Grouped Variable Selection via Nested Spike and Slab Priors

Tso-Jung Yen[*]
Institute of Statistical Science
Academia Sinica

Yu-Min Yen[†]
Department of Finance
London School of Economics and Political Science

June 30, 2011

Abstract

Abstract. In this paper we study grouped variable selection problems by proposing a specified prior, called the nested spike and slab prior, to model collective behavior of regression coefficients. At the group level, the nested spike and slab prior puts positive mass on the event that the l_2-norm of the grouped coefficients is equal to zero. At the individual level, each coefficient is assumed to follow a spike and slab prior. We carry out maximum a posteriori estimation for the model by applying blockwise coordinate descent algorithms to solve an optimization problem involving an approximate objective modified by majorization-minimization techniques. Simulation studies show that the proposed estimator performs relatively well in the situations in which the true and redundant covariates are both covered by the same group. Asymptotic analysis under a frequentist's framework further shows that the l_2 estimation error of the proposed estimator can have a better upper bound if the group that covers the true covariates does not cover too many redundant covariates. In addition, given some regular conditions hold, the proposed estimator is asymptotically invariant to group structures, and its model selection consistency can be established without imposing irrepresentable-type conditions.

Keywords: Log-sum approximation; Majorization-minimization algorithms; Subgradients; Group sparsity.

[*] Postdoctoral Fellow, Institute of Statistical Science, Academia Sinica. E-mail: tjyen@stat.sinica.edu.tw

[†] PhD Candidate, Department of Finance, London School of Economics and Political Science, Houghton Street, London WC2A 2AE, UK. E-mail: Y.YEN@lse.ac.uk.

1 Introduction

Variable selection has long been an important issue in regression-based statistical analysis. Recently, many efficient methods have been developed to tackle the problems in the situation when the number of covariates is large. At the same time, many efforts have also been made in understanding the statistical properties of these methods. In this paper we focus on grouped variable selection problems. More specifically, we study variable selection in the following regression model:

$$y_i = \left(\sum_{j \in G_1} x_{ij}\beta_j\right) + \left(\sum_{j \in G_2} x_{ij}\beta_j\right) + \cdots + \left(\sum_{j \in G_m} x_{ij}\beta_j\right) + \epsilon_i, \tag{1.1}$$

where y_i is the response variable for subject i, $G_k \subseteq \{1, 2, \cdots, p\}$ is the index set associated to the kth group, and ϵ_i is the corresponding error term following some specified distribution. Throughout the paper, we focus on non-overlapping cases, i.e. for two index sets G_k and $G_{k'}$ with $k, k' \in \{1, 2, \cdots, m\}$, we assume $G_k \cap G_{k'} = \emptyset$ for $k \neq k'$. Now let β_{G_k} denote the regression vector with entries indexed by G_k. Grouped variable selection aims to select covariates groupwisely, that is, entries in β_{G_k} are either estimated with non-zero values or they are all estimated with zero values. In grouped variable selection, one benchmark method for estimating $\beta = (\beta_{G_1}, \beta_{G_2}, \cdots, \beta_{G_m})$ is the group lasso [32]:

$$\widehat{\beta}_{\text{GL}} \quad = \quad \arg\min_{\beta} \left\{ \frac{1}{2} \left|\left| y - \sum_{k=1}^{m} X_{G_k}\beta_{G_k} \right|\right|_2^2 + \lambda \sum_{k=1}^{m} w_k ||\beta_{G_k}||_2 \right\}, \tag{1.2}$$

where X_{G_k} is an $n \times |G_k|$ matrix representing the covariates indexed by G_k, $\lambda \geq 0$ is the tuning parameter, and w_k is a specified weight corresponding to the kth group.

The group lasso estimator (1.2) has several advantages over the lasso in dealing with the variable selection problem associated with model (1.1). First, since the l_2-norm $||\beta_{G_k}||_2$ is not separable in β_{G_k}, the group lasso provides a more suitable way for regression coefficient estimation when either covariates have meaningful interpretations as a whole [19, 22, 8], or they can be expressed as a group of dummy variables [32], or they are represented as linear combinations of basis functions [2, 23, 14]. In addition, as shown in [13, 16], given some regular conditions hold, the l_2 estimation error of (1.2) can have an order of magnitude similar or even smaller than that of the lasso estimator. Moreover, like the lasso, (1.2) can also enjoy model selection consistency if some irrepresentable-type conditions [33] are satisfied [2, 23, 22, 16].

Note that the group lasso estimator (1.2) is only able to produce between-group-sparsity, that is, once the l_2-norm $||\beta_{G_k}||_2$ is estimated with a non-zero value, all entries in β_{G_k} will be estimated with non-zero values. However, sometimes the pre-specified group structure may not exactly cover the true covariates. As a result of that, redundant covariates may be wrongly selected in the model, along with the true covariates. To correct this, one need to consider within-group-sparsity. Friedman et al. [10] proposed the sparse group lasso estimation by adding an l_1 penalty to the objective function stated in (1.2). Under the sparse group lasso estimation, within-group-sparsity can be reached, since with the l_1 penalty the regression coefficients in the active groups are allowed to have zero-valued estimates.

In this paper we will study the grouped variable selection problem by developing a specified spike and slab prior [21], called the nested spike and slab prior, to model the group regression coefficient vector β_{G_k}. The nested spike and slab prior assigns positive mass on events $\{||\beta_{G_k}||_2 = 0\}$ and $\{||\beta_{G_k}||_2 \neq 0\}$ to represent the sparsity between group coefficient vectors $\beta_{G_1}, \beta_{G2}, \cdots, \beta_{G_m}$. Given that $||\beta_{G_k}||_2 \neq 0$, it further assigns each entry in β_{G_k} with a spike and slab prior [21]. Under the nested spike and slab prior, sparsity between groups and sparsity within a group can be achieved simultaneously with a positive probability.

We then develop a method to carry out maximum a posteriori (MAP) estimation for the model. More specifically, we formulate the estimation problem as an optimization problem in which the objective function is approximated by the majorization-minimization algorithms [15, 30]. We then solve the optimization problem by proposing blockwise coordinate descent algorithms based on the ideas developed in [10, 9]. Simulation studies show that the proposed estimator performs relatively well in the situations in which the within-group-sparsity is present. However, its performance may get deteriorated if the true covariates are scattered over a large number of groups that contain many redundant covariates.

Further we will show that under a frequentist's framework, the proposed MAP estimator can have a better l_2 estimation error bound if the number of groups that cover the true covariates and the numbers of redundant covariates in such groups are small. In addition, if some regular conditions on tuning parameters hold, the values of the proposed estimates will be asymptotically invariant to group structures. We will also establish model selection consistency for the proposed estimator. The result does not require one to impose the irrepresentable-type conditions.

The paper is organized as follows. In Section 3 we develop the nested spike and slab prior and construct a Bayesian hierarchical model based on the proposed prior. We then present a method to carry out maximum a posteriori estimation for the model. In Section 4 we conduct a simulation study to demonstrate finite sample properties of the proposed estimator. In Section 5 we establish asymptotic results for the proposed estimator under a frequentist's framework. Section 6 contains two real data examples. Section 7 is the discussion.

2 Notation

For the kth index set G_k, we let q_k denote the number of elements in it, i.e. $q_k = |G_k|$. For the jth covariate, we let k_j denote the index for the group that j belongs to, that is, if $j \in G_{k'}$, then $k_j = k'$. For a p-dimensional vector $b = (b_1, b_2, \cdots, b_p)$, we define $(b)_j = b_j$ and b_{G_k} be the vector whose entries are those indexed by G_k in b. For the vector b, we define the associated l_1-norm by $||b||_1 = \sum_{j=1}^{p} |b_j|$ and l_2-norm by $||b||_2 = (\sum_{j=1}^{p} |b_j|^2)^{1/2}$. We define the sign function of z by $\text{sign}(z) = 1$ if $z > 0$; $\text{sign}(z) = -1$ if $z < 0$; $\text{sign}(z) = 0$ if $z = 0$. Finally, we define the soft-thresholding operator ST_λ by

$$ST_\lambda(z) = \text{sign}(z)(|z| - \lambda)_+. \tag{2.1}$$

3 Nested spike and slab prior

Since our aim is to jointly select covariates indexed by G_k, therefore the information about whether β_{G_k} is a zero vector or not is crucial. Practically, we assign probability mass on event $\{||\beta_{G_k}||_2 \neq 0\}$ to express our belief that β_{G_k} is not a zero vector. Let θ_k denote the probability. With θ_k, we further assume β_{G_k} follows a distribution which has a density given by

$$\begin{aligned} f(\beta_{G_k}) &= \theta_k \left\{ \prod_{j \in G_k} \Big[\omega_j g(\beta_j) + (1 - \omega_j)\delta_{(-\xi,\xi)}(\beta_j) \Big] \right\} \\ &\quad + (1 - \theta_k)\delta_0(||\beta_{G_k}||_2), \end{aligned} \tag{3.1}$$

where $\omega_j \in [0, 1]$, $g(\beta_j)$ is some specified density defined on $\mathbb{R}\backslash(-\xi, \xi)$, and $\delta_0(||\beta_{G_k}||_2)$ is the Dirac delta function centered at event $\{||\beta_{G_k}||_2 = 0\}$. The density (3.1) is called the nested spike and slab prior, since the joint spike and slab prior assigned on entries in β_{G_k} at the individual level is wrapped by a spike and slab prior assigned at the group level. The nested spike and slab prior (3.1) implies that β_{G_k} has probability θ_k to be a non-zero vector. In addition, given that β_{G_k} is not a zero vector, the entries in β_{G_k} are independently distributed, and each entry will have probability ω_j to follow a distribution with density $g(\beta_j)$ and probability $1 - \omega_j$ to fall uniformly in the region $(-\xi, \xi)$.

For practical purposes, we introduce two sets of Bernoulli variables $\gamma = (\gamma_1, \gamma_2, \cdots, \gamma_m)$ and $\alpha = (\alpha_1, \alpha_2, \cdots, \alpha_p)$. The former will be used to model regression coefficients at the group level while the latter will be used to model regression coefficients at the individual level. Below we reformulate the nested spike and slab prior (3.1) in terms of α and γ. For group k, we let $\gamma_k \sim \text{Bernoulli}(\theta_k)$. For $j \in G_k$, we assume $\alpha_j | \gamma_k = 1 \sim \text{Bernoulli}(\omega_j)$. Here α_j is defined conditional on $\gamma_k = 1$, reflecting the nested structure of (3.1). Now conditional on γ_k and α_{G_k}, the density $f(\beta_{G_k}| \gamma_k, \alpha_{G_k})$ has the same format as the nested spike and slab prior (3.1) with θ_k replaced by γ_k and ω_j replaced by α_j. Further it can be shown that the expectation $\mathbb{E}_{\gamma_k,\alpha_{G_k}}[f(\beta_{G_k}|\gamma_k, \alpha_{G_k})]$ is the nested spike and slab prior (3.1). In addition, given γ_k and α_{G_k} are known, the prior density $f(\beta_{G_k}| \gamma_k, \alpha_{G_k})$ has an equivalent representation:

$$f(\beta_{G_k}| \gamma_k, \alpha_{G_k}) = \left\{ \prod_{j \in G_k} g(\beta_j)^{\alpha_j} \delta_{(-\xi,\xi)}(\beta_j)^{1-\alpha_j} \right\}^{\gamma_k} \delta_0(||\beta_{G_k}||_2)^{1-\gamma_k}. \tag{3.2}$$

Below we will use the augmented form (3.2) to derive the joint posterior density of β, α and γ.

