|Z_1,...,Z_k)/P_C(\boldsymbol{\gamma}_0|Z_1,...,Z_k) \leqslant \widetilde{C}^k\{p_n^{kr_n}(1+C'^k\psi_n^2)^{-(kn+\sum v_i)/4}\} \xrightarrow{p} 0.$$

Second, we consider $\boldsymbol{\gamma}_0 = \mathbf{0}$. In this situation, S_2 is empty, we only need to consider S_1. In this case $T_4 = 0$, T_1 and T_3 are bounded as before. Similar to the argument above, $T_2 \geqslant 2^{-1}|\boldsymbol{\gamma}|\log(1+C_3n^{1-\delta}\underline{\phi}_n)$, and $T_5 \geqslant -2^{-1}|\boldsymbol{\gamma}|\alpha_0\log p_n$. Therefore, the proof is completed. □

The composite posterior probability of the true model can be formulated as

$$P_C(\boldsymbol{\gamma}_0|Z_1,Z_2,...,Z_k) = \frac{1}{1+\sum_{\boldsymbol{\gamma}\neq\boldsymbol{\gamma}_0}P_C(\boldsymbol{\gamma}|Z_1,Z_2,...,Z_k)/P_C(\boldsymbol{\gamma}_0|Z_1,Z_2,...,Z_k)}.$$

Following Theorem 2.2.1, we can get

$$\inf_{c_{i,1},...,c_{i,p_n}\in[\underline{\phi}_n,\overline{\phi}_n],i\in\{1,...,k\}} P_C(\boldsymbol{\gamma}_0|Z_1,Z_2,...,Z_k) \xrightarrow{p} 1.$$

This implies that the composite posterior probability of the true model converges to 1, thereby establishing the strong consistency of the model selection procedure. Achieving this result requires demonstrating that the sum of the Bayes factors tends to zero, as shown by Shang and Clayton (2011). This condition is stronger than the weaker consistency, which only requires each individual Bayes factor to approach zero.

2.2.2 Model selection with sub-exponential random errors

The sub-Gaussian error distributions discussed in the previous subsection exhibit tail behavior similar to that of Gaussian distributions, with tail probabilities decaying exponentially fast. In contrast, sub-exponential distributions have heavier tails, resulting in slower decay of their tail probabilities compared to Gaussian distributions.

In this subsection, we extend our model selection results to accommodate sub-exponential random errors. Although exponential deviation inequalities for weighted sums of independent sub-exponential random variables are available, the control over large deviations is governed by weaker bounds. To address the challenges posed by the heavier tails of sub-exponential distributions, we introduce modified assumptions tailored to this setting.

Assumption 2.2.5 *It is assumed that* $s_n \leqslant r_n \leqslant n$, *and*

$$\max\{r_n^2 \log^2 p_n, r_n \log n\} = o(n \log(1 + \min\{1, \psi_n^2\})).$$

Remark 2.2.5 *Similar to Assumption 2.2.2, this assumption limits the considered model size to be smaller than the data size. Compared to Assumption 2.2.3 for sub-Gaussian errors, Assumption*

2.2.5 for sub-exponential errors is more stringent, requiring a lower growth rate of r_n, with respect to the sample size n. When $\psi_n > 0$, Assumption 2.2.5 can be simplified to $\max\{r_n^2 \log^2 p_n, r_n \log n\} = o(n)$. This indicates that, for the same r_n, the sub-exponential distributions require a larger sample size compared to sub-Gaussian distributions to achieve similar guarantees.

Theorem 2.2.2 *Under Assumptions 2.2.1, 2.2.2, 2.2.4, 2.2.5, if for some $\alpha_0 > 7$, $p_n^{\alpha_0 - 2M/k} = o(n^{1-\delta}\underline{\phi}_n)$, then*

$$\sup_{c_{i,1},...,c_{i,p_n} \in [\underline{\phi}_n, \overline{\phi}_n], i \in \{1,...,k\}} \max_{\boldsymbol{\gamma} \neq \boldsymbol{\gamma}_0} P_C(\boldsymbol{\gamma}|Z_1, Z_2, ..., Z_k)/P_C(\boldsymbol{\gamma}_0|Z_1, Z_2, ..., Z_k) \xrightarrow{p} 0.$$

If $p_n^{\alpha_0 - \frac{2(M-1)}{k}} = o(n^{1-\delta}\underline{\phi}_n)$, then

$$\sup_{c_{i,1},...,c_{i,p_n} \in [\underline{\phi}_n, \overline{\phi}_n], i \in \{1,...,k\}} \sum_{\boldsymbol{\gamma} \neq \boldsymbol{\gamma}_0} P_C(\boldsymbol{\gamma}|Z_1, Z_2, ..., Z_k) \xrightarrow{p} 0.$$

Lemma 2.2.4 *Suppose that $\epsilon_i \sim SubE(\sigma^2, b)$, that is for $\forall t : |t| < 1/b, \mathbb{E}e^{t\epsilon_i} \leqslant e^{t^2\sigma^2/2}$, for $i \in \{1, \ldots, n\}$, X is $n \times p_n$ dimensional matrix satisfies the assumptions of design matrix, $X_{\boldsymbol{\gamma}}$ is constructed by the columns of X, which are indexed by 1 in $\boldsymbol{\gamma}$.*

(a). Let $\boldsymbol{v}_{\boldsymbol{\gamma}} = (I_n - P_{\boldsymbol{\gamma}})X_{\boldsymbol{\gamma}_0 \setminus \boldsymbol{\gamma}}\boldsymbol{\beta}^0_{\boldsymbol{\gamma}_0 \setminus \boldsymbol{\gamma}}$. If S_2 is not empty, then

$$\max_{\boldsymbol{\gamma} \in S_2} |\boldsymbol{v}_{\boldsymbol{\gamma}}^\top \boldsymbol{\epsilon}|/||\boldsymbol{v}_{\boldsymbol{\gamma}}||_2 \leqslant 4r_n \log p_n/\lambda$$

with probability tending to 1, for some $0 < \lambda \leqslant 1/b$, where we adopt the convention that $|\boldsymbol{v}_{\boldsymbol{\gamma}}^\top \boldsymbol{\epsilon}|/||\boldsymbol{v}_{\boldsymbol{\gamma}}||_2 = 0$ when $\boldsymbol{v}_{\boldsymbol{\gamma}} = 0$.

(b). If S_1 is not empty, with probability approaching 1,

$$\max_{\boldsymbol{\gamma} \in S_1} \boldsymbol{\epsilon}^\top (P_{\boldsymbol{\gamma}} - P_{\boldsymbol{\gamma}_0})\boldsymbol{\epsilon} \leqslant 7\alpha\sigma^2(|\gamma| - s_n)\log p_n),$$

for some $\alpha > 1$.

(c). If S_2 is not empty,

$$\max_{\boldsymbol{\gamma} \in S_2} \boldsymbol{\epsilon}^\top P_{\boldsymbol{\gamma}}\boldsymbol{\epsilon} \leqslant 7\alpha\sigma^2|\boldsymbol{\gamma}| \log p_n$$

with probability approaching 1, for some $\alpha > 1$. We adopt the convention that $\boldsymbol{\epsilon}^\top P_{\boldsymbol{\gamma}}\boldsymbol{\epsilon}/|\boldsymbol{\gamma}| = 0$, when $\boldsymbol{\gamma}$ is empty.

Proof of Lemma 2.2.4. (a) Denote $\boldsymbol{v}_{\boldsymbol{\gamma}}/||\boldsymbol{v}_{\boldsymbol{\gamma}}|| = \boldsymbol{V} = (V_1, ..., V_n)^\top, \boldsymbol{\epsilon} = (\epsilon_1, ..., \epsilon_n)^\top$. Then

$$\mathbb{E}(e^{\lambda \boldsymbol{V}^\top \boldsymbol{\epsilon}}) = \mathbb{E}(e^{\lambda \sum_{i=1}^n V_i \epsilon_i}) = \prod_{i=1}^n \mathbb{E}(e^{\lambda V_i \epsilon_i}).$$

Using the definition, we obtain the following inequality. When $|\lambda V_i| \leqslant |\lambda| \leqslant 1/b$,

$$\mathbb{E}(e^{\lambda V_i \epsilon_i}) \leqslant e^{\frac{\lambda^2 V_i^2 \sigma^2}{2}}.$$

Thus,

$$\mathbb{E}(e^{\lambda \boldsymbol{V}^\top \boldsymbol{\epsilon}}) = \prod_{i=1}^n \mathbb{E}(e^{\lambda V_i \epsilon_i}) \leqslant e^{\frac{\lambda^2 \sigma^2 \sum_{i=1}^n V_i^2}{2}} = e^{\frac{\lambda^2 \sigma^2}{2}}.$$

Then by Markov's inequality,

$$\Pr(\boldsymbol{V}^\top \boldsymbol{\epsilon} \geqslant t) = \Pr(e^{\lambda \boldsymbol{V}^\top \boldsymbol{\epsilon}} \geqslant e^{\lambda t}) \leqslant \frac{\mathbb{E}(e^{\lambda \boldsymbol{V}^\top \boldsymbol{\epsilon}})}{e^{\lambda t}} \leqslant e^{\frac{\lambda^2 \sigma^2}{2} - \lambda t} \leqslant e^{-\frac{\lambda t}{2}},$$

when $|\lambda| \leqslant 1/b$, $t \geqslant \sigma^2/b$.

Denote the size of S_2 by $|S_2|$. Then we have

$$|S_2| \leqslant \binom{p_n}{1} + \binom{p_n}{2} + ... + \binom{p_n}{r_n} \leqslant \sum_{\ell \leqslant r_n} p_n^\ell / \ell! \leqslant p_n^{r_n}.$$

Choose a λ such that $|\lambda| \leqslant 1/b$. With sufficient large r_n such that $2r_n \geqslant \sigma^2 \lambda / b$, we have

$$\begin{aligned}
&\Pr(\max_{\boldsymbol{\gamma} \in S_2} \frac{|\boldsymbol{v}_{\boldsymbol{\gamma}}^\top \boldsymbol{\epsilon}|}{||\boldsymbol{v}_{\boldsymbol{\gamma}}||} \geqslant 4 r_n \log p_n / \lambda) \leqslant \sum_{\boldsymbol{\gamma} \in S_2} \Pr(\frac{|\boldsymbol{v}_{\boldsymbol{\gamma}}^\top \boldsymbol{\epsilon}|}{||\boldsymbol{v}_{\boldsymbol{\gamma}}||} \geqslant 4 r_n \log p_n / \lambda) \\
&\leqslant p_n^{r_n} \exp(-4\lambda r_n \log p_n / (2\lambda)) = p_n^{-r_n} \to 0,
\end{aligned}$$

as $n \to \infty$.

(b). When $\boldsymbol{\gamma} \in S_1$, $P_{\boldsymbol{\gamma}} = X_{\boldsymbol{\gamma}} \left(X_{\boldsymbol{\gamma}}^\top X_{\boldsymbol{\gamma}} \right)^{-1} X_{\boldsymbol{\gamma}}^\top$ is a projection matrix with dimension $|\boldsymbol{\gamma}|$. The true model $\boldsymbol{\gamma}_0$ is nested in model $\boldsymbol{\gamma}$, so

$$P_{\boldsymbol{\gamma}} - P_{\boldsymbol{\gamma}_0} = X_{\boldsymbol{\gamma}} \left(X_{\boldsymbol{\gamma}}^\top X_{\boldsymbol{\gamma}} \right)^{-1} X_{\boldsymbol{\gamma}}^\top - X_{\boldsymbol{\gamma}_0} \left(X_{\boldsymbol{\gamma}_0}^\top X_{\boldsymbol{\gamma}_0} \right)^{-1} X_{\boldsymbol{\gamma}_0}^\top$$

and

$$P_{\boldsymbol{\gamma}\backslash\boldsymbol{\gamma}_0} = X_{\boldsymbol{\gamma}\backslash\boldsymbol{\gamma}_0} \left(X_{\boldsymbol{\gamma}\backslash\boldsymbol{\gamma}_0}^\top X_{\boldsymbol{\gamma}\backslash\boldsymbol{\gamma}_0} \right)^{-1} X_{\boldsymbol{\gamma}\backslash\boldsymbol{\gamma}_0}^\top$$

are both projection matrices with dimension $|\boldsymbol{\gamma}| - |\boldsymbol{\gamma}_0|$. By eigenvalue decomposition,

$$X_{\boldsymbol{\gamma}\backslash\boldsymbol{\gamma}_0} \left(X_{\boldsymbol{\gamma}\backslash\boldsymbol{\gamma}_0}^\top X_{\boldsymbol{\gamma}\backslash\boldsymbol{\gamma}_0} \right)^{-1} X_{\boldsymbol{\gamma}\backslash\boldsymbol{\gamma}_0}^\top = R \begin{pmatrix} I_{\boldsymbol{\gamma}\backslash\boldsymbol{\gamma}_0} & \\ & O \end{pmatrix} R^\top,$$

$$X_{\boldsymbol{\gamma}} \left(X_{\boldsymbol{\gamma}}^\top X_{\boldsymbol{\gamma}} \right)^{-1} X_{\boldsymbol{\gamma}}^\top - X_{\boldsymbol{\gamma}_0} \left(X_{\boldsymbol{\gamma}_0}^\top X_{\boldsymbol{\gamma}_0} \right)^{-1} X_{\boldsymbol{\gamma}_0}^\top = Q \begin{pmatrix} I_{\boldsymbol{\gamma}\backslash\boldsymbol{\gamma}_0} & \\ & O \end{pmatrix} Q^\top,$$

where Q and R are orthogonal matrix containing all eigenvectors of $P_{\boldsymbol{\gamma}} - P_{\boldsymbol{\gamma}_0}$ and $P_{\boldsymbol{\gamma}\backslash\boldsymbol{\gamma}_0}$ respectively. So, we have

$$\begin{aligned} X_{\boldsymbol{\gamma}} \left(X_{\boldsymbol{\gamma}}^\top X_{\boldsymbol{\gamma}} \right)^{-1} X_{\boldsymbol{\gamma}}^\top - X_{\boldsymbol{\gamma}_0} \left(X_{\boldsymbol{\gamma}_0}^\top X_{\boldsymbol{\gamma}_0} \right)^{-1} X_{\boldsymbol{\gamma}_0}^\top &= Q R^\top X_{\boldsymbol{\gamma}\backslash\boldsymbol{\gamma}_0} \left(X_{\boldsymbol{\gamma}\backslash\boldsymbol{\gamma}_0}^\top X_{\boldsymbol{\gamma}\backslash\boldsymbol{\gamma}_0} \right)^{-1} X_{\boldsymbol{\gamma}\backslash\boldsymbol{\gamma}_0}^\top R Q^\top \\ &= \widetilde{V} A (A^\top A)^{-1} A^\top \widetilde{V}^\top, \end{aligned}$$

where $\widetilde{V} = QR^\top, A = X_{\boldsymbol{\gamma}\backslash\boldsymbol{\gamma}_0}$. Next,

$$\begin{aligned} \boldsymbol{\epsilon}^\top \left(P_{\boldsymbol{\gamma}} - P_{\boldsymbol{\gamma}_0} \right) \boldsymbol{\epsilon} &= \boldsymbol{\epsilon}^\top \widetilde{V} A (A^\top A)^{-1} A^\top \widetilde{V}^\top \boldsymbol{\epsilon} \\ &= \frac{\boldsymbol{\epsilon}^\top \widetilde{V} A}{\sqrt{n}} (\frac{A^\top A}{n})^{-1} \frac{A^\top \widetilde{V}^\top}{\sqrt{n}} \boldsymbol{\epsilon} \\ &= \frac{\boldsymbol{\epsilon}^\top \widetilde{V} A}{\sqrt{n}} (\frac{A^\top A}{n})^{-\frac{1}{2}} (\frac{A^\top A}{n})^{-\frac{1}{2}} \frac{A^\top \widetilde{V}^\top}{\sqrt{n}} \boldsymbol{\epsilon} \\ &= \boldsymbol{U}^\top \boldsymbol{U}, \end{aligned} \tag{2.3}$$

where $\boldsymbol{U} = (A^\top A/n)^{-1/2} A^\top \widetilde{V}^\top \boldsymbol{\epsilon}/\sqrt{n} = H^{-1/2} A^\top \widetilde{V}^\top \boldsymbol{\epsilon}/\sqrt{n}$, and $H = A^\top A/n$. Denote $\widetilde{V} = (\widetilde{\boldsymbol{V}}_1, \widetilde{\boldsymbol{V}}_2, ...\widetilde{\boldsymbol{V}}_n)^\top$, where $\widetilde{\boldsymbol{V}}_i$'s are column vectors. Let $\boldsymbol{\epsilon} = (\epsilon_1, ..., \epsilon_n)^\top$. Then

$$\begin{aligned} \boldsymbol{U} &= H^{-1/2} A^\top \widetilde{V}^\top \boldsymbol{\epsilon}/\sqrt{n} = 1/\sqrt{n} H^{-1/2} A^\top (\widetilde{\boldsymbol{V}}_1, \widetilde{\boldsymbol{V}}_2, ...\widetilde{\boldsymbol{V}}_n)(\epsilon_1, ..., \epsilon_n)^\top \\ &= 1/\sqrt{n} H^{-1/2} A^\top (\widetilde{\boldsymbol{V}}_1 \epsilon_1 + \cdots + \widetilde{\boldsymbol{V}}_n \epsilon_n). \end{aligned}$$

Thus the moment generating function of $\boldsymbol{U}$ is

$$\mathbb{E}(e^{\boldsymbol{t}^\top \boldsymbol{U}}) = \mathbb{E}(e^{\boldsymbol{t}^\top \frac{1}{\sqrt{n}} H^{-\frac{1}{2}} A^\top \sum_{i=1}^n \widetilde{\boldsymbol{V}}_i \epsilon_i}) = \prod_{i=1}^n \mathbb{E}(e^{\frac{\boldsymbol{t}^\top}{\sqrt{n}} H^{-\frac{1}{2}} A^\top \widetilde{\boldsymbol{V}}_i \epsilon_i}) = \prod_{i=1}^n \mathbb{E}(e^{\frac{s_i}{\sqrt{n}} \epsilon_i})$$

by setting $s_i = \boldsymbol{t}^\top H^{-1/2} A^\top \widetilde{\boldsymbol{V}}_i$. Using the definition of sub-exponential distribution, when $|s_i| \leqslant \sqrt{n}/b$, $\mathbb{E}(e^{\frac{s_i}{\sqrt{n}} \epsilon_i}) \leqslant \exp^{\frac{s_i^2 \sigma^2}{2n}}$, then

$$\mathbb{E}(e^{\boldsymbol{t}^\top \boldsymbol{U}}) \leqslant \exp\left(\sum_{i=1}^n \frac{\sigma^2 s_i^2}{2n} \right).$$

By the fact that

$$\begin{aligned} \sum_{i=1}^n s_i^2 &= \sum_{i=1}^n \boldsymbol{t}^\top H^{-\frac{1}{2}} A^\top \widetilde{\boldsymbol{V}}_i \widetilde{\boldsymbol{V}}_i^\top A H^{-\frac{1}{2}} \boldsymbol{t} = \boldsymbol{t}^\top H^{-\frac{1}{2}} A^\top \left(\sum_{i=1}^n \widetilde{\boldsymbol{V}}_i \widetilde{\boldsymbol{V}}_i^\top \right) A H^{-\frac{1}{2}} \boldsymbol{t} \\ &= \boldsymbol{t}^\top H^{-\frac{1}{2}} A^\top \widetilde{V}^\top \widetilde{V} A H^{-\frac{1}{2}} \boldsymbol{t} = \boldsymbol{t}^\top H^{-\frac{1}{2}} A^\top A H^{-\frac{1}{2}} \boldsymbol{t} = n \boldsymbol{t}^\top \boldsymbol{t}, \end{aligned}$$

so $\mathbb{E}(e^{\boldsymbol{t}^\top \boldsymbol{U}}) \leqslant \exp\left(\sigma^2 \boldsymbol{t}^\top \boldsymbol{t}/2\right)$. For s_i, we have

$$\begin{aligned} |s_i| &= |\boldsymbol{t}^\top H^{-\frac{1}{2}} A^\top \widetilde{\boldsymbol{V}}_i| \\ &\leqslant ||\boldsymbol{t}|| \sqrt{\widetilde{\boldsymbol{V}}_i^\top A H^{-1} A^\top \widetilde{\boldsymbol{V}}_i}. \end{aligned}$$

Denote $Q = (\boldsymbol{Q}_1, ..., \boldsymbol{Q}_n)^\top$. Then

$$\widetilde{V} = QR^\top = (RQ^\top)^\top = (R(\boldsymbol{Q}_1, ..., \boldsymbol{Q}_n))^\top = (R\boldsymbol{Q}_1, ..., R\boldsymbol{Q}_n)^\top.$$

By the properties of semi-positive definite matrix and Assumption 2.2.1, we have

$$\begin{aligned}
&\widetilde{\boldsymbol{V}}_i^\top AH^{-1}A^\top \widetilde{\boldsymbol{V}}_i = \boldsymbol{Q}_i^\top R^\top AH^{-1}A^\top R\boldsymbol{Q}_i = n\boldsymbol{Q}_i^\top R^\top A(A^\top A)^{-1}A^\top R\boldsymbol{Q}_i \\
&= n\boldsymbol{Q}_i^\top \begin{pmatrix} I_{\boldsymbol{\gamma}\setminus\boldsymbol{\gamma}_0} & \\ & O \end{pmatrix} \boldsymbol{Q}_i = n\left(X_{\boldsymbol{\gamma}} \left(X_{\boldsymbol{\gamma}}^\top X_{\boldsymbol{\gamma}} \right)^{-1} X_{\boldsymbol{\gamma}}^\top - X_{\boldsymbol{\gamma}_0} \left(X_{\boldsymbol{\gamma}_0}^\top X_{\boldsymbol{\gamma}_0} \right)^{-1} X_{\boldsymbol{\gamma}_0}^\top \right)_{ii} \\
&= \left(X_{\boldsymbol{\gamma}} \left(X_{\boldsymbol{\gamma}}^\top X_{\boldsymbol{\gamma}}/n \right)^{-1} X_{\boldsymbol{\gamma}}^\top \right)_{ii} - \left(X_{\boldsymbol{\gamma}_0} \left(X_{\boldsymbol{\gamma}_0}^\top X_{\boldsymbol{\gamma}_0}/n \right)^{-1} X_{\boldsymbol{\gamma}_0}^\top \right)_{ii} \\
&\leqslant \left(X_{\boldsymbol{\gamma}} \left(X_{\boldsymbol{\gamma}}^\top X_{\boldsymbol{\gamma}}/n \right)^{-1} X_{\boldsymbol{\gamma}}^\top \right)_{ii} \leqslant C_5^{-1}(X_{\boldsymbol{\gamma}} X_{\boldsymbol{\gamma}}^\top)_{ii} \leqslant C_4^2 C_5^{-1}|\boldsymbol{\gamma}| \leqslant C_4^2 C_5^{-1} r_n.
\end{aligned}$$

Therefore, $|s_i| \leqslant \sqrt{C_4^2 C_5^{-1} r_n}||\boldsymbol{t}||$. When $||\boldsymbol{t}|| \leqslant \sqrt{\frac{n}{C_4^2 C_5^{-1} b^2 r_n}}$, we have $s_i \leqslant \sqrt{n}/b, \mathbb{E}(e^{\boldsymbol{t}^\top \boldsymbol{U}/\sigma}) \leqslant e^{\boldsymbol{t}^\top \boldsymbol{t}/2}$, which means that $\boldsymbol{U}/\sigma$ satisfies the exponential moment condition. Using Corollary 4.2 in Spokoiny (2013), we have

$$\begin{aligned}
&\Pr\big(\boldsymbol{\epsilon}^\top (P_{\boldsymbol{\gamma}} - P_{\boldsymbol{\gamma_0}})\boldsymbol{\epsilon}/\sigma^2 \geqslant (|\boldsymbol{\gamma}| - s_n + 6x)/\sigma^2\big) \\
&= \Pr\big(\boldsymbol{\epsilon}^\top (P_{\boldsymbol{\gamma}} - P_{\boldsymbol{\gamma_0}})\boldsymbol{\epsilon}/\sigma^2 \geqslant (|\boldsymbol{\gamma}| - s_n + 6\alpha\sigma^2(|\gamma| - s_n)\log p_n)/\sigma^2\big) \\
&\leqslant 10.4\exp(-(\alpha(|\gamma| - s_n)\log p_n)) \\
&\leqslant 10.4\exp(-\alpha(|\gamma| - s_n)\log p_n).
\end{aligned}$$

Thus,

$$\begin{aligned}
&\Pr(\max_{\boldsymbol{\gamma}\in S_1} \boldsymbol{\epsilon}^\top (P_{\boldsymbol{\gamma}} - P_{\boldsymbol{\gamma_0}})\boldsymbol{\epsilon} \geqslant (7\alpha\sigma^2(|\gamma| - s_n)\log p_n)) \\
&\leqslant \Pr(\max_{\boldsymbol{\gamma}\in S_1} \boldsymbol{\epsilon}^\top (P_{\boldsymbol{\gamma}} - P_{\boldsymbol{\gamma_0}})\boldsymbol{\epsilon} \geqslant (|\boldsymbol{\gamma}| - s_n + 6\alpha\sigma^2(|\gamma| - s_n)\log p_n))
\end{aligned}$$

$$\leqslant \sum_{\boldsymbol{\gamma}\in S_1} \Pr\big(\boldsymbol{\epsilon}^\top(P_{\boldsymbol{\gamma}} - P_{\boldsymbol{\gamma_0}})\boldsymbol{\epsilon} \geqslant (|\boldsymbol{\gamma}| - s_n + 6\alpha\sigma^2(|\gamma| - s_n)\log p_n)\big)$$

$$= \sum_{\boldsymbol{\gamma}\in S_1} 10.4 p_n^{-\alpha(|\gamma|-s_n)} \leqslant \sum_{r=1}^{r_n - s_n} \begin{pmatrix} p_n - s_n \\ r \end{pmatrix} p_n^{-\alpha/r}$$

$$\leqslant (1 + p_n^{-\alpha})^{p_n - s_n} - 1 \to 0.$$

(c). We reformulate the following quadratic form

$$\begin{aligned}
\boldsymbol{\epsilon}^\top X_{\boldsymbol{\gamma}}(X_{\boldsymbol{\gamma}}^\top X_{\boldsymbol{\gamma}})^{-1} X_{\boldsymbol{\gamma}}^\top \boldsymbol{\epsilon} &= \frac{\boldsymbol{\epsilon}^\top X_{\boldsymbol{\gamma}}}{\sqrt{n}} (\frac{X_{\boldsymbol{\gamma}}^\top X_{\boldsymbol{\gamma}}}{n})^{-1} \frac{X_{\boldsymbol{\gamma}}^\top \boldsymbol{\epsilon}}{\sqrt{n}} \\
&= \frac{\boldsymbol{\epsilon}^\top X_{\boldsymbol{\gamma}}}{\sqrt{n}} (\frac{X_{\boldsymbol{\gamma}}^\top X_{\boldsymbol{\gamma}}}{n})^{-\frac{1}{2}} (\frac{X_{\boldsymbol{\gamma}}^\top X_{\boldsymbol{\gamma}}}{n})^{-\frac{1}{2}} \frac{X_{\boldsymbol{\gamma}}^\top \boldsymbol{\epsilon}}{\sqrt{n}} \\
&= \boldsymbol{U}^\top \boldsymbol{U},
\end{aligned}$$

where $\boldsymbol{U} = (X_{\boldsymbol{\gamma}}^\top X_{\boldsymbol{\gamma}}/n)^{-1/2} X_{\boldsymbol{\gamma}}^\top \boldsymbol{\epsilon}/\sqrt{n} = H^{-1/2} X_{\boldsymbol{\gamma}}^\top \boldsymbol{\epsilon}/\sqrt{n}$, and $H = X_{\boldsymbol{\gamma}}^\top X_{\boldsymbol{\gamma}}/n$. Denote $X_{\boldsymbol{\gamma}} = (\boldsymbol{X}_1, \boldsymbol{X}_2, ...\boldsymbol{X}_n)^\top$, where $\boldsymbol{X}_i$'s are column vectors. Denote $\boldsymbol{\epsilon} = (\epsilon_1, ..., \epsilon_n)^\top$. Then

$$\boldsymbol{U} = H^{-1/2} X_{\boldsymbol{\gamma}}^\top \boldsymbol{\epsilon}/\sqrt{n} = H^{-1/2}(\boldsymbol{X}_1, \boldsymbol{X}_2, ..., \boldsymbol{X}_n)(\epsilon_1, ..., \epsilon_n)^\top/\sqrt{n} = H^{-1/2} \sum_{i=1}^n \boldsymbol{X}_i \epsilon_i/\sqrt{n}.$$