3.1 Model

We now turn back to regression model (1.1). With the prior setting given above, we can construct a hierarchical Bayesian model and carry out inference on parameters in (1.1). For practical purposes, we will only focuses on a situation in which the region $(-\xi, \xi)$ is a small region concentrating around 0, that is, $\xi \to 0$. Under this situation, we can represent (1.1) in terms of Bernoulli variables α and γ by $y_i = \sum_{k=1}^m \gamma_k (\sum_{j \in G_k} x_{ij} \alpha_j \beta_j) + \epsilon_i$. Given that there are n subjects, we assume

$$\begin{aligned}
y_i |\, X, \beta, \alpha, \gamma, \sigma^2 &\sim \text{Normal}\left\{ \sum_{k=1}^m \gamma_k \left(\sum_{j \in G_k} x_{ij} \alpha_j \beta_j \right), \sigma^2 \right\}, \text{ for } i = 1, 2, \cdots, n, \\
\beta_{G_k} |\, \alpha_{G_k}, \gamma_k, \sigma^2, \lambda &\sim \gamma_k \Bigg[\prod_{j \in G_k} \Big\{ \alpha_j \text{Normal}(0, \sigma^2 \lambda^{-1}) \mathbb{I}\{\mathbb{R} \setminus (-\xi, \xi)\} \\
&\qquad + (1 - \alpha_j) \delta_{(-\xi,\xi)}(\beta_j) \Big\} \Bigg] \\
&\qquad + (1 - \gamma_k) \delta_0(||\beta_{G_k}||_2), \text{ for } k = 1, 2, \cdots, m, \\
\alpha_{G_k} |\, \gamma_k, \omega_{G_k} &\sim \left\{ \prod_{j \in G_k} \text{Bernoulli}(\omega_j) \right\}^{\gamma_k} \delta_0(\alpha_{G_k})^{1-\gamma_k}, \text{ for } k = 1, 2, \cdots, m, \\
\gamma_k |\, \theta_k &\sim \text{Bernoulli}(\theta_k), \text{ for } k = 1, 2, \cdots, m. \qquad (3.3)
\end{aligned}$$

Under hierarchical Bayesian model (3.3), the joint posterior density of β, α and γ is given by

$$\begin{aligned}
&f(\beta, \alpha, \gamma |\, y, X, \lambda, \sigma^2, \omega, \theta) \\
&\qquad \propto f(y |\, X, \beta, \alpha, \gamma, \sigma^2) f(\beta |\, \alpha, \gamma, \sigma^2, \lambda) f(\alpha |\, \gamma, \omega) f(\gamma |\, \theta), \qquad (3.4)
\end{aligned}$$

where $y = (y_1, y_2, \cdots, y_n)$, and for notational simplicity, similar definitions are applied to ω and θ. With the joint posterior density (3.4), various methods can be proposed to make inference on the parameters. Here we adopt the maximum a posteriori (MAP) approach to carrying out the parameter estimation. We define the maximum a posteriori estimator for β, α and γ by

$$(\widehat{\beta}, \widehat{\alpha}, \widehat{\gamma}) = \arg\min_{\beta, \alpha, \gamma} -2 \log f(\beta, \alpha, \gamma |\, y, X, \lambda, \sigma^2, \omega, \theta),$$

where

$$\begin{aligned}
&-2 \log f(\beta, \alpha, \gamma |\, y, X, \lambda, \sigma^2, \omega, \theta) \\
&\qquad = -2 \log f(y |\, X, \beta, \alpha, \gamma, \sigma^2) \\
&\qquad\quad -2 \log\{ f(\beta |\, \alpha, \gamma, \sigma^2, \lambda) f(\alpha |\, \gamma, \omega) f(\gamma |\, \theta) \} \\
&\qquad\quad -2 \log\{\text{normalizing constant}\}. \qquad (3.5)
\end{aligned}$$

3.2 Parameter estimation

By definition, we can write $\gamma_k = \mathbb{I}\{||\beta_{G_k}||_2 \neq 0\}$ and $\alpha_j = \mathbb{I}\{\beta_j \notin (-\xi, \xi)| \; ||\beta_{G_{k_j}}||_2 \neq 0\}$, where k_j is the index for the group that j belongs to. With argumented representation (3.2), the second term on the right hand side of (3.5) can be expressed as

$$\begin{aligned}
&-2\log\{f(\beta|\; \alpha, \gamma, \sigma^2, \lambda) f(\alpha|\; \gamma, \omega) f(\gamma|\; \theta)\} \\
&\qquad = \frac{\lambda}{\sigma^2}\sum_{k=1}^{m}\gamma_k\left(\sum_{j\in G_k}\alpha_j\beta_j^2\right) + \log\left(\frac{2\pi\sigma^2}{\lambda}\right)\sum_{k=1}^{m}\sum_{j\in G_k}\gamma_k\alpha_j \\
&\qquad + \sum_{k=1}^{m}\sum_{j\in G_k}\left[\log\left(\frac{1-\omega_j}{\omega_j}\right)^2\right]\gamma_k\alpha_j \\
&\qquad + \sum_{k=1}^{m}\left\{\log\left[\left(\frac{1-\theta_k}{\theta_k}\right)^2\prod_{j\in G_k}\left(\frac{1}{1-\omega_j}\right)^2\right]\right\}\gamma_k. \qquad (3.6)
\end{aligned}$$

Here we have used the facts that $(1-\alpha_j)\log\delta_{(-\xi,\xi)}(\beta_j) = 0$, $(1-\gamma_k)\log\delta_0(||\beta_{G_k}||_2) = 0$, and $(1-\gamma_k)\log\delta_0(\alpha_{G_k}) = 0$ in deriving (3.6).

In addition, given that $\xi \to 0$, we have $\alpha_j \approx \mathbb{I}\{\beta_j \neq 0| \; ||\beta_{G_{k_j}}||_2 \neq 0\}$. Further by a direct calculation, we have $\gamma_{k_j}\alpha_j = \mathbb{I}\{\beta_j \neq 0 \cap ||\beta_{G_{k_j}}||_2 \neq 0\} = \mathbb{I}\{||\beta_{G_{k_j}}||_2 \neq 0| \; \beta_j \neq 0\}\mathbb{I}\{\beta_j \neq 0\}$. Note that the expectation of the index $\mathbb{I}\{||\beta_{G_{k_j}}||_2 \neq 0| \; \beta_j \neq 0\}$ is $\mathbb{P}(||\beta_{G_{k_j}}||_2 \neq 0| \; \beta_j \neq 0)$, which is obviously equal to 1 since $j \in G_{k_j}$ and $\beta_j \neq 0$ implies $||\beta_{G_{k_j}}||_2 \neq 0$ almost surely. This further implies that $\mathbb{I}\{||\beta_{G_{k_j}}||_2 \neq 0| \; \beta_j \neq 0\}$ is equal to 1 almost surely. Therefore we have

$$\gamma_{k_j}\alpha_j = \mathbb{I}\{\beta_j \neq 0\}. \qquad (3.7)$$

Now consider the hyperparameters λ, σ^2, θ, ω. Since there is no easy way to determine values of these hyperparameters, therefore for practical purposes, we will impose some constraints on these hyperparameters. We assume $\omega_j = \omega_1$ for all j. Further we define $\rho_1 = \sigma^2\log\{[(2\pi\sigma^2)/\lambda][(1-\omega_1)/\omega_1]^2\}$ and assume $\rho_1 \geq 0$. For the fourth term on the right hand side of (3.6) that involves θ_k's, we adopt the following parametrization. We will assume all γ_k's in the fourth term on the right hand side of (3.6) have an equal weight. Given that $\omega_j = \omega_1$ for all j, we can choose appropriate θ_k's from interval $[0, 1]$ to make the weights of γ_k's the same for all k. Let θ_k^* be such appropriate value of θ_k. With values of θ_k^*'s, we define $\rho_2 = \sigma^2\log\{[(1-\theta_k^*)/\theta_k^*]^{2/\sqrt{q_k}}(1-\omega_1)^{-2\sqrt{q_k}}\}$, where $q_k = |G_k|$. We assume $\rho_2 \geq 0$.

With (3.7) and the definitions of ρ_1 and ρ_2, minimizing (3.5) with respect to β, α and γ is equivalent to minimizing the function

$$\begin{aligned}
V(\beta) \;&=\; \left|\left|y - \sum_{k=1}^{m}X_{G_k}\beta_{G_k}\right|\right|_2^2 + \lambda\sum_{k=1}^{m}||\beta_{G_k}||_2^2 \\
&\quad + \rho_1\sum_{k=1}^{m}\sum_{j\in G_k}\mathbb{I}\{\beta_j \neq 0\} + \rho_2\sum_{k=1}^{m}\sqrt{q_k}\mathbb{I}\{||\beta_{G_k}||_2 \neq 0\} \qquad (3.8)
\end{aligned}$$

with respect to β. Here we define the gvsnss estimator (**G**rouped **V**ariable **S**election via **N**ested **S**pike and **S**lab Priors) as the one that minimizes (3.8). Below we provide a numerical procedure to calculate the gvsnss estimator.

3.2.1 Majorization-minimization algorithms

Since the last two terms in (3.8) are discrete in their domain, the minimization problem involving (3.8) is combinatorial and in general is considered to be difficult. Here we adopt a continuous relaxation procedure to modify (3.8). More specifically, we use the function

$$g_\tau(a) = \frac{\log(1+\tau^{-1}|a|)}{\log(1+\tau^{-1})} \tag{3.9}$$

to approximate index function $\mathbb{I}\{a \neq 0\}$. It can be shown that $g_\tau(a) \to \mathbb{I}\{a \neq 0\}$ as $\tau \to 0$ [26, 31]. Figure 1 shows $\mathbb{I}\{a \neq 0\}$ and $g_\tau(a)$ and the absolute difference between the two functions as a function of $-\log\tau$. Since (3.9) is continuous on $\mathbb{R}$, the combinatorial nature of $\mathbb{I}\{a \neq 0\}$ is relaxed. However, (3.9) is not convex in a, and using (3.9) for continuous relaxation on (3.8) still makes (3.8) remain non-convex. We adopt a majorization-minimization approach to tackling this problem. Majorization-minimization (MM) algorithms [15, 30] aim to solve difficult minimization problems by modifying the corresponding objective functions so that solution spaces of the modified ones are easier to explore. For an objective function $V^*(a)$, the modification procedure relies on finding a function $V^{**}(a;a^{(d)})$ that satisfies the following properties:

$$\begin{aligned} V^{**}(a;a^{(d)}) &\geq V^*(a) \quad \text{for all } a, \\ V^{**}(a^{(d)};a^{(d)}) &= V^*(a^{(d)}). \end{aligned} \tag{3.10}$$

In (3.10), the objective function $V^*(a)$ is said to be majorized by $V^{**}(a;a^{(d)})$. In this sense, $V^{**}(a;a^{(d)})$ is called the majorization function. In addition, (3.10) implies that $V^{**}(a;a^{(d)})$ is tangent to $V^*(a)$ at $a^{(d)}$. Moreover if $a^{(d+1)}$ is a minimizer of $V^{**}(a;a^{(d)})$, then (3.10) further implies that $V^*(a^{(d)}) = V^{**}(a^{(d)};a^{(d)}) \geq V^{**}(a^{(d+1)};a^{(d)}) \geq V^*(a^{(d+1)})$, which means that the iteration procedure $a^{(d)}$ pushes $V^*(a)$ toward its minimum.

Now we turn back to function (3.9). Note that, since $\log(a)$ is a concave function of a for $a > 0$, therefore the inequality

$$\log(a') + \frac{a}{a'} - 1 \geq \log(a) \tag{3.11}$$

holds for all $a > 0$ and $a' > 0$. Note that the left hand side of (3.11) is convex in a. In addition, if we let $a = a'$, then (3.11) becomes an equality, which implies that the left hand side of (3.11) satisfies the properties stated in (3.10), therefore is a valid function for majorizing $\log(a)$.

Now by applying (3.9) and the left hand side of (3.11) to $\sum_{k=1}^m \sum_{j\in G_k} \mathbb{I}\{\beta_j \neq 0\}$, we can establish the following inequality:

$$\begin{aligned}
&\sum_{k=1}^m \sum_{j\in G_k} \mathbb{I}\{\beta_j \neq 0\} \\
= \ & \lim_{\tau\to 0} \sum_{k=1}^m \sum_{j\in G_k} \frac{\log(1+\tau^{-1}|\beta_j|)}{\log(1+\tau^{-1})} \\
\leq \ & \lim_{\tau\to 0} \frac{1}{\log(1+\tau^{-1})} \sum_{j=1}^p \left(\log\left(1+\tau^{-1}|\beta_j'|\right) + \frac{\tau+|\beta_j|}{\tau+|\beta_j'|} - 1 \right). \qquad (3.12)
\end{aligned}$$

Similarly for $\sum_{k=1}^m \{||\beta_{G_k}||_2 \neq 0\}$, we have

$$\begin{aligned}
\sum_{k=1}^m \sqrt{q_k}\mathbb{I}\{||\beta_{G_k}||_2 \neq 0\} \ \leq \ & \lim_{\tau\to 0} \frac{1}{\log(1+\tau^{-1})} \\
& \times \sum_{k=1}^m \sqrt{q_k} \left(\log\left(1+\tau^{-1}||\beta_{G_k}'||_2\right) + \frac{\tau+||\beta_{G_k}||_2}{\tau+||\beta_{G_k}'||_2} - 1 \right). \\
& \qquad (3.13)
\end{aligned}$$

3.2.2 Blockwise coordinate descent algorithms

With the majorization-minimization results (3.12) and (3.13), we can establish an iterative scheme to find the minimizer of (3.8). In practice, we use the blockwise iterative scheme

$$\begin{aligned}
\widehat{\beta}_{G_k}^{(d+1)} \ = \ & \arg\min_{\beta_{G_k}} \Bigg\{ \big|\big|r_{-G_k} - X_{G_k}\beta_{G_k}\big|\big|_2^2 + \lambda||\beta_{G_k}||_2^2 \\
& + \lambda_1||\widehat{\nu}_{G_k}^{(d)}\beta_{G_k}||_1 + \lambda_2\widehat{\phi}_k^{(d)}||\beta_{G_k}||_2 \Bigg\} \qquad (3.14)
\end{aligned}$$

to find the solution that minimizes (3.8), where $\lambda_1 = \rho_1 \lim_{\tau\to 0}[\log(1+\tau^{-1})]^{-1}$, $\lambda_2 = \rho_2 \lim_{\tau\to 0}[\log(1+\tau^{-1})]^{-1}$, and $r_{-G_k} = y - \sum_{k'\neq k} X_{G_{k'}}\beta_{G_{k'}}$. In addition, for $j \in G_k$, $\widehat{\nu}_j^{(d)} = \lim_{\tau\to 0}(\tau + |\widehat{\beta}_j^{(d)}|)^{-1}$, and $\widehat{\phi}_k^{(d)} = \lim_{\tau\to 0}\sqrt{q_k}(\tau + ||\widehat{\beta}_{G_k}^{(d)}||_2)^{-1}$.