Thus the moment generating function of $\boldsymbol{U}$ is

$$\begin{aligned}
\mathbb{E}(e^{\boldsymbol{t}^\top \boldsymbol{U}}) &= \mathbb{E}(e^{\boldsymbol{t}^\top/\sqrt{n} H^{-1/2} \sum_{i=1}^n \boldsymbol{X}_i \epsilon_i}) \\
&= \prod_{i=1}^n \mathbb{E}(e^{\boldsymbol{t}^\top/\sqrt{n} H^{-1/2} \boldsymbol{X}_i \epsilon_i}) = \prod_{i=1}^n \mathbb{E}(e^{s_i/\sqrt{n}\boldsymbol{\epsilon}_i}) \leqslant e^{\sum_{i=1}^n s_i^2 \sigma^2/(2n)}
\end{aligned}$$

by setting $s_i = \boldsymbol{t}^\top H^{-\frac{1}{2}} \boldsymbol{X}_i$. When $|s_i| \leqslant \sqrt{n}/b$,

$$\mathbb{E}(e^{\boldsymbol{t}^\top \boldsymbol{U}}) \leqslant \exp\left(\frac{\sigma^2}{2n} \sum_{i=1}^n s_i^2\right).$$

By the fact that

$$\begin{aligned}
\sum_{i=1}^n s_i^2 &= \sum_{i=1}^n \boldsymbol{t}^\top H^{-\frac{1}{2}} \boldsymbol{X}_i \boldsymbol{X}_i^\top H^{-\frac{1}{2}} \boldsymbol{t} = \boldsymbol{t}^\top H^{-\frac{1}{2}} \left(\sum_{i=1}^n \boldsymbol{X}_i \boldsymbol{X}_i^\top\right) H^{-\frac{1}{2}} \boldsymbol{t} \\
&= \boldsymbol{t}^\top H^{-\frac{1}{2}} X_{\boldsymbol{\gamma}}^\top X_{\boldsymbol{\gamma}} H^{-\frac{1}{2}} \boldsymbol{t} = n \boldsymbol{t}^\top \boldsymbol{t},
\end{aligned}$$

we have $\mathbb{E}(e^{\boldsymbol{t}^\top \boldsymbol{U}}) \leqslant \exp\left(\sigma^2 \boldsymbol{t}^\top \boldsymbol{t}/2\right)$. For s_i, by Assumption 2.2.1, we have

$$\begin{aligned}
|s_i| = |\boldsymbol{t}^\top H^{-\frac{1}{2}} \boldsymbol{X}_i| &\leqslant ||\boldsymbol{t}|| \sqrt{\boldsymbol{X}_i^\top H^{-1} \boldsymbol{X}_i} \\
&\leqslant ||\boldsymbol{t}|| ||\boldsymbol{X}_i|| C_5^{-\frac{1}{2}} \leqslant C_4 C_5^{-\frac{1}{2}} \sqrt{|\boldsymbol{\gamma}|} ||\boldsymbol{t}|| \\
&\leqslant C_4 C_5^{-\frac{1}{2}} \sqrt{r_n} ||\boldsymbol{t}||.
\end{aligned}$$

When $||\boldsymbol{t}|| \leqslant \sqrt{\frac{n}{b^2 C_4^2 C_5^{-1} r_n}}$, we have $s_i \leqslant \sqrt{n}/b$, and $\mathbb{E}(e^{\boldsymbol{t}^\top \boldsymbol{U}/\sigma}) \leqslant e^{\boldsymbol{t}^\top \boldsymbol{t}/2}$, which means that $\boldsymbol{U}/\sigma$ satisfies the exponential moment condition.

$$\begin{aligned}
&\Pr\left(\boldsymbol{\epsilon}^\top X_{\boldsymbol{\gamma}} \left(X_{\boldsymbol{\gamma}}^\top X_{\boldsymbol{\gamma}}\right)^{-1} X_{\boldsymbol{\gamma}}^\top \boldsymbol{\epsilon} \geqslant t\right) \\
&= \Pr\Big(\frac{\boldsymbol{\epsilon}^\top X_{\boldsymbol{\gamma}}}{\sqrt{n}} \left(\frac{X_{\boldsymbol{\gamma}}^\top X_{\boldsymbol{\gamma}}}{n}\right)^{-\frac{1}{2}} I_{|\boldsymbol{\gamma}|} \left(\frac{X_{\boldsymbol{\gamma}}^\top X_{\boldsymbol{\gamma}}}{n}\right)^{-\frac{1}{2}} \frac{X_{\boldsymbol{\gamma}}^\top \boldsymbol{\epsilon}}{\sqrt{n}} \geqslant t\Big) \\
&= \Pr\left(\boldsymbol{U}^\top I_{|\boldsymbol{\gamma}|} \boldsymbol{U} \geqslant t\right) \\
&= \Pr\left(\boldsymbol{U}^\top I_{|\boldsymbol{\gamma}|} \boldsymbol{U}/\sigma^2 \geqslant t/\sigma^2\right).
\end{aligned}$$

By Corollary 4.2 in Spokoiny (2013), we have

$$\begin{aligned}
&\Pr\left(\boldsymbol{\epsilon}^\top X_{\boldsymbol{\gamma}} \left(X_{\boldsymbol{\gamma}}^\top X_{\boldsymbol{\gamma}}\right)^{-1} X_{\boldsymbol{\gamma}}^\top \boldsymbol{\epsilon} \geqslant |\boldsymbol{\gamma}| + 6x\right) \\
&\leqslant 10.4 \exp(-x/\sigma^2) = 10.4 \exp(-\alpha |\boldsymbol{\gamma}| \log p_n) = 10.4 p_n^{-\alpha|\boldsymbol{\gamma}|}.
\end{aligned}$$

This entails

$$\begin{aligned}
&\Pr(\max_{\boldsymbol{\gamma} \in S_2} \boldsymbol{\epsilon}^\top P_{\boldsymbol{\gamma}} \boldsymbol{\epsilon} \geqslant 7x) = \Pr(\max_{\boldsymbol{\gamma} \in S_2 \setminus \varnothing} \boldsymbol{\epsilon}^\top P_{\boldsymbol{\gamma}} \boldsymbol{\epsilon} \geqslant 7x) \\
&\leqslant \sum_{\boldsymbol{\gamma} \in S_2 \setminus \varnothing} \Pr\left(\boldsymbol{\epsilon}^\top P_{\boldsymbol{\gamma}} \boldsymbol{\epsilon} \geqslant |\boldsymbol{\gamma}| + 6x\right) \leqslant \sum_{\boldsymbol{\gamma} \in S_2 \setminus \varnothing} 10.4 p_n^{-\alpha|\boldsymbol{\gamma}|} \\
&\leqslant \sum_{r=1}^{r_n} \binom{p_n}{r} 10.4 p_n^{-\alpha r} \leqslant 10.4((1 + p_n^{-\alpha})^{p_n} - 1) \to 0.
\end{aligned}$$

□

Lemma 2.2.5 *When n tends to infinity, for $\boldsymbol{\gamma} \in S_1$, there exists a constant $\alpha_0 > 7$, such that*

$$T_5^{(i)} \geqslant -2^{-1}(|\boldsymbol{\gamma}| - s_n)\alpha_0 \log p_n$$

with probability tending to 1. When $\boldsymbol{\gamma}_0 \neq \mathbf{0}$, for $\boldsymbol{\gamma} \in S_2$, there exists a constant C' such that

$$T_5^{(i)} \geqslant (n + v_i)/2 \log(1 + C'\psi_n^2)$$

with probability tending to 1. When $\boldsymbol{\gamma}_0 = \mathbf{0}$, S_2 is empty. Then $T_5^{(i)}$ is not considered.

Proof of Lemma 2.2.5. We consider two cases of $\boldsymbol{\gamma} \in S_1$ and $\boldsymbol{\gamma} \in S_2$ separately. Suppose $\epsilon_i \sim SubE(\sigma_i^2, b), i \in \{1, \ldots, k\}$. According to Assumption 2.2.5, $r_n^2 \log^2 p_n = o(n \log(1+\psi_n^2))$. Therefore, $r_n^2 \log^2 p_n = o(n\psi_n^2)$. Let $\boldsymbol{v}_{i,\boldsymbol{\gamma}} = (I_n - P_{i,\boldsymbol{\gamma}})X_{i,\boldsymbol{\gamma}_0\setminus\boldsymbol{\gamma}}\boldsymbol{\beta}^0_{i,\boldsymbol{\gamma}_0\setminus\boldsymbol{\gamma}}$. For any $\boldsymbol{\gamma} \in S_2$, according to Lemma 2.2.4 (a) and (c), there exists $C > 0$ such that when n goes to infinity, with probability tending to 1,

$$\begin{aligned}
\boldsymbol{y}_i^\top (I_n - P_{i,\boldsymbol{\gamma}})\boldsymbol{y}_i &= ||\boldsymbol{v}_{i,\boldsymbol{\gamma}}||_2^2 + 2v_{i,\boldsymbol{\gamma}}^\top \boldsymbol{\epsilon}_i + \boldsymbol{\epsilon}_i^\top (I_n - P_{i,\boldsymbol{\gamma}})\boldsymbol{\epsilon}_i \\
&\geqslant ||\boldsymbol{v}_{i,\boldsymbol{\gamma}}||_2^2 - 2Cr_n \log p_n ||\boldsymbol{v}_{i,\boldsymbol{\gamma}}||_2 + \boldsymbol{\epsilon}_i^\top \boldsymbol{\epsilon}_i - 7\alpha\sigma_i^2|\gamma| \log p_n \\
&\geqslant ||\boldsymbol{v}_{i,\boldsymbol{\gamma}}||_2^2 \Big(1 - 2C\frac{r_n \log p_n}{||\boldsymbol{v}_{i,\boldsymbol{\gamma}}||_2} - \frac{7\alpha|\gamma| \log p_n}{n\varphi_{\min}(n)\psi_n^2}\Big) + \boldsymbol{\epsilon}_i^\top \boldsymbol{\epsilon}_i \\
&= ||\boldsymbol{v}_{i,\boldsymbol{\gamma}}||_2^2 (1 + o(1)) + \boldsymbol{\epsilon}_i^\top \boldsymbol{\epsilon}_i \\
&\geqslant n\varphi_{\min}(n)\psi_n^2(1 + o(1)) + \boldsymbol{\epsilon}_i^\top \boldsymbol{\epsilon}_i.
\end{aligned}$$

Note that Assumption 2.2.5 implies that $s_n = o(n)$, and therefore, $\boldsymbol{y}_i^\top (I_n - P_{i,\boldsymbol{\gamma}_0})\boldsymbol{y}_i = \boldsymbol{\epsilon}_i^\top (I_n - P_{i,\boldsymbol{\gamma}_0})\boldsymbol{\epsilon}_i = (n - s_n)\sigma_i^2(1 + o_p(1))$. Thus, for $\boldsymbol{\gamma} \in S_2$, there exists a C' such that with probability tending to 1,

$$T_5 \geqslant \frac{n + v_i}{2} \log\Big(\frac{1 + n\varphi_{\min}(n)\psi_n^2(1 + o(1)) + \boldsymbol{\epsilon}_i^\top \boldsymbol{\epsilon}_i}{1 + (n - s_n)\sigma_i^2(1 + o_p(1))}\Big) \geqslant \frac{n + v_i}{2} \log(1 + C'\psi_n^2).$$

On the other hand, for $\boldsymbol{\gamma} \in S_1$, by the properties of projection matrices and Lemma 2.2.4 (b), with

probability tending to 1,

$$\begin{aligned}
&\frac{1+\boldsymbol{y}_i^\top(I_n-P_{i,\boldsymbol{\gamma}})\boldsymbol{y}_i}{1+\boldsymbol{y}_i^\top(I_n-P_{i,\boldsymbol{\gamma}_0})\boldsymbol{y}_i}\\
=\;& 1-\frac{\boldsymbol{y}_i^\top(P_{i,\boldsymbol{\gamma}}-P_{i,\boldsymbol{\gamma}_0})\boldsymbol{y}_i}{1+\boldsymbol{y}_i^\top(I_n-P_{i,\boldsymbol{\gamma}_0})\boldsymbol{y}_i}\\
=\;& 1-\frac{\boldsymbol{\beta}_{\boldsymbol{\gamma}_0}^{0}X_{\boldsymbol{\gamma}_0}^\top(P_{i,\boldsymbol{\gamma}}-P_{i,\boldsymbol{\gamma}_0})X_{\boldsymbol{\gamma}_0}\boldsymbol{\beta}_{\boldsymbol{\gamma}_0}+2\boldsymbol{\beta}_{\boldsymbol{\gamma}_0}^{0\top}X_{\boldsymbol{\gamma}_0}^\top(P_{i,\boldsymbol{\gamma}}-P_{i,\boldsymbol{\gamma}_0})\boldsymbol{\epsilon}_i+\boldsymbol{\epsilon}_i^\top(P_{i,\boldsymbol{\gamma}}-P_{i,\boldsymbol{\gamma}_0})\boldsymbol{\epsilon}_i}{1+\boldsymbol{y}_i^\top(I_n-P_{i,\boldsymbol{\gamma}_0})\boldsymbol{y}_i}\\
=\;& 1-\frac{\boldsymbol{\epsilon}_i^\top(P_{i,\boldsymbol{\gamma}}-P_{i,\boldsymbol{\gamma}_0})\boldsymbol{\epsilon}_i}{1+\boldsymbol{\epsilon}_i^\top(I_n-P_{i,\boldsymbol{\gamma}_0})\boldsymbol{\epsilon}_i}\geqslant 1-\frac{7\alpha(|\boldsymbol{\gamma}|-s_n)\log p_n}{n},
\end{aligned}$$

where $\alpha_0>7\alpha\leqslant\alpha_0$. Using the fact that $\log(1-x)\geqslant -x$ for $x\in(0,1/2)$, with probability tending to 1, $T_5^{(i)}\geqslant\frac{n+v_i}{2}\log\left(1-\frac{7\alpha(|\boldsymbol{\gamma}|-s_n)\log p_n}{n}\right)\geqslant -2^{-1}\alpha_0(|\boldsymbol{\gamma}|-s_n)\log p_n$.

□

Proof of Theorem 2.2.2. First, we consider $\boldsymbol{\gamma}_0\neq\mathbf{0}$. Combining the results from the lemmas, we obtain the bounds for each of the terms, T_1, T_2, T_3, T_4, and T_5. Using the fact that $p_n^{\alpha_0-\frac{2M}{k}}=o(\rho_n)$ with $\rho_n\equiv n^{1-\sigma}\underline{\phi}_n$, uniformly for $\boldsymbol{\gamma}\in S_1$ with probability tending to 1,

$$\begin{aligned}
&P_C(\boldsymbol{\gamma}|Z_1,...,Z_k)/P_C(\boldsymbol{\gamma}_0|Z_1,...,Z_k)\\
&\leqslant\widetilde{C}\exp\left(-\frac{|\boldsymbol{\gamma}|-s_n}{2}(\log(1+C_3\rho_n)-\alpha_0\log p_n)\right)^k\left(\frac{q}{1-q}\right)^{|\boldsymbol{\gamma}|-s_n}\\
&=\widetilde{C}\left(\frac{1+C_3\rho_n}{p_n^{\alpha_0}}\right)^{-\frac{k(|\boldsymbol{\gamma}|-s_n)}{2}}\left(\frac{q}{1-q}\right)^{|\boldsymbol{\gamma}|-s_n}\\
&=\widetilde{C}\left(\frac{p_n^{\alpha_0}}{1+C_3\rho_n}\left(\frac{p_n^{-M}}{1-p_n^{-M}}\right)^{\frac{2}{k}}\right)^{\frac{k(|\boldsymbol{\gamma}|-s_n)}{2}}\\
&=\left(\frac{p_n^{\alpha_0-\frac{2M}{k}}}{(1+C_3\rho_n)(1-p_n^{-M})^{\frac{2}{k}}}\right)^{\frac{k(|\boldsymbol{\gamma}|-s_n)}{2}}\xrightarrow{p}0,\text{when } n\to\infty,
\end{aligned}$$

where $\widetilde{C}$ is a constant depending on the lower bound of $T_4^{(i)}$. By Assumptions 2.2.2 and 2.2.5,

$s_n \log(1 + C_2 n\overline{\phi}_n) \ll (n + v_i)/2 \log(1 + C'\psi_n^2)$. Therefore, with probability tending to 1, uniformly for $\boldsymbol{\gamma} \in S_2$,

$$\begin{aligned}
&P_C(\boldsymbol{\gamma}|Z_1, ..., Z_k)/P_C(\boldsymbol{\gamma}_0|Z_1, ..., Z_k) \\
&\leqslant \widetilde{C} \exp\left(\frac{s_n k}{2} \log(1 + C_2 n\overline{\phi}_n) - \frac{kn + \sum v_i}{2} \log(1 + C'\psi_n^2)\right) \left(\frac{q}{1-q}\right)^{|\boldsymbol{\gamma}|-s_n} \\
&\leqslant \widetilde{C}(1 + C'\psi_n^2)^{-\frac{nk+\sum v_i}{4}} \left(\frac{q}{1-q}\right)^{|\boldsymbol{\gamma}|-s_n} \xrightarrow{p} 0,
\end{aligned}$$

when $n \to \infty$.

To establish the stronger version of the model selection consistency, which demonstrates that the probability of selecting true model approaches 1, we consider two cases of overfitting models and wrong models. For overfitting models, we have

$$\begin{aligned}
&\sum_{\boldsymbol{\gamma} \in S_1} P_C(\boldsymbol{\gamma}|Z_1, ..., Z_k)/P_C(\boldsymbol{\gamma}_0|Z_1, ..., Z_k) \leqslant \widetilde{C} \sum_{\boldsymbol{\gamma} \in S_1} \left(\frac{1 + C_3\rho_n}{p_n^{\alpha_0 - \frac{2M}{k}}}\right)^{-2^{-1}k(|\boldsymbol{\gamma}|-s_n)} \\
= \;& \widetilde{C} \sum_{r=1}^{r_n - s_n} \binom{p_n - s_n}{r} \left(\frac{1 + C_3\rho_n}{p_n^{\alpha_0 - \frac{2M}{k}}}\right)^{-rk/2} \leqslant \widetilde{C} \sum_{r=1}^{r_n - s_n} \frac{p_n^r}{r!} \left(\frac{p_n^{\alpha_0 - \frac{2M}{k}}}{1 + C_3\rho_n}\right)^{rk/2} \\
= \;& \widetilde{C} \sum_{r=1}^{r_n - s_n} \frac{1}{r!} \left(\frac{p_n^{\alpha_0 - \frac{2M-2}{k}}}{1 + C_3\rho_n}\right)^{rk/2} \leqslant \widetilde{C} \sum_{r=1}^{r_n - s_n} \left(\frac{p_n^{\alpha_0 - \frac{2M-2}{k}}}{1 + C_3\rho_n}\right)^{kr/2} \\
\leqslant \;& \widetilde{C} \frac{\left(\frac{p_n^{\alpha_0 - \frac{2M-2}{k}}}{1 + C_3\rho_n}\right)^{k/2}}{1 - \left(\frac{p_n^{\alpha_0 - \frac{2M-2}{k}}}{1 + C_3\rho_n}\right)^{k/2}} \xrightarrow{p} 0, \text{ when } n \to \infty,
\end{aligned}$$

where the last limit result comes from the assumption that $p_n^{\alpha_0 + (2-2M)/k} = o(\rho_n)$. Note that

$$\left(\frac{q}{1-q}\right)^{|\boldsymbol{\gamma}|-s_n} \leqslant \left(\frac{1-q}{q}\right)^{s_n} \leqslant q^{-s_n} = p_n^{Ms_n}.$$

Using the fact that $r_n^2 \log p_n = o(n \log(1 + \psi_n^2))$, for $\boldsymbol{\gamma} \in S_2$, we have

$$\sum_{\boldsymbol{\gamma} \in S_2} P_C(\boldsymbol{\gamma}|Z_1, ..., Z_k)/P_C(\boldsymbol{\gamma}_0|Z_1, ..., Z_k) \leqslant \widetilde{C} p_n^{Ms_n} (1 + C'\psi_n^2)^{-(kn + \sum v_i)/4} \xrightarrow{p} 0.$$

When $\boldsymbol{\gamma}_0$ is null, similar to the null case in the sub-Gaussian section, we can prove the theorem as well. Thus

$$P_C(\boldsymbol{\gamma}_0|Z_1, ..., Z_k) \xrightarrow{p} 1.$$

□

Remark 2.2.6 *Theorem 2.2.2 implies the strong model selection consistency result under sub-exponential error distributions. For both Theorems 2.2.1 and 2.2.2, if a larger M is chosen in the prior probability $q = p_n^{-M}$, it will make the requirement on p_n less stringent and allow p_n to grow at a higher rate with the sample size n.*

2.3 Union support recovery

In the previous sections, we focued on the setting that the true model is identical across different platforms. However, in real-world applications, different platforms may share a common set of important predictors while still having distinct true models. For example, in the analysis of handwritten digit images, chirography can vary among individuals. For the same digit, the pixel-based features are similar but not identical across different people (Ramesh et al. (2022)). When handwriting samples are collected from different individuals, viewed as separate platforms, the underlying models may exhibit broad concordance while differing in finer details. Another example involves multitask learning, as considered by Park and D'Amico (2024), where three tasks—predicting satellite key points, direct pose regression, and binary segmentation of the satellite foreground—were jointly modelled. While these tasks are distinct, they are fundamentally connected by the spacecraft's core features, including its geometry and pose.

Selecting the predictors that are important in at least one platform corresponds to identifying

the union of the true models across different platforms, referred to as union support recovery. In this section, we demonstrate that when the true model varies across platforms, the model selection criterion in Equation (2.1) is consistent in selecting the union of the true models.

From earlier discussions, it is evident that the ratio of the model selection criterion between the true model and an under-fitting model decreases exponentially with the sample size, while the ratio between the true model and an over-fitting model decreases at a much slower rate. The union support model is either the true model or an over-fitting model for all platforms, whereas any other competing model is an under-fitting model for at least one platform. Consequently, the union support model is preferred by the model selection criterion.

Assume that the true models for each platform are $\boldsymbol{\gamma}_1^0, ..., \boldsymbol{\gamma}_k^0$, which are not necessarily identical. The union of the true models is denoted by $\boldsymbol{\gamma}_o$, whose jth element is defined by $\boldsymbol{\gamma}_{oj} = I\{(\gamma_{1j}^0 + \cdots + \gamma_{kj}^0) \neq 0\}, j \in \{1, \ldots, p_n\}$. Denote $S_3 = \{\boldsymbol{\gamma} \mid \boldsymbol{\gamma}_o \backslash \boldsymbol{\gamma} = \mathbf{0}, \boldsymbol{\gamma}_o \neq \boldsymbol{\gamma}\}$ and $S_4 = \{\boldsymbol{\gamma} \mid \boldsymbol{\gamma}_o \backslash \boldsymbol{\gamma} \neq \mathbf{0}\}$. Then, $S_3 \cup S_4 \cup \{\boldsymbol{\gamma}_o\}$ is the whole model space. Denote

$$p_c(\boldsymbol{\gamma}|Z_i) \propto |W_{i,\boldsymbol{\gamma}}|^{-\frac{1}{2}} (1 + \boldsymbol{y_i}^\top (I_n - X_{i,\boldsymbol{\gamma}} (X_{i,\boldsymbol{\gamma}}^\top X_{i,\boldsymbol{\gamma}} + \Sigma_{i,\boldsymbol{\gamma}}^{-1})^{-1} X_{i,\boldsymbol{\gamma}}^\top) \boldsymbol{y_i})^{-\frac{n+v_i}{2}},$$

and

$$P_C(\boldsymbol{\gamma}|Z_1, ..., Z_k) / P_C(\boldsymbol{\gamma}_o|Z_1, ..., Z_k) = \frac{p(\boldsymbol{\gamma})}{p(\boldsymbol{\gamma}_o)} \prod_{i=1}^{k} \{\frac{p_c(\boldsymbol{\gamma}|Z_i)}{p_c(\boldsymbol{\gamma}_o|Z_i)}\}.$$

Assumption 2.3.1 *It is assumed that for some constant $C' > 0$,*

$$r_n \log(1 + n\overline{\phi}_n) = o(n \log(1 + C'\psi_n^2)),$$

where ψ_n is defined in Assumption 2.2.2.

Remark 2.3.1 *When the true models differ across platforms, a single candidate model may simultaneously overfit some platforms while underfitting others. This scenario necessitates calculating*

the lower bound of the Bayes factors between an overfitting model and an underfitting model. To establish this lower bound, Assumption 2.3.1 is required.

Theorem 2.3.1 *Under Assumptions 2.2.1, 2.2.3, 2.2.4, 2.2.5, 2.3.1 , given sub-Gaussian random errors with some $\alpha_0 > 2$ or given sub-exponential random errors with some $\alpha_0 > 7$, if $p_n^{\alpha_0 - 2M/k} = o(n^{1-\delta}\underline{\phi}_n)$, then*

$$\sup_{c_{i,1},...,c_{i,p_n} \in [\underline{\phi}_n, \overline{\phi}_n], i \in \{1,...,k\}} \max_{\boldsymbol{\gamma} \neq \boldsymbol{\gamma}_o} P_C(\boldsymbol{\gamma}|Z_1, Z_2, ..., Z_k)/P_C(\boldsymbol{\gamma}_o|Z_1, Z_2, ..., Z_k) \xrightarrow{p} 0.$$

If $p_n^{\alpha_0 - 2(M-1)/k} = o(n^{1-\delta}\underline{\phi}_n)$, then

$$\sup_{c_{i,1},...,c_{i,p_n} \in [\underline{\phi}_n, \overline{\phi}_n], i \in \{1,...,k\}} \sum_{\boldsymbol{\gamma} \neq \boldsymbol{\gamma}_o} P_C(\boldsymbol{\gamma}|Z_1, Z_2, ..., Z_k) \xrightarrow{p} 0.$$

To prove Theorem 2.3.1, we need to introduce two lemmas.

Lemma 2.3.1 *Suppose that $\boldsymbol{\gamma}_i^0$ is the true model for Z_i, and $\boldsymbol{\gamma}_o$ is an over-fitting model for Z_i. Under Assumptions 2.2.1, 2.2.2, 2.2.5, 2.3.1, for model $\boldsymbol{\gamma}$, with $\boldsymbol{\gamma}_i^0 \subseteq \boldsymbol{\gamma} \subseteq \boldsymbol{\gamma}_o$, there exists a constant C_7 such that*

$$p_c(\boldsymbol{\gamma}|Z_i)/p_c(\boldsymbol{\gamma}_o|Z_i) \leqslant C_7(1 + nC_6\ \overline{\phi}_n)^{|\boldsymbol{\gamma}_o|/2}$$

with probability tending to 1.

Lemma 2.3.2 *Under Assumptions 2.2.1, 2.2.2, 2.2.5, 2.3.1, if $\boldsymbol{\gamma} \in S_3$, there exists a constant C_8 such that*

$$p_c(\boldsymbol{\gamma}|Z_i)/p_c(\boldsymbol{\gamma}_o|Z_i) \leqslant C_8\Big(\exp(-2^{-1}(|\boldsymbol{\gamma}| - |\boldsymbol{\gamma}_o|))\log((1 + C_3\rho_n)/p_n^{\alpha_0})\Big)$$

with probability tending to 1. If $\boldsymbol{\gamma} \in S_4$, and $\boldsymbol{\gamma}_i^0 \nsubseteq \boldsymbol{\gamma}$, then

$$p_c(\boldsymbol{\gamma}|Z_i)/p_c(\boldsymbol{\gamma}_o|Z_i) \leqslant C_9(1 + C'\psi_n^2)^{-\frac{n+v_i}{4}}$$

with probability tending to 1.

Proof of Lemma 2.3.1. According to Equation 2.2, we have

$$\begin{aligned}
-\log\Big(\frac{p_c(\boldsymbol{\gamma}_o|Z_i)}{p_c(\boldsymbol{\gamma}|Z_i)}\Big) &= \frac{1}{2}\log\Big(\frac{|W_{i,\boldsymbol{\gamma}_o}|}{|W_{i,\boldsymbol{\gamma}}|}\Big) + \frac{n+v_i}{2}\log\Big(\frac{1+\boldsymbol{y}_i^\top(I_n - X_{i,\boldsymbol{\gamma}_o}U_{i,\boldsymbol{\gamma}_o}^{-1}X_{i,\boldsymbol{\gamma}_o}^\top)\boldsymbol{y}_i}{1+\boldsymbol{y}_i^\top(I_n-P_{i,\boldsymbol{\gamma}_o})\boldsymbol{y}_i}\Big) \\
&- \frac{n+v_i}{2}\log\Big(\frac{1+\boldsymbol{y}_i^\top(I_n - X_{i,\boldsymbol{\gamma}}U_{i,\boldsymbol{\gamma}}^{-1}X_{i,\boldsymbol{\gamma}}^\top)\boldsymbol{y}_i}{1+\boldsymbol{y}_i^\top(I_n-P_{i,\boldsymbol{\gamma}})\boldsymbol{y}_i}\Big) \\
&+ \frac{n+v_i}{2}\log\Big(\frac{1+\boldsymbol{y}_i^\top(I_n - P_{i,\boldsymbol{\gamma}_o})\boldsymbol{y}_i}{1+\boldsymbol{y}_i^\top(I_n-P_{i,\boldsymbol{\gamma}})\boldsymbol{y}_i}\Big)\Big\} \\
&= T_2^{(i)} + T_3^{(i)} - T_4^{(i)} + T_5^{(i)}. \qquad (2.4)
\end{aligned}$$

Note that $|W_{i,\boldsymbol{\gamma}}| \geqslant 1$. For $T_2^{(i)}$, according to Assumption 2.2.1,

$$\begin{aligned}
T_2^{(i)} &= 1/2\log(\frac{|W_{i,\boldsymbol{\gamma}_o}|}{|W_{i,\boldsymbol{\gamma}}|}) \leqslant 1/2\log|W_{i,\boldsymbol{\gamma}_o}| = 1/2\log(|U_{i,\boldsymbol{\gamma}_o}||\Sigma_{i,\boldsymbol{\gamma}_o}|) \\
&= 1/2\log(|I_{|\boldsymbol{\gamma}_o|} + X_{i,\boldsymbol{\gamma}_o}^\top X_{i,\boldsymbol{\gamma}_o}\Sigma_{i,\boldsymbol{\gamma}_o}|) \leqslant 1/2|\boldsymbol{\gamma}_o|\log(1+C_2 n\overline{\phi}_n).
\end{aligned}$$

Since

$$U_{i,\boldsymbol{\gamma}} = X_{i,\boldsymbol{\gamma}}^\top X_{i,\boldsymbol{\gamma}} + \Sigma_{i,\boldsymbol{\gamma}}^{-1} \geqslant X_{i,\boldsymbol{\gamma}}^\top X_{i,\boldsymbol{\gamma}},$$

we have $T_4^{(i)} \geqslant 0$ for any n. Next, we establish a bound for $T_3^{(i)}$. Let

$$\Delta = \boldsymbol{y}_i^\top X_{i,\boldsymbol{\gamma}_o}(X_{i,\boldsymbol{\gamma}_o}^\top X_{i,\boldsymbol{\gamma}_o})^{-1}(\Sigma_{\boldsymbol{\gamma}_o} + (X_{i,\boldsymbol{\gamma}_o}^\top X_{i,\boldsymbol{\gamma}_o})^{-1})^{-1}(X_{i,\boldsymbol{\gamma}_o}^\top X_{i,\boldsymbol{\gamma}_o})^{-1}X_{i,\boldsymbol{\gamma}_o}^\top \boldsymbol{y}_i.$$

By the Sherman-Morrison-Woodbury formula,

$$\begin{aligned}
U_{i,\boldsymbol{\gamma}_o}^{-1} =& (X_{i,\boldsymbol{\gamma}_o}^\top X_{i,\boldsymbol{\gamma}_o} + \Sigma_{i,\boldsymbol{\gamma}_o}^{-1})^{-1} = (X_{i,\boldsymbol{\gamma}_o}^\top X_{i,\boldsymbol{\gamma}_o})^{-1} \\
&- (X_{i,\boldsymbol{\gamma}_o}^\top X_{i,\boldsymbol{\gamma}_o})^{-1}(\Sigma_{\boldsymbol{\gamma}_o} + (X_{i,\boldsymbol{\gamma}_o}^\top X_{i,\boldsymbol{\gamma}_o})^{-1})^{-1}(X_{i,\boldsymbol{\gamma}_o}^\top X_{i,\boldsymbol{\gamma}_o})^{-1}.
\end{aligned}$$

This implies that

$$U_{i,\boldsymbol{\gamma}_o}^{-1} - (X_{i,\boldsymbol{\gamma}_o}^\top X_{i,\boldsymbol{\gamma}_o})^{-1} = -(X_{i,\boldsymbol{\gamma}_o}^\top X_{i,\boldsymbol{\gamma}_o})^{-1}(\Sigma_{\boldsymbol{\gamma}_o} + (X_{i,\boldsymbol{\gamma}_o}^\top X_{i,\boldsymbol{\gamma}_o})^{-1})^{-1}(X_{i,\boldsymbol{\gamma}_o}^\top X_{i,\boldsymbol{\gamma}_o})^{-1}.$$

Thus,

$$\begin{aligned}
&\frac{1+\boldsymbol{y}_i^\top(I_n - X_{i,\boldsymbol{\gamma}_o}U_{i,\boldsymbol{\gamma}_o}^{-1}X_{i,\boldsymbol{\gamma}_o}^\top)\boldsymbol{y}_i}{1+\boldsymbol{y}_i^\top(I_n-P_{i,\boldsymbol{\gamma}_o})\boldsymbol{y}_i} = 1 + \frac{\boldsymbol{y}_i^\top(P_{i,\boldsymbol{\gamma}_o} - X_{i,\boldsymbol{\gamma}_o}U_{i,\boldsymbol{\gamma}_o}^{-1}X_{i,\boldsymbol{\gamma}_o}^\top)\boldsymbol{y}_i}{1+\boldsymbol{y}_i^\top(I_n-P_{i,\boldsymbol{\gamma}_o})\boldsymbol{y}_i} \\
&= 1 + \frac{\boldsymbol{y}_i^\top X_{i,\boldsymbol{\gamma}_o}((X_{i,\boldsymbol{\gamma}_o}^\top X_{i,\boldsymbol{\gamma}_o})^{-1} - U_{i,\boldsymbol{\gamma}_o}^{-1})X_{i,\boldsymbol{\gamma}_o}^\top \boldsymbol{y}_i}{1+\boldsymbol{y}_i^\top(I_n-P_{i,\boldsymbol{\gamma}_o})\boldsymbol{y}_i} = 1 + \frac{\Delta}{1+\boldsymbol{y}_i^\top(I_n-P_{i,\boldsymbol{\gamma}_o})\boldsymbol{y}_i}.
\end{aligned}$$

Using the facts that $(a+b)^2 \leqslant 2(a^2+b^2)$, and $\boldsymbol{y}_i = X_{i,\gamma_o}\boldsymbol{\beta}^0_{i,\gamma_i^0} + \boldsymbol{\epsilon}_i$, and $(\Sigma_{i,\gamma_o} + (X^\top_{i,\gamma_o}X_{i,\gamma_o})^{-1})^{-1} \leqslant \Sigma^{-1}_{i,\gamma_o}$, we have

$$1 + \frac{\Delta}{1+\boldsymbol{y}_i^\top(I_n - P_{i,\gamma_o})\boldsymbol{y}_i} \leqslant$$
$$1 + 2\frac{\boldsymbol{\beta}^{0\top}_{i,\gamma_i^0}\Sigma^{-1}_{i,\gamma_o}\boldsymbol{\beta}^0_{i,\gamma_i^0} + \boldsymbol{\epsilon}_i^\top X_{i,\gamma_o}(X^\top_{i,\gamma_o}X_{i,\gamma_o})^{-1}\Sigma^{-1}_{i,\gamma_o}(X^\top_{i,\gamma_o}X_{i,\gamma_o})^{-1}X^\top_{i,\gamma_o}\boldsymbol{\epsilon}_i}{1+\boldsymbol{y}_i^\top(I_n - P_{i,\gamma_o})\boldsymbol{y}_i}.$$

By Assumption 2.2.2, $\underline{\phi}_n I_{\gamma_o} \leqslant \Sigma_{i,\gamma_o}$, therefore

$$1 + \frac{\Delta}{1+\boldsymbol{y}_i^\top(I_n - P_{i,\gamma_o})\boldsymbol{y}_i} \leqslant 1 + 2\underline{\phi}_n^{-1}\frac{\|\boldsymbol{\beta}^0_{i,\gamma_i^0}\|_o^2 + \boldsymbol{\epsilon}_i^\top X_{i,\gamma_o}(X^\top_{i,\gamma_o}X_{i,\gamma_o})^{-2}X^\top_{i,\gamma_o}\boldsymbol{\epsilon}_i}{1+\boldsymbol{y}_i^\top(I_n - P_{i,\gamma_o})\boldsymbol{y}_i}.$$

Next, we have

$$\begin{aligned}\boldsymbol{y}_i^\top(I_n - P_{i,\gamma_o})\boldsymbol{y}_i/n &\leqslant \boldsymbol{y}_i^\top(I_n - P_{i,\gamma_i^0})\boldsymbol{y}_i/n = \boldsymbol{\epsilon}_i^\top(I_n - P_{i,\gamma_i^0})\boldsymbol{\epsilon}_i/n \\ &= \frac{\boldsymbol{\epsilon}_i^\top(I_n - P_{i,\gamma_i^0})\boldsymbol{\epsilon}_i}{(n-|\boldsymbol{\gamma}_i^0|)}\frac{n-|\boldsymbol{\gamma}_i^0|}{n} \xrightarrow{p} \sigma_i^2,\end{aligned}$$

as $n \to \infty$. Note that $C_5 \leqslant \lambda_-(\frac{1}{n}X^\top_{i,\gamma_o}X_{i,\gamma_o})$ and the eigenvalues of $(X^\top_{i,\gamma_o}X_{i,\gamma_o})^{-1}$ are the same as these of $X_{i,\gamma_o}(X^\top_{i,\gamma_o}X_{i,\gamma_o})^{-2}X^\top_{i,\gamma_o}$. Therefore,

$$\boldsymbol{\epsilon}_i^\top X_{i,\gamma_o}(X^\top_{i,\gamma_o}X_{i,\gamma_o})^{-2}X^\top_{i,\gamma_o}\boldsymbol{\epsilon}_i \leqslant \boldsymbol{\epsilon}_i^\top I_{|\boldsymbol{\gamma}_o|}\boldsymbol{\epsilon}_i\lambda_{\max}((X^\top_{i,\gamma_o}X_{i,\gamma_o})^{-1})$$
$$\leqslant \boldsymbol{\epsilon}_i^\top I_{|\boldsymbol{\gamma}_o|}\boldsymbol{\epsilon}_i\lambda_-(X^\top_{i,\gamma_o}X_{i,\gamma_o})^{-1} \leqslant \boldsymbol{\epsilon}_i^\top I_{|\boldsymbol{\gamma}_o|}\boldsymbol{\epsilon}_i(nC_5)^{-1}.$$

Suppose for platform $i,\ i \in \{1,\ldots,k\}$, the variace of $\epsilon_{i1},...,\epsilon_{in}$ are the same, the value is denoted as $\text{var}(\boldsymbol{\epsilon}_i)$. Next, we have

$$\mathbb{E}(\boldsymbol{\epsilon}_i^\top X_{i,\gamma_o}(X^\top_{i,\gamma_o}X_{i,\gamma_o})^{-2}X^\top_{i,\gamma_o}\boldsymbol{\epsilon}_i) \leqslant \text{var}(\boldsymbol{\epsilon}_i)|\boldsymbol{\gamma}_o|(nC_5)^{-1},$$

$$\boldsymbol{\epsilon}_i^\top X_{i,\gamma_o}(X^\top_{i,\gamma_o}X_{i,\gamma_o})^{-2}X^\top_{i,\gamma_o}\boldsymbol{\epsilon}_i = O_p(\text{var}(\boldsymbol{\epsilon}_i)|\boldsymbol{\gamma}_o|(nC_5)^{-1}).$$

This entails that

$$\begin{aligned}
&\frac{1+\boldsymbol{y}_i^\top(I_n - X_{i,\boldsymbol{\gamma}_o}U_{i,\boldsymbol{\gamma}_o}^{-1}X_{i,\boldsymbol{\gamma}_o}^\top)\boldsymbol{y}_i}{1+\boldsymbol{y}_i^\top(I_n - P_{i,\boldsymbol{\gamma}_o})\boldsymbol{y}_i} \leqslant 1 + 2\underline{\phi}_n^{-1}\frac{k_n^2|\boldsymbol{\gamma}_i^0| + O_p(\mathrm{var}(\boldsymbol{\epsilon}_i)|\boldsymbol{\gamma}_o|(nC_5)^{-1})}{1+n\mathrm{var}(\boldsymbol{\epsilon}_i)} \\
&\leqslant 1 + 2\underline{\phi}_n^{-1}\frac{k_n^2|\boldsymbol{\gamma}_i^0| + O_p(\mathrm{var}(\boldsymbol{\epsilon}_i)|\boldsymbol{\gamma}_o|(nC_5)^{-1})}{1+n\mathrm{var}(\boldsymbol{\epsilon}_i)} \\
&\leqslant 1 + 2\underline{\phi}_n^{-1}\frac{k_n^2|\boldsymbol{\gamma}_i^0| + O_p(\mathrm{var}(\boldsymbol{\epsilon}_i)|\boldsymbol{\gamma}_o|(nC_5)^{-1})}{n\mathrm{var}(\boldsymbol{\epsilon}_i)} \\
&= 1 + \frac{2k_n^2|\boldsymbol{\gamma}_i^0|}{n\underline{\phi}_n\mathrm{var}(\boldsymbol{\epsilon}_i)}(1 + O_p(\frac{\mathrm{var}(\boldsymbol{\epsilon}_i)|\boldsymbol{\gamma}_o|}{n\varphi_{i,\min}(n)k_n^2|\boldsymbol{\gamma}_i^0|})) \leqslant 1 + \frac{2k_n^2|\boldsymbol{\gamma}_i^0|}{n\underline{\phi}_n\mathrm{var}(\boldsymbol{\epsilon}_i)}(1+o_p(1)).
\end{aligned}$$

Consequently, $T_3^{(i)} \leqslant C_7$ with probability approaching 1, where C_7 is a constant. As $\gamma \subset \gamma_0$,

$$\frac{1+\boldsymbol{y}_i^\top(I_n - P_{i,\boldsymbol{\gamma}_o})\boldsymbol{y}_i}{1+\boldsymbol{y}_i^\top(I_n - P_{i,\boldsymbol{\gamma}})\boldsymbol{y}_i} = 1 + \frac{\boldsymbol{y}_i^\top(P_{i,\boldsymbol{\gamma}} - P_{i,\boldsymbol{\gamma}_o})\boldsymbol{y}_i}{1+\boldsymbol{y}_i^\top(I_n - P_{i,\boldsymbol{\gamma}})\boldsymbol{y}_i} \leqslant 1.$$

Therefore, $T_5^{(i)} \leqslant 0$. Combining all the results above, with probability tending to 1,

$$p_c(\boldsymbol{\gamma}|Z_i)/p_c(\boldsymbol{\gamma}_o|Z_i) \leqslant C_7(1 + nC_6\ \overline{\phi}_n)^{|\boldsymbol{\gamma}_o|/2}.$$

□

Lemma 2.3.3 *If* $\boldsymbol{\gamma} \in S_3$,

$$T_2^{(i)} \geqslant 1/2(|\boldsymbol{\gamma}| - |\boldsymbol{\gamma}_o|)\log(1 + C_3 n^{1-\delta}\underline{\phi}_n)$$

with probability tending to 1. If $\boldsymbol{\gamma} \in S_4$, *and* $\boldsymbol{\gamma}_i^0 \nsubseteq \boldsymbol{\gamma}$, *then*

$$T_2^{(i)} \geqslant -1/2|\boldsymbol{\gamma}_o|\log(1 + C_2 n\overline{\phi}_n)$$

with probability tending to 1.

Proof of Lemma 2.3.3. If $\boldsymbol{\gamma} \in S_3$, then using the result in Lemma 2.2.4, we get $T_2^{(i)} \geqslant 1/2(|\boldsymbol{\gamma}| - s_n)\log(1 + C_3 n^{1-\delta}\underline{\phi}_n)$. If $\boldsymbol{\gamma} \in S_4, \boldsymbol{\gamma}_i^0 \nsubseteq \boldsymbol{\gamma}$, following similar arguments as in the proof of Lemma 2.2.1, we obtain that $T_2^{(i)} \geqslant -1/2|\boldsymbol{\gamma}_o|\log(1 + C_2 n\overline{\phi}_n)$. □

Lemma 2.3.4 *For $\boldsymbol{\gamma} \in S_3$,*

$$T_5^{(i)} \geqslant -1/2(|\boldsymbol{\gamma}| - s_n)\alpha_0 \log p$$

with probability tending to 1, where α_0 is a constant. For $\boldsymbol{\gamma} \in S_4$, $\boldsymbol{\gamma}_i^0 \nsubseteq \boldsymbol{\gamma}$,

$$T_5^{(i)} \geqslant -1/2(n + v_i) \log(1 + C'\psi_n^2)$$

with probability tending to 1, where C' is a constant.

Proof of Lemma 2.3.4. Following the similar arguments as in Lemma 2.2.3 and Lemma 2.2.5, we obtain the same result for $T_5^{(i)}$. □

We are not in position to proceed with the proof of Lemma 2.3.2

Proof of Lemma 2.3.2. Combining the results above, we have $T_3^{(i)} \geqslant 0$, $T_4^{(i)} \leqslant C_8$ with probability tending to 1, where C_8 is a constant. By Lemmas 2.3.3 and 2.3.4, for $\boldsymbol{\gamma} \in S_3$, with probability tending to 1,

$$\begin{aligned} & p_c(\boldsymbol{\gamma}|Z_i)/p_c(\boldsymbol{\gamma}_o|Z_i) \\ \leqslant \;& C_8\Big(\exp(-2^{-1}(|\boldsymbol{\gamma}| - |\boldsymbol{\gamma}_o|)) \log((1 + C_3\rho_n)/p_n^{\alpha_0})\Big). \end{aligned}$$

For $\boldsymbol{\gamma} \in S_4$ and $\boldsymbol{\gamma}_i^0 \nsubseteq \boldsymbol{\gamma}$, with probability tending to 1,

$$\begin{aligned} p_c(\boldsymbol{\gamma}|Z_i)/p_c(\boldsymbol{\gamma}_o|Z_i) \leqslant & C_8 \exp(2^{-1}|\boldsymbol{\gamma}_o| \log(1 + C_2 n\overline{\phi}_n) - \frac{n + v_i}{2} \log(1 + C'\psi_n^2)) \\ \leqslant & C_8(1 + C'\psi_n^2)^{-\frac{n+v_i}{4}}. \end{aligned}$$

□

Proof of Theorem 2.3.1. We consider three scenarios. Let $\hat{C} = \max\{C_7, C_8\}$. First, for all the

overfitting models,

$$\begin{aligned}
&\sum_{\boldsymbol{\gamma}\in S_3} P_C(\boldsymbol{\gamma}|Z_1,...,Z_k)/P_C(\boldsymbol{\gamma}_o|Z_1,...,Z_k) = \sum_{\boldsymbol{\gamma}\in S_3}\prod_{i=1}^{k}\{\frac{p_c(\boldsymbol{\gamma}|Z_i)}{p_c(\boldsymbol{\gamma}_o|Z_i)}\}\frac{p(\boldsymbol{\gamma})}{p(\boldsymbol{\gamma}_o)}\\
\leqslant\ & \hat{C}^k\sum_{\boldsymbol{\gamma}\in S_3}(\frac{1+C_3\rho_n}{p_n^{\alpha_0-\frac{2C}{k}}})^{-2^{-1}k(|\boldsymbol{\gamma}|-s_n)} = \hat{C}^k\sum_{r=1}^{r_n-s_n}\binom{p_n-s_n}{r}(\frac{1+C_3\rho_n}{p_n^{\alpha_0-\frac{2M}{k}}})^{-rk/2}\\
\leqslant\ & \hat{C}^k\sum_{r=1}^{r_n-s_n}\frac{p_n^r}{r!}(\frac{p_n^{\alpha_0-\frac{2M}{k}}}{1+C_3\rho_n})^{rk/2} = \hat{C}^k\sum_{r=1}^{r_n-s_n}\frac{1}{r!}(\frac{p_n^{\alpha_0-\frac{2M-2}{k}}}{1+C_3\rho_n})^{rk/2}\\
\leqslant\ & \hat{C}^k\sum_{r=1}^{r_n-s_n}(\frac{p_n^{\alpha_0-\frac{2M-2}{k}}}{1+C_3\rho_n})^{kr/2} \leqslant \hat{C}^k\frac{(\frac{p_n^{\alpha_0-\frac{2M-2}{k}}}{1+C_3\rho_n})^{k/2}}{1-(\frac{p_n^{\alpha_0-\frac{2M-2}{k}}}{1+C_3\rho_n})^{k/2}} \xrightarrow{p} 0,
\end{aligned}$$

when $n\to\infty$. The last limit result above follows from the assumption that $p_n^{\alpha_0-2(M-1)/k} = o(\rho_n)$. Secondly, we consider the scenario when $\boldsymbol{\gamma}_i\subseteq\boldsymbol{\gamma}\subseteq\boldsymbol{\gamma}_o$ for some $i\in D\subset\{1,...,k\}$. Then by Lemma 2.3.1 and Assumption 2.3.1,

$$\begin{aligned}
\frac{P_C(\boldsymbol{\gamma}|Z_1,...,Z_k)}{P_C(\boldsymbol{\gamma}_o|Z_1,...,Z_k)} &= \prod_{j\in D}\frac{p_c(\boldsymbol{\gamma}|Z_j)}{p_c(\boldsymbol{\gamma}_o|Z_j)}\prod_{j=1,j\notin D}^{k}\{\frac{p_c(\boldsymbol{\gamma}|Z_j)}{p_c(\boldsymbol{\gamma}_o|Z_j)}\}\frac{p(\boldsymbol{\gamma})}{p(\boldsymbol{\gamma}_o)}\\
&\leqslant \hat{C}^{k-1}(1+nC_6\ \overline{\phi}_n)^{(k-1)|\boldsymbol{\gamma}_o|/2}(1+C'\psi_n^2)^{-(n+v_i)/4}\\
&\xrightarrow{p} 0 \text{ when } n\to\infty.
\end{aligned}$$

Lastly, for other model in S_4, $P_C(\boldsymbol{\gamma}|Z_1,...,Z_k)\leqslant P_C(\boldsymbol{\gamma}_i|Z_1,...,Z_k)$. Using the assumption that $r_n\log p_n = o(n\log(1+\psi_n^2))$, $r_n\log(1+n\overline{\phi}_n) = o(n\log(1+C'\psi_n^2))$, we show that

$$\begin{aligned}
&\sum_{\boldsymbol{\gamma}\in S_4} P_C(\boldsymbol{\gamma}|Z_1,...,Z_k)/P_C(\boldsymbol{\gamma}_o|Z_1,...,Z_k)\\
\leqslant\ & \hat{C}^{k-1}p_n^{r_n}(1+nC_6\ \overline{\phi}_n)^{(k-1)|\boldsymbol{\gamma}_o|/2}(1+C'\psi_n^2)^{-(n+v_i)/4}.
\end{aligned}$$

The right-hand side converges to 0 in probability, as $n\to\infty$. □

Theorem 2.3.1 implies

$$\inf_{c_{i,1},...,c_{i,p_n}\in[\underline{\phi}_n,\overline{\phi}_n],i\in\{1,...,k\}} P_C(\boldsymbol{\gamma}_o|Z_1,Z_2,...,Z_k)\xrightarrow{p}1,$$

which establishes the strong model selection consistency for the union model.

2.4 Gibbs sampling and simulation

Initialize: $\boldsymbol{\gamma}^{(0)}$ where the superscript (t) means the tth iteration

for $t = 1, 2, \ldots$ **do**

$$\gamma_{\cdot 1}^{(t)} \sim p(\gamma_{\cdot 1}|\gamma_{\cdot 2} = \gamma_{\cdot 2}^{(t-1)}, \gamma_{\cdot 3} = \gamma_{\cdot 3}^{(t-1)}, ..., \gamma_{\cdot p} = \gamma_{\cdot p}^{(t-1)})$$

$$\gamma_{\cdot 2}^{(t)} \sim p(\gamma_{\cdot 2}|\gamma_{\cdot 1} = \gamma_{\cdot 1}^{(t)}, \gamma_{\cdot 3} = \gamma_{\cdot 3}^{(t-1)}, ..., \gamma_{\cdot p} = \gamma_{\cdot p}^{(t-1)})$$

$$\vdots$$

$$\gamma_{\cdot p}^{(t)} \sim p(\gamma_{\cdot p}|\gamma_{\cdot 1} = \gamma_{\cdot 1}^{(t)}, \gamma_{\cdot 2} = \gamma_{\cdot 2}^{(t)}, ..., \gamma_{\cdot p-1} = \gamma_{\cdot p-1}^{(t)})$$

end for

noend 1: Gibbs sampler

We use $P_C(\boldsymbol{\gamma}|Z_1, ..., Z_k)$ as the model selection criterion and search for the optimal model that maximizes $P_C(\boldsymbol{\gamma}|Z_1, ..., Z_k)$. After incorporating the normalizing constant, $P_C(\boldsymbol{\gamma}|Z_1, ..., Z_k)$ represents a valid probability distribution over the model space. To explore this probability distribution and select the model with the highest probability, we employ Gibbs sampling procedure. The algorithm is outlined in the table above and the conditional probabilities used in the Gibbs sampling procedure are:

$$p(\gamma_{\cdot j}|\gamma_{\cdot 1}, ..., \gamma_{\cdot j-1}, \gamma_{\cdot j+1}, ..., \gamma_{\cdot p}) = \frac{P_C(\gamma_{\cdot 1}, ..., \gamma_{\cdot p}|Z_1, ..., Z_k)}{P_C(\gamma_{\cdot,-j}, \gamma_{\cdot,j} = 0|Z_1, ..., Z_k) + P_C(\gamma_{\cdot,-j}, \gamma_{\cdot,j} = 1|Z_1, ..., Z_k)},$$

where $\gamma_{\cdot,-j} = (\gamma_{\cdot 1}, ..., \gamma_{\cdot j-1}, \gamma_{\cdot j+1}..., \gamma_{\cdot p})$.