With the objective function stated in (3.14), one can derive associated KKT conditions and solve them for the minimizer $\widehat{\beta}_{G_k}^{(d+1)}$. However, the third and fourth terms on the right hand side of (3.14) are not smooth, therefore special attention is needed to obtain a gradient-like vector for (3.14). Here we adopt a subgradient-based approach to tackling this problem. For the idea of subgradients and related theoretical properties, please see Section B.5 of [4]. By applying the subgradient calculus to the objective function in (3.14) with respect to β_{G_k}, we can obtain a gradient-like vector for the objective function. Then by setting the vector to zero, we obtain the subgradient equations

$$2X_{G_k}^T r_{-G_k} - 2X_{G_k}^T X_{G_k}\beta_{G_k} - 2\lambda\beta_{G_k} - \lambda_1\widehat{\nu}_{G_k}^{(d)} h_{G_k} - \lambda_2\widehat{\phi}_k^{(d)} v_{G_k} = 0, \qquad (3.15)$$

where h_{G_k} is a subgradient vector of the l_1-norm $||\beta_{G_k}||_1$, and its entry is defined as that, for $j \in G_k$ $h_j = 1$ if $\beta_j > 0$; $h_j = h_j^* \in [-1, 1]$ if $\beta_j = 0$; and $h_j = -1$ if $\beta_j < 0$. In addition, v_{G_k} is a subgradient vector of the l_2-norm $||\beta_{G_k}||_2$ and is defined as

$$v_{G_k} = \begin{cases} \beta_{G_k}/||\beta_{G_k}||_2 & \text{if } ||\beta_{G_k}||_2 \neq 0, \\ v^*_{G_k} \text{ such that } ||v^*_{G_k}||_2^2 \leq 1 & \text{if } ||\beta_{G_k}||_2 = 0. \end{cases} \tag{3.16}$$

Below we adopt a method provided by Friedman et al. [10] to solve the subgradient equations (3.15). The method uses a testing procedure to identify whether β_{G_k} is a zero vector or not. First note that, if $\beta_{G_k} = 0$, then the subgradient equations (3.15) becomes

$$2X_{G_k}^T r_{-G_k} - \lambda_1 \widehat{\nu}_{G_k}^{(d)} h_{G_k} = \lambda_2 \widehat{\phi}_k^{(d)} v_{G_k}. \tag{3.17}$$

Now by definition (3.16), if $||\beta_{G_k}||_2 = 0$, i.e. β_{G_k} is a zero vector, then $||v_{G_k}||_2 \leq 1$, therefore (3.17) implies that

$$||2X_{G_k}^T r_{-G_k} - \lambda_1 \widehat{\nu}_{G_k}^{(d)} h_{G_k}||_2 \leq \lambda_2 \widehat{\phi}_k^{(d)}. \tag{3.18}$$

To numerically verify the condition (3.18), we need to know h_{G_k}. Friedman et al. [10] provided a practical way to estimate h_{G_k} by solving the least squares problem $\min_{h_{G_k}} ||2X_{G_k}^T r_{-G_k} - \lambda_1 \widehat{\nu}_{G_k}^{(d)} h_{G_k}||_2^2$ subject to $-1 \leq h_j \leq 1$ for $j \in G_k$. The resulting estimate takes a soft-thresholding form, and by plugging it into (3.18), one obtains

$$\left|\left| ST_{\lambda_1 \widehat{\nu}_{G_k}^{(d)}} (2X_{G_k}^T r_{-G_k}) \right|\right|_2 \leq \lambda_2 \widehat{\phi}_k^{(d)}, \tag{3.19}$$

Note that if condition (3.19) holds, we let $\widehat{\beta}_{G_k}^{(d+1)} = 0$, otherwise we go further to estimate entries in β_{G_k} with other values.

Below we describe a numerical procedure for estimating non-zero entries in β_{G_k}. First note that, as shown in [29], the l_2-norm $||\beta_{G_k}||_2$ on the right hand side of (3.14) can be bounded in a way such that

$$||\beta'_{G_k}||_2 + \frac{1}{2||\beta'_{G_k}||_2}(||\beta_{G_k}||_2^2 - ||\beta'_{G_k}||_2^2) \geq ||\beta_{G_k}||_2. \tag{3.20}$$

Here the function on the left hand side is convex in β_{G_k}. Now if we let $\beta_{G_k} = \beta'_{G_k}$, then the equality will hold between the two sides of (3.20). Therefore the function on the left hand side of (3.20) majorizes $||\beta_{G_k}||_2$. With the majorization result (3.20), we construct the following iterative scheme:

$$\begin{aligned} \widehat{\beta}_{G_k}^{(d_1+1,d_2+1)} \quad = \quad & \arg\min_{\beta_j} \Bigg\{ \left|\left| r_{-G_k} - X_{G_k}\beta_{G_k} \right|\right|_2^2 \\ & + \lambda_1 ||\widehat{\nu}_{G_k}^{(d_1)} \beta_{G_k}||_1 + \left(\lambda + \frac{\lambda_2 \widehat{\phi}_k^{(d_1)}}{2||\widehat{\beta}_k^{(d_1+1,d_2)}||_2} \right) ||\beta_{G_k}||_2^2 \Bigg\} \end{aligned} \tag{3.21}$$

to obtain $\widehat{\beta}_{G_k}^{(d_1+1)}$. The scheme (3.21) can be approximated by the following iterative least squares procedure:

$$\widehat{\beta}_{G_k}^{(d_1+1,d_2+1)} = \left[X_{G_k}^T X_{G_k} + \left(\lambda + \frac{\lambda_2 \widehat{\phi}_k^{(d_1)}}{2||\widehat{\beta}_{G_k}^{(d_1+1,d_2)}||_2}\right) I_{q_k \times q_k}\right]^{-1} ST_{\lambda_1 \widehat{\nu}_{G_k}^{(d_1)}/2}(X_{G_k}^T r_{-G_k}), \tag{3.22}$$

where $ST_{\lambda_1 \widehat{\nu}_{G_k}^{(d_1)}/2}(X_{G_k}^T r_{-G_k})$ is the soft thresholding operator defined in (2.1). A least squares result similar to (3.22) can be found in [9]. For large-scale problems, we construct a one dimensional soft thresholding scheme to approximate (3.21). The soft-thresholding scheme is given by

$$\widehat{\beta}_j^{(d_1+1,d_2+1)} = \left(\sum_{i=1}^n x_{ij}^2 + \lambda + \frac{\lambda_2 \widehat{\phi}_{k_j}^{(d_1)}}{2||\widehat{\beta}_{k_j}^{(d_1+1,d_2)}||_2}\right)^{-1} ST_{\lambda_1 \widehat{\nu}_j^{(d_1)}/2}\left(\sum_{i=1}^n x_{ij} r_{i,-j}^{(*)}\right), \tag{3.23}$$

where $r_{i,-j}^{(*)} = r_{i,-G_k} - \sum_{j' \neq j; j', j \in G_k} x_{ij'} \beta_{j'}^{(*)}$ with $\beta_{j'}^{(*)} = \beta_{j'}^{(d_1+1,d_2+1)}$ for $j' < j$ and $\beta_{j'}^{(*)} = \beta_{j'}^{(d_1+1,d_2)}$ for $j' > j$.

3.3 Determining tuning parameter values

For λ_1, λ_2 and λ, we adopt a grid search strategy to find their optimal values. Here we assume that each column of design matrix X is standardized. To find optimal λ_1, we search along a grid of candidate values in the interval $[0, \lambda_1^*]$, where λ_1^* is defined as $\lambda_1^* = 2.05\tau \times \max_{j \in \{1,2,\cdots,p\}} |x_j^T y|$. To find optimal λ_2, we search along a grid of candidate values in the interval $[0, \lambda_2^*]$, where $\lambda_2^* = 1.1\tau \times \max_{k \in \{1,2,\cdots,m\}} ||2X_{G_k}^T y||_2/\sqrt{q_k}$. For λ, we assume it decreases with sample size n and is proportional to λ_2. More specifically, we let $\lambda = \lambda_2/(10n)$. With the reparametrization on λ given above, we only need to do grid searches for λ_1 and λ_2. For parameter τ, we let $\tau = 5 \times 10^{-4}$.

3.4 Connection with other approaches

Recent research on variable selection using maximum a posteriori estimation includes [12, 31]. Armagan et al. [1] developed a shrinkage-based method for variable selection based on the generalized double Pareto priors. The idea of using spike and slab priors in grouped variable selection has also been adopted by Scheipl et al. [25], who developed an MCMC-based approach to carrying out posterior inference on additive regression models. The idea of using (3.9) in approximating an index function has been mentioned in [7, 26, 18, 31]. Tipping [28] has pointed out a connection between the log function (3.9) and the improper Student's t density.

4 Simulation study

In this section we study finite sample properties of the gvsnss estimator by fitting regression models with simulated data. In the simulation study, we assume the covariates are randomly divided into m groups, and the true covariates, i.e. the covariates with non-zero coefficients, are covered by $r \leq m$ groups. We will focus on the following two situations:

i. The true covariates are covered by the r groups, but at the same time, some redundant covariates, i.e. the covariates with zero coefficients, are also covered by the r groups.

ii. The true covariates are re-assigned with different group labels. In this situation, r, the number of groups that covers the true covariates, will change.

To create the first situation, we focuses on varying the level of sparsity in the groups that contain the true covariates. To create the second situation, we focuses on re-assigning covariates to other groups according to some group switching probabilities. Under the two situations, each simulation experiment is characterized by the pair (spr, mis-labeled), where "spr" denotes the level of within-group-sparsity and "mis-labeled" denotes the group switching probability. For a covariate in an active group, spr = 0.3 means that the value of its coefficient will have probability 0.3 to be coerced to zero, and mis-labeled = 0.3 means that it will be re-assigned with a different group label with probability 0.3.

Below we introduce the basic simulation scheme. For the $n \times p$ design matrix X, we generate its rows i.i.d. from $\text{MVN}(0, I_{p\times p})$. For regression coefficients $\beta = (\beta_1, \beta_2, \cdots, \beta_p)$, we first randomly assign the corresponding covariates into m groups. We then choose $r \leq m$ groups of covariates and generate their coefficients i.i.d. from $\text{Normal}(0, 1)$. We further set coefficients of the covariates in the rest of $m - r$ groups to zero. We then re-proceed each coefficient by either coercing its value to zero or re-assigning its covariate with a different group label according to the pre-specified values in (spr, mis-labeled). For the error vector ϵ, we generate its entries i.i.d. from $\text{Normal}(0, 1)$. Finally, we compute the response vector $y = X\beta + \epsilon$.