In the numeric studies, we compare the proposed Bayesian model selection method with the group LASSO algorithm implemented by `R` package `FusionLearn` (Gao and Zhong (2019)). Additionally, we evaluate the performance of the Bayesian method applied separately to each dataset,

as well as two alternative approaches: the union method, which combines all predictors selected from the four marginal models, and the intersection method, which selects only the predictors common across the four marginal models. These comparisons provide a comprehensive assessment of the proposed Bayesian method's effectiveness relative to established techniques and alternative strategies.

We simulate data from k different platforms. We let $k = 4$, and generate an $n \times 4$ response matrix $(\boldsymbol{y}_1, \boldsymbol{y}_2, \boldsymbol{y}_3, \boldsymbol{y}_4)$. The row vector $(y_{i,1}, y_{i,2}, y_{i,3}, y_{i,4})$ follows a multivariate heavy-tailed distribution with means equal to $(X_{i1}^\top \boldsymbol{\beta}_1^0, X_{i2}^\top \boldsymbol{\beta}_2^0, X_{i3}^\top \boldsymbol{\beta}_3^0, X_{i4}^\top \boldsymbol{\beta}_4^0)$, variances equal to one and covariances of 0.6 or 0.8. To demonstrate the robust performance in handling heavy-tailed distributions, we selected the multivariate t-distribution, which exhibits heavier tails compared to the typical sub-exponential distribution. We choose these two levels of covariance to represent both medium and high correlation scenarios across the platforms. The row of design matrix X_i is generated from a p-dimensional multivariate normal distribution with mean equal to zero, variances equal to one, and covariances equal to 0.2 or 0.5. These two covariance levels represent low and medium correlations among the predictors. The consistency of LASSO method (Zhao and Yu (2006)) requires the irrepresentable condition, which will be violated under high correlations among the predictors. The consistency of our proposed method also requires the eigenvalues of the design matrix to be bounded away from zero. So the performance of our method is also affected if the correlations are too high. In addition, we consider the situation of large correlations within blocks of predictors. We partition the predictors into blocks of 20 members and all predictors within each block have correlations 0.8, whereas predictors from different blocks have zero correlations.

The sample size n is set to 500 or 800 and the number of predictors p is set to 1000. The true model size s_n is set to 10 and the maximum model size r_n is set to 20. The nonzero parameters

Table 2.1: The summary of different simulation settings.

Case	corr(x)	corr(y)	n	Case	corr(x)	corr(y)	n
A	0.2	0.8	500	D	0.2	0.8	800
B	0.5	0.8	500	E	0.8	0.8	500
C	0.2	0.6	800	F	0.8	0.8	800

The symbol corr(x) denotes the correlation among the covariates; the symbol corr(y) denotes the correlation among the responses from different platforms. For Case E and F, corr(x) denotes the blockwise correlation among the covariates.

are simulated from $\mathcal{U}(0.5, 1)$. We choose this range because the lower bound is sufficiently away from zero so that the methods are able to detect these nonzero parameters. We choose the upper bounds to not be too high so that different methods can compete in terms of their performance. If the nonzero parameters are set too high, all the competing methods would exhibit similarly good performance. In total, we consider six different simulation settings which are summarized in Table 2.1. Under each setting, we simulate 100 datasets.

To initialize the Markov chain at a starting model, we use a ranked t-test to sort the predictors and select the highest r_n predictors in the initial model. We set the hyper-parameters $v_i = 2$, $c_{1j} = c_{2j} = 4$, and $c_{3j} = c_{4j} = 10$. After initializing the Markov chain, we use Gibbs sampling to update the models and calculate the composite posterior probability of each model. For one trial, we simulate 50000 runs, burn in the first 20000 runs and record all the models selected in each run. The final model is the one that occurs most frequently. The results of 100 simulations are provided in Figure 2.1. We use positive selection rate (psr) and false discovery rate (fdr) to

compare the performance of the methods:

$$psr = \frac{\sum_{j=1}^{p_n} I(\hat{\gamma}_{.j} \neq 0) I(\gamma_{oj} \neq 0)}{\sum_{j=1}^{p_n} I(\gamma_{oj} \neq 0)},$$
$$fdr = \frac{\sum_{j=1}^{p_n} I(\hat{\gamma}_{.j} \neq 0) I(\gamma_{oj} = 0)}{\sum_{j=1}^{p_n} I(\hat{\gamma}_{.j} \neq 0)},$$

where $\hat{\gamma}_{.j}$ and γ_{oj} denote the indicators of the jth predictor for the selected model and the true model, respectively. From Figure 2.1, we observe that all the methods improve their performance with the increase in sample size. The proposed Bayesian method greatly outperforms the marginal methods, the intersection method, and the union method. The proposed Bayesian method has comparable performance over the group LASSO method, with the Bayesian method exhibiting a higher positive selection rate but also a higher false discovery rate. We evaluate the performance of each method when the correlations among the predictors increase. The group LASSO method's performance deteriorates as the correlation among the predictors increases, whereas the proposed Bayesian method is less affected. We examine the performance of the two methods in the presence of high correlations within blocks of the predictors. The proposed method continues to outperform than the LASSO method. We demonstrate the results of all methods when the correlation among different platforms increases. Both the proposed Bayesian method and the group LASSO maintain similar satisfactory performance in the presence of high correlations among the platforms.

We further evaluate the performance of the proposed method for the union support recovery when the true models for four platforms are not identical. The union of four true models contains a total of twelve predictors and the difference between the union model and each of the true models for different platforms is a set of two predictors. We compare the performance of the proposed Bayesian method, the group LASSO and the union method which combines the selected models from each marginal model. The number of predictors is set to $p = 600$. The nonzero parameters

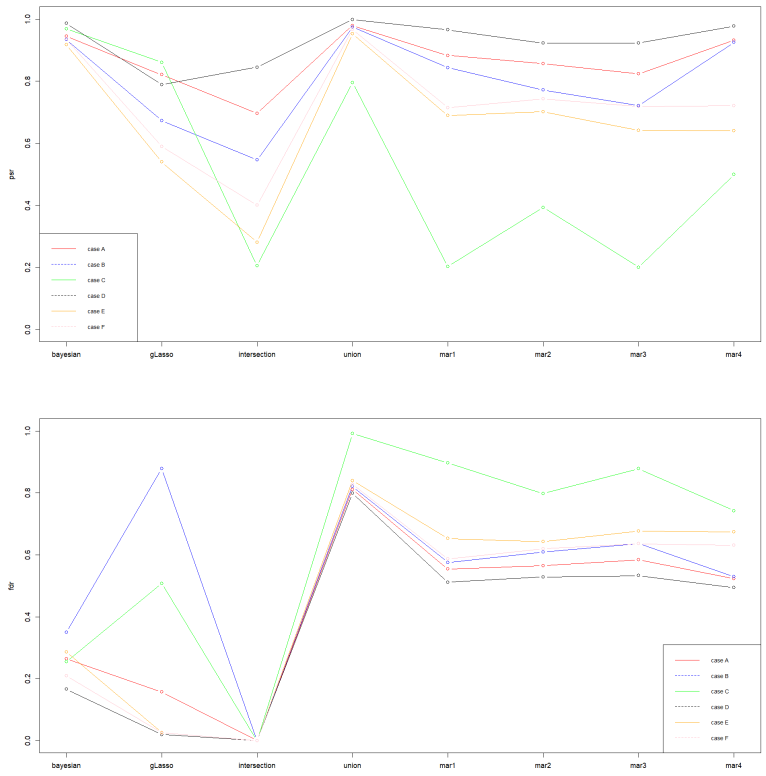

Figure 2.1: The model selection performance of eight different methods. There are six different simulation settings: Case A, B, C, D, E and F. Symbol 'mar xi" denotes the result when the Bayesian method is applied on the ith marginal platform; "intersection" denotes the method that selects the commonly nonzero parameters of four marginal models; "union" denotes the method that combines all nonzero parameters of four marginal models.

are simulated from $\mathcal{U}(1,2)$. The sparsity pattern of union support recovery differs across different platforms. The signal of nonzero parameters may appear on one of the platforms. So compared to the simulation in the setting that nonzero parameters show up in all of the platforms, achieving similar performance requires larger nonzero parameters in the union support recovery simulation. So, we select a larger lower bound of the uniform distribution. The result is summarized in Figure 2.2. It is observed that the proposed Bayesian method performs best among all the methods, and

it achieves a high positive selection rate and a low false discovery rate.

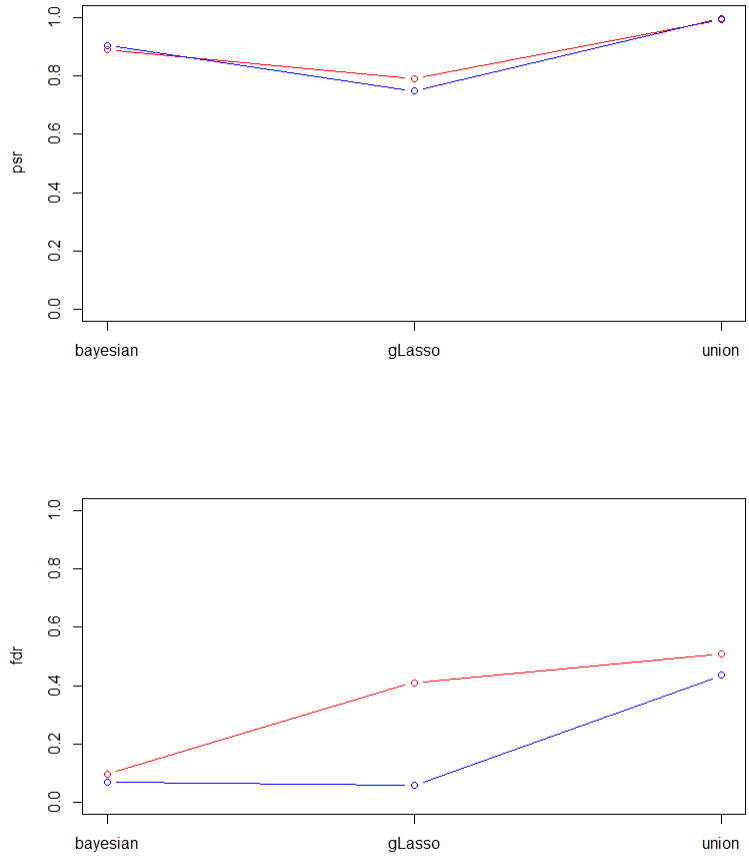

Figure 2.2: The union support recovery performance of three different methods. Symbol "Bayesian" and "Glasso" represent the proposed Bayesian method and the group LASSO method. Symbol "union" denotes the method that combines all nonzero parameters of four marginal models. The simulations are conducted with $\text{var}(y) = 1$, $\text{corr}(x) = 0.2$, and $\text{corr}(y) = 0.6$. Red line and blue line represent the setting of $n = 200$ and $n = 500$, respectively.

2.5 Real data analysis

We consider a data integration problem in which a collection of worldwide equity indices, North American bond indices, and major stocks indices are included to model a panel of three different

financial indices such as S&P 500 Index, the Dow Jones Index, and the VIX index. The analysis aims to select a subset of covariates which are true predictors for at least one of the indices in the panel. The data set includes three-year performance of the S&P 500 Index (GSPC), Dow Jones Index (DJI) and VIX index (VIX) and the 46 covariates. To reduce the autocorrelation and satisfy the independent assumption, we select the index values every three days between September 2013 to June 2016. For each index, the value used in the analysis is the logarithmic returem calculated as the loganithm of today's value divided by yesterday's value multiplied by 100. The transformed values of the GSPC, DJI and VIX indices are not autocorrelated at the 5% significance level. We use cross-validation to verify the performance of our proposed method and the group LASSO method. The training set contains 232 records for all indexes. For each response, we built a linear regression model based on the set of 46 covariates and use our method and group LASSO method to select the predictors. The subset selected by group LASSO contains covariates 'DJA', 'NYA', 'OEX', and 'RUI'; and the subset selected by the proposed Bayesian method contains 'DJT', 'DJU', 'DJA', 'OEX', 'XMI', 'RUI'. The two methods identify three covariates in common. We used the models built from the training data set to predict the response variables on a validation data set of 232 records, which consists of indexes values two days after each observation in the training set. The mean squared prediction error for each platform (MSE) is defined as $\text{MSE}_i = (1/n)\{(y_{i1} - \hat{y}_{i1})^2 + \cdots + (y_{in} - \hat{y}_{in})^2\}$. As shown by Table 2.2, the models built by the Bayesian method have lower prediction errors than the models built by the group LASSO for all three responses, which demonstrating the advantage of the proposed method in model selection and prediction.

Table 2.2: The Mean Squared Prediction Error (MSE) of each competing method on the financial data.

indices	group LASSO	Bayesian	full model
VIX	17.0710	**16.8442**	24.2533
GSPC	0.0006	**0.0005**	0.0008
DJI	0.0221	**0.0006**	0.0011

The full model is the model contains all of the covariates.

2.6 Summary

We propose a Bayesian model selection method tailored for high-dimensional model selection when data are collected from multiple platforms. The method is shown to achieve model selection consistency for a broad class of error distributions, including sub-Gaussian and sub-exponential errors. Moreover, we demonstrate that the method is consistent for union support recovery even when the true models across different platforms are not identical. This significantly broadens the scope of applications, as many real-world tasks involve heterogeneous models across platforms. Simulations and real data analyses illustrate that the proposed data integration method outperforms competing methods in model selection performance. Furthermore, the models built using this integrated approach exhibit higher predictive power compared to models derived from individual data sources.

The current work focuses on integrating linear models from different tasks. As future research, the method could be extended to handle more complex models from multiple platforms, such

as generalized linear models and survival models. A key challenge in such extensions lies in deriving a closed-form expression for the composite posterior distribution across different models. Investigating the use of Laplace approximations for composite posterior probabilities to combine multiple nonlinear models is a promising direction for further exploration.

3 Bayesian Grouping-Gibbs Sampling Estimation of High-dimensional Linear Model with Non-sparsity

3.1 Model setup

Notations:

We use $\boldsymbol{X}$ to denote a $n \times p$ design matrix. We denote the i-th row and j-th column of $\boldsymbol{X}$ by $\boldsymbol{x}_i, i = 1, \ldots, n$ and $\boldsymbol{x}_{(j)}, j = 1, \ldots, p$, respectively. Let $\mathcal{C}$ be a subset of $\{1, \ldots, p\}$, and $|\mathcal{C}|$ is the length of $\mathcal{C}$. $\boldsymbol{X}_{\mathcal{C}}$ denotes a $n \times |\mathcal{C}|$ submatrix of $\boldsymbol{X}$, containing the columns of $\boldsymbol{X}$ indexed by $\mathcal{C}$. Moreover, we use $\boldsymbol{\beta}_{\mathcal{C}}$ to denote the corresponding regression coefficients associated with $\boldsymbol{X}_{\mathcal{C}}$. Let $I_{\{\cdot\}}$ and $\mathbf{1}_d$ be an indicator function and a $d \times 1$ vector with all elements 1, respectively. Consider a linear regression model,

$$y_i = \boldsymbol{x}_i^\top \boldsymbol{\beta} + \epsilon_i, \ i = 1, \ldots, n,$$

where $y_i \in \mathbb{R}$ are response observations, $\boldsymbol{x}_i \in \mathbb{R}^p$ are predictors, $\epsilon_i \in \mathbb{R}$ are random errors with mean zero and unknown constant variance, $\boldsymbol{\beta} \in \mathbb{R}^p$ is an unknown regression vector that needs to be estimated. Unlike previous works that usually impose sparse assumptions on $\boldsymbol{\beta}$, in this dissertation, we consider the non-sparse situation where $\boldsymbol{\beta}$ is composed of both large and small components, but possibly not zeros—that is, the majority, if not all, of the regression coefficients have non-zero values. To tackle the ill-posed problem, we assume that $\boldsymbol{\beta}$ may carry a grouping structure. Let

$\Pi_k = \{\mathcal{C}_1, \mathcal{C}_2, \ldots, \mathcal{C}_k\}$ be a partition of $\{1, 2, \ldots, p\}$, where k is the number of subgroups, $k < n$. Denote $\boldsymbol{\beta} = (\boldsymbol{\beta}_{\mathcal{C}_1}^\top, \boldsymbol{\beta}_{\mathcal{C}_2}^\top, \ldots, \boldsymbol{\beta}_{\mathcal{C}_k}^\top)^\top$, $\boldsymbol{\beta}_{\mathcal{C}_\ell} = \{\beta_j, j \in \mathcal{C}_\ell\}$ for $\ell = 1, 2, \ldots, k$. Within a subgroup, say $\mathcal{C}_\ell$, the components of $\boldsymbol{\beta}_{\mathcal{C}_\ell}$ are similar or identical. Notably, the number of non-zero elements of $\boldsymbol{\beta}$ has a magnitude comparable to the sample size n or larger than n, but the number of subgroups does not exceed n. Furthermore, denote $\boldsymbol{X} = (\boldsymbol{X}_{\mathcal{C}_1}, \boldsymbol{X}_{\mathcal{C}_2}, \ldots, \boldsymbol{X}_{\mathcal{C}_k})$. Most sparse regularized methods work well on weakly dependent predictors. In our work, the strength of correlation among predictors follows the framework of Witten et al. (2014), assuming that $\boldsymbol{x}_{(j)}$ and $\boldsymbol{x}_{(j')}$ are highly correlated if $j, j' \in \mathcal{C}_\ell$, but weakly correlated if $j \in \mathcal{C}_\ell, j' \in \mathcal{C}_{\ell'}, \ell \neq \ell'$.

Compared to the sparsity assumption, where only a small subset of predictors significantly influences the responses, the group structure allows for the inclusion of more predictors by clustering weakly influential predictors together and treating them as a group to demonstrate their collective effect. If the true underlying model considers all predictors important, the group structure may be more suitable than the sparsity assumption, as it includes more predictors and better approximates the real model. In cases where most predictors are unnecessary, the group structure can still be effective by grouping all relevant predictors into different subgroups while placing the remaining non-influential predictors into another subgroup. The primary drawback, however, is the increased computational load and model complexity compared to the sparsity assumption.

3.1.1 Prior specification

We propose to provide grouping prior information via a random indicator vector, $\boldsymbol{v} = (v_1, \ldots, v_p)^\top$, enhancing the representational capacity of the priors in terms of grouping predictors, where $v_j = \ell$ means that β_j belongs to the ℓ-th subgroup, $j = 1, \ldots, p, \ell = 1, \ldots, k$, where k is pre-specified and

smaller than n. The priors on β_j given v_j is specified as

$$\beta_j|v_j = \ell \sim N(\mu_\ell, \tau_n^2) \text{ with } P(v_j = \ell) = d_{j,\ell} \text{ and } \mu_\ell \sim N(0, c^2),$$

where μ_ℓ refers the ℓ-th grouping mean, $d_{j,\ell} \geqslant 0$, and $\sum_{\ell=1}^{k} d_{j,\ell} = 1$. Note that $P(j \in \mathcal{C}_\ell) = P(v_j = \ell) = d_{j,\ell}$. In practice, β_j may belong to any group equally if there is no any prior information. Therefore, we further assume $d_{j,\ell} = 1/k$. The regression coefficients might be identical or similar within each group, whereas they are different between groups. We assume both the grouping variance of each group τ_n^2 and the variance of μ_ℓ satisfy $\tau_n^2 \to 0,\ c^2 \to 0$. We remark that the order of τ_n^2 is flexible. In Narisetty et al. (2019), $\tau_n^2 = 1/p$. Herein, we take $\tau_n^2 = 1/(n^a), 0 < a < 1$. More assumptions on τ_n^2, c^2 are given in Assumption 3.2.1 in Section 3.2. As for the random error, we assume $\epsilon_i \sim N(0, \sigma^2), i = 1, \ldots, n$, and the conjugate prior for the normal variance σ^2 is inverse Gamma (IGamma), i.e., $\sigma^2 \sim \text{IGamma}(\alpha, \beta)$, where $\alpha > 0$ and $\beta > 0$. Following the process outlined in Li and Zhang (2010), when $\alpha \to 0$ and $\boldsymbol{\beta} \to 0$, it simplifies to a flat prior distribution, which is adopted in this paper.

3.1.2 Grouping-Gibbs sampler

Given $\boldsymbol{X}, \boldsymbol{\beta}, \sigma^2, \boldsymbol{\mu}, \boldsymbol{v}$, we have

$$y_i|\boldsymbol{X}, \boldsymbol{\beta}, \sigma^2, \boldsymbol{\mu}, \boldsymbol{v} \sim N(\boldsymbol{x}_i^\top \boldsymbol{\beta}, \sigma^2), i = 1, \ldots, n.$$

The joint distribution of $\boldsymbol{\beta}, \boldsymbol{v}, \boldsymbol{\mu}, \boldsymbol{y}, \sigma^2|\boldsymbol{X}$ is given by

$$\begin{aligned}\boldsymbol{\beta}, \boldsymbol{v}, \boldsymbol{\mu}, \boldsymbol{y}, \sigma^2|\boldsymbol{X} \propto & \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(y_i - \boldsymbol{x}_i^\top \boldsymbol{\beta})^2\right) \prod_{j=1}^{p} \prod_{\ell=1}^{k} \left\{ d_\ell \frac{1}{\sqrt{2\pi\tau_n^2}} \exp\left(-\frac{1}{2\tau_n^2}(\beta_j - \mu_\ell)^2\right) \right\}^{I_{\{v_j=\ell\}}} \\ & \times \prod_{\ell=1}^{k} \frac{1}{\sqrt{2\pi c^2}} \exp\left(-\frac{1}{2c^2}\mu_\ell^2\right) \frac{1}{\sigma^2}. \end{aligned} \tag{3.1}$$

By some computation, we obtain the posterior distributions of $\boldsymbol{\beta}_{\mathcal{C}_\ell}, \mu_\ell$, v_j and σ^2 theoretically for $\ell = 1, \ldots, k$, $j = 1, \ldots, p$ as follows.

(i) Conditional distribution of $\boldsymbol{\beta}_{\mathcal{C}_\ell}, \ell = 1, \ldots, k$ for giving $\boldsymbol{v}, \boldsymbol{\mu}, \boldsymbol{y}, \sigma^2, \boldsymbol{X}$:

$$\boldsymbol{\beta}_{\mathcal{C}_\ell}|\boldsymbol{v}, \boldsymbol{\mu}, \boldsymbol{y}, \sigma^2, \boldsymbol{X} \propto \exp\left\{ -\frac{1}{2}\left(\boldsymbol{\beta}_{\mathcal{C}_\ell}^\top \boldsymbol{\Sigma}_\ell^{-1} \boldsymbol{\beta}_{\mathcal{C}_\ell} - 2(\sigma^{-2}(\boldsymbol{y}^\top \boldsymbol{X}_{\mathcal{C}_\ell} - \sum_{\ell' \neq \ell} \boldsymbol{\beta}_{\mathcal{C}_{\ell'}}^\top \boldsymbol{X}_{\mathcal{C}_{\ell'}}^\top \boldsymbol{X}_{\mathcal{C}_\ell}) + \tau_n^{-2}\mu_\ell \mathbf{1}_{|\mathcal{C}_\ell|})\boldsymbol{\beta}_{\mathcal{C}_\ell} \right) \right\},$$

where $\boldsymbol{\Sigma}_\ell^{-1} = \sigma^{-2}\boldsymbol{X}_{\mathcal{C}_\ell}^\top \boldsymbol{X}_{\mathcal{C}_\ell} + \tau_n^{-2} I_{|\mathcal{C}_\ell|}$. Thus, the conditional distribution of $\boldsymbol{\beta}_{\mathcal{C}_\ell}$ for giving $\boldsymbol{v}, \boldsymbol{\mu}, \boldsymbol{y}, \sigma^2, \boldsymbol{X}$ follows a $|\mathcal{C}_\ell|$-multivariate normal distribution, i.e.,

$$\boldsymbol{\beta}_{\mathcal{C}_\ell}|\boldsymbol{v}, \boldsymbol{\mu}, \boldsymbol{y}, \sigma^2, \boldsymbol{X} \sim N(\boldsymbol{m}_\ell, \boldsymbol{\Sigma}_\ell), \tag{3.2}$$

where the grouping mean and covariance are respectively given by

$$\boldsymbol{m}_\ell = \boldsymbol{\Sigma}_\ell(\sigma^{-2}(\boldsymbol{X}_{\mathcal{C}_\ell}^\top \boldsymbol{y} - \boldsymbol{X}_{\mathcal{C}_\ell}^\top \sum_{\ell' \neq \ell} \boldsymbol{X}_{\mathcal{C}_{\ell'}} \boldsymbol{\beta}_{\mathcal{C}_{\ell'}}) + \tau_n^{-2}\mu_\ell \mathbf{1}_{|\mathcal{C}_\ell|}),\ \boldsymbol{\Sigma}_\ell = (\sigma^{-2}\boldsymbol{X}_{\mathcal{C}_\ell}^\top \boldsymbol{X}_{\mathcal{C}_\ell} + \tau_n^{-2} I_{|\mathcal{C}_\ell|})^{-1}.$$

The posterior of $\boldsymbol{\beta}_{\mathcal{C}_\ell}$ follows a multivariate normal distribution, enabling the grouping Gibbs-sampling of $\boldsymbol{\beta}_{\mathcal{C}_\ell}$. Clearly, the maximum variances of $\beta_j s, j \in \mathcal{C}_\ell$ are proportional to τ_n^2, which ensures that the posterior variance approaches zero.

(ii) Conditional distribution of $\mu_\ell, \ell = 1, \ldots, k$ for giving $\boldsymbol{\beta}, \boldsymbol{v}, \boldsymbol{y}, \sigma^2, \boldsymbol{X}$:

$$\mu_\ell|\boldsymbol{\beta}, \boldsymbol{v}, \boldsymbol{y}, \sigma^2, \boldsymbol{X} \propto \exp\left\{ -\frac{1}{2}\left((\tau_n^{-2}|\mathcal{C}_\ell| + c^{-2})\mu_\ell^2 - 2\tau_n^{-2} \sum_{j \in \mathcal{C}_\ell} \beta_j \mu_\ell \right) \right\}.$$

Then, the conditional distribution of μ_ℓ for giving $\boldsymbol{\beta}, \boldsymbol{v}, \boldsymbol{y}, \sigma^2, \boldsymbol{X}$ follows a normal distribution, i.e.,

$$\mu_\ell|\boldsymbol{\beta}, \boldsymbol{v}, \boldsymbol{y}, \sigma^2, \boldsymbol{X} \sim N\left(\frac{c^2}{c^2|\mathcal{C}_\ell| + \tau_n^2} \sum_{j \in \mathcal{C}_\ell} \beta_j, \frac{\tau_n^2 c^2}{c^2|\mathcal{C}_\ell| + \tau_n^2} \right). \tag{3.3}$$

In the ℓ-th subgroup, the posterior mean of μ_ℓ is proportional to $\frac{1}{|\mathcal{C}_\ell|}\sum_{j \in \mathcal{C}_\ell} \beta_j$ if the ratio of τ_n^2/c^2 tends to zero. The posterior variance of μ_ℓ is equal to $(|\mathcal{C}_\ell|/\tau_n^2 + 1/c^2)^{-1}$, which approaches to zero, guaranteeing that the grouping mean μ_ℓ is equal to the $\frac{1}{|\mathcal{C}_\ell|}\sum_{j \in \mathcal{C}_\ell} \beta_j$ with large probability.