4.1 Methods for comparisons

We conducted two gvsnss estimations for the regression model. The first one used five fold cross validation for tuning parameter selection. The second one used the following logarithm of the Bayes factor:

$$\log \text{BF}(\widehat{S}, \text{null}; y) = \frac{n}{2} \log \left\{ \frac{y^T y}{y^T(\lambda^{-1} X_{\widehat{S}} X_{\widehat{S}}^T + I_{n\times n})^{-1} y} \right\} - \frac{1}{2} \log \left| \lambda^{-1} X_{\widehat{S}} X_{\widehat{S}}^T + I_{n\times n} \right| \tag{4.1}$$

for tuning parameter selection, where $\widehat{S} = \{j : \widehat{\beta}_{\text{gvsnss},j} \neq 0\}$. The logarithm Bayes factor (4.1) corresponds to the model that assigns $\text{Normal}(0, \sigma^2/\lambda)$ on β_j and Inverse-Gamma(τ_1, τ_2) on σ^2 with τ_1 and τ_2 both approaching to zero. For tuning parameter selection, we searched optimal λ_1 along a grid of 20 candidate values and optimal λ_2 along a grid of another 20 candidate values.

We also conducted three other estimations for the regression model. The first one is the group lasso using five fold cross validation for tuning parameter selection. The second one is also the group lasso but using a naive AIC for tuning parameter selection. The naive AIC is given by nAIC $= ||y - X\widehat{\beta}_{\mathrm{GL}}||_2^2/\widehat{\sigma}^2 + 2\widehat{s}_{\mathrm{GL}}$, where $\widehat{\sigma}^2$ is estimated from the null model and $\widehat{s}_{\mathrm{GL}}$ is the number of non-zero entries in $\widehat{\beta}_{\mathrm{GL}}$. Numerical calculations for the two group lasso estimations were done by using R package grplasso [19]. The third one is the lasso using ten fold cross validation for tuning parameter estimation. We used R package glmnet [11] to carry out numerical computations for the lasso estimation. For all the three estimations, we searched optimal tuning parameters along a grid of 100 candidate values.

We collected three performance measures at each simulation run. The first one is the sign-adjusted false positive rate, which is defined as

$$\mathrm{SFPR} = \frac{\#\{j \in \widehat{S} : \mathrm{sign}(\widehat{\beta}_j) \neq \mathrm{sign}(\beta_{\mathrm{true},j})\}}{|\widehat{S}|}.$$

The second one is the squared l_2 estimation error, which is defined as

$$l_2\text{-dis} = \frac{\sum_{j=1}^{p}(\widehat{\beta}_j - \beta_{\mathrm{true},j})^2}{p}.$$

The third one is the predictive mean squared error, which is defined as

$$\mathrm{PMSE} = \frac{\sum_{i=1}^{n'}(y_{i,\mathrm{new}} - x_{i,\mathrm{new}}^T\widehat{\beta})^2}{n'},$$

where $n' = 10 \times n$, $y_{i,\mathrm{new}}$ and $x_{i,\mathrm{new}}$ are new data points generated under the same simulation scheme.

4.2 Results

In practice, we let $p = 200$, $m = 10$, and $r = 2$. We considered different values of sample size n and the pair (spr, mis-labeled) in generating data points.

We first considered the scenario in which the group switching probability is zero. The results are shown in Figure 2, with the first, second and third rows being the plots of SFPR, l_2-dis and PMSE, respectively and the first, second and third columns being the plots for cases with spr = 0, 0.3, 0.6, respectively. Each point in the plot is an average over 100 simulation runs. The results show that the gvsnss estimator has relatively good performances over the group lasso in variable selection when the level of within-group-sparsity is increasing. In addition, among the five estimations, the gvsnss estimation using the Bayes factor has relatively small values in squared l_2 estimation error and PMSE. However, we also noticed that the advantages of using group-based estimations such as the group lasso or gvsnss estimations over the lasso estimation will gradually disappear as the level of within-group-sparsity increases.

We then considered scenarios under different group switching probabilities. The results are given in Figures 3 and 4 for group switching probability equal to 0.1 and

0.5, respectively. The results show that the gvsnss estimator can still have relatively good performances over other benchmark estimation methods in variable selection. However, we also noticed the lasso estimation almost dominates performances in l_2 estimation error and PMSE over group-based estimation methods in these scenarios, especially when the group switching probability is high. A high group switching probability will lead to an increase in r, the number of groups that cover the true covariates. In Section 5 we will give a theoretical explanation to these simulation results by deriving an upper bound for the l_2 estimation error.

5 Asymptotic analysis

In this section we investigate asymptotic behavior of the gvsnss estimator. Before presenting these results, we give some notation definitions. For simplicity, we define $\beta = \beta_{\text{true}}$ throughout this section. Further define $S = \{j : \beta_j \neq 0\}$ and $G_R = \{G_k : k \in R\}$, a collection of disjoint index sets G_k's indexed by R that covers S, i.e. $S \subseteq G_R$. Define $s = |S|$, the number of non-zero coefficients, $q_R = | \cup_{k \in R} G_k|$, the number of indices covered by G_R, and $r = |R|$, the number of groups that cover indices for covariates with non-zero coefficients.

Now consider the following function:

$$\begin{aligned} V_\tau(w', \beta', G') &= ||\epsilon' - Xw'||_2^2 + \lambda ||w' + \beta'||_2^2 \\ &+ \rho_1 \sum_{j=1}^{p} \frac{\log(1 + \tau^{-1}|w'_j + \beta'_j|)}{\log(1+\tau^{-1})} \\ &+ \rho_2 \sum_{k=1}^{m'} \sqrt{q'_k} \frac{\log(1 + \tau^{-1}||w'_{G'_k} + \beta'_{G'_k}||_2)}{\log(1+\tau^{-1})}, \end{aligned} \tag{5.1}$$

where $\epsilon' = y - X\beta'$, $G' = \{G'_k : k = 1, 2, \cdots, m\}$ and $q'_k = |G'_k|$. At a fixed τ', we define $\widehat{\beta}^{\tau'}$ by

$$\widehat{\beta}^{\tau'} = \arg\min_{\beta'} \lim_{\tau \to \tau'} V_\tau(0, \beta', G). \tag{5.2}$$

Further define $\widehat{S}^{\tau'} = \{j : \widehat{\beta}_j^{\tau'} \neq 0\}$ and $\widehat{s}^{\tau'} = |\widehat{S}^{\tau'}|$. Note that if we let $\tau \to 0$, then $V_\tau(0, \beta', G')$ will approach to the objective function (3.8). Therefore technically we can express the gvsnss estimator as

$$\widehat{\beta}_{\text{gvsnss}} = \arg\min_{\beta'} \lim_{\tau \to 0} V_\tau(0, \beta', G). \tag{5.3}$$

We further define $\widehat{S} = \{j : \widehat{\beta}_{\text{gvsnss}} \neq 0\}$ and $\widehat{s} = |\widehat{S}|$. Note that by definition, as $\tau \to 0$, (5.2) becomes $\widehat{\beta}^0 = \arg\min_{\beta'} \lim_{\tau \to 0} V_\tau(0, \beta', G) = \widehat{\beta}_{\text{gvsnss}}$. As a result of that, we have $\widehat{S}^\tau \to \widehat{S}$ and $\widehat{s}^\tau \to \widehat{s}$ as $\tau \to 0$.

5.1 l_2 estimation error

One useful concept to justify the advantage of group-based estimation is the strong group sparsity [13]. We say the true coefficient vector β is (s_0, r_0) strongly group-sparse if there exists a collection of index sets $G_R = \{G_k : k \in R\}$ such that $S \subseteq G_R$ with $q_R = |G_R| \leq s_0$ and $r = |R| \leq r_0$. For group lasso $\widehat{\beta}_{\text{GL}}$ defined in (1.2), Huang and Zhang [13] showed that if β is (s_0, r_0) strongly group-sparse, then given some regular conditions hold, with $1 - \alpha$ probability, the l_2 estimation error $||\widehat{\beta}_{\text{GL}} - \beta||_2 = O(n^{-1/2}\sqrt{s_0 + r_0 \log(m/\alpha)})$. The order of magnitude implies that the group lasso estimation can be beneficial if q_R, the number of indices in G_R, and r, the number of index sets that cover S, are small.

Here we have to note that directly comparing rates of the l_2 estimation error between the lasso and group lasso is not easy since it requires one to derive the rates under the same assumptions. Lounici et al. [16] provided such comparisons for multi-task learning cases and showed that the upper bound for the l_2 estimation error of the group lasso can have an order of magnitude smaller than the lower bound for the l_2 estimation error of the lasso.

Below we start our investigation on the l_2 estimation error $||\widehat{\beta}_{\text{gvsnss}} - \beta||_2$ by deriving a deterministic upper bound for $||\widehat{\beta}^{\tau} - \beta||_2$.

Theorem 5.1. *For $\epsilon = y - X\beta$, $\tau \in [0, 1)$, and $1 \leq \max(q_R, \widehat{s}^{\tau}) \leq p$, we have,*

$$\begin{aligned} ||\widehat{\beta}^{\tau} - \beta||_2 \;\leq\; & \frac{q_R^{1/2}}{\kappa_n + \lambda n^{-1}} \left\{ 4\left[1 + \left(\frac{\widehat{s}^{\tau}}{4s}\right)^{1/2}\right] \frac{||X^T\epsilon||_\infty}{n} + 2 \max_{j \in S} |\beta_j| \frac{\lambda}{n} \right. \\ & \left. + \left[\frac{2c_2^{-1} + 1}{\log(\tau^{-1})} + c_3^{-1}\right]\left(\frac{\rho_1 + \rho_2}{n}\right)\right\}, \end{aligned} \tag{5.4}$$

where $\kappa_n = n^{-1} \min_w w^T X^T X w$, $c_2 = \min_{j \in \widehat{S}^{\tau}} |\widehat{\beta}_j^{\tau}|$, and $c_3 = \min_{j \in S} |\beta_j|$.

Theorem 5.1 does not rely on any distribution assumption on the error vector ϵ. It is stated in a deterministic way and does not have any probabilistic interpretation.

Below we will give some conditions that are useful in deriving upper bounds for $||\widehat{\beta}_{\text{gvsnss}} - \beta||_2$ in a situation in which some distribution assumption is imposed on ϵ.

Assumption 1. Let κ_n be the same as the one defined in Theorem 5.4. We assume $\kappa_n + \lambda n^{-1} > 0$ as $n \to \infty$.

Assumption 1 is similar to Condition A1 in [34]. It mainly serves as a statement to guarantee that the minimum eigenvalue of the matrix $n^{-1}(X^T X + \lambda I_{p \times p})$ is positive when $n \to \infty$. Note that without Assumption 1, κ_n will be equal to zero when $n < p < \infty$, but the minimum eigenvalue value $\kappa_n + n^{-1}\lambda = n^{-1}\lambda$ will remain positive if $\lambda > 0$. Assumption 1 further implies that $\sqrt{n}(\kappa_n + \lambda n^{-1}) \to \infty$ when $n \to \infty$.

Theorem 5.2. *Assume that ϵ_i's are i.i.d. as Normal$(0, \sigma^2)$. Further assume that $n^{-1}\sum_{i=1}^n x_{ij}^2 = \zeta_j$, for $j = 1, 2, \cdots, p$, $\tau = n^{-1}$, $\lambda = A_1\psi_n$, $\rho_1 = A_2\psi_n$, $\rho_2 = A_3\psi_n$*

with A_1, A_2 and A_3 being some positive constants, and

$$\psi_n = 2\sigma\sqrt{2n\max_j \zeta_j\left[\log\left(\frac{m}{\alpha}\right)+\log\overline{q}\right]}, \tag{5.5}$$

where α is a non-negative constant and $\overline{q} = m^{-1}\sum_{k=1}^{m} q_k$. Then given that Assumption 1 holds, for $1 \leq \max(q_R, \widehat{s}) \leq p$, with $1-\alpha$ probability, we have

$$||\widehat{\beta}_{\text{gvsnss}} - \beta||_2 \;\leq\; \frac{2\sqrt{2}\sigma\Lambda_n \max_j \zeta_j^{1/2}}{\sqrt{n}(\kappa_n + \Omega_n)}\sqrt{q_R\left[\log\left(\frac{m}{\alpha}\right)+\log\overline{q}\right]}, \tag{5.6}$$

as $n \to \infty$, where

$$\begin{aligned}\Lambda_n &= \left\{2\left[1+\left(\frac{\widehat{s}}{4s}\right)^{1/2}\right] + 2\max_{j\in S}|\beta_j|A_1\right.\\ &\quad \left. +(A_2+A_3)\left[\frac{2c_2^{-1}+1}{\log(n)} + c_3^{-1}\right]\right\}, &(5.7)\\ \Omega_n &= 2A_1\sigma\sqrt{\frac{2\max_j \zeta_j}{n}\left[\log\left(\frac{m}{\alpha}\right)+\log\overline{q}\right]}, &(5.8)\end{aligned}$$

where $\widehat{s} = |\widehat{S}|$, c_2 and c_3 are defined in Theorem 5.4.