(iii) Conditional distribution of $v_j, j = 1, \ldots, p$ for giving $\boldsymbol{\beta}, \boldsymbol{\mu}, \boldsymbol{y}, \sigma^2, \boldsymbol{X}$ follows a multi-nominal distribution, i.e.,

$$v_j|\boldsymbol{\beta}, \boldsymbol{\mu}, \boldsymbol{y}, \sigma^2, \boldsymbol{X} \sim \text{multinomial}(k, \boldsymbol{p}_j), \tag{3.4}$$

where $\boldsymbol{p}_j = (p_{j,1}, \ldots, p_{j,k})^\top$ with

$$p_{j,\ell} = \frac{\exp\{-(2\tau_n^2)^{-1}(\beta_j - \mu_\ell)^2\}}{\sum_{\ell=1}^{k} \exp\{-(2\tau_n^2)^{-1}(\beta_j - \mu_\ell)^2\}}, \ell = 1, \ldots, k.$$

In the ℓ-th subgroup, the posterior probability of $v_j = \ell$ is proportional to the Euclidean distance of β_j and μ_ℓ for $j \in \mathcal{C}_\ell$. When τ_n^2 is given, the closer of β_j and μ_ℓ, the higher probability of $v_j = \ell$.
(iv) Conditional distribution of σ^2 for giving $\boldsymbol{\beta}, \boldsymbol{y}, \boldsymbol{X}$:

$$\sigma^2|\boldsymbol{\beta}, \boldsymbol{y}, \boldsymbol{X} \propto \frac{1}{\sigma^{n+2}} \exp\left\{-\frac{(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^\top(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})}{2\sigma^2}\right\}.$$

Then, the posterior on σ^2 is given by $\sigma^2|\boldsymbol{\beta}, \boldsymbol{y}, \boldsymbol{X} \sim \text{IG}(n/2, (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^\top(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})/2)$. The scale parameter is positively related to the mean squared error of residuals. When n is given, the residual variance is more likely centered at zero as the response $\boldsymbol{y}$ is well fitted.

3.1.3 Grouping-Gibbs algorithm

Narisetty et al. (2019) proposed a Skinny Gibbs algorithm under a sparse assumption for the case of large p, which is a simple yet effective modification of the Gibbs sampler. Herein, we develop the grouping-Gibbs algorithm, an enhanced modification of the Gibbs or Skinny Gibbs sampler, designed to address large p with non-sparsity. In each Gibbs sampling iteration, $\boldsymbol{\beta}$ is partitioned into k different segments, $\boldsymbol{\beta}_{\mathcal{C}_1}, \boldsymbol{\beta}_{\mathcal{C}_2}, \ldots, \boldsymbol{\beta}_{\mathcal{C}_k}$, where k is pre-specified. We use a index indicator variable to express the group to which each component of $\boldsymbol{\beta}$ belongs to, i.e., $v_j = \ell, j = 1, \ldots, p; \ell = 1, \ldots, k$, implying that β_j belongs to the ℓ-th group. We limit k to be less

than n, but it can vary with n. Therefore, the components of $\boldsymbol{\beta}_{\mathcal{C}_\ell}$ are sampled from a multi-normal distribution with a common mean and variance. Specifically, the grouping-Gibbs algorithm is detailed as follows.

- ***Step 1***: Initialization: the index indicator variables $v_j^{(0)}, j = 1, \ldots, p$, are sampled from 1 to k with equally probabilities. Then denote $\mathcal{C}_\ell^{(0)} = \{j, v_j^{(0)} = \ell\}$ for $\ell = 1, \ldots, k, j = 1, ..., p$, and the initial partition of $\boldsymbol{X}^{(0)} = (\boldsymbol{X}_{\mathcal{C}_1^{(0)}}, ..., \boldsymbol{X}_{\mathcal{C}_k^{(0)}})$ and $\boldsymbol{\beta}^{(0)} = (\boldsymbol{\beta}_{\mathcal{C}_1^{(0)}}^{(0)\top}, \ldots, \boldsymbol{\beta}_{\mathcal{C}_k^{(0)}}^{(0)\top})^\top$. The grouping means $\{\mu_1^{(0)}, \mu_2^{(0)}, \ldots, \mu_k^{(0)}\}$ are i.i.d. generated from $N(0, c^2)$. The components of $\boldsymbol{\beta}$ corresponding to $v_j^{(0)} = \ell$, are i.i.d. sampled from the normal distribution, i.e., $\beta_j^{(0)} | v_j^{(0)} = \ell \sim N(\mu_\ell^{(0)}, \tau_n^2)$, for $j = 1, ..., p, \ell = 1, ..., k$, where k, c^2, σ^2, τ_n^2 are pre-specified.

- ***Step 2***: Given the m-th sampling results, $\boldsymbol{v}^{(m)}, \mathcal{C}_\ell^{(m)}, \mu_\ell^{(m)}, \boldsymbol{X}_{\mathcal{C}_\ell^{(m)}}, \boldsymbol{\beta}_{\mathcal{C}_\ell}^{(m)}, \ell = 1, \ldots, k$, generate $\boldsymbol{\beta}_{\mathcal{C}_\ell}^{(m+1)}$ for $(m+1)$-th iteration,

$$\boldsymbol{\beta}_{\mathcal{C}_\ell^{(m)}}^{(m+1)} | \mathcal{C}_\ell^{(m)}, \mu_\ell^{(m)}, \boldsymbol{X}_{\mathcal{C}_\ell^{(m)}}, \boldsymbol{\beta}_{\mathcal{C}_\ell^{(m)}}^{(m)} \sim N(\boldsymbol{m}_\ell^{(m)}, \boldsymbol{\Sigma}_\ell^{(m)}), \ell = 1, ..., k, \tag{3.5}$$

where

$$\boldsymbol{\Sigma}_\ell^{(m)} = (\sigma^{-2} \boldsymbol{X}_{\mathcal{C}_\ell^{(m)}}^\top \boldsymbol{X}_{\mathcal{C}_\ell^{(m)}} + \tau_n^{-2} I_{|\mathcal{C}_\ell^{(m)}|})^{-1},$$

$$\boldsymbol{m}_\ell^{(m)} = \boldsymbol{\Sigma}_\ell^{(m)} (\sigma^{-2} (\boldsymbol{X}_{\mathcal{C}_\ell^{(m)}}^\top \boldsymbol{y} - \boldsymbol{X}_{\mathcal{C}_\ell^{(m)}}^\top \sum_{\ell' \neq \ell} \boldsymbol{X}_{\mathcal{C}_{\ell'}^{(m)}} \boldsymbol{\beta}_{\mathcal{C}_{\ell'}^{(m)}}^{(m)}) + \tau_n^{-2} \mu_\ell^{(m)} \mathbf{1}_{|\mathcal{C}_\ell^{(m)}|}).$$

- ***Step 3***: Given the m-th sampling results, $\boldsymbol{v}^{(m)}, \mathcal{C}_\ell^{(m)}, \mu_\ell^{(m)}, \boldsymbol{X}_{\mathcal{C}_\ell^{(m)}}$, and the $(m+1)$-th samplers $\boldsymbol{\beta}_{\mathcal{C}_\ell^{(m)}}^{(m+1)}, \ell = 1, \ldots, k$, generate $\mu_\ell^{(m+1)}, \ell = 1, ..., k$,

$$\mu_\ell^{(m+1)} | \boldsymbol{\beta}_{\mathcal{C}_\ell}^{(m+1)}, \mathcal{C}_\ell^{(m)}, \boldsymbol{y}, \boldsymbol{X}_{\mathcal{C}_\ell^{(m)}} \sim N(\frac{c^2}{c^2 |\mathcal{C}_\ell^{(m)}| + \tau_n^2} \sum_{j \in \mathcal{C}_\ell^{(m)}} \beta_j^{(m+1)}, \frac{\tau_n^2 c^2}{c^2 |\mathcal{C}_\ell^{(m)}| + \tau_n^2}). \tag{3.6}$$

- ***Step 4***: Given the m-th sampling results $\boldsymbol{v}^{(m)}$ and the $(m+1)$-th samplers $\mu_\ell^{(m+1)}$, $\boldsymbol{\beta}^{(m+1)}, \ell =$

$1, \ldots, k$, generate $v_j^{(m+1)}, j = 1, ..., p$ by

$$v_j^{(m+1)} | \boldsymbol{\beta}^{(m+1)}, \mu_\ell^{(m+1)} \sim \text{multinomial}(k, \boldsymbol{p}_j^{(m)}), \tag{3.7}$$

where $\boldsymbol{p}_j^{(m)} = (p_{j,1}^{(m)}, \ldots, p_{j,k}^{(m)})^\top$ with

$$p_{j,\ell}^{(m)} = \frac{\exp\{-(2\tau_n^2)^{-1} \sum_{j=1}^{p} \sum_{\ell=1}^{k} (\beta_j^{(m+1)} - \mu_\ell^{(m+1)})^2 I_{\{v_j^{(m)}=\ell\}}\}}{\sum_{\ell=1}^{k} \exp\{-(2\tau_n^2)^{-1} \sum_{j=1}^{p} \sum_{\ell=1}^{k} (\beta_j^{(m+1)} - \mu_\ell^{(m+1)})^2 I_{\{v_j^{(m)}=\ell\}}\}}.$$

Then update $\mathcal{C}_\ell^{(m+1)} = \{j, v_j^{(m+1)} = \ell\}$.

- ***Step 5***: Repeat steps 2-4 for $m = 1, \ldots, 2M$ to generate a sequence of samples of $\boldsymbol{\beta}$, i.e., $\{\boldsymbol{\beta}^{(1)}, \ldots, \boldsymbol{\beta}^{(2M)}\}$, where $2M$ is a pre-specified maximum length of the Gibbs Chains.

Thus, $\hat{\boldsymbol{\beta}}$ can be obtained by taking average of the Gibbs chains with a burn-in of M, i.e.,

$$\hat{\boldsymbol{\beta}} = \frac{1}{M} \sum_{m=M+1}^{2M} \boldsymbol{\beta}^{(m)}. \tag{3.8}$$

One should note that $\boldsymbol{\beta}$ undergoes low-dimensional, grouping-wise sampling, which efficiently reduces the computational load. The computational complexity for each iteration is $n(kp)$, similar to the complexity of the Skinny Gibbs algorithm proposed in Narisetty et al. (2019) that is designed for a sparse $\boldsymbol{\beta}$.

Remark 3.1.1 *Theoretically, we take $\tau_n^2 = 1/(n^a), c^2 = n^{-a'}$ for some $0 < a' < a < 1$ in Assumption 3.2.1 to guarantee asymptotic properties of parameter estimation and model selection (see Section 3.2 for more details). For finite samples, one can also choose other values, such as $\tau_n^2 = 1/p, c^2 = 1/\log(p)$, or others. As for the residual variance σ^2, although we have made an inference on the posterior distribution of σ^2 in Section 3.1.2, the sampling process is inefficient for numerical studies. In order to reduce the computational load, we set σ^2 as a small constant value.*

3.1.4 Selection of the grouping number k

When performing the grouping-Gibbs sampling algorithm, the grouping number k should be pre-specified. Suppose that k^0 is the underlying true grouping number. Intuitively, if $k = k^0 + 1$, then one of the true subgroups is partitioned into two segments. Without loss of generality, if the ℓ-th subgroup is over-segmented, i.e., $\boldsymbol{\beta}_{\mathcal{C}_\ell} = (\boldsymbol{\beta}_{\mathcal{C}_{\ell,1}}^\top, \boldsymbol{\beta}_{\mathcal{C}_{\ell,2}}^\top)^\top$, the posterior distributions of $\boldsymbol{\beta}_{\mathcal{C}_{\ell,1}}$ and $\boldsymbol{\beta}_{\mathcal{C}_{\ell,2}}$ will have similar means and covariance. Therefore, the over-segmentation of $\boldsymbol{\beta}$ may rarely affect the estimation accuracy of $\hat{\boldsymbol{\beta}}$. However, as k increases, the bias of $\hat{\boldsymbol{\beta}}$ may increase. Therefore, k should be chosen within an appropriate range. We use numerical simulations to choose a proper range of k. Given a candidate set of k, i.e., $k \in \{2, \ldots, n\}$, we calculate their mean squared errors of $\boldsymbol{\beta}$ (MSE_β) for all k. The details are given in Section 3.3.1. It is recommended that one may take $k_{\max} = n/4$ for a real problem, where $k_{\max}$ is the maximum grouping number. In addition to specifying an appropriate range of k, one may also be interested in determining the optimal value of k. Herein, we also propose an approach to select k provided that a proper range is given.

- ***Step 1***: Take a rough k value, say, $k = n/4$.

- ***Step 2***: Perform the grouping-Gibbs algorithm in Section 3.1.3, and obtain the parameter estimate, $\hat{\boldsymbol{\beta}}^{[k]}$.

- ***Step 3***: Sort the components of $\hat{\boldsymbol{\beta}}^{[k]}$ by an increasing order, i.e., $\hat{\beta}_{(1)}^{[k]}, \ldots, \hat{\beta}_{(p)}^{[k]}$ and implement first-order difference, denoted as $\text{diff}(\hat{\boldsymbol{\beta}}^{[k]}) = (d\beta_1, \ldots, d\beta_{p-1})^\top$, where $d\beta_j = \hat{\beta}_{(j+1)}^{[k]} - \hat{\beta}_{(j)}^{[k]}$ for $j = 1, \ldots, p-1$.

- ***Step 4***: Denote $S_\alpha = \{j | d\beta_j > \alpha, j = 1, \ldots, p-1\}$, where $\alpha > 0$ is a threshold. Then $\hat{k} = |S_\alpha|$.

Remark 3.1.2 *Theoretically, the threshold α measures the degree of difference between subgroups, i.e., $\alpha = \min_{j\in\mathcal{C}_\ell, j'\in\mathcal{C}_{\ell'}, \ell\neq\ell'} |\beta_j^0 - \beta_{j'}^0|$ for $\ell, \ell' = 1, \ldots, k^0$, where β_j^0's are true. However, α is unknown for a real problem. We can provide a candidate set of α, such as $\alpha \in \{\alpha_1, \ldots, \alpha_{\max}\}$. Intuitively, as α increases, $|S_\alpha|$ will decrease. We use the "Elbow plot" to choose $\hat{k}$ quantitatively. One may take $\hat{k}$ as the elbow appears. Numerical simulations are performed to illustrate the selection of $\hat{k}$.The details are provided in Section 3.3.1.*

3.2 Theoretical results

Denote $\boldsymbol{V} \in \mathbb{R}^{p\times k}$ as a generic model, where k is the grouping number (or the size of the model $\boldsymbol{V}$). For each row of $\boldsymbol{V}$, only one element equals one, i.e., $v_{j\ell} = 1$ and $v_{j\ell'} = 0$ for $\ell' \neq \ell$, indicating that β_j belongs to the ℓ-th group, $j = 1, \ldots, p, \ell = 1, \ldots, k$. Let $\boldsymbol{\beta}(\boldsymbol{V}) = (\boldsymbol{\beta}_{\mathcal{C}_1}^\top, \ldots, \boldsymbol{\beta}_{\mathcal{C}_k}^\top)^\top \in \mathbb{R}^p$ be the regression vector corresponding to model $\boldsymbol{V}$. We use $\boldsymbol{V}_t \in \mathbb{R}^{p\times k^0}$ to denote the underlying true model and $\boldsymbol{V}_{\tilde{t}} \in \mathbb{R}^{p\times\tilde{k}}$ to denote an over-segmented true model for $\tilde{k} \geqslant k^0$. Now we provide a simple example to explain the over-segmented true model. Let $\boldsymbol{\beta}(\boldsymbol{V}_t) = (\boldsymbol{\beta}_{\mathcal{C}_1}^\top, \ldots, \boldsymbol{\beta}_{\mathcal{C}_{k^0}}^\top)^\top$, $\tilde{k} = k^0 + 1$, implying that for any $\ell \in \{1, \ldots, k^0\}$, the ℓ-th subgroup is partitioned into two segments, i.e., $\boldsymbol{\beta}_{\mathcal{C}_\ell} = (\boldsymbol{\beta}_{\mathcal{C}_{\ell,1}}^\top, \boldsymbol{\beta}_{\mathcal{C}_{\ell,2}}^\top)^\top$, where the distributions of $\boldsymbol{\beta}_{\mathcal{C}_{\ell,1}}$ and $\boldsymbol{\beta}_{\mathcal{C}_{\ell,2}}$ are similar or identical. Rewrite $\boldsymbol{\beta}(\boldsymbol{V}_t) = \boldsymbol{\beta}(\boldsymbol{V}_{\tilde{t}}) = (\boldsymbol{\beta}_{\mathcal{C}_1}^\top, \ldots, \boldsymbol{\beta}_{\mathcal{C}_{\ell,1}}^\top, \boldsymbol{\beta}_{\mathcal{C}_{\ell,2}}^\top, \ldots, \boldsymbol{\beta}_{\mathcal{C}_{k^0}}^\top)^\top$. Therefore, the parameter estimation of the over-segmented true model will rarely be affected if $\tilde{k}$ is given properly. Note that the selection of $\tilde{k}$ has been discussed in the Section 3.1.4. For simplicity, we use $\boldsymbol{V}_t \in \mathbb{R}^{p\times k}, k^0 \leqslant k \leqslant k_{\max}$, to denote the true model and the over-segmented true models in the followings. We use $\boldsymbol{V}_f \in \mathbb{R}^{p\times k}$ to denote a wrong model such that $\boldsymbol{V}_f \neq \boldsymbol{V}_t$. For a given k, denote a collection of $\boldsymbol{V}_t$ and $\boldsymbol{V}_f$ by $\mathbb{V}_t$ and $\mathbb{V}_f$, respectively. Clearly, if $k = k^0$, $\mathbb{V}_t$ contains one element, which is the underling true

model. While if $k^0 < k \leqslant k_{\max}$, $\mathbb{V}_t$ contains more elements. Now we study the model selection consistency, parameter estimation consistency of $\hat{\boldsymbol{\beta}}$, and bounds for the prediction error.

First, we need to obtain the posterior distribution of $\boldsymbol{V}$. Since σ^2 is assumed to be a constant, we omit it in the formulas for simple notations hereafter. (3.1), the joint posterior distribution of $\boldsymbol{\beta}, \boldsymbol{V}, \boldsymbol{\mu}|\boldsymbol{X}, \boldsymbol{y}$ can be rewritten as

$$f(\boldsymbol{\beta}, \boldsymbol{V}, \boldsymbol{\mu}|\boldsymbol{X}, \boldsymbol{y}) \propto \exp\Big\{ -\frac{1}{2}\Big(\boldsymbol{\beta}^\top \boldsymbol{\Lambda}_1 \boldsymbol{\beta} - 2\boldsymbol{\beta}^\top \boldsymbol{\Lambda}_2 + \frac{\boldsymbol{y}^\top \boldsymbol{y}}{\sigma^2} + \frac{\boldsymbol{\mu}^\top \boldsymbol{V}^\top \boldsymbol{V} \boldsymbol{\mu}}{\tau^2} + \frac{\boldsymbol{\mu}^\top \boldsymbol{\mu}}{c^2}\Big)\Big\}, \tag{3.9}$$

where $\boldsymbol{\Lambda}_1 = \frac{\boldsymbol{X}^\top \boldsymbol{X}}{\sigma^2} + \frac{\boldsymbol{I}_p}{\tau^2}$, and $\boldsymbol{\Lambda}_2 = \frac{\boldsymbol{V}\boldsymbol{\mu}}{\tau^2} + \frac{\boldsymbol{X}^\top \boldsymbol{y}}{\sigma^2}$. By integrating $f(\boldsymbol{\beta}, \boldsymbol{V}, \boldsymbol{\mu}|\boldsymbol{X}, \boldsymbol{y})$ with respect to $\boldsymbol{\beta}$, we have the joint posterior distribution of $\boldsymbol{V}, \boldsymbol{\mu}|\boldsymbol{X}, \boldsymbol{y}$,

$$f(\boldsymbol{V}, \boldsymbol{\mu}|\boldsymbol{X}, \boldsymbol{y}) \propto \exp\Big\{ -\frac{1}{2}\Big(\frac{\boldsymbol{y}^\top \boldsymbol{y}}{\sigma^2} + \frac{\boldsymbol{\mu}^\top \boldsymbol{V}^\top \boldsymbol{V} \boldsymbol{\mu}}{\tau^2} + \frac{\boldsymbol{\mu}^\top \boldsymbol{\mu}}{c^2} - \boldsymbol{\Lambda}_2^\top \boldsymbol{\Lambda}_1^{-1} \boldsymbol{\Lambda}_2\Big)\Big\}. \tag{3.10}$$

Denote $\boldsymbol{A}(\boldsymbol{V}) = \frac{\boldsymbol{V}^\top \boldsymbol{V}}{\tau^2} + \frac{\boldsymbol{I}_k}{c^2} - \frac{\boldsymbol{V}^\top \boldsymbol{\Lambda}_1^{-1} \boldsymbol{V}}{\tau^4}$. By integrating $f(\boldsymbol{V}, \boldsymbol{\mu}|\boldsymbol{X}, \boldsymbol{y})$ with respect to $\boldsymbol{\mu}$, the posterior distribution of $\boldsymbol{V}$ is given by

$$f(\boldsymbol{V}|\boldsymbol{X}, \boldsymbol{y}) \propto \big|\boldsymbol{A}(\boldsymbol{V})\big|^{-\frac{1}{2}} \exp\Big\{ -\frac{1}{2}\Big(\frac{\boldsymbol{y}^\top \boldsymbol{X}}{\tau^2 \sigma^2} \boldsymbol{\Lambda}_1^{-1} \boldsymbol{V} \boldsymbol{A}(\boldsymbol{V})^{-1} \boldsymbol{V}^\top \boldsymbol{\Lambda}_1^{-1} \frac{\boldsymbol{X}^\top \boldsymbol{y}}{\tau^2 \sigma^2}\Big)\Big\}. \tag{3.11}$$

Next, we study the model selection consistency under the Bayesian framework. Before proceeding, we review two types of consistency central to this work, i.e., the posterior odds ratio consistency (weak model selection consistency) and the posterior model consistency (PMC: strong model selection consistency). According to Zellner (1978), the posterior odds ratio is defined as $f(\boldsymbol{V}_f|\boldsymbol{X}, \boldsymbol{y})/f(\boldsymbol{V}_t|\boldsymbol{X}, \boldsymbol{y})$. Generally speaking, to make a correct model selection, it holds that

$$\max_{\boldsymbol{V}_f} f(\boldsymbol{V}_f|\boldsymbol{X}, \boldsymbol{y})/f(\boldsymbol{V}_t|\boldsymbol{X}, \boldsymbol{y}) \to 0, \text{ as } n \to \infty, \tag{3.12}$$

implying that the posterior probability of the true model asymptotically dominates that of any incorrect model, which guarantees the posterior odds ratio consistency. While for the strong model

selection consistency (PMC), it holds that $f(\boldsymbol{V}_t|\boldsymbol{X},\boldsymbol{y}) \to 1$ as $n \to \infty$. Representative references include assessment of posterior odds ratio and PMC can be found in Zellner (1978) and Liang et al. (2008). Under our model settings, the strong model selection consistency holds if

$$\sum_{\boldsymbol{V}_t \in \mathbb{V}_t} f(\boldsymbol{V}_t|\boldsymbol{X},\boldsymbol{y}) \to 1, \ as \ n \to \infty, \tag{3.13}$$

since there are more than one alternative true models if $k > k^0$. Meanwhile, the posterior odds ratio consistency follows if

$$\frac{\sum_{\boldsymbol{V}_t \in \mathbb{V}_t} f(\boldsymbol{V}_t|\boldsymbol{X},\boldsymbol{y})}{\max_{\boldsymbol{V}_f \in \mathbb{V}_f} f(\boldsymbol{V}_f|\boldsymbol{X},\boldsymbol{y})} \to \infty, \quad \text{as} \quad n \to \infty. \tag{3.14}$$

Now, we make the following assumptions.

Assumption 3.2.1 *Assume that* $k < n, \tau^2 = O(n^{-a}), c^2 = O(n^{-a'}), 0 < b \leqslant a' \leqslant a < 1$, $\log p = o(n^{2/3})$, *and* $n = o(p^{2/3})$.

Recall that $A(\boldsymbol{V}) = \frac{\boldsymbol{I}_k}{c^2} + \boldsymbol{V}^\top \frac{\boldsymbol{X}^\top \boldsymbol{X}}{\sigma^2} (\frac{\tau^2 \boldsymbol{X}^\top \boldsymbol{X}}{\sigma^2} + \boldsymbol{I}_p)^{-1} \boldsymbol{V}$. Let $M(\boldsymbol{V}_1, \boldsymbol{V}_2) = \boldsymbol{V}_1 A(\boldsymbol{V}_1) \boldsymbol{V}_1^\top - \boldsymbol{V}_2 A(\boldsymbol{V}_2) \boldsymbol{V}_2^\top$, which measures the difference of any two models $\boldsymbol{V}_1$, $\boldsymbol{V}_2 \in \mathbb{R}^{p \times k}$.

Assumption 3.2.2 *Assume that, for any* $\boldsymbol{V}_t$ *and* $\boldsymbol{V}_f$,

$$\frac{\boldsymbol{Y}^\top \boldsymbol{X}}{\tau^2 \sigma^2} \boldsymbol{\Lambda}_1^{-1} M(\boldsymbol{V}_t, \boldsymbol{V}_f) \boldsymbol{\Lambda}_1^{-1} \frac{\boldsymbol{X}^\top \boldsymbol{Y}}{\tau^2 \sigma^2} \geqslant \frac{k}{2} \log \left(1 + \frac{\lambda_{\max} c^2}{\lambda_{\max} \tau^2 + \sigma^2 / n} \right) + \log n, \tag{3.15}$$

where $\boldsymbol{\Lambda}_1 = \frac{\boldsymbol{X}^\top \boldsymbol{X}}{\sigma^2} + \frac{\boldsymbol{I}_p}{\tau^2}$.

Remark 3.2.1 *Assumption 3.2.1 gives common conditions used to constrain the ranges of constant parameters in the priors. To relax the sparsity assumption, we consider the scenario where the number of groups,* k, *satisfies* $k < n$. *In this setting, the number of nonzero parameters can exceed* n. *Assumption 3.2.2 is used to show the influence of the structure of* $\boldsymbol{\beta}$, *which bounds the difference*

between the true and incorrect models. The left-hand side can be viewed as the difference of two norms of $(\frac{\boldsymbol{X}^\top \boldsymbol{X}}{\sigma^2} + \frac{\boldsymbol{I}_p}{\tau^2})^{-1} \frac{\boldsymbol{X}^\top \boldsymbol{Y}}{\tau^2 \sigma^2}$ *with respect to* $\boldsymbol{V}_t A(\boldsymbol{V}_t) \boldsymbol{V}_t^\top$ *and* $\boldsymbol{V}_f A(\boldsymbol{V}_f) \boldsymbol{V}_f^\top$*. It assumes that the true model has the greatest norm among all models.*

Based on the above notations and assumptions, we now establish the weak model selection consistency of the proposed BGGS method.

Theorem 3.2.1 *Under the Assumptions 3.2.1-3.2.2, with probability at least* $1 - \frac{1}{2 \log n}$*, it follows that*

$$\frac{\sum_{\boldsymbol{V}_t \in \mathbb{V}_t} f(\boldsymbol{V}_t | \boldsymbol{X}, \boldsymbol{y})}{\max_{\boldsymbol{V}_f \in \mathbb{V}_f} f(\boldsymbol{V}_f | \boldsymbol{X}, \boldsymbol{y})} \to \infty, \quad as \quad n \to \infty. \tag{3.16}$$

Theorem 3.2.1 shows that the posterior probability of selecting the true model is significantly greater than that of a false model, which implies weak model selection consistency. If Assumption 3.2.2 is further strengthened by following assumption 3.2.3, then the strong model selection (PMC) will be guaranteed.

Assumption 3.2.3 *Assume that, for any* $\boldsymbol{V}_t$ *and* $\boldsymbol{V}_f$*,*

$$\frac{\boldsymbol{Y}^\top \boldsymbol{X}}{\tau^2 \sigma^2} \boldsymbol{\Lambda}_1^{-1} M(\boldsymbol{V}_t, \boldsymbol{V}_f) \boldsymbol{\Lambda}_1^{-1} \frac{\boldsymbol{X}^\top \boldsymbol{Y}}{\tau^2 \sigma^2} \geqslant \frac{k}{2} \log \left(1 + \frac{\lambda_{\max} c^2}{\lambda_{\max} \tau^2 + \sigma^2 / n}\right) + 2n \log p. \tag{3.17}$$

Theorem 3.2.2 *Under the Assumptions 3.2.1 and 3.2.3, with probability at least* $1 - \frac{1}{2 \log n}$*, it follows that*

$$\frac{\max_{\boldsymbol{V}_f \in \mathbb{V}_f} f(\boldsymbol{V}_f | \boldsymbol{X}, \boldsymbol{y})}{\sum_{\boldsymbol{V}_t \in \mathbb{V}_t} f(\boldsymbol{V}_t | \boldsymbol{X}, \boldsymbol{y})} \leqslant \frac{1}{p^n}, \quad as \quad n \to \infty. \tag{3.18}$$

Furthermore,

$$\sum_{\boldsymbol{V}_t \in \mathbb{V}_t} f(\boldsymbol{V}_t | \boldsymbol{X}, \boldsymbol{y}) \to 1, \quad as \quad n \to \infty. \tag{3.19}$$

Theorem 3.2.2 states that under certain conditions, the posterior probability $f(\boldsymbol{V}_t|\boldsymbol{X},\boldsymbol{y})$ dominates $f(\boldsymbol{V}_f|\boldsymbol{X},\boldsymbol{y})$ for any model $\boldsymbol{V}_f \neq \boldsymbol{V}_t$. This implies that the Bayesian model selection procedure will correctly identify the underlying true model with high probability.

By the grouping-Gibbs algorithm, we also provide the estimate of $\boldsymbol{\beta}$ in Eq.(3.8). In the following, we study the estimation consistency of $\hat{\boldsymbol{\beta}}$ and the bounds of the prediction error in L_2 norm, i.e., $\frac{1}{n}\|\boldsymbol{X}\boldsymbol{\beta}-\boldsymbol{X}\hat{\boldsymbol{\beta}}\|_2^2$. The proof of Theorem 3.2.1 and 3.2.2 are provided below.

Proof. The posterior distribution of $\boldsymbol{V}$ is given by

$$f(\boldsymbol{V}|\boldsymbol{X},\boldsymbol{y})\propto \big|\boldsymbol{A}(\boldsymbol{V})\big|^{-\frac{1}{2}}\exp\Big\{-\frac{1}{2}\Big(\frac{\boldsymbol{y}^\top\boldsymbol{X}}{\tau^2\sigma^2}\boldsymbol{\Lambda}_1^{-1}\boldsymbol{V}\boldsymbol{A}(\boldsymbol{V})^{-1}\boldsymbol{V}^\top\boldsymbol{\Lambda}_1^{-1}\frac{\boldsymbol{X}^\top\boldsymbol{y}}{\tau^2\sigma^2}\Big)\Big\}, \tag{3.20}$$

where $\boldsymbol{\Lambda}_1=\frac{\boldsymbol{X}^\top\boldsymbol{X}}{\sigma^2}+\frac{\boldsymbol{I}_p}{\tau^2}$, $\boldsymbol{A}(\boldsymbol{V})=\frac{\boldsymbol{V}^\top\boldsymbol{V}}{\tau^2}+\frac{\boldsymbol{I}_k}{c^2}-\frac{\boldsymbol{V}^\top\boldsymbol{\Lambda}_1^{-1}\boldsymbol{V}}{\tau^4}=\frac{\boldsymbol{I}_k}{c^2}+\boldsymbol{V}^\top\frac{\boldsymbol{X}^\top\boldsymbol{X}}{n}(\frac{\tau^2\boldsymbol{X}^\top\boldsymbol{X}}{n}+\frac{\sigma^2\boldsymbol{I}_p}{n})^{-1}\boldsymbol{V}$. In order to prove theorem 3.2.1, we need to find the lower bound of the posterior odds ratio, i.e.,

$$\frac{f(\boldsymbol{V}_t|\boldsymbol{X},\boldsymbol{y})}{f(\boldsymbol{V}_f|\boldsymbol{X},\boldsymbol{y})}=\frac{|\boldsymbol{A}(\boldsymbol{V}_f)|^{1/2}}{|\boldsymbol{A}(\boldsymbol{V}_t)|^{1/2}}\exp\Big\{\frac{1}{2}\Big(\frac{\boldsymbol{y}^\top\boldsymbol{X}}{\tau^2\sigma^2}\boldsymbol{\Lambda}_1^{-1}M(\boldsymbol{V}_t,\boldsymbol{V}_f)\boldsymbol{\Lambda}_1^{-1}\frac{\boldsymbol{X}^\top\boldsymbol{y}}{\tau^2\sigma^2}\Big)\Big\}.$$

By the boundness of nonzero eigenvalues of $\frac{\boldsymbol{X}^\top\boldsymbol{X}}{n}$ in Assumption 3.2.4, it follows that

$$\boldsymbol{0}\leqslant\frac{\boldsymbol{X}^\top\boldsymbol{X}}{n}\Big(\frac{\tau^2\boldsymbol{X}^\top\boldsymbol{X}}{n}+\frac{\sigma^2\boldsymbol{I}_p}{n}\Big)^{-1}\leqslant\frac{\lambda_{\max}}{\lambda_{\max}\tau^2+\sigma^2/n}\boldsymbol{I}_p.$$

By the definition of $\boldsymbol{A}(\boldsymbol{V})$, we obtain that $\boldsymbol{I}_k/c^2\leqslant\boldsymbol{A}(\boldsymbol{V})\leqslant\boldsymbol{I}_k/c^2+\frac{\lambda_{\max}}{\lambda_{\max}\tau^2+\sigma^2/n}\boldsymbol{V}_t^\top\boldsymbol{V}_t$. Then, for any $\boldsymbol{V}_t\in\mathbb{V}_t$, $\boldsymbol{V}_f\in\mathbb{V}_f$, it holds that

$$\frac{f(\boldsymbol{V}_t|\boldsymbol{X},\boldsymbol{y})}{f(\boldsymbol{V}_f|\boldsymbol{X},\boldsymbol{y})}\geqslant\frac{|\boldsymbol{I}_k/c^2|^{1/2}}{|\boldsymbol{I}_k/c^2+\frac{\lambda_{\max}}{\lambda_{\max}\tau^2+\sigma^2/n}\boldsymbol{V}_t^\top\boldsymbol{V}_t|^{1/2}}\exp\Big\{\frac{1}{2}\Big(\frac{\boldsymbol{y}^\top\boldsymbol{X}}{\tau^2\sigma^2}\boldsymbol{\Lambda}_1^{-1}M(\boldsymbol{V}_t,\boldsymbol{V}_f)\boldsymbol{\Lambda}_1^{-1}\frac{\boldsymbol{X}^\top\boldsymbol{y}}{\tau^2\sigma^2}\Big)\Big\}. \tag{3.21}$$

For the first term of Eq.(3.21), we can obtain that

$$\frac{|\boldsymbol{I}_k/c^2|^{1/2}}{|\boldsymbol{I}_k/c^2+\frac{\lambda_{\max}}{\lambda_{\max}\tau^2+\sigma^2/n}\boldsymbol{V}_t^\top\boldsymbol{V}_t|^{1/2}}\geqslant\sqrt{\frac{1}{(1+\frac{\lambda_{\max}c^2}{\lambda_{\max}\tau^2+\sigma^2/n})^k}}=\Big(1+\frac{\lambda_{\max}c^2}{\lambda_{\max}\tau^2+\sigma^2/n}\Big)^{-\frac{k}{2}}. \tag{3.22}$$

While for the second term of Eq.(3.21), by Assumption 3.2.2, it follows that

$$\exp\Big\{\frac{1}{2}\Big(\frac{\boldsymbol{y}^\top \boldsymbol{X}}{\tau^2\sigma^2}\boldsymbol{\Lambda}_1^{-1}M(\boldsymbol{V}_t,\boldsymbol{V}_f)\boldsymbol{\Lambda}_1^{-1}\frac{\boldsymbol{X}^\top \boldsymbol{y}}{\tau^2\sigma^2}\Big)\Big\} \geqslant \sqrt{n}\Big(1+\frac{\lambda_{\max}c^2}{\lambda_{\max}\tau^2+\sigma^2/n}\Big)^{\frac{k}{2}}. \tag{3.23}$$

Putting (3.21)-(3.23) together, therefore, we have

$$\frac{\sum_{\boldsymbol{V}_t\in\mathbb{V}_t} f(\boldsymbol{V}_t|\boldsymbol{X},\boldsymbol{y})}{f(\boldsymbol{V}_f|\boldsymbol{X},\boldsymbol{y})} \geqslant \sqrt{n}\to\infty, as\ n\to\infty. \tag{3.24}$$

This completes the proof of Theorem 3.2.1.

Under Assumption 3.2.3, we have that

$$\frac{f(\boldsymbol{V}_f|\boldsymbol{X},\boldsymbol{Y})}{\sum_{\boldsymbol{V}_t\in\mathbb{V}_t} f(\boldsymbol{V}_t|\boldsymbol{X},\boldsymbol{Y})} \leqslant p^{-n},$$

$$\sum_{\boldsymbol{V}_t\in\mathbb{V}_t} f(\boldsymbol{V}_t|\boldsymbol{X},\boldsymbol{Y}) = \frac{1}{1+\frac{\sum_{\boldsymbol{V}_f\in\mathbb{V}_f} f(\boldsymbol{V}_f|\boldsymbol{X},\boldsymbol{Y})}{\sum_{\boldsymbol{V}_t\in\mathbb{V}_t} f(\boldsymbol{V}_t|\boldsymbol{X},\boldsymbol{Y})}} \geqslant \frac{1}{1+\frac{(p-1)!}{(k-1)!(p-k)!p^n}} \to 1, as\ p\to\infty.$$

This completes the proof of Theorem 3.2.2. □

Assumption 3.2.4 *The rank of $\boldsymbol{X}^\top\boldsymbol{X}$ is r satisfying $r < n^b/\log n$. Let the nonzero eigenvalues of $\frac{\boldsymbol{X}^\top\boldsymbol{X}}{n}$ be $\lambda_1,\lambda_2,...,\lambda_r$ bounded by $\lambda_{\max}$ and $\lambda_{\min}$, where $\lambda_{\min} > \sqrt{\log n}/n^{1-a}$ and $\lambda_{\max}$ is bounded by a constant. Furthermore, assume that for the real value of $\boldsymbol{\beta}$, denoted as $\boldsymbol{\beta}^0$, satisfies $\boldsymbol{\beta}^{0\top}\boldsymbol{\beta}^0 \leqslant C < \infty$ and $\sum_{j=r+1}^p \beta_j^2 = O(1/\log n)$, where $\beta_j, j = r+1,\ldots,p$, are associated with zero eigenvalues of $\frac{\boldsymbol{X}^\top\boldsymbol{X}}{n}$.*

Theorem 3.2.3 *Under the Assumptions 3.2.1 and 3.2.4, with probability greater than $(1-e^{-\sqrt{n}})(1-e^{-n^b/\log n})$, it follows that*

$$\|\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}^0\|_2^2 = O\left(\frac{1}{\log n}\right). \tag{3.25}$$

The following Lemma, derivated by Hsu et al. (2012), is needed in the proof of Theorem 3.2.3.

Lemma 3.2.1 *Assume $\boldsymbol{\beta}$ follows a multivariate normal distribution $N(\boldsymbol{\beta}^*, \boldsymbol{\Sigma})$, with mean vector $\boldsymbol{\beta}^*$ and covariance $\boldsymbol{\Sigma}$. Let $\boldsymbol{\xi} = \boldsymbol{\Sigma}^{-1/2}(\boldsymbol{\beta} - \boldsymbol{\beta}^*) \sim N(\mathbf{0}, \boldsymbol{I}_p)$. For all $t > 0$, it follows that*

$$P(\boldsymbol{\xi}^\top \boldsymbol{\Sigma} \boldsymbol{\xi} > tr(\boldsymbol{\Sigma}) + 2\sqrt{tr(\boldsymbol{\Sigma}^2)t} + 2||\boldsymbol{\Sigma}||t) \leqslant e^{-t}, \tag{3.26}$$

where $\|\boldsymbol{\Sigma}\|$ denotes the spectral norm of the matrix $\boldsymbol{\Sigma}$.

Now we begin to prove Theorem 3.2.3.

Proof. By integrating $f(\boldsymbol{\beta}, \boldsymbol{V}, \boldsymbol{\mu}|\boldsymbol{X}, \boldsymbol{y})$ with respect to $\boldsymbol{\mu}$, we get the posterior probability of $\boldsymbol{\beta}|\boldsymbol{V}, \boldsymbol{X}, \boldsymbol{Y}$, i.e.,

$$f(\boldsymbol{\beta}|\boldsymbol{V}, \boldsymbol{X}, \boldsymbol{Y}) \propto \exp\left\{ -\frac{1}{2}\Big(- \frac{2\boldsymbol{Y}^\top \boldsymbol{X}\boldsymbol{\beta}}{\sigma^2} + \boldsymbol{\beta}^\top (\frac{\boldsymbol{X}^\top \boldsymbol{X}}{\sigma^2} + \frac{\boldsymbol{I}_p}{\tau^2} - \frac{\boldsymbol{V}}{\tau^2}(\frac{\boldsymbol{I}_k}{c^2} + \frac{\boldsymbol{V}^\top \boldsymbol{V}}{\tau^2})^{-1}\frac{\boldsymbol{V}^\top}{\tau^2})\boldsymbol{\beta}\Big)\right\}. \tag{3.27}$$

Clearly, $\boldsymbol{\beta}$ follows a multivariate normal distribution, $N(\boldsymbol{\beta}^*, \boldsymbol{\Sigma})$, where $\boldsymbol{\beta}^* = \boldsymbol{\Sigma}\frac{\boldsymbol{X}^\top \boldsymbol{Y}}{\sigma^2}, \boldsymbol{\Sigma} = (\frac{\boldsymbol{I}_p}{\tau^2} + \frac{\boldsymbol{X}^\top \boldsymbol{X}}{\sigma^2} - \frac{\boldsymbol{V}}{\tau^2}(\frac{\boldsymbol{I}_k}{c^2} + \frac{\boldsymbol{V}^\top \boldsymbol{V}}{\tau^2})^{-1}\frac{\boldsymbol{V}^\top}{\tau^2})^{-1}$. Since $\hat{\boldsymbol{\beta}} = \sum_{m=1}^{M} \boldsymbol{\beta}^{(m)}/M$, then $\hat{\boldsymbol{\beta}} \sim N(\boldsymbol{\beta}^*, \boldsymbol{\Sigma}/M)$. Considering the fact that

$$(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0)^\top (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0) \leqslant 2(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)^\top (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*) + 2(\boldsymbol{\beta}^* - \boldsymbol{\beta}^0)^\top (\boldsymbol{\beta}^* - \boldsymbol{\beta}^0), \tag{3.28}$$

we now work on the two terms in the right hand of (3.28). Before proceeding, we need to approximate the eigenvalues and trace of $\boldsymbol{\Sigma}$. We have the following inequalities,

$$\frac{\boldsymbol{I}_p}{\tau^2} - \frac{\boldsymbol{V}}{\tau^2}(\frac{\boldsymbol{I}_k}{c^2} + \frac{\boldsymbol{V}^\top \boldsymbol{V}}{\tau^2})^{-1}\frac{\boldsymbol{V}^\top}{\tau^2} \geqslant \frac{\boldsymbol{I}_p}{\tau^2} - \frac{\boldsymbol{V}}{\tau^2}(\frac{\boldsymbol{I}_k}{c^2} + \frac{\boldsymbol{I}_k}{\tau^2})^{-1}\frac{\boldsymbol{V}^\top}{\tau^2} \geqslant (\frac{1}{\tau^2} - \frac{1}{\tau^4}(\frac{1}{c^2} + \frac{1}{\tau^2})^{-1})\boldsymbol{I}_p = \frac{\boldsymbol{I}_p}{c^2 + \tau^2}.$$

By Assumption 3.2.4, it holds that

$$\boldsymbol{\Sigma} \leqslant (\frac{\boldsymbol{X}^\top \boldsymbol{X}}{\sigma^2} + \frac{\boldsymbol{I}_p}{c^2 + \tau^2})^{-1} \leqslant (c^2 + \tau^2)\boldsymbol{I}_p,$$

$$tr(\boldsymbol{\Sigma}) \leqslant p(c^2 + \tau^2).$$

Next, we provide the boundness of the first term in the right hand of (3.28). Taking $t = \sqrt{n}$, by Lemma 3.2.1, we have

$$
\begin{aligned}
e^{-\sqrt{n}} &\geqslant P\{\boldsymbol{\xi}^\top \boldsymbol{\Sigma} \boldsymbol{\xi}/M > tr(\boldsymbol{\Sigma})/M + 2\sqrt{tr(\boldsymbol{\Sigma}^2)t}/M + 2||\boldsymbol{\Sigma}||t/M\} \qquad (3.29)\\
&= P\{(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)^\top (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*) > tr(\boldsymbol{\Sigma})/M + 2\sqrt{tr(\boldsymbol{\Sigma}^2)t}/M + 2||\boldsymbol{\Sigma}||t/M\}\\
&\geqslant P\{(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)^\top (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*) > \frac{p(c^2+\tau^2)}{M} + 2\frac{\sqrt{p}(c^2+\tau^2)}{M} n^{1/4} + \frac{2\sqrt{n}(c^2+\tau^2)}{M}\}.
\end{aligned}
$$

With probability greater than $1 - e^{-\sqrt{n}}$, therefore, it holds that,

$$
(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)^\top (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*) \leqslant \frac{p(c^2+\tau^2)}{M} + 2\frac{\sqrt{p}(c^2+\tau^2)}{M} n^{1/4} + \frac{2\sqrt{n}(c^2+\tau^2)}{M} \to 0, \ as\ n \to \infty. \qquad (3.30)
$$

Then we provide the boundness of the term $(\boldsymbol{\beta}^* - \boldsymbol{\beta}^0)^\top (\boldsymbol{\beta}^* - \boldsymbol{\beta}^0)$. By plugging in $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta}^0 + \boldsymbol{\epsilon}$, we get that

$$
(\boldsymbol{\beta}^* - \boldsymbol{\beta}^0)^\top (\boldsymbol{\beta}^* - \boldsymbol{\beta}^0) = \left((\frac{\boldsymbol{\Sigma}\boldsymbol{X}^\top \boldsymbol{X}}{\sigma^2} - \boldsymbol{I}_p)\boldsymbol{\beta}^0 + \frac{\boldsymbol{\Sigma}\boldsymbol{X}^\top \boldsymbol{\epsilon}}{\sigma^2} \right)^\top \left((\frac{\boldsymbol{\Sigma}\boldsymbol{X}^\top \boldsymbol{X}}{\sigma^2} - \boldsymbol{I}_p)\boldsymbol{\beta}^0 + \frac{\boldsymbol{\Sigma}\boldsymbol{X}^\top \boldsymbol{\epsilon}}{\sigma^2} \right).
$$

It is easy to verify the following inequalities hold, i.e.,

$$
\begin{aligned}
\boldsymbol{0} \leqslant \boldsymbol{I}_p - \frac{\boldsymbol{\Sigma}\boldsymbol{X}^\top \boldsymbol{X}}{\sigma^2} &= \boldsymbol{I}_p - \left(\frac{\boldsymbol{I}_p}{\tau^2} + \frac{\boldsymbol{X}^\top \boldsymbol{X}}{\sigma^2} - \frac{\boldsymbol{V}}{\tau^2} (\frac{\boldsymbol{I}_k}{c^2} + \frac{\boldsymbol{V}^\top \boldsymbol{V}}{\tau^2})^{-1} \frac{\boldsymbol{V}^\top}{\tau^2} \right)^{-1} \frac{\boldsymbol{X}^\top \boldsymbol{X}}{\sigma^2}\\
&\leqslant \boldsymbol{I}_p - \left(\frac{\boldsymbol{I}_p}{\tau^2} + \frac{\boldsymbol{X}^\top \boldsymbol{X}}{\sigma^2} - \frac{\boldsymbol{V}}{\tau^2} (\frac{1}{c^2} + \frac{p-k+1}{\tau^2})^{-1} \frac{\boldsymbol{V}^\top}{\tau^2} \right)^{-1} \frac{\boldsymbol{X}^\top \boldsymbol{X}}{\sigma^2}\\
&\leqslant \boldsymbol{I}_p - \left(\frac{\tau^2 + c^2(p-k)}{\tau^4 + c^2\tau^2(p-k+1)} \boldsymbol{I}_p + \frac{\boldsymbol{X}^\top \boldsymbol{X}}{\sigma^2} \right)^{-1} \frac{\boldsymbol{X}^\top \boldsymbol{X}}{\sigma^2}\\
&= \left(\frac{\tau^2 + c^2(p-k)}{\tau^4 + c^2\tau^2(p-k+1)} \boldsymbol{I}_p + \frac{\boldsymbol{X}^\top \boldsymbol{X}}{\sigma^2} \right)^{-1} \frac{\tau^2 + c^2(p-k)}{\tau^4 + c^2\tau^2(p-k+1)} \boldsymbol{I}_p\\
&= \left(\boldsymbol{I}_p + \frac{\tau^4 + c^2\tau^2(p-k+1)}{\tau^2 + c^2(p-k)} \frac{\boldsymbol{X}^\top \boldsymbol{X}}{\sigma^2} \right)^{-1}\\
&\leqslant \left(\boldsymbol{I}_p + \tau^2 \frac{\boldsymbol{X}^\top \boldsymbol{X}}{\sigma^2} \right)^{-1}.
\end{aligned}
$$

Then we have

$$
\begin{aligned}
(\boldsymbol{\beta}^* - \boldsymbol{\beta}^0)^\top(\boldsymbol{\beta}^* - \boldsymbol{\beta}^0) =& \left((\frac{\boldsymbol{\Sigma}\boldsymbol{X}^\top\boldsymbol{X}}{\sigma^2} - \boldsymbol{I}_p)\boldsymbol{\beta}^0 + \frac{\boldsymbol{\Sigma}\boldsymbol{X}^\top\boldsymbol{\epsilon}}{\sigma^2}\right)^\top \left((\frac{\boldsymbol{\Sigma}\boldsymbol{X}^\top\boldsymbol{X}}{\sigma^2} - \boldsymbol{I}_p)\boldsymbol{\beta}^0 + \frac{\boldsymbol{\Sigma}\boldsymbol{X}^\top\boldsymbol{\epsilon}}{\sigma^2}\right) \\
\leqslant& 2\boldsymbol{\beta}^{0\top}\left(\frac{\boldsymbol{\Sigma}\boldsymbol{X}^\top\boldsymbol{X}}{\sigma^2} - \boldsymbol{I}_p\right)^\top\left(\frac{\boldsymbol{\Sigma}\boldsymbol{X}^\top\boldsymbol{X}}{\sigma^2} - \boldsymbol{I}_p\right)\boldsymbol{\beta}^0 + \frac{2\boldsymbol{\epsilon}^\top\boldsymbol{X}\boldsymbol{\Sigma}^2\boldsymbol{X}^\top\boldsymbol{\epsilon}}{\sigma^4} \\
\leqslant& \frac{2\sigma^4\boldsymbol{\beta}^{0\top}\boldsymbol{\beta}^0}{(n\tau^2\lambda_{\min} + \sigma^2)^2} + 2\sum_{i=j}^{p}\beta_i^2 + 2\boldsymbol{\epsilon}^\top\boldsymbol{X}\left(\boldsymbol{X}^\top\boldsymbol{X} + \frac{\sigma^2\boldsymbol{I}_p}{c^2+\tau^2}\right)^{-2}\boldsymbol{X}^\top\boldsymbol{\epsilon} \\
\leqslant& \frac{2\sigma^4 C}{(n\tau^2\lambda_{\min} + \sigma^2)^2} + \frac{2}{\log n} + \frac{5\sigma_0^2 n^b c^2}{\sigma^2\log n},
\end{aligned} \tag{3.31}
$$

where $\lambda_{\min} > n^{a-1}\sqrt{\log n}$. The left-hand side convergence to zero when $c^2 = O(n^{-b})$. The nonzero eigenvalues of $\boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X} + \frac{\sigma^2\boldsymbol{I}_p}{c^2+\tau^2})^{-2}\boldsymbol{X}^\top$ are $n\lambda_i(n\lambda_i + \sigma^2/(c^2+\tau^2))^{-2}, i = 1, ..., r$, and

$$
\begin{aligned}
tr(\boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X} + \frac{\sigma^2\boldsymbol{I}_p}{c^2+\tau^2})^{-2}\boldsymbol{X}^\top) \leqslant& n^{1+b}\lambda_{\min}(n\lambda_{\min} + \sigma^2/(c^2+\tau^2))^{-2}/\log n \\
\leqslant& n^{1+b}\lambda_{\min}(n\lambda_{\min} + \frac{\sigma^2}{2c^2})^{-2}/\log n \\
\leqslant& \frac{Cn^b c^2}{2\sigma^2\log n},
\end{aligned} \tag{3.32}
$$

for some constant C. The second inequality is derived from the fact that $P(\boldsymbol{\epsilon}^\top\boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X} + \frac{\sigma^2\boldsymbol{I}_p}{c^2+\tau^2})^{-2}\boldsymbol{X}^\top\boldsymbol{\epsilon}/\sigma_0^2 \geqslant \frac{5n^b c^2}{2\sigma^2\log n}) \leqslant e^{-n^b/\log n}$, where σ_0^2 is the variance of $\boldsymbol{\epsilon}$, which is assumed to be a constant. Putting Eqs.(3.30)-(3.31) together, we thus have $(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0)^\top(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0) = O(\frac{1}{\log n})$, with probability greater than $(1 - e^{-\sqrt{n}})(1 - e^{-n^b/\log n})$. This completes the proof. □

Theorem 3.2.3 demonstrates that the distance of $\boldsymbol{\beta}_0$ and $\hat{\boldsymbol{\beta}}$ in the L_2 norm goes to zero, implying the consistency of $\hat{\boldsymbol{\beta}}$. Aligned with Assumption 3.2.4, the bounds on the prediction error can be obtained directly from Theorem 3.2.3, i.e.,

$$
\frac{1}{n}\|\boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{X}\hat{\boldsymbol{\beta}}\|_2^2 \leqslant \lambda_{\max}\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0\|_2^2 = O_p\left(\frac{1}{\log n}\right). \tag{3.33}
$$

Remark 3.2.2 *Assumption 3.2.4 provides a bounded support for $\boldsymbol{\beta}^0$ in the L_2 norm. Denote the components of $\boldsymbol{\beta}$ associated with zero eigenvalues as $\boldsymbol{\beta}_{\mathcal{G}_{II}}$. Since $p \to \infty$, we limit the elements of $\boldsymbol{\beta}_{\mathcal{G}_{II}}$ to have small values, such that $\|\boldsymbol{\beta}_{\mathcal{G}_{II}}\|_2^2 = O\left(\frac{1}{\log n}\right)$. This assumption is in line with Assumption (A2) in Ding et al. (2021). A more stringent constraint on the order of $\|\boldsymbol{\beta}_{\mathcal{G}_{II}}\|_2^2$, such as $\|\boldsymbol{\beta}_{\mathcal{G}_{II}}\|_2^2 = O\left(\frac{1}{n^{a'-b}\log n}\right)$, leads to a less restrictive degree of non-sparsity. Specifically, when $\|\boldsymbol{\beta}_{\mathcal{G}_{II}}\|_2^2 = 0$, it reduces to sparsity. Furthermore, we assume that $\lambda_{\min} > n^{(3a'-b)/2-1}\sqrt{\log n}$. Then, the order of Eq.*(3.25) *becomes $O\left(\frac{1}{n^{a'-b}\log n}\right)$. To ensure that $\lambda_{\min}$ tends to zero, we assume that $0 < b < a' < (2+b)/3$. Consequently, the minimum convergence rate of the BGGS estimator exceeds $O(\frac{1}{n^{2/3}\log n})$. In comparison with the convergence rate of Lasso, which is $O_p(n^{-1/2})$ (Fan and Li, 2001), the BGGS estimator achieves a faster convergence rate.*

3.3 Simulations

In this section, we perform simulations to illustrate the selection of k numerically. Following this, we compare our proposed BGGS method with several classic frequentist methods, such as Lasso (R package `glmnet` by Friedman et al. (2021)), ncTLF and Gflasso (R package `FGSG` by Shen et al. (2015)) in terms of parameter estimation and prediction accuracy via two examples. Lasso is designed for a sparse model, whereas both ncTLF and Gflasso are designed for sparse and grouping models. Although fused Lasso (Tibshirani et al., 2005) is proposed for estimating homogeneous subgroups, its major limitation is that the predictors should be given in a serial order associated with the regression coefficients. For this reason, we exclude fused Lasso from our comparison.