The deterministic result stated in Theorem 5.1 will serve as a bone for deriving upper bound (5.6). Note that since we have assumed $\tau = n^{-1}$, therefore effectively we have $\widehat{\beta}^{\tau} \to \widehat{\beta}_{\text{gvsnss}}$ and $\widehat{S}^{\tau} \to \widehat{S}$ as $n \to \infty$. Detailed derivations of Theorem 5.1 and Theorem 5.2 are given in Appendix A.

Note that the bound (5.6) is proportional to $q_R^{1/2}$ and by definition

$$q_R = s + \sum_{k\in R} \#\{j \in G_k : \beta_j = 0\}.$$

Given that s is fixed, the result implies that, if groups that contain the true covariates also contain large numbers of redundant covariates, or if the true covariates are scattered over a large number of groups, like the scenarios with high group switching probabilities we have seen in Section 4, then the gvsnss estimator will not perform well.

Now if we adopt an equal group setting, i.e. $q_1 = q_2, \cdots, = q_m$, and let $\zeta_j = 1$ for $j = 1, 2, \cdots, p$, then $q_R = |G_R| = |R| \times |G_k| = rq_1$, and the right hand side of (5.6) will have an order of magnitude equal to $n^{-1/2}\sqrt{r\log q_1 + r\log(m/\alpha)}$. Further note that $\log q_1 \leq q_1$. Therefore with $1-\alpha$ probability, as $n \to \infty$, we have $||\widehat{\beta}_{\text{gvsnss}} - \beta||_2 = O(n^{-1/2}\sqrt{s_0 + r_0\log(m/\alpha)})$, where $r_0 = r$ and $s_0 = q_R$. The result given above implies that the gvsnss estimator can achieve an l_2 estimation error with an order of magnitude proportional to that of the group lasso established in [13].

The following corollary states that if the maximum size of groups is equal to one, then the gvsnss estimator can have an l_2 estimation error with an order of magnitude similar to that of the lasso established in [20, 5].

Corollary 5.1. *Assume that* $\max_k q_k = 1$ *and* $\zeta_j = 1$ *for* $j = 1, 2, \cdots, p$. *Then given that all assumptions stated in Theorem 5.2 hold, with* $1 - \alpha$ *probability, we have*

$$||\widehat{\beta}_{\text{gvsnss}} - \beta||_2 \ \leq \ \frac{2\sqrt{2}\sigma\Lambda_n}{\sqrt{n}(\kappa_n + \Omega_n)}\sqrt{s\log\left(\frac{p}{\alpha}\right)} \tag{5.9}$$

as $n \to \infty$, *where* Λ_n *is the same as the one defined in (5.7) and*

$$\Omega_n \ = \ 2A_1\sigma\sqrt{\frac{2}{n}\log\left(\frac{p}{\alpha}\right)}.$$

Proof of Corollary 5.1. Obviously given that the maximum group size is one, $q_R = s$. In addition, the number of groups is $m = p$. Then by inserting the results given above into the right hand side of (5.6), we obtain (5.9), which completes the proof. □

5.2 Label-invariance property

Here we show that the gvsnss estimator (5.3) is asymptotically invariant to group structures. We consider two collections of index sets $G^* = \{G_k^* : k = 1, 2, \cdots, m^*\}$ and $G^{**} = \{G_l^{**} : l = 1, 2, \cdots, m^{**}\}$. In the following discussion as well as in the proof we will see $*$ and $**$ attached to various vector-valued quantities and the presence of $*$ (or $**$) in a given vector means that the entries of the vector are indexed by G_k^* (or G_k^{**}) in the original vector.

Our result relies on the fact that the third term in $V_\tau(0, \beta', G')$ allows the gvsnss estimation to produce zero estimates for coefficients whose covariates are in active groups. Without this setting, we would be unable to establish the label-invariance property for some cases, and $\widehat{\beta}^*_{\text{gvsnss}} = \arg\min_{\beta'} \lim_{\tau\to 0} V_\tau(0, \beta', G^*)$ might never be a solution to the subgradient equations of $\lim_{\tau\to 0} V_\tau(0, \beta', G^{**})$, where G^{**} is an arbitrary collection of index sets. Therefore we assume $\rho_1 > 0$. In addition, our result relies on evaluating the difference between the log-sum penalties involving l_2-norms in $V_\tau(0, \beta', G^*)$ and $V_\tau(0, \beta', G^{**})$. Since ρ_2 and the size of a group play a crucial role in the evaluation process, we will also impose an assumption on their orders of magnitude.

Theorem 5.3. *Assume that*

$$\widehat{\beta}^{\tau *} = \arg\min_{\beta'} V_\tau(0, \beta', G^*)$$

is the unique solution to the subgradient equations of $V_\tau(0, \beta', G^*)$ *for all* $\tau \in [0, 1)$. *Further assume that* $\rho_1 > 0$, $\rho_2 \max_k \sqrt{q_k} = o(\log n)$, *and* $\tau = n^{-1}$. *Then as* $n \to \infty$, $\widehat{\beta}^*_{\text{gvsnss}} = \arg\min_{\beta'} \lim_{\tau\to 0} V_\tau(0, \beta', G^*)$ *is the minimizer of* $\lim_{\tau\to 0} V_\tau(0, \beta', G^{**})$, *where* G^{**} *is an arbitrary collection of index sets.*

5.3 Variable selection and sign consistency

Here we study asymptotic behavior of the gvsnss estimator in variable selection. In particular, we focus on sign consistency of the estimated coefficients. We explain the idea of sign consistency first. An estimator $\widehat{\beta}(n)$ is said to be sign consistent in estimating β if probability $\mathbb{P}\{\text{sign}(\widehat{\beta}(n)) = \text{ sign}(\beta)\}$ approaches to one as $n \to \infty$. Given the sign consistency holds, the estimated index set $\widehat{S}(n) = \{j : \widehat{\beta}_j(n) \neq 0\}$ will be the same as the true index set S, therefore the sign consistency implies variable selection consistency, that is, asymptotically with probability one, non-zero valued coefficients will have non-zero estimated values, and zero-valued coefficients will be estimated with zero values.

Below we derive a lower bound for $\mathbb{P}\{\text{sign}(\widehat{\beta}^{\tau}) = \text{ sign}(\beta)\}$. Then with $\tau = n^{-1}$, we have $\widehat{\beta}^{\tau} \to \widehat{\beta}_{\text{gvsnss}}$ as $n \to \infty$, and in turn, the lower bound for $\mathbb{P}\{\text{sign}(\widehat{\beta}_{\text{gvsnss}}) = \text{ sign}(\beta)\}$ can be established asymptotically. The following assumptions on eigenvalues of matrices are useful in deriving the lower bound.

Assumption 2. Define $C_{SS} = n^{-1}(X_S^T X_S + \lambda I_{s\times s})$. Define $\kappa_{\min} = \min_w wC_{SS}w$. We assume $0 < \kappa_{\min} < \infty$ as $n \to \infty$.

Assumption 3. Define $\varsigma_{\max} = \max_w n^{-1} wX_S X_S^T w$. We assume $0 < \varsigma_{\max} < \infty$ as $n \to \infty$.

Assumption 4. Define $\nu_{\max,k} = \max_w n^{-1} wX_{G_k} X_{G_k}^T w$ and $\nu_{\max} = \max_k \nu_{\max,k}$. For $k = 1, 2, \cdots, m$, we assume $0 < \nu_{\max,k} < \infty$ as $n \to \infty$.

Theorem 5.4. *Assume that ϵ_i's are i.i.d. as Normal$(0, \sigma^2)$. Further assume that $n^{-1}\sum_{i=1}^n x_{ij}^2 = 1$ for $j = 1, 2, \cdots, p$, $\tau = n^{-1}$, $\lambda = O(n^{1/2})$, $\rho_1 = O(n^{1/2})$, $\rho_2 = O(n^{1/2})$, and $p = o(n(\log(n+1))^{-2})$. Then given that Assumptions 2, 3 and 4 hold, the probability $\mathbb{P}\big\{\text{sign}(\widehat{\beta}^{\tau}) = \text{sign}(\beta)\big\}$ can be bounded from below in a way such that*

$$\begin{aligned}
&\mathbb{P}\big\{\text{sign}(\widehat{\beta}^{\tau}) = \text{sign}(\beta)\big\}\\
&\geq 1 - \exp\bigg\{-n\bigg(\frac{\psi_{1,n}^2\kappa_{\min}^2}{2\sigma^2} - \frac{\log s}{n}\bigg)\bigg\}\\
&-\exp\bigg\{-n\bigg[\frac{\psi_{2,n}^2\kappa_{\min}^2}{8n(\varsigma_{\max}+\kappa_{\min})^2\sigma^2} - \frac{\log s_1^c}{n}\bigg]\bigg\}\\
&-\exp\bigg\{-n\bigg[\frac{\kappa_{\min}^2\psi_{3,n}^2}{16n^2\nu_{\max}(\varsigma_{\max}+\kappa_{\min})^2\sigma^2} - 0.35 - \frac{\log r^c}{n}\bigg]\bigg\},
\end{aligned} \tag{5.10}$$

where $s_1^c = |S_1^c|$ with $S_1^c = S^c \cap G_R$, $r^c = |R^c|$, $\psi_{1,n}$, $\psi_{2,n}$ and $\psi_{3,n}$ are non-negative constants and as $n \to \infty$, $\psi_{1,n} = O(1)$, $\psi_{2,n} = O(n^{3/2}(\log n)^{-1})$ and $\psi_{3,n} = O(n^{3/2}(\log n)^{-1})$.

The proof can be found in Appendix C. The proof will start by exploring the KKT conditions associated to the minimization problem involving objective function (5.1). Note that in Theorem 5.4 we do not assume that the irrepresentable-type conditions [33] should hold.

Corollary 5.2. *Assume that all assumptions and results stated in Theorem 5.4 hold. Then*

$$\mathbb{P}\big\{\text{sign}(\widehat{\beta}_{\text{gvsnss}}) = \text{sign}(\beta)\big\} \to 1$$

as $n \to \infty$.

Proof of Corollary 5.2. Note that $s \leq p = o(n(\log(1+n))^{-2})$, therefore $n^{-1}\log s \to 0$ as $n \to \infty$. In addition, $\psi_{1,n} = O(1)$, therefore $(2\sigma^2)^{-1}\psi_{1,n}^2\kappa_{\min}^2 > 0$. Then as $n \to \infty$, the first exponential term in (5.10) will approach to zero. For the second exponential term in (5.10), since $\psi_{2,n} = O(n^{3/2}(\log(n))^{-1})$, therefore we have $n^{-1}\psi_{2,n}^2 = O(n^2(\log(n))^{-2}) \to \infty$ as $n \to \infty$. In addition, $s_1^c \leq p = o(n(\log(1+n))^{-2})$, therefore $n^{-1}\log s_1^c \to 0$ as $n \to \infty$. Then as $n \to \infty$, the second exponential term in (5.10) will approach to zero. Furthermore, since $n^{-2}\psi_{3,n}^2 = O(n(\log n)^{-2}) \to \infty$ and $n^{-1}\log r^c \to 0$ as $n \to \infty$, therefore the third exponential term in (5.10) will approach to zero as $n \to \infty$. Finally note that since $\tau = n^{-1}$, therefore $\widehat{\beta}^{\tau} \to \widehat{\beta}_{\text{gvsnss}}$ as $n \to \infty$. The results given above imply that $\mathbb{P}\big\{\text{sign}(\widehat{\beta}_{\text{gvsnss}}) = \text{sign}(\beta)\big\} \to 1$ as $n \to \infty$, which completes the proof.

6 Real data examples

6.1 The U.S. industrial product index

The data set we consider here contains the monthly-based U.S. industrial production index and 125 macroeconomic variables, spanning from July 1964 to December 2010. The industrial production index is an important indicator for economic policymaking. Our aim here is to predict the growth rate of the industrial production index from the 125 macroeconomic variables. Similar data set was used in [27, 3, 17]. The 125 macroeconomic variables are essentially a subset of the 132 variables used by Bai and Ng [3]. For the 125 macroeconomic variables, we follow a benchmark categorization to divide them into 8 groups: 1) output and income (OI), 2) labor market (LM), 3) housing (H), 4) consumption, orders and inventories (COI), 5) money and credits (MC), 6) bond and exchange rates (BE), 7) prices (P), 8) stock market (SM).