We emphasize that our study focuses on the estimation of $\boldsymbol{\beta}$ and the model's prediction accuracy rather than on identifying the groups of $\boldsymbol{\beta}$. The grouping technique serves as a tool to achieve the

estimation objective. As a result, the evaluation metrics we adopt are mean squared error (MSE) and mean absolute error (MAE) to measure the prediction accuracy, i.e.,

$$\text{MSE} = \frac{1}{n}\|\boldsymbol{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0)\|_2^2, \ \text{MAE} = \frac{1}{n}\|\boldsymbol{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0)\|_1,$$

where $\hat{\boldsymbol{\beta}}$ is an estimate of $\boldsymbol{\beta}$, $\boldsymbol{\beta}^0$ is the underlying true regression vector, $\boldsymbol{X} \in \mathbb{R}^{n\times p}$ is p predictors with a sample size n. Moreover, we apply the following two evaluation metrics to measure parameter estimation accuracy, i.e.,

$$\text{MSE}_\beta = \frac{1}{p}\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0\|_2^2, \ \text{MAE}_\beta = \frac{1}{p}\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0\|_1.$$

3.3.1 Numerical study for selection of k

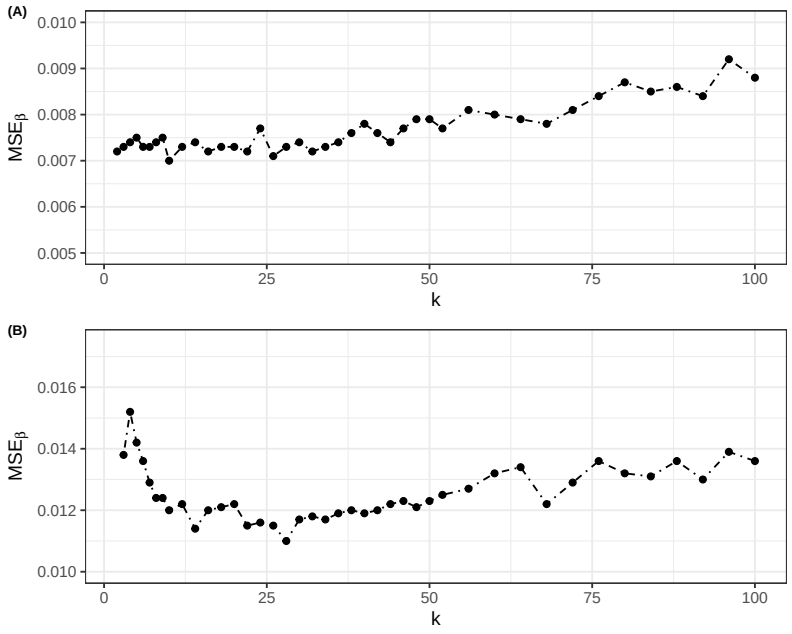

Figure 3.1: Plots of MSE_β given different k for (A) Example 1: $k^0 = 2, \sigma^2 = 1$, and (B) Example 2: $k^0 = 5, \sigma^2 = 1$.

In this subsection, we perform simulations to illustrate that when k is chosen within an appropriate range, it rarely affects the estimation of $\boldsymbol{\beta}$. Figure 3.1 displays the plots of MSE_β for different

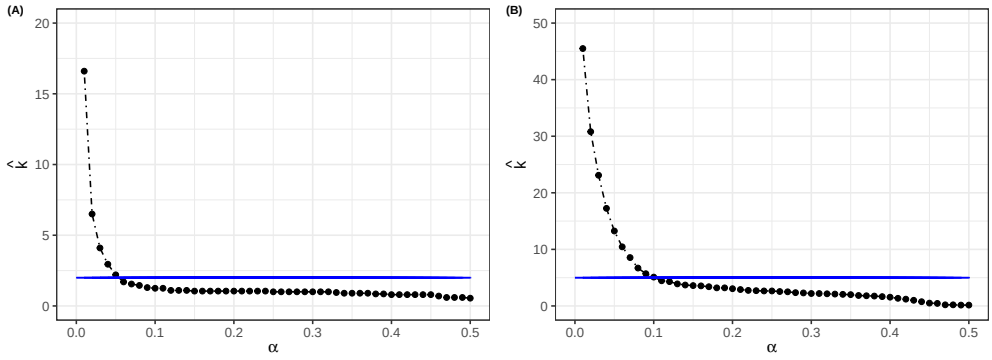

Figure 3.2: The elbow plots the estimated $\hat{k}$ given different α for (A) Example 1: $k^0 = 2, \sigma^2 = 1$, and (B) Example 2: $k^0 = 5, \sigma^2 = 1$. The blue horizontal lines show the underlying true grouing numbers.

k values in Example 1 and Example 2. Clearly, as k increases, MSE_β initially decreases, then stabilizes when $k > k^0$, and increases when k continues to grow. If $k < k^0$, $\boldsymbol{\beta}$ is under-segmented, resulting in a large bias of $\hat{\boldsymbol{\beta}}$. Conversely, if $k > K$, $\boldsymbol{\beta}$ becomes overly segmented, increasing number of groups and computational loads. Due to the limited processing time, the convergence of model selection may not be guaranteed, which can amplify the variance of $\hat{\boldsymbol{\beta}}$. However, if k falls within an appropriate range, the effects of k on the parameter estimation is negligible.

Additionally, we conduct two more simulations to illustrate the approach for selecting k. Figure 3.2 shows the elbow plots of the estimated $\hat{k}$ for different α, which is helpful for determining the grouping number $\hat{k}$. Clearly, as α increases, $\hat{k}$ decreases and approaches a constant value. From the two subplots in Figure 3.2, the elbows appear around the true k^0 for both examples. Although this approach can be somewhat subjective when identifying the elbow, it provides at least an estimate of $\hat{k}$. Even if $\hat{k} > k^0$, it tends towards the true k^0, rather than deviating from it. The elbow plot is a well-established quantitative approach for selecting principal components, regularization parameters, the K value in the K-Nearest Neighbors method, and others. Finally, we note that

the data generation for example 1 and example 2 is provided in the following two subsections.

3.3.2 Example 1

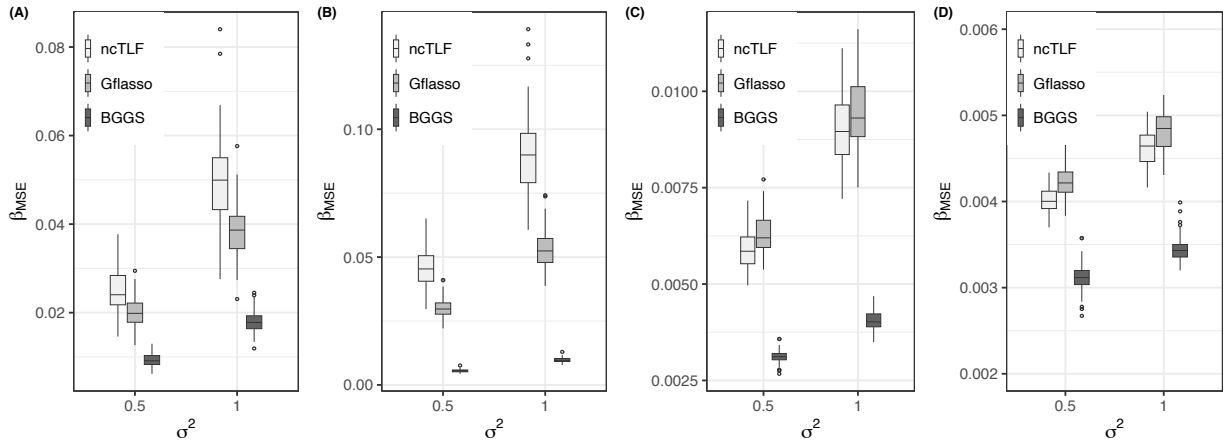

Figure 3.3: Example 1:boxplots of MSE_β estimated via ncTLF, Gflasso, and BGGS over 100 simulations under the settings of $n = 200$, $\sigma^2 = 0.5$, 1, and (A) $p = 100$, (B) $p = 200$, (C) $p = 500$, (D) $p = 1000$.

In this example, we consider an interesting situation in which $\boldsymbol{\beta} \in \mathbb{R}^p$ consists of two groups: components in group I are large, while components in group II are small but possibly nonzero. Ding et al. (2021) studied the asymptotic behavior of the ridge regularized high-dimensional multivariate M-estimator of $\boldsymbol{\beta}$ in group II by applying the double leave-one-out method. The components in group I are generated from a uniform distribution, specifically $U(1, 1.2)$, while the components in group II are generated from $U(0.1, 0.3)$. We further assume that $|\mathcal{C}_1| = p/4$, and $|\mathcal{C}_2| = 3p/4$. A total of 100 simulations were conducted.

Table 3.1 lists the average MSEs and MAEs (standard deviation in parentheses) of Lasso, ncTLF, Gflasso and BGGS under the settings $n = 200$, $p = 100, 200, 500, 1000$, and $\sigma^2 = 0.5$, 1. The results are alos presented in Table 3.1 and Figures 3.3-3.4. As observed from Table 3.1, for

Table 3.1: Example 1: comparison of average MSEs and MAEs (standard deviation in parentheses) among Lasso, ncTLF, Gflasso and BGGS under the settings of $n = 200$, $p = 100, 200, 500, 1000$, $\sigma^2 = 0.5$, 1.

p	Method	$\sigma^2 = 0.5$		$\sigma^2 = 1$	
		MSE	MAE	MSE	MAE
100	Lasso	1.5464(0.26)	0.9871(0.09)	1.3101(0.30)	0.9095(0.11)
	ncTLF	0.5241(0.12)	0.5760(0.07)	1.0235(0.21)	0.8034(0.08)
	Gflasso	0.4257(0.08)	0.5204(0.05)	0.7881(0.14)	0.7053(0.06)
	BGGS	**0.2014**(0.03)	**0.3585**(0.03)	**0.3799**(0.06)	**0.4901**(0.04)
200	Lasso	11.8387(1.92)	2.7368(0.22)	11.6605(1.96)	2.7199(0.23)
	ncTLF	1.8725(0.37)	1.0921(0.11)	3.7226(3.72)	1.5364(0.16)
	Gflasso	1.2116(0.21)	0.8802(0.08)	2.1836(0.39)	1.1758(0.11)
	BGGS	**0.2610**(0.04)	**0.4071**(0.03)	**0.4457**(0.08)	**0.5314**(0.05)
500	Lasso	135.0585(15.54)	9.2941(0.53)	133.4525(18.30)	9.2112(0.64)
	ncTLF	0.6469(0.09)	0.6377(0.05)	0.9852(0.15)	0.7922(0.06)
	Gflasso	0.7514(0.11)	0.6889(0.05)	1.0726(0.17)	0.8246(0.06)
	BGGS	**0.3733**(0.07)	**0.4870**(0.04)	**0.4616**(0.08)	**0.5408**(0.05)
1000	Lasso	693.1763(82.68)	20.9944(1.33)	697.9479(86.889)	21.0823(1.37)
	ncTLF	1.3283(0.24)	0.9181(0.08)	1.4367(0.25)	0.9551(0.08)
	Gflasso	1.5584(0.31)	0.9934(0.10)	1.6442(0.31)	1.0213(0.10)
	BGGS	**1.0002**(0.28)	**0.7939**(0.11)	**1.0345**(0.30)	**0.8055**(0.11)

Table 3.2: Example 1: average MSEs and MAEs (standard deviation in parentheses) of Lasso, ncTLF, Gflasso and BGGS under the settings of $\mathcal{C}_1 = \varnothing$, $n = 200$, $p = 100, 200, 500, 1000$, $\sigma^2 = 0.5, 1$.

p	Method	σ^2= 0.5		$\sigma^2 = 1$	
		MSE	MAE	MSE	MAE
100	Lasso	0.4535(0.10)	0.5357(0.06)	0.5816(0.11)	0.6062(0.06)
	ncTLF	0.5055(0.11)	0.5653(0.06)	0.9935(0.21)	0.7909(0.08)
	Gflasso	0.3887(0.07)	0.4966(0.05)	0.7148(0.13)	0.6719(0.06)
	BGGS	**0.1900**(0.04)	**0.3473**(0.03)	**0.3624**(0.07)	**0.4790**(0.04)
200	Lasso	3.3618(0.49)	1.4594(0.11)	3.1977(0.54)	1.4286(0.12)
	ncTLF	1.8957(0.37)	1.0958(0.11)	3.6019(0.61)	1.5096(0.13)
	Gflasso	1.0855(0.18)	0.8303(0.07)	1.7618(0.26)	1.0605(0.08)
	BGGS	**0.2228**(0.03)	**0.3772**(0.03)	**0.3794**(0.05)	**0.4924**(0.04)
500	Lasso	32.5542(4.89)	4.5517(0.34)	30.9022(4.35)	4.4190(0.31)
	ncTLF	0.5327(0.07)	0.5806(0.04)	0.9024(0.15)	0.7581(0.06)
	Gflasso	0.5704(0.08)	0.6003(0.05)	0.9369(0.15)	0.7719(0.77)
	BGGS	**0.3165**(0.04)	**0.4490**(0.03)	**0.4194**(0.05)	**0.5178**(0.03)
1000	Lasso	156.6529(18.61)	9.9879(0.62)	152.5972(21.22)	9.8274(0.71)
	ncTLF	0.6811(0.07)	0.6584(0.04)	0.7868(0.09)	0.7065(0.04)
	Gflasso	0.7187(0.08)	0.6764(0.04)	0.8235(0.09)	0.7227(0.04)
	BGGS	**0.6213**(0.06)	**0.6284**(0.03)	**0.6431**(0.06)	**0.6385**(0.03)

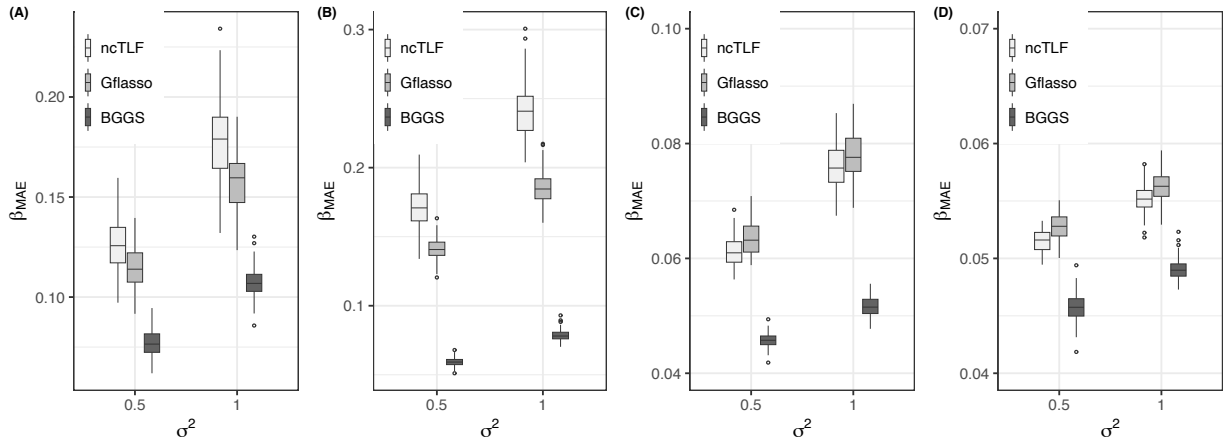

Figure 3.4: Example 1:boxplots of MAE_β estimated via ncTLF, Gflasso, and BGGS over 100 simulations under the settings of $n = 200$, $\sigma^2 = 0.5$, 1, and (A) $p = 100$, (B) $p = 200$, (C) $p = 500$, (D) $p = 1000$.

the Lasso method tailored for sparsity, there is a notable increase in both MSEs and MAEs as p increases. In contrast, for BGGS, MSEs and MAEs show a gradual increase with the augmentation of p. In contrast, for ncTLF and Gflasso, both MSEs and MAEs exhibit slight fluctuations as p varies. Among Lasso, ncTLF, and Gflasso, BGGS demonstrates superior prediction performance, achieving the smallest MSEs and MAEs. Figures 3.3-3.4 show boxplots of MSE_β and MAE_β respectively, estimated via ncTLF, Gflasso, and BGGS over 100 simulations under the settings $n = 200,\ p = 100, 200, 500, 1000$, and $\sigma^2 = 0.5$, 1. Given that Lasso yields the largest MSEs and MAEs across all cases, we omit MSE_β and MAE_β of Lasso in the figures. These findings align with observations reported for both MSEs and MAEs.

Additionlly, we consider a special case where all components of $\boldsymbol{\beta}$ are small, i.e., $\mathcal{C}_1 = \varnothing$. The corresponding MSEs and MAEs are presented in Table 3.2. For a given p and σ^2, the MSEs and MAEs are comparatively smaller than those reported in Table 3.1. Similar trends are evident across both tables. Clearly, BGGS consistently achieves the highest prediction accuracy across all

cases.

3.3.3 Example 2

.

Table 3.3: Example 2: comparison of average MSEs and MAEs (standard deviation in parentheses) among Lasso, ncTLF, Gflasso and BGGS over 100 simulations under different pairs of (n, p) and the settings of $\sigma^2 = 1,2,5$.

(n,p)	Method	$\sigma^2=1$		$\sigma^2=2$		$\sigma^2=5$	
		MSE	MAE	MSE	MAE	MSE	MAE
(50, 100)	Lasso	9.1256(3.05)	2.3982(0.39)	9.1676(2.67)	2.3963(0.36)	9.1840(2.64)	2.3983(0.37)
	ncTLF	1.8998(0.72)	1.0876(0.21)	2.7598(0.93)	1.3137(0.23)	6.4442(2.26)	2.0043(0.35)
	Gflasso	2.5988(0.97)	1.2729(0.24)	3.2564(0.91)	1.4356(0.21)	6.1936(2.04)	1.9582(0.33)
	BGGS	**1.4564**(0.57)	**0.9542**(0.18)	**2.1742**(0.75)	**1.1641**(0.20)	**4.9450**(1.65)	**1.7595**(0.29)
(50, 200)	Lasso	56.4708(14.43)	5.9651(0.80)	59.1229(14.71)	6.0791(0.81)	54.2584(13.92)	5.8492(0.75)
	ncTLF	4.9759(2.29)	1.7506(0.40)	5.4278(2.00)	1.8381(0.35)	6.6057(2.86)	2.0122(0.42)
	Gflasso	8.5559(3.17)	2.3119(0.43)	8.6024(3.20)	2.3117(0.42)	8.5537(3.31)	2.2955(0.43)
	BGGS	**3.7482**(1.80)	**1.5159**(0.35)	**4.1878**(1.56)	**1.6147**(0.31)	**5.4928**(2.45)	**1.8380**(0.40)
(100, 200)	Lasso	22.4695(4.53)	3.7745(0.39)	21.3346(4.73)	3.6631(0.43)	21.2851(4.21)	3.6667(0.39)
	ncTLF	1.7965(0.48)	1.0628(0.14)	2.7913(0.67)	1.3300(0.17)	6.1035(1.44)	1.9595(0.25)
	Gflasso	2.4767(0.60)	1.2477(0.15)	3.1407(0.68)	1.4103(0.15)	6.0758(1.30)	1.9612(0.23)
	BGGS	**1.3764**(0.35)	**0.9325**(0.12)	**2.1232**(0.52)	**1.1585**(0.15)	**4.6314**(1.05)	**1.7110**(0.21)
(100, 500)	Lasso	241.1811(44.35)	12.3069(1.16)	236.9172(46.60)	12.2308(1.20)	234.0847(41.70)	12.2058(1.12)
	ncTLF	7.9784(2.48)	2.2346(0.36)	8.0775(2.27)	2.245(0.33)	8.9332(2.96)	2.3705(0.39)
	Gflasso	11.5257(3.22)	2.6887(0.38)	11.1470(3.07)	2.6427(0.37)	11.2446(3.42)	2.6672(0.42)
	BGGS	**5.8641**(1.85)	**1.9180**(0.31)	**6.0245**(1.79)	**1.9367**(0.29)	**6.6712**(2.31)	**2.0444**(0.35)
(200, 500)	Lasso	97.5593(12.92)	7.8508(0.52)	99.6449(14.18)	7.9360(0.59)	96.9504(12.47)	7.8677(0.56)
	ncTLF	2.0526(0.35)	1.1375(0.10)	2.7838(0.54)	1.3242(0.13)	4.8786(0.87)	1.7569(0.15)
	Gflasso	2.8889(0.53)	1.3492(0.13)	3.4613(0.66)	1.4774(0.14)	5.1933(0.88)	1.8151(0.16)
	BGGS	**1.5494**(0.26)	**0.9896**(0.09)	**2.1513**(0.43)	**1.1639**(0.11)	**3.8383**(0.68)	**1.5581**(0.14)
(200, 1000)	Lasso	557.5391(70.05)	18.8198(1.24)	565.0438(76.35)	18.9851(1.28)	552.8262(68.37)	18.7333(1.23)
	ncTLF	7.0901(1.57)	2.1162(0.24)	7.7320(1.80)	2.2125(0.26)	8.6426(2.10)	2.3343(0.28)
	Gflasso	9.3120(2.00)	2.4246(0.27)	9.8988(2.22)	2.5045(0.28)	10.3957(2.44)	2.5621(0.30)
	BGGS	**6.3771**(1.53)	**2.0058**(0.25)	**6.7427**(1.53)	**2.0670**(0.24)	**7.1612**(1.75)	**2.1227**(0.25)

In the second example, we consider a more complex situation where $\boldsymbol{\beta}$ contains five clusters. The components of $\boldsymbol{\beta}$ are generated as follows, $\{\beta_j \sim U(-1.5, -1.4), j \in \mathcal{C}_1\}$, $\{\beta_j \sim U(-0.7, -0.6), j \in \mathcal{C}_2\}$, $\{\beta_j \sim U(0.6, 0.7), j \in \mathcal{C}_4\}$, $\{\beta_j \sim U(1.4, 1.5), j \in \mathcal{C}_5\}$ and $|\mathcal{C}_1| = |\mathcal{C}_2| = |\mathcal{C}_4| = |\mathcal{C}_5| = p/16$,

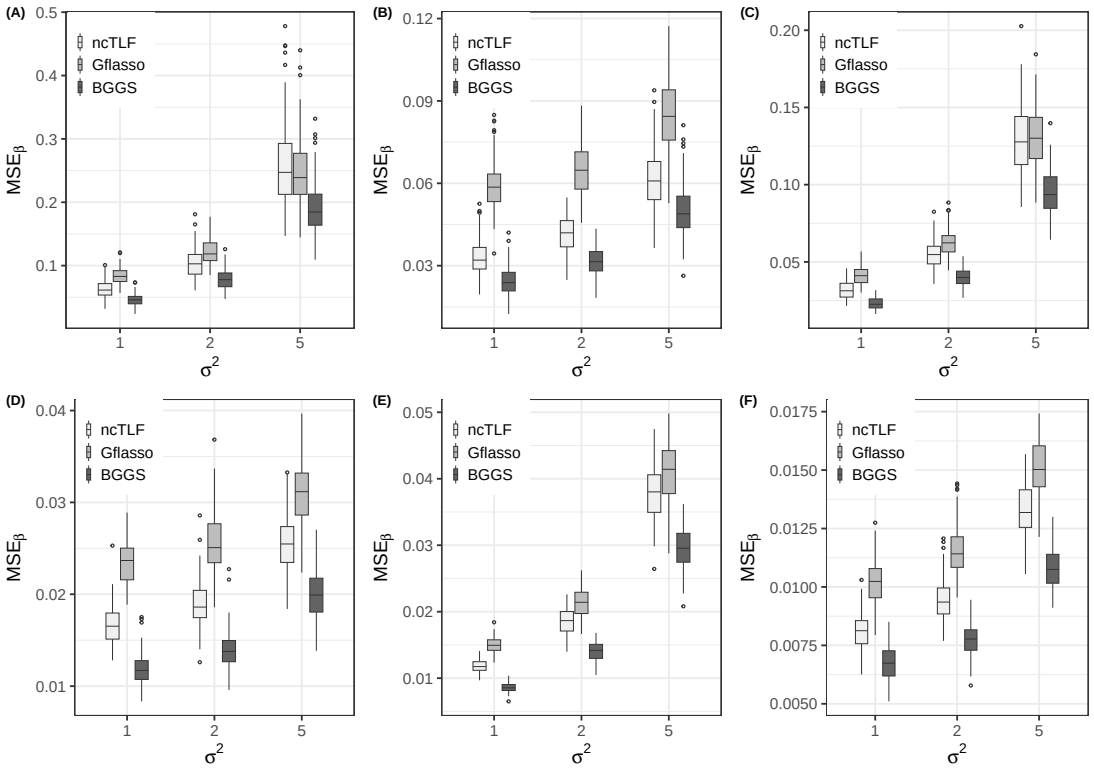

Figure 3.5: Example 2: boxplots of MSE_β estimated via ncTLF, Gflasso, and BGGS over 100 simulations with different pairs of (n, p): (A) $(n, p) = (50, 100)$, (B) $(n, p) = (50, 200)$, (C) $(n, p) = (100, 200)$, (D) $(n, p) = (100, 500)$, (E) $(n, p) = (200, 500)$, and (F) $(n, p) = (200, 1000)$.

$\{\beta_j = 0.2, j \in \mathcal{C}_3\}, |\mathcal{C}_3| = 3p/4$. We consider different pairs of (n, p) and σ^2. The simulation results are presented in Table 3.3 and Figures 3.5-3.6. Under complex scenario involving positive, negative, and small values of β_j, the BGGS method achieves the smallest MSEs and MAEs among the Lasso, ncTLF, Gflasso and BGGS. Furthermore, the results of MAE_β and MSE_β shown in Figures 3.5-3.6 also demonstrate the superior performance of BGGS.

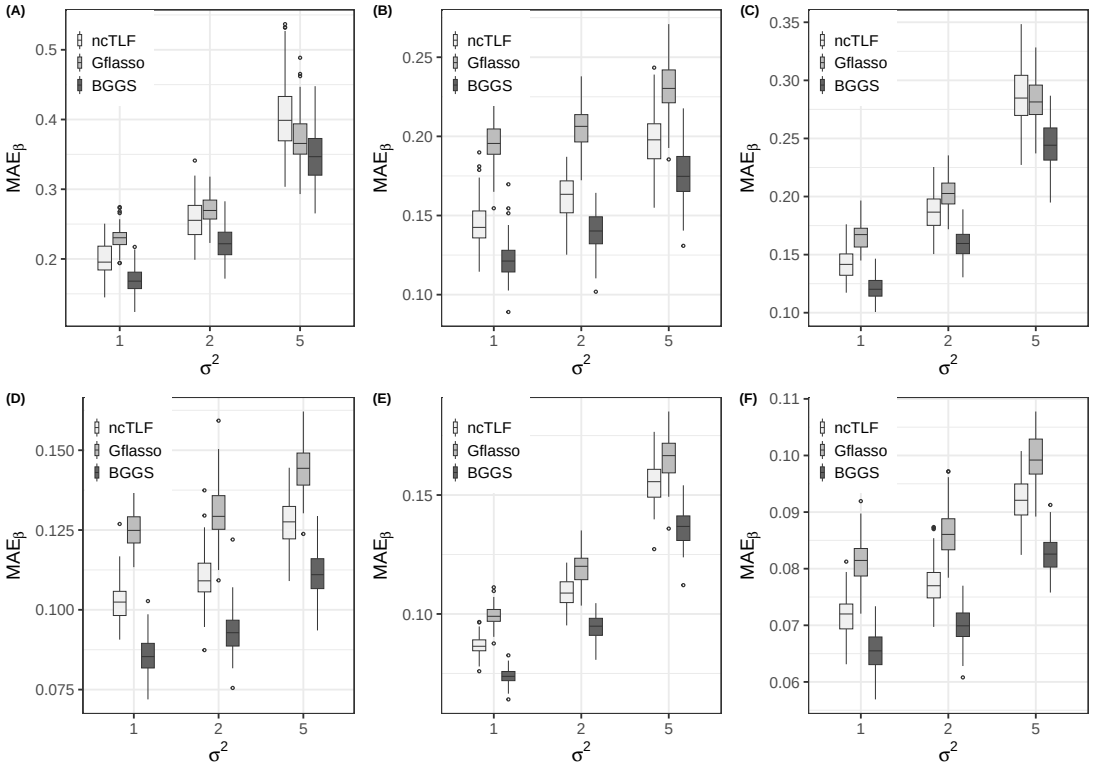

Figure 3.6: Example 2: boxplots of MAE_β estimated via ncTLF, Gflasso, and BGGS over 100 simulations with different pairs of (n, p): (A) $(n, p) = (50, 100)$, (B) $(n, p) = (50, 200)$, (C) $(n, p) = (100, 200)$, (D) $(n, p) = (100, 500)$, (E) $(n, p) = (200, 500)$, and (F) $(n, p) = (200, 1000)$.

3.4 Real data analysis

We further demonstrate the practical performance of the proposed BGGS method using a financial dataset. Our objective is to build a linear regression model that relates the S&P 500 index to the stocks within the S&P 500 list. We assess the predictive accuracy of various methods, including Ordinary Least Squares (OLS), Lasso, ncTLF, Gflasso, and the newly introduced BGGS. Specifically, we collect and transform the weekly closing prices of the S&P 500 index and the stocks listed within it into a time series of weekly log returns. We conduct our study over two recent three-year periods: from January 3, 2017 to December 30, 2019 (Market I), and from January 3,

Table 3.4: Predicted MSE of test sets under different pairs of (n, p) for two three-year periods of January 3, 2017 to December 30, 2019 (Market I) and January 3, 2020 to December 30, 2022 (Market II).

(n, p)	OLS	Lasso	ncTLF	Gflasso	BGGS
Market I					
(100, 50)	0.1665	0.3127	0.1665	0.1565	**0.1432**
(100, 75)	0.2569	0.1402	0.2577	0.1273	**0.1250**
(100, 98)	6.7170	0.1182	0.7740	**0.0885**	0.0930
(100, 158)	–	0.1838	0.1491	**0.0945**	0.1044
(100, 229)	–	0.1440	0.1000	0.1145	**0.0960**
Market II					
(100, 50)	44.9960	41.7844	44.9960	28.4911	**23.2088**
(100, 75)	43.9987	41.9822	43.9982	17.2896	**14.9495**
(100, 98)	75.5661	24.8826	48.6707	14.9381	**8.6421**
(100, 158)	–	19.8430	16.7482	18.4435	**4.1263**
(100, 229)	–	9.4832	5.9774	2.8768	**2.0414**

2020 to December 30, 2022 (Market II). These periods are selected due to the significant impact of the global outbreak of COVID-19 on the financial market, resulting in a change in the underlying model. We focus on stocks belonging to five sectors: Technology, Healthcare, Financial Services, Energy, and Communication Services, with each sector forming a subgroup. Stocks within the same sector often exhibit strong correlations, while those from different sectors tend to have weaker correlations. We anticipate that the selected stocks within each sector have similar contributions

to the S&P 500 index. Naturally, each stock's contribution to the S&P 500 may be relatively small but not negligible. In other words, the regression coefficients do not exhibit sparsity. Relying on methods assuming sparsity may result in misleading model estimates and substantial prediction errors.

After data preprocessing, which involves the removal of stocks or samples with excessive missing values, we are left with a dataset comprising 156 samples and 229 stocks for two distinct time periods. To effectively showcase the practical applicability of various models, we divide the datasets into in-sample and out-sample sets. In the in-sample dataset, the sample size n is fixed at 100. Additionally, we investigate different scenarios with varying numbers of stocks, $p = 50$, 75, 98, 158, and 229. The performance of different methods is compared as p increases. The results of our analysis are presented in Table 3.4, which details the Mean Squared Errors (MSEs) of out-sample sets for various combinations of (n, p) over two distinct three-year periods: January 3, 2017, to December 30, 2019 (Market I) and January 3, 2020, to December 30, 2022 (Market II). The smallest MSEs are highlighted in bold.

In general, we observe that the MSEs for Market II are larger than those for Market I, primarily due to the increased volatility in Market II. It is evident that the predictive accuracy of OLS diminishes as p increases. However, when $p > n$, OLS becomes infeasible. Among the other four methods considered, the proposed BGGS method consistently yields the smallest MSEs in the majority of cases (8 out of 10), with the exceptions occurring when p equals 98 and 158 in Market I. Nonetheless, the MSEs for these two cases (0.0930 and 0.1044) are remarkably close to the two smallest MSEs (0.0885 and 0.0945), which are achieved by the Gflasso method. In essence, both Gflasso and BGGS outperform OLS, Lasso, and ncTLF in terms of prediction accuracy for both financial markets. In Market I, BGGS exhibits behavior similar to Gflasso, as evidenced

by their closely matched MSEs. However, in Market II, characterized by higher volatility, BGGS significantly outperforms Gflasso in terms of prediction accuracy, highlighting the robustness of the BGGS method. These empirical findings showcase the practical utility of BGGS in the context of high-dimensional linear regression models with non-sparse assumptions.

3.5 Summary

This dissertation delves into crucial aspects, such as parameter estimation and prediction, for high-dimensional linear regression models under non-sparse assumptions. To tackle these challenges, we introduce the BGGS method within the Bayesian framework. This method employs a grouping strategy to categorize regression coefficients into k groups, facilitating rapid sampling in high-dimensional space by avoiding the computation of high-dimensional covariance matrices. The computational complexity for each iteration is $n(kp)$, as is the complexity of Skinny Gibbs in Narisetty et al. (2019) that is designed for sparsity. Furthermore, through simulations exploring the selection of the grouping number (k), we find that within an appropriate range, the impact of k on parameter estimation is negligible. More precisely, we advocate the use of the "Elbow plot" approach for selecting k. Theoretical analysis, contingent upon regular conditions, reveals that BGGS achieves both weak and strong model selection consistency, parameter estimation consistency and mean squared prediction error bounds. In numerical analysis, our proposed method exhibits superior predictive accuracy across various simulation seetings. Our simulations demonstrate that the BGGS method consistently yields the smallest Mean Squared Error (MSE) and Mean Absolute Error (MAE) compared to alternative methods under diverse scenarios. Finally, the analysis of a financial dataset illustrates the outperformance of BGGS in prediction under com-

plex real-world situations where sparsity is often unknown. This suggests that when confronted with problems featuring more predictors, it may be more advantageous to utilize the entirely of the available data information, rather than disregarding a portion of predictors.

In practice, justifying the group structure is challenging without background knowledge or additional information. The high correlation and the approximate eigenvalues of covariate matrix may indicate group structure. But we can't say the group structure estimated by our BGGS methods is correct, even we have a better estimation results than other methods. But we provide a new attempt towards extension the limits of the size of non-zero parameters. In practice, justifying the group structure is challenging without background knowledge or additional information. High correlations and approximate eigenvalues in the covariate matrix may suggest a group structure, but we cannot assert that the group structure estimated by our BGGS methods is entirely accurate, even if our estimation outperforms other methods. However, our approach offers a novel attempt to extend the limits on the size of non-zero parameters, contributing to advancements in this area.

4 Conclusions and Future Work

4.1 Conclusions

Chapter 2 of this dissertation addresses the challenge of data integration for correlated data collected across multiple platforms, modeling the linear relationships between responses and predictors within each platform. By incorporating random errors from a broader class of sub-Gaussian and sub-exponential distributions, we extend traditional frameworks to identify key predictors across platforms, even as the number of predictors and observations grows infinitely large.

The proposed method integrates marginal response densities from multiple platforms into a composite likelihood and introduces a Bayesian model selection criterion based on composite posterior probabilities. We demonstrate the consistency of this criterion under regularity conditions, ensuring reliable model selection even with diverging model sizes. Furthermore, the method effectively recovers the union support of predictors, capturing those relevant in at least one platform, even when models differ across platforms. To achieve this, an efficient Monte Carlo Markov Chain (MCMC) algorithm is implemented.

Simulation studies highlight that data integration across platforms significantly enhances model selection accuracy compared to single-source analyses. A real-world application to financial data further supports the utility of the method, combining information from three financial indices to identify a shared set of important predictors. The resulting predictive model demonstrates higher

accuracy and lower mean squared error than models relying on individual data sources.

Chapter 3 mainly discusses that in high-dimensional linear regression models, the commonly assumed sparsity of regression coefficients $\boldsymbol{\beta} \in \Re^p$ can lead to significant bias when most or all coefficients are nonzero. To address this limitation, we propose the Bayesian Grouping-Gibbs Sampling (BGGS) method, which relaxes the sparsity assumption by partitioning $\boldsymbol{\beta}$ into k distinct groups. This grouping framework facilitates efficient sampling and modeling in high-dimensional spaces where sparsity-based methods may fail.

Through simulations, we provide guidance on selecting k, recommending the use of an "elbow plot" for reliable determination. Theoretical guarantees affirm that the BGGS method achieves model selection consistency, parameter estimation consistency, and bounded prediction error under regularity conditions. Numerical experiments on finite examples further highlight the competitive edge of BGGS in both parameter estimation and predictive accuracy.

In a practical application to financial data, the BGGS method demonstrated its effectiveness by identifying robust predictive models without relying on sparsity assumptions. This result underscores the method's utility in scenarios where traditional approaches are prone to bias or fail to capture the underlying data structure.

In conclusion, Chapter 2 provides a robust framework for data integration in high-dimensional settings, offering consistent model selection and improved prediction accuracy through the union of information across platforms. This approach has broad applicability and offers significant benefits in multi-source predictive modeling scenarios. In Chapter 3, the BGGS method offers a powerful alternative to sparsity-based approaches, providing flexibility and accuracy in high-dimensional regression settings. Its ability to accommodate dense coefficient structures and deliver consistent results makes it a valuable tool for a wide range of applications, particularly in fields where

predictive performance and model robustness are critical.

4.2 Future work

This dissertation has introduced a framework for multivariate statistical modeling, particularly in the context of high-dimensional data with group structures. While the proposed methods provide significant advancements, several extensions and refinements remain to be explored in future research.

One key direction involves extending the current framework to more complex scenarios where response variables follow multivariate distributions and exhibit structured interdependencies. Existing methodologies often treat response variables independently, neglecting intrinsic group structures that arise naturally in biological, medical, and financial applications. Developing models that simultaneously account for both predictor and response group structures remains an open challenge. A promising approach is to incorporate a grouped penalty into our method to better capture these dependencies and improve interpretability.

High-dimensional statistical analyses, such as genomic association studies, frequently encounter challenges related to sparsity, where only a small subset of predictors is truly relevant to the response. Regularization techniques, including the lasso (Tibshirani (1996)), adaptive lasso (Zou (2006)), elastic net (Zou and Hastie (2005)), and smoothly clipped absolute deviation (SCAD) penalties (Fan and Li (2001)), have been widely employed to enhance model interpretability and predictive performance. While these methods have been extensively studied for univariate response settings, their extension to multivariate responses, particularly with complex group structures, warrants further investigation. Developing new regularization strategies that leverage both response

and predictor dependencies is a critical area for future work.

Several methodologies have been proposed to address these challenges in multivariate regression. For example, mixed l_1/l_2 penalties have been employed to identify hub covariates in gene regulatory networks (Peng, J. et al. (2010)), while Lounici et al. (2011) established oracle inequalities in the multitask learning framework. Despite these advancements, there has been limited focus, to our knowledge, on cases where the responses themselves exhibit group structures, even though such cases are common in biological studies. Future research should focus on refining these techniques to incorporate biologically meaningful groupings in a statistically rigorous manner.

A common approach for analyzing multivariate responses is to perform covariate selection for each response variable independently. While this strategy can account for the predictor group structure, it often neglects the intrinsic group structure present in the responses. This oversight can lead to suboptimal results, particularly when the response group structure is critical for understanding the underlying biological or functional relationships. Addressing this gap by developing methods that simultaneously consider both predictor and response group structures remains a pressing challenge in multivariate statistical analysis.

Li et al. (2015) proposed a multivariate sparse group lasso method for variable selection and estimation, designed to handle data with both high-dimensional predictors and high-dimensional response variables. This method was implemented within the framework of a penalized multivariate multiple linear regression model, allowing for an arbitrary group structure to be imposed on the regression coefficients. By incorporating both sparsity and group-level structure into the penalty design, the proposed approach enables simultaneous selection of relevant predictors while accounting for the inherent group relationships within the data. This dual consideration is particularly advantageous in applications where predictors and responses exhibit complex interdependencies,

such as genomic studies or functional neuroimaging analyses. The flexibility of our model ensures its applicability to a wide range of high-dimensional problems, providing a robust tool for uncovering meaningful associations in data with group structures.

Beyond the previously discussed models, future work could also consider more complex situations, such as formulating the full likelihood for multivariate clustered data, which can be a complex and computationally infeasible task. The Generalized Estimating Equation (GEE) framework provides a viable alternative, ensuring consistency even when the working correlation matrix is misspecified. Li et al. (2015) introduced a multivariate sparse group lasso method for variable selection and estimation in scenarios involving both high-dimensional predictors and high-dimensional response variables within the GEE framework. Extending this framework by introducing an additional dimension, referred to as the platform dimension, presents a significant computational and theoretical challenge that warrants further exploration. This extension accommodates multiple data sources and accounts for the possibility of varying underlying influential covariates across platforms.

In conclusion, future research will aim to enhance the robustness, flexibility, and applicability of high-dimensional statistical models by incorporating multivariate responses, complex group structures, and more general assumptions. Theoretical advancements, coupled with empirical validation through large-scale simulations and real-world applications, will be essential for refining these methodologies and expanding their impact across various domains, including biology, finance, and healthcare.

Bibliography

Balan, R. M. and I. Schiopu-Kratina (2005). Asymptotic results with generalized estimating equations for longitudinal data. The Annals of Statistics 33(2), 522–541.

Berger, J.O. , Ghosh, J.K. , and Mukhopadhyay, N. Approximations and consistency of bayes factors as model dimension grows. Journal of Statistical Planning and Inference, 112(1):241–258, 2003.

Bondell, H. D. and Reich, B. J.(2012), Consistent high-dimensional Bayesian variable selection via penalized credible regions. Journal of the American Statistical Association, 107(500):1610–1624.

Bunea, F., She, Y., and Wegkamp, M. (2011). Optimal selection of reduced rank estimators of high-dimensional matrices. An nals of Statistics, 39, 1282—1309.

Caruana, R. Multitask Learning. *Machine Learning*, 28, 41–75, 1997.

Casella, G., Girón, F.J., Martínez, M.L., and Moreno, E. Consistency of Bayesian procedures for variable selection. *The Annals of Statistics*, 37(3):1207 – 1228, 2009.

Chandler, R.E. and Bate, S. Inference for clustered data using the independence loglikelihood. *Biometrika*, 94(1):167–183, 03 2007.

Chen, R.B., Chu,C.H.. Yuan, S. and Wu, Y.(2015) Bayesian sparse group selection. Journal of Computational and Graphical Statistics, 25:00–00, 05.

Clyde, M., Parmigiani, G., and Vidakovic, B. Multiple shrinkage and subset selection in wavelets. *Biometrika*, 85(2):391–401, 06 1998.

Clyde, M. and George, E.I. Flexible empirical bayes estimation for wavelets. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 62(4):681–698, 2000.

Cooley, D., Nychka, D., and Naveau, P. Bayesian spatial modeling of extreme precipitation return levels. *Journal of the American Statistical Association*, 102(479):824–840, 2007.

Cox, D.R. and Reid, N. A note on pseudolikelihood constructed from marginal densities. *Biometrika*, 91(3):729–737, 2004.

Dai, G., U Müller, U., and Carroll, R.J. Data integration in high dimension with multiple quantiles. *arXiv preprint arXiv:2006.16357*, 2020.

Ding, H., Qin, S., Wu, Y., and Wu, Y. (2021). Asymptotic properties on high-dimensional multivariate regression M-estimation. Journal of Multivariate Analysis, 183, 104730.

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

Fernández, C., Ley, E., and Steel, M.F.J. Model uncertainty in cross-country growth regressions. *Journal of Applied Econometrics*, 16(5):563–576, 2001.

Friedman, J., Hastie, T., Tibshirani, R., Narasimhan, B., Tay, K., Simon, N., & Qian, J. (2021). Package 'glmnet'. CRAN R Repository, 595.

Gao, X. and Carroll, R.J. Data integration with high dimensionality. *Biometrika*, 104(2):251–272, 05 2017.

Gao, X. and Zhong, Y. FusionLearn: a biomarker selection algorithm on cross-platform data. *Bioinformatics*,35(21):4465-4468, 2019.

George, E.I. and McCulloch, R.E. Variable selection via gibbs sampling. *Journal of the American Statistical Association*, 88(423):881–889, 1993.

George, E.I. and Foster,D.P.(2000). Calibration and empirical bayes variable selection. Biometrika, 87(4):731-747.

Götze, F., Sambale, H., and Sinulis, A. Concentration inequalities for polynomials in a-sub-exponential random variables. *Electronic Journal of Probability*, 26(none):1 – 22, 2021.

Hsu, D., Kakade, S., and Tong Zhang, T. (2012). A tail inequality for quadratic forms of subgaussian random vectors. Electron. Commun. Probab. 17, 1-6, 2012.

Ishwaran, H. and Rao, J.S.(2005), Spike and slab gene selection for multigroup microarray data. Journal of the American Statistical Association, 100(471):764–780.

Johnson, V.E. and Rossell, D.(2012) Bayesian model selection in high-dimensional settings. Journal of the American Statistical Association, 107(498):649–660.

Kent, J. Robust properties of likelihood ratio tests. *Biometrika*, 69: 19-27, 1982.

Lee, J. and Goh, G. (2023) A hybrid deterministic–deterministic approach for high-dimensional Bayesian variable selection with a default prior. Computational Statistics, 05.

Lee, K. and Cao, X. (2020) Bayesian group selection in logistic regression with application to mri data analysis. Biometrics, 77.

Li, B. (1996). A minimax approach to consistency and efficiency for estimating equations. The Annals of Statistics 24(3), 1283–1297.

Li, F., and Zhang, N.R. (2010). Bayesian variable selection in structured high-dimensional covariate spaces with applications in genomics. Journal of the American statistical association, 105(491), 1202-1214.

Li, X., Wang, J., Tan, J. et al. A graph neural network-based stock forecasting method utilizing multi-source heterogeneous data fusion. *Multimed Tools Appl*, 2022.

Li, Y., Nan, B., & Zhu, J. (2015). Multivariate Sparse Group Lasso for the Multivariate Multiple Linear Regression with an Arbitrary Group Structure. Biometrics, 71(2), 354–363.

Li, Y., Gao, X. and Xu.,W. (2023) Statistical Consistency of High Dimensional Generalized Estimating Equation with L1 regularization. Canadian Journal of Statistics. under review.

Liang, F., Paulo, R., Molina, G., Clyde, M. A. & Berger, J. O. (2008). Mixtures of g-Priors for Bayesian Variable Selection, Journal of the American Statistical Association, 103(481), 410-423.

Liang, K.Y. and S. L. Zeger (1986). Longitudinal data analysis using generalized linear models. Biometrika, 73(1), 13–22.

Lindsay, B.G. Composite likelihood methods. *Contemporary Mathematics*, 80(1):221–239, 1988.

Lounici, K., Pontil, M., Tsybakov, A. B., and van de Geer, S. (2011). Oracle inequalities and optimal inference under group sparsity. Annals of Statistics, 39, 2164—2204.

Maity, A.K. , Carroll, R.J. , and Mallick, B.K. Integration of survival and binary data for variable selection and prediction: a bayesian approach. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 68(5):1577–1595, 2019.

McCullagh, P. and J. A. Nelder (1989). Generalized Linear Models (2nd ed.), Routledge

Moreno, E. and Girón, F.J. Consistency of bayes factors for intrinsic priors in normal linear models. *Comptes Rendus Mathematique*, 340(12):911–914, 2005.

Moreno, E., Girón, F.J., and Casella, G. Consistency of objective Bayes factors as the model dimension grows. *The Annals of Statistics*, 38(4):1937 – 1952, 2010.

Narisetty, N.N., Shen, J., and He, X. (2019). Skinny Gibbs: A consistent and scalable Gibbs sampler for model selection. Journal of the American Statistical Association, 114(527), 1205-1217.

Obozinski, G., Wainwright, M., and Jordan, M. (2011). Support union recovery in high-dimensional multivariate regression. Annals of Statistics, 39, 1—47.

Park, T. and D'Amico, S. (2024) Robust Multi-Task Learning and Online Refinement for Spacecraft Pose Estimation across Domain Gap. Advances in Space Research, 73, 11,5726-5740,2024.

Peng, J., Zhu, J., Bergamaschi, A., Han, W., Noh, D.-Y., Pollack, J. R., and Wang, P. (2010). Newblock regularized multivari ate regression for identifying master predictors with applica tion to integrative genomics study of breast cancer. Annals of Applied Statistics, 4, 53—77.

Ramesh, L. and Murthy, C. R. and Tyagi, H. Sample-measurement tradeoff in support recovery under a subgaussian prior. *Proc. IEEE Int. Symp. Inf. Theory,*, 2709–2713, 2019.

Ramesh, L. and Murthy, C. R. and Tyagi, H. Multiple Support Recovery Using Very Few Measurements Per Sample. *IEEE Transactions on Signal Processing*, 70:2193-2206, 2022.

Ribatet, M., Cooley, D., and Davison, A.C. Bayesian inference from composite likelihoods, with an application to spatial extremes. *Statistica Sinica*, 22(2):813–845, 2012.

Shang, Z. and Clayton, M.K. Consistency of bayesian linear model selection with a growing number of parameters. *Journal of Statistical Planning and Inference*, 141(11):3463–3474, 2011.

Shang, Z. and Li, P. Bayesian high-dimensional screening via mcmc. *Journal of Statistical Planning and Inference*, 155:54–78, 2014.

Shankar, J., Szpakowski, S., Solis, N. V., Mounaud, S., Liu, H., Losada, L., Nierman, W. & Filler, S. G. (2015). A systematic evaluation of high-dimensional, ensemble-based regression for exploring large model spaces in microbiome analyses. BMC bioinformatics, 16, 1-18.

Shen, X., Huang, H. C., & Pan, W. (2012). Simultaneous supervised clustering and feature selection over a graph. Biometrika, 99(4), 899-914.

Shen, X., Sun, Y. and Julie Langou, J. (2015) Package 'FGSG'. Available at http://cran.nexr.com/web/packages/FGSG/

Spokoiny, V. Sharp deviation bounds for quadratic forms, Math. Meth. Stat. 22, 100–113 (2013).

Tang, G. and Nehorai, A. Performance analysis for sparse support recovery. *IEEE Trans. Inf. Theory*, 56(3):1383–1399, 2010.

Tian, Q., Stepaniants, S., Mao, M., Weng, L., Feetham, M., Doyle, M., Yi, E., Dai, H., Thorsson, V., Eng, J., Goodlett, D., Berger, J., Gunter, B., Linseley, P., Stoughton, R., Aebersold, R., Collins, S., Hanlon, W., Hood, L. Integrated genomic and proteomic analyses of gene expression in mammalian cells. *Mol Cell Proteomics*, 3: 960—969, 2004.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. J. R. Stat. Soc. Ser. B Stat. Methodol. 58 267–288.

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

Varin, C., Reid, N., and Firth, D. An overview of composite likelihood methods. *Statistica Sinica*, 21(1):5–42, 2011.

Wainwright, M.J. Information-theoretic limits on sparsity recovery in the high-dimensional and noisy setting. *IEEE transactions on information theory*, 55(12):5728–5741, 2009.

Wang, J., Cai, X.Z. and Li, R.Z. (2021) Variable selection for partially linear models via Bayesian subset modeling with diffusing prior. Journal of Multivariate Analysis, 183:104733.

Witten, D. M., Shojaie, A., and Zhang, F. (2014). The cluster elastic net for high-dimensional regression with unknown variable grouping. Technometrics, 56(1), 112-122.

Wolfe, P.J., Godsill, S.J., and Ng, W. Bayesian variable selection and regularization for time–frequency surface estimation. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 66(3):575–589, 2004.

Wu, S., Xu, Y., Feng, Z. et al. Multiple-platform data integration method with application to combined analysis of microarray and proteomic data. *BMC Bioinformatics*, 13, 320, 2012.

Xie, M. and Y. Yang (2003). Asymptotics for eneralized estimating equations with large cluster sizes. The Annals of statistics 31(1), 310–347.

Yi, J.T. and Tang, N.S.(2022) Variational Bayesian inference in high-dimensional linear mixed models. Mathematics, 10:463.

Yuan, M., & Lin, Y. (2006). Model selection and estimation in regression with grouped variables. Journal of the Royal Statistical Society Series B: Statistical Methodology, 68(1), 49-67.

Zellner, A. (1978). Jeffreys-Bayes posterior odds ratio and the Akaike information criterion for discriminating between models. Economics Letters, 1(4), 337-342.

Zhang, C. H. (2010), Nearly Unbiased Variable Selection Under MinimaxConcave Penalty, Annals of Statistics, 38, 894–942.

Zhao, J., Zhou, Y., & Liu, Y. (2023). Estimation of Linear Functionals in High Dimensional Linear Models: From Sparsity to Non-sparsity. Journal of the American Statistical Association, (just-accepted), 1-13.

Zhao, P. and Yu, B. On Model Selection Consistency of Lasso. *Journal of Machine Learning Research*, 7(90):: 2541–2563, 2006.

Zhu, Y., Shen, X., & Pan, W. (2013). Simultaneous grouping pursuit and feature selection over an undirected graph. Journal of the American Statistical Association, 108(502), 713-725.

Zou, H. and Hastie, T. (2005). Regularization and Variable Selection Via the Elastic Net. Journal of the Royal Statistical Society Series B: Statistical Methodology, 67(2), 301–320.

Zou, H. (2006). The adaptive lasso and its oracle properties. Journal of the American statistical association, 101(476), 1418-1429.