Now let IP_t denote the level of the industrial production index at time t. We define the growth rate at time $t+t'$ by $y_{t+t'} = (t')^{-1}1200[\log(IP_{t+t'}) - \log(IP_t)]$. The plot in the top left panel of Figure 5 shows the corresponding time series trend. We further model the growth rate $y_{t+t'}$ by

$$y_{t+t'} = \eta_0 + \sum_{l=0}^{3} z_{t-l}\eta_{l+1} + \sum_{k=1}^{8}\sum_{j\in G_k} x_{tj}\beta_j + \varepsilon_{t+t'}, \tag{6.1}$$

where $z_{t-l} = 1200[\log(IP_{t-l}) - \log(IP_{t-l-1})]$ is the lth lag term, x_{tj} is the jth macroeconomic variable at time t, G_k is the index set corresponding to the kth macroeconomic group, and $\varepsilon_{t+t'}$ is the error term.

We adopt an expanding window scheme to carry out real time estimation for model (6.1). That is, we estimate parameters η_l's and β_j's with information from time 1 to time t. Note that in such setting, at time t, dependent variable $y_{t''+t'}$ is only available for $t'' = 1, \dots, t-t'$. Let $\widehat{\eta}_l^{1,t-t'}$'s and $\widehat{\beta}_j^{1,t-t'}$'s denote the corresponding estimates. With model (6.1) and the estimates, at time t, we predict $y_{t+t'}$ by

$$\widehat{y}_{t+t'} = \widehat{\eta}_0^{1,t-t'} + \sum_{l=0}^{3} z_{t-l}\widehat{\eta}_{l+1}^{1,t-t'} + \sum_{k=1}^{8}\sum_{j\in G_k} x_{tj}\widehat{\beta}_j^{1,t-t'}. \tag{6.2}$$

In practice, we let $t' = 12$, which corresponds to one year change. The prediction is started from $t = 132$ (June 1975) and ended at $t = 546$ (December 2009). Under this setting, there are 415 time blocks. For each time block, we applied two methods to estimate parameters in model (6.1). The first method used the gvsnss to select the 125 macroeconomic variables and then re-estimate regression coefficients of the selected variables with the ordinary least squares method. For the gvsnss estimation, we used five fold cross validation to select the tuning parameter. The second method is similar to the first one but using the lasso for variable selection. For the lasso estimation, we also used five fold cross validation to select the tuning parameter.

In addition, we also used principal components (PCs) of the selected variables to construct models for prediction. For simplicity, we use the first four PCs for the prediction. If the number of selected variables is less than four, we use the selected variables as the predictors.

The plot in the top right panel of Figure 5 shows the number of selected variables for the 415 time blocks while plots in the bottom panel of Figure 5 show frequencies of selected variables for each macroeconomic group under the gvsnss and the lasso, respectively. The results show that the gvsnss estimation selected less variables and produced stronger between-group-sparsity and within-group-sparsity than the lasso.

In addition, we also reported the out-of-sample mean squared error under the two estimation methods. The out-of-sample mean squared error is defined as

$$MSE_{OS}^{t'} = \frac{1}{T-t'}\sum_{t=1}^{T-t'}(y_{t+t'} - \widehat{y}_{t+t'})^2. \tag{6.3}$$

The results are shown in Table 1 and Figure 6, where Model 1 is the model without the lag terms, Model 2 is the model with the lag terms, PC is the model using the first four PCs of all macroeconomic variables, and AR is the model with the lag terms but without the grouped variable terms. The results suggest that including the macroeconomic variables can slightly improve the prediction results.

6.2 Retirement plan data

The data set, adopted from [6, 24], contains information about employee retirement plans of 92 firms. The retirement plans are managed by a company called Best Retirement Inc. (BRI). The response variable is the contribution to retirement plan at the end of the first year. It is measured at the logarithm scale. Let y_i denote the response variable corresponding to the ith retirement plan. Our aim here is to help the company to assess whether the presence of a specially trained sales, named Susan Shepard, has a positive effect on y_i. For the ith retirement plan, we define $x_{i9} = 1$ if Susan Shepard is present and $x_{i9} = 0$ otherwise. The data set also contains eight other variables. To fully assess the presence of Susan Shepard on y_i, we will consider interactions between x_{i9} and the eight variables in the regression model. We call the collection of x_{i9} and the interaction terms the "Susan Shepard Effect" group. Let G_{SSE} denote the set that contains indices of covariates in the Susan Shepard Effect group. We will jointly estimate regression coefficients of the covariates with indices in G_{SSE}. After some calculations, we excluded one interaction variable that has the same value for all retirement plans. The set G_{SSE} therefore only contains indices of eight variables.

We model the expectation of the response variable $\mu_i = \mathbb{E}(y_i|\ \beta, x_i)$ by

$$\mu_i = \sum_{j=1}^{8} x_{ij}\beta_j + \sum_{j \in G_{\text{SSE}}} x_{ij}\beta_j. \tag{6.4}$$

We applied three methods, the gvsnss with five fold cross validation, the gvsnss with the Bayes factor, and the lasso with ten fold cross validation to estimate parameters in model (6.4). To carry out the parameter estimations, each column of design matrix X was standardized to have mean zero and variance one. The results are shown in Figure 7. The estimation results under the lasso suggest that covariates in the Susan Shepard Effect group do have positive effects on the response variable while the results under the two gvsnss estimations imply that covariates in the Susan Shepard Effect group do not have such effects.

We also carried out 100 sub-sampling estimations for the model. At each sub-sampling instance, we randomly split two thirds of the data into the training set and one third of the data into the test set. We used data from the training set to estimate parameters in model (6.4) and data from the test set to compute the predictive mean squared error. We also computed the number of covariates with non-zero estimated coefficients and the number of covariates with positive estimated coefficients in the Susan Shepard Effect group. The results are shown in Table 2.

7 Discussion

We have proposed a specified prior, called the nested spike and slab prior, to model collective behavior of regression coefficients in grouped variable selection. We have developed numerical procedures for solving the optimization problem related to maximum a posteriori estimation for the model. Simulation studies showed that the proposed estimator performs relatively well in variable selection when within-group-sparsity is present. However, we have found the proposed estimator will loss its advantage in parameter estimation if groups that contain the true covariates also contain too many redundant covariates. Subsequent asymptotic analysis also confirmed our findings.

With suitable modifications, the nested spike and slab prior can be extended to tackle grouped variable selection problems in the generalized linear models, time series models such as autoregressive and moving average models, or graphical models in covariance matrix estimation.

Acknowledgments

Tso-Jung Yen is supported by grants NSC 97-3112-B-001-020 and NSC 98-3112-B-001-027 in the National Research Program for Genomic Medicine and Academia Sinica grant AS-100-TP2-C01. Yu-Min Yen would like to thank Professor Oliver Linton for his encouragement and helpful suggestions.

A Proof of Theorems 5.1 and 5.2

Proof of Theorem 5.1. Now define $w = \widehat{\beta}^{\tau} - \beta$. It can be shown that w is the minimizer of the objective function $V_\tau(w^*, \beta, G)$ defined in (5.1) with respect to w^*. Therefore $V_\tau(w, \beta, G) \leq V_\tau(0, \beta, G)$. Here $V_\tau(0, \beta, G)$ can be explicitly expressed as

$$\begin{aligned} V_\tau(0, \beta, G) &= ||\epsilon||_2^2 + \lambda||\beta||_2^2 \\ &\quad + \rho_1 \sum_{j=1}^{p} \frac{\log(1 + \tau^{-1}|\beta_j|)}{\log(1 + \tau^{-1})} + \rho_2 \sum_{k=1}^{m} \sqrt{q_k} \frac{\log(1 + \tau^{-1}||\beta_{G_k}||_2)}{\log(1 + \tau^{-1})}. \end{aligned}$$

where $\epsilon = y - X\beta$. Further note that

$$\begin{aligned} & \big|\big|\epsilon - Xw\big|\big|_2^2 + \lambda||w + \beta||_2^2 \\ = \;& \epsilon^T\epsilon + w^T X^T X w - 2w^T X^T \epsilon + \lambda(w^T w + 2w^T \beta + \beta^T \beta) \\ = \;& ||\epsilon||_2^2 + w^T(X^T X + \lambda)w - 2w^T(X^T\epsilon - \lambda\beta) + \lambda||\beta||_2^2. \end{aligned}$$

With the results given above, we can compute $V_\tau(w, \beta, G) - V_\tau(0, \beta, G)$. In addition, since $V_\tau(w, \beta, G) - V_\tau(0, \beta, G) \leq 0$, therefore by rearranging the terms in $V_\tau(w, \beta, G) - V_\tau(0, \beta, G)$, we obtain

$$\begin{aligned} & w^T(X^T X + \lambda)w & \text{(A.1)} \\ \leq\; & 2w^T(X^T\epsilon - \lambda\beta) & \text{(A.2)} \\ & + \rho_1 \sum_{j=1}^{p} \left[\frac{\log(1 + \tau^{-1}|\beta_j|)}{\log(1 + \tau^{-1})} - \frac{\log(1 + \tau^{-1}|w_j + \beta_j|)}{\log(1 + \tau^{-1})} \right] & \text{(A.3)} \\ & + \rho_2 \sum_{k=1}^{m} \sqrt{q_k} \left[\frac{\log(1 + \tau^{-1}||\beta_{G_k}||_2)}{\log(1 + \tau^{-1})} - \frac{\log(1 + \tau^{-1}||w_{G_k} + \beta_{G_k}||_2)}{\log(1 + \tau^{-1})} \right]. & \text{(A.4)} \end{aligned}$$

Note that by Assumption 1, (A.1) can be bounded from below in a way such that

$$w^T(X^T X + \lambda I)w \geq n(\kappa_n + \lambda n^{-1})||w||_2^2. \tag{A.5}$$

In the following discussion we derive inequalities to bound (A.2), (A.3) and (A.4).

Deriving an upper bound for (A.3). We first derive an inequality to bound the difference $\sum_{j=1}^{p}[\log(1 + \tau^{-1}|\beta_j|) - \log(1 + \tau^{-1}|w_j + \beta_j|)$. For $j \in \widehat{S}^{\tau} = \{j : \widehat{\beta}_j^{\tau} \neq 0\}$, $|w_j + \beta_j| = |\widehat{\beta}_j^{\tau}| > 0$. Then given that $\tau \in [0, 1)$, for $j \in \widehat{S}^{\tau}$, we have

$$\begin{aligned} \log\left(\frac{1 + \tau^{-1}|\beta_j|}{1 + \tau^{-1}|w_j + \beta_j|} \right) &= \log\left(1 + \frac{|\beta_j| - |w_j + \beta_j|}{\tau + |w_j + \beta_j|} \right) \\ &\leq \frac{|\beta_j| - |w_j + \beta_j|}{\tau + |w_j + \beta_j|} \\ &\leq \frac{|\beta_j| - |w_j + \beta_j| + |w_j|}{\tau + |w_j + \beta_j|}. \end{aligned} \tag{A.6}$$

Now for $j \in \widehat{S}^\tau \cap S^c$, we have $\beta_j = 0$, therefore for $j \in \widehat{S}^\tau \cap S^c$, the right hand side of (A.6) is zero. For $j \in \widehat{S}^\tau \cap S$, note that $|\beta_j| - |w_j + \beta_j| \le |\beta_j - w_j - \beta_j| = |w_j|$. Then with the result given above, we have

$$\begin{aligned}
\sum_{j \in \widehat{S}^\tau} \log\left(\frac{1+\tau^{-1}|\beta_j|}{1+\tau^{-1}|w_j+\beta_j|}\right) &\le \sum_{j \in \widehat{S}^\tau \cap S} \frac{|\beta_j - w_j - \beta_j| + |w_j|}{\tau + |w_j + \beta_j|} \\
&\le 2c_2^{-1} \sum_{j \in \widehat{S}^\tau \cap S} |w_j| \\
&\le 2c_2^{-1} \sum_{j \in S} |w_j| \\
&\le 2c_2^{-1} s^{1/2} ||w||_2, \qquad \text{(A.7)}
\end{aligned}$$

where $c_2 = \min_{j \in \widehat{S}^\tau} |\widehat{\beta}_j|$.

Now consider the summation over indices $j \in (\widehat{S}^\tau)^c$. Note that for $j \in (\widehat{S}^\tau)^c \cap S^c$, we have $\widehat{\beta}_j^\tau = \beta_j = 0$, therefore the difference $\log(1+\tau^{-1}|\beta_j|) - \log(1+\tau^{-1}|w_j+\beta_j|) = 0$. On the other hand, for $j \in (\widehat{S}^\tau)^c \cap S$, we have $|w_j + \beta_j| = |\widehat{\beta}_j^\tau - \beta_j + \beta_j| = 0$ and $|\beta_j| = |\widehat{\beta}_j^\tau - \beta_j| = |w_j|$. Therefore for $j \in (\widehat{S}^\tau)^c \cap S$, we have

$$\log(1+\tau^{-1}|\beta_j|) - \log(1+\tau^{-1}|w_j+\beta_j|) = \log(\tau + |w_j|) + \log(\tau^{-1}),$$

In addition, for $\tau \in [0,1)$, $\log(\tau + |w_j|) \le \log(1 + |w_j|) \le |w_j|$. Now with $c_3 = \min_{j \in S} |\beta_j|$, we have $c_3 \le \min_{j \in (\widehat{S}^\tau)^c \cap S} |\beta_j| = \min_{j \in (\widehat{S}^\tau)^c \cap S} |w_j| \le |w_j|$ for any $j \in (\widehat{S}^\tau)^c \cap S$. Therefore with the results given above, we have

$$\begin{aligned}
&\sum_{j \in (\widehat{S}^\tau)^c} \log(1+\tau^{-1}|\beta_j|) - \log(1+\tau^{-1}|w_j+\beta_j|) \\
\le\ & \sum_{j \in \widehat{S}^c \cap S} |w_j| \left[1 + c_3^{-1} \log(\tau^{-1})\right] \\
\le\ & \left[1 + c_3^{-1} \log(\tau^{-1})\right] \sum_{j \in S} |w_j| \\
\le\ & \left[1 + c_3^{-1} \log(\tau^{-1})\right] s^{1/2} ||w||_2. \qquad \text{(A.8)}
\end{aligned}$$

For $\tau \in [0,1)$, we have $[\log(1+\tau^{-1})]^{-1} \le [\log(\tau^{-1})]^{-1}$. Now combining results in (A.7) and (A.8), we can bound (A.3) in a way such that

$$\begin{aligned}
&\rho_1 \sum_{j=1}^p \left[\frac{\log(1+\tau^{-1}|\beta_j|)}{\log(1+\tau^{-1})} - \frac{\log(1+\tau^{-1}|w_j+\beta_j|)}{\log(1+\tau^{-1})}\right] \\
\le\ & \frac{\rho_1}{\log(\tau^{-1})} \left[\sum_{j \in \widehat{S}} \log\left(\frac{1+\tau^{-1}|\beta_j|}{1+\tau^{-1}|w_j+\beta_j|}\right) + \sum_{j \in \widehat{S}^c} \log\left(\frac{1+\tau^{-1}|\beta_j|}{1+\tau^{-1}|w_j+\beta_j|}\right)\right] \\
\le\ & \frac{\rho_1}{\log(\tau^{-1})} \left\{2c_2^{-1} s^{1/2} ||w||_2 + \left[1 + c_3^{-1} \log(\tau^{-1})\right] s^{1/2} ||w||_2\right\} \\
=\ & \rho_1 \left[\frac{2c_2^{-1} + 1}{\log(\tau^{-1})} + c_3^{-1}\right] s^{1/2} ||w||_2. \qquad \text{(A.9)}
\end{aligned}$$

Deriving an upper bound for (A.4). Similarly, for $k \in \widehat{R}^\tau = \{k : ||\widehat{\beta}^\tau_{G_k}||_2 > 0\}$, we have $||w_{G_k} + \beta_{G_k}||_2 = ||\widehat{\beta}^\tau_{G_k}||_2 > 0$. In turn, we have

$$\log\left(\frac{1+\tau^{-1}||\beta_{G_k}||_2}{1+\tau^{-1}||w_{G_k}+\beta_{G_k}||_2}\right) \;\leq\; \frac{||\beta_{G_k}||_2 - ||w_{G_k}+\beta_{G_k}||_2 + ||w_{G_k}||_2}{\tau + ||w_{G_k}+\beta_{G_k}||_2}. \quad (A.10)$$

for $k \in \widehat{R}^\tau$.

Now if $k \in \widehat{R}^\tau \cap R^c$, where $R^c = \{k : ||\beta_{G_k}||_2 = 0\}$, then the right hand side of (A.10) is zero. On the other hand, for $j \in G_{\widehat{R}^\tau} \cap S$, we have $c_2 = \min_{j\in\widehat{S}^\tau} |\widehat{\beta}_j| \leq \min_{k\in\widehat{R}^\tau} ||\widehat{\beta}_{G_k}||_2 \leq ||\widehat{\beta}_{G_k}||_2$. In addition, $||\beta_{G_k}||_2 - ||w_{G_k} + \beta_{G_k}||_2 \leq ||\beta_{G_k} - w_{G_k} - \beta_{G_k}||_2 = ||w_{G_k}||_2$. Then with the results given above, we can further obtain

$$\begin{aligned}
\sum_{k\in\widehat{R}^\tau} \sqrt{q_k}\log\left(\frac{1+\tau^{-1}||\beta_{G_k}||_2}{1+\tau^{-1}||w_{G_k}+\beta_{G_k}||_2}\right) &\leq \sum_{k\in\widehat{R}^\tau\cap R} \sqrt{q_k}\frac{2||w_{G_k}||_2}{\tau+||w_{G_k}+\beta_{G_k}||_2} \\
&\leq \sum_{k\in\widehat{R}^\tau\cap R} \sqrt{q_k}\frac{2||w_{G_k}||_2}{c_2} \\
&\leq 2c_2^{-1}\Bigg(\sum_{k\in R}\sqrt{q_k}^2\Bigg)^{1/2}\Bigg(\sum_{k\in R}||w_{G_k}||_2^2\Bigg)^{1/2} \\
&\leq 2c_2^{-1}q_R^{1/2}||w||_2, \qquad (A.11)
\end{aligned}$$

where $q_R = |G_R| = \sum_{k\in R} q_k$ is the number of indices covered by G_R. We now consider the summation over indices $k \in (\widehat{R}^\tau)^c$. If $k \in (\widehat{R}^\tau)^c$, $||\widehat{\beta}^\tau_{G_k}||_2 = 0$. Therefore, we have $||w_{G_k} + \beta_{G_k}||_2 = ||\widehat{\beta}^\tau_{G_k} - \beta_{G_k} + \beta_{G_k}||_2 = 0$ and $||\beta_{G_k}||_2 = ||\widehat{\beta}^\tau_{G_k} - \beta_{G_k}||_2 = ||w_{G_k}||_2$. In turn,

$$\log(1+\tau^{-1}||\beta_{G_k}||_2) - \log(1+\tau^{-1}||w_{G_k}+\beta_{G_k}||_2) = \log(\tau + ||w_{G_k}||_2) + \log(\tau^{-1})$$

for $k \in (\widehat{R}^\tau)^c$. In addition, for $\tau \in [0,1)$, $\log(\tau+||\beta_{G_k}||_2) \leq \log(1+||\beta_{G_k}||_2) \leq ||\beta_{G_k}||_2$. Further note that

$$c_3 = \min_{j\in S}|\beta_j| \leq \min_{k\in(\widehat{R}^\tau)^c, ||\beta_{G_k}||_2\neq 0} ||\beta_{G_k}||_2 = \min_{k\in(\widehat{R}^\tau)^c, ||w_{G_k}||_2\neq 0} ||w_{G_k}||_2.$$

Moreover, for an arbitrary index $k \in (\widehat{R}^\tau)^c$, $||w_{G_k}||_2 = ||\beta_{G_k}||_2$, therefore $||w_{G_k}||_2 \neq 0$ implies $||\beta_{G_k}||_2 \neq 0$ and the index $k \in R$. Now by applying the results given above, we have

$$\begin{aligned}
& \sum_{k\in(\widehat{R}^\tau)^c} \sqrt{q_k}\big[\log(1+\tau^{-1}||\beta_{G_k}||_2) - \log(1+\tau^{-1}||w_{G_k}+\beta_{G_k}||_2)\big] \\
= & \sum_{k\in(\widehat{R}^\tau)^c, ||w_{G_k}||_2\neq 0} \sqrt{q_k}\big[\log(\tau+||w_{G_k}||_2)+\log(\tau^{-1})\big] \\
& + \sum_{k\in(\widehat{R}^\tau)^c, ||w_{G_k}||_2 = 0} \sqrt{q_k}\big[\log(\tau+||w_{G_k}||_2)+\log(\tau^{-1})\big] \\
\leq & \big[1+c_3^{-1}\log(\tau^{-1})\big] \sum_{k\in(\widehat{R}^\tau)^c, k\in R} \sqrt{q_k}||w_{G_k}||_2 \\
\leq & \big[1+c_3^{-1}\log(\tau^{-1})\big] q_R^{1/2}||w||_2. \qquad \text{(A.12)}
\end{aligned}$$

Combining the results in (A.11) and (A.12), we can bound (A.4) in a way such that

$$\begin{aligned}
& \rho_2 \sum_{k=1}^{m} \sqrt{q_k}\left[\frac{\log(1+\tau^{-1}||\beta_{G_k}||_2)}{\log(1+\tau^{-1})} - \frac{\log(1+\tau^{-1}||w_{G_k}+\beta_{G_k}||_2)}{\log(1+\tau^{-1})}\right] \\
\leq & \frac{\rho_2}{\log(\tau^{-1})}\left[\sum_{k\in\widehat{R^\tau}} \sqrt{q_k}\log\left(\frac{1+\tau^{-1}||\beta_{G_k}||_2}{1+\tau^{-1}||w_{G_k}+\beta_{G_k}||_2}\right)\right. \\
& \left. + \sum_{k\in(\widehat{R}^\tau)^c} \sqrt{q_k}\log\left(\frac{1+\tau^{-1}||\beta_{G_k}||_2}{1+\tau^{-1}||w_{G_k}+\beta_{G_k}||_2}\right)\right] \\
\leq & \frac{\rho_2}{\log(\tau^{-1})}\Big\{2c_2^{-1}q_R^{1/2}||w||_2 + \big[1+c_3^{-1}\log(\tau^{-1})\big]q_R^{1/2}||w||_2\Big\} \\
= & \rho_2\left[\frac{2c_2^{-1}+1}{\log(\tau^{-1})}+c_3^{-1}\right]q_R^{1/2}||w||_2. \qquad \text{(A.13)}
\end{aligned}$$

Deriving an upper bound for (A.2). First note that

$$w^T X^T \epsilon \leq ||w||_1 ||X^T\epsilon||_\infty.$$

Now for $||w||_1$, we can decompose it as

$$||w||_1 = ||w_{\widehat{S}^\tau\cap S}||_1 + ||w_{\widehat{S}^\tau\cap S^c}||_1 + ||w_{(\widehat{S}^\tau)^c\cap S}||_1 + ||w_{(\widehat{S}^\tau)^c\cap S^c}||_1. \qquad \text{(A.14)}$$

Note that for the first and third terms on the right hand side of (A.14), we have $||w_{\widehat{S}^\tau\cap S}||_1 \leq ||w_S||_1$ and $||w_{(\widehat{S}^\tau)^c\cap S}||_1 \leq ||w_S||_1$. For the second term on the right hand side of (A.14), we have $||w_{\widehat{S}^\tau\cap S^c}||_1 \leq ||w_{\widehat{S}^\tau}||_1$. The fourth term on the right hand side of (A.14) is zero since $(\widehat{S}^\tau)^c \cap S^c$ is an intersection of indices for entries with zero values in β and entries with zero values in $\widehat{\beta}^\tau$. With the results given above, we can further bound $||w||_1$ in a way such that

$$\begin{aligned}
||w||_1 & \leq 2||w_S||_1 + ||w_{\widehat{S}^\tau}||_1 \\
& \leq s^{1/2}\left[2+\left(\frac{\widehat{S}^\tau}{s}\right)^{1/2}\right]||w||_2. \qquad \text{(A.15)}
\end{aligned}$$

With the result in (A.15), we can bound (A.2) in a way such that

$$\begin{aligned}2w^T(X^T\epsilon - \lambda\beta) &\leq 2||w||_1||X^T\epsilon||_\infty + 2\lambda|w^T\beta| \\ &\leq 2s^{1/2}\left[2 + \left(\frac{\widehat{s}^\tau}{s}\right)^{1/2}\right]||w||_2||X^T\epsilon||_\infty \\ &\quad +2\lambda||w||_2 s^{1/2}\max_{j\in S}|\beta_j|. \qquad \text{(A.16)}\end{aligned}$$

Combining the results (A.9), (A.13) and (A.16), we obtain

$$\begin{aligned}n(\kappa_n + \lambda n^{-1})||w||_2^2 &\leq 2s^{1/2}\left[2 + \left(\frac{\widehat{s}^\tau}{s}\right)^{1/2}\right]||w||_2||X^T\epsilon||_\infty + 2\lambda s^{1/2}\max_{j\in S}|\beta_j|||w||_2 \\ &\quad +\rho_1\left[\frac{2c_2^{-1}+1}{\log(\tau^{-1})} + c_3^{-1}\right]s^{1/2}||w||_2 \\ &\quad +\rho_2\left[\frac{2c_2^{-1}+1}{\log(\tau^{-1})} + c_3^{-1}\right]q_R^{1/2}||w||_2. \qquad \text{(A.17)}\end{aligned}$$

Then by using the fact that $s = |S| \leq |G_R| = q_R$ and doing some rearrangement in (A.17), we obtain the inequality (5.4), which completes the proof. □

Proof of Theorem 5.2. We start our proof by showing that with at least $1-\alpha$ probability, the inequality $2||X^T\epsilon||_\infty < \psi_n$ will hold, where ψ_n is defined in (5.5). Note that $\{2||X^T\epsilon||_\infty < \psi_n\}$ is equivalent to the following event:

$$\mathcal{A} = \bigcap_{k=1}^m \left\{2||X_{G_k}^T\epsilon||_\infty < \psi_n\right\}.$$

We will establish the inequality $\mathbb{P}(\mathcal{A}) = 1 - \mathbb{P}(\mathcal{A}^c) \geq 1 - \alpha$ by showing that given ψ_n is defined in (5.5), $\mathbb{P}(\mathcal{A}^c) \leq \alpha$. The technique we use to derive the inequality $\mathbb{P}(\mathcal{A}^c) \leq \alpha$ is borrowed from Lemma B.1 of [5]. Note that the tail probability $\mathbb{P}(\mathcal{A}^c)$ can be bounded in a way such that

$$\begin{aligned}\mathbb{P}(\mathcal{A}^c) &= \mathbb{P}\left(\bigcup_{k=1}^m \left\{2||X_{G_k}^T\epsilon||_\infty \geq \psi_n\right\}\right) \\ &\leq \sum_{k=1}^m \mathbb{P}\left(||X_{G_k}^T\epsilon||_\infty \geq \frac{\psi_n}{2}\right) \leq \sum_{k=1}^m\sum_{j\in G_k}\mathbb{P}\left(\left|\sum_{i=1}^n x_{ij}\epsilon_i\right| \geq \frac{\psi_n}{2}\right). \qquad \text{(A.18)}\end{aligned}$$

Under assumptions given in Theorem 5.2, ϵ_i's are i.i.d. normal variables with mean zero and variance σ^2, therefore $\sum_{i=1}^n x_{ij}\epsilon_i$ is a normal variable with mean zero and variance $\sigma^2\sum_{i=1}^n x_{ij}^2 = n\zeta_j\sigma^2$. In turn, we can express $|\sum_{i=1}^n x_{ij}\epsilon_i| = \sqrt{n\zeta_j}\sigma|Z|$, where Z is a standard normal variable. By using the Chernoff bound argument on the tail probability of a standard normal variable, we can bound the right hand side of (A.18) in a way such that

$$\begin{aligned}\sum_{k=1}^{m}\sum_{j\in G_k}\mathbb{P}\Big(\Big|\sum_{i=1}^{n}x_{ij}\epsilon_i\Big|\geq\frac{\psi_n}{2}\Big) &\leq \sum_{k=1}^{m}q_k\mathbb{P}\Big(|Z|\geq\frac{\psi_n}{2\sqrt{n\max_j\zeta_j}\sigma}\Big)\\ &\leq m\sum_{k=1}^{m}\frac{q_k}{m}\exp\Big(-\frac{\psi_n^2}{8n\max_j\zeta_j\sigma^2}\Big)\\ &\leq m\exp\Big(-\frac{\psi_n^2}{8n\sigma^2\max_j\zeta_j}+\log\overline{q}\Big),\end{aligned} \tag{A.19}$$

where $\overline{q}=m^{-1}\sum_{k=1}^{k}q_k$. With ψ_n defined in (5.5), the right hand side of (A.19) is equal to α, and further with (A.18), we obtain $\mathbb{P}(\mathcal{A}^c)\leq\alpha$, which implies that with ψ_n defined in (5.5), $\mathbb{P}(\mathcal{A})=1-\mathbb{P}(\mathcal{A}^c)\geq 1-\alpha$.

To complete the proof, note that since we have assumed $\tau=n^{-1}$, therefore effectively we have $\widehat{\beta}^{\tau}\rightarrow\widehat{\beta}_{\text{gvsnss}}$ and $\widehat{s}^{\tau}\rightarrow\widehat{s}$ as $n\rightarrow\infty$. Therefore with the result from Theorem 5.1 and the assumptions on λ, ρ_1, ρ_2 and τ, as $n\rightarrow\infty$, the inequality

$$||\widehat{\beta}_{\text{gvsnss}}-\beta||_2\leq\frac{q_R^{1/2}}{(\kappa_n+\Omega_n)}\frac{\Lambda_n\psi_n}{n} \tag{A.20}$$

will hold with $1-\alpha$ probability, where Λ_n is defined in (5.7) and $\Omega_n=n^{-1}\lambda$ is defined in (5.8) and ψ_n defined in (5.5), which completes the proof. □

B Proof of Theorem 5.3

Proof of Theorem 5.3. Now define

$$\begin{aligned}U_\tau(\beta',G^*,G^{**}) &= \rho_2\sum_{k=1}^{m^*}\sqrt{q_k^*}\frac{\log(1+\tau^{-1}||\beta'_{G_k^*}||_2)}{\log(1+\tau^{-1})}\\ &\quad -\rho_2\sum_{l=1}^{m^{**}}\sqrt{q_l^{**}}\frac{\log(1+\tau^{-1}||\beta'_{G_l^{**}}||_2)}{\log(1+\tau^{-1})}.\end{aligned} \tag{B.1}$$

where $\beta'_{G_k^*}$ is the coefficient vector in which the elements are those indexed by G_k^* in the vector β'. The vector $\beta'_{G_l^{**}}$ follows a similar definition. The function (B.1) is the difference between the log-sum penalties involving l_2-norms indexed by G^* and G^{**}. Note that, with (B.1), the objective function $V_\tau(0,\beta',G^*)$ in (5.10) can be re-expressed as

$$V_\tau(0,\beta',G^*)=V_\tau(0,\beta',G^{**})+U_\tau(\beta',G^*,G^{**}). \tag{B.2}$$

Since $\widehat{\beta}^{\tau*}$ is the minimizer of $V_\tau(0,\beta',G^*)$, therefore it must be the solution to the following subgradient equations:

$$2X^T(y-X\beta')-2\lambda\beta'-\rho_1 g'-\rho_2\sqrt{q^{**}}u^{**}-\rho_2(\sqrt{q^*}u^*-\sqrt{q^{**}}u^{**})=0, \tag{B.3}$$

where

$$(g')_j = \frac{h'_j}{[\tau \log(1+\tau^{-1})](1+\tau^{-1}|\beta'_j|)},$$

with $h'_j = \mathrm{sign}(\beta'_j)$ if $\beta'_j \neq 0$ and $-1 \leq h'_j \leq 1$ if $\beta'_j = 0$, and

$$(u^*)_j = \frac{v^*_j}{[\tau \log(1+\tau^{-1})](1+\tau^{-1}||\beta'_{G^*_{k_j}}||_2)}$$

with $v^*_j = \beta'_j/||\beta'_{G^*_{k_j}}||_2$ if $||\beta'_{G^*_{k_j}}||_2 > 0$ and $\sum_{j\in G^*_{k_j}}(v^*_j)^2 \leq 1$ if $||\beta'_{G^*_{k_j}}||_2 = 0$, where k_j is the index for the group that j belongs to, i.e. if $j \in G^*_{k'}$, then $k_j = k'$. The quantity $(u^{**})_j$ follows a similar definition. In addition, $(q^*)_j = q^*_{k_j}$ and $(q^{**})_j = q^{**}_{l_j}$.

Note that the derivation of the subgradient equations (B.3) has explicitly used representation (B.2), and after some simple arrangement, (B.3) becomes

$$2X^T(y - X\beta) - 2\lambda\beta - \rho_1 g' - \rho_2\sqrt{q^{**}}u^{**} = \rho_2(\sqrt{q^*}u^* - \sqrt{q^{**}}u^{**}), \tag{B.4}$$

where

$$\begin{aligned}(\sqrt{q^*}u^* - \sqrt{q^{**}}u^{**})_j \;&=\; \frac{1}{\log(1+\tau^{-1})}\Bigg[\frac{(\tau + ||\beta'_{G^{**}_{l_j}}||_2)\sqrt{q^*_{k_j}}v^*_j}{(\tau + ||\beta'_{G^*_{k_j}}||_2)(\tau + ||\beta'_{G^{**}_{l_j}}||_2)} \\ &\quad - \frac{(\tau + ||\beta'_{G^*_{k_j}}||_2)\sqrt{q^{**}_{l_j}}v^{**}_j}{(\tau + ||\beta'_{G^*_{k_j}}||_2)(\tau + ||\beta'_{G^{**}_{l_j}}||_2)}\Bigg].\end{aligned} \tag{B.5}$$

For each j, one of the following four cases will occur: (i) $||\beta'_{G^*_{k_j}}||_2 = 0$ and $||\beta'_{G^{**}_{l_j}}||_2 = 0$; (ii) $||\beta'_{G^*_{k_j}}||_2 > 0$ and $||\beta'_{G^{**}_{l_j}}||_2 = 0$; (iii) $||\beta'_{G^*_{k_j}}||_2 = 0$ and $||\beta'_{G^{**}_{l_j}}||_2 > 0$; and (iv) $||\beta'_{G^*_{k_j}}||_2 > 0$ and $||\beta'_{G^{**}_{l_j}}||_2 > 0$. In the following discussion, we will evaluate (B.5) under the four cases.

We consider case (i) first. If (i) occurs, then all regression coefficients with indices in $G^*_{k_j}$ or $G^{**}_{l_j}$ will be zero. It implies that $\beta'_j = 0$ and by definitions, v^*_j is an arbitrary quantity such that $0 \leq (v^*_j)^2 \leq \sum_{j\in G^*_{k_j}}(v^*_j)^2 \leq 1$. The same property applies to v^{**}_j. For practical purposes, we choose $v^*_j = \tau$ and $v^{**}_j = \tau$. Then under case (i),

$$\begin{aligned}(\sqrt{q^*}u^* - \sqrt{q^{**}}u^{**})_j \;&=\; \frac{1}{\tau\log(1+\tau^{-1})}\left[\frac{\tau\sqrt{q^*_{k_j}}v^*_j}{\tau} - \frac{\tau\sqrt{q^{**}_{l_j}}v^{**}_j}{\tau}\right] \\ &=\; \frac{\left(\sqrt{q^*_{k_j}} - \sqrt{q^{**}_{l_j}}\right)}{\log(1+\tau^{-1})}.\end{aligned} \tag{B.6}$$

Now consider case (ii). If (ii) holds, then by definition, $v^*_j = \beta'_j/||\beta_{G^*_{k_j}}||_2$. In addition, since $||\beta'_{G^{**}_{l_j}}||_2 = 0$, therefore v^{**}_j is an arbitrary quantity such that $0 \leq$

$(v_j^{**})^2 \leq \sum_{j \in G_{l_j}^*}(v_j^{**})^2 \leq 1$. For practical purposes, we choose $v_j^{**} = \tau$. Moreover, $||\beta'_{G_{l_j}^{**}}||_2 = 0$ implies that all coefficients with indices in $G_{l_j}^{**}$ are zero. Therefore $\beta'_j = 0$ and $v_j^* = \beta'_j/||\beta'_{G_{k_j}^*}||_2 = 0$. Then under case (ii),

$$\begin{aligned}
(\sqrt{q^*}u^* - \sqrt{q^{**}}u^{**})_j &= \frac{1}{\tau\log(1+\tau^{-1})}\left[\frac{\tau\sqrt{q_{k_j}^*}v_j^*}{\tau + ||\beta'_{G_{k_j}^*}||_2} - \frac{(\tau + ||\beta'_{G_{k_j}^*}||_2)\sqrt{q_{l_j}^{**}}v_j^{**}}{\tau + ||\beta'_{G_{k_j}^*}||_2}\right] \\
&= -\frac{\sqrt{q_{l_j}^{**}}}{\log(1+\tau^{-1})}. \qquad \text{(B.7)}
\end{aligned}$$

Now consider case (iii). Under case (iii), since $||\beta'_{G_{l_j}^{**}}||_2 > 0$, therefore $v_j^{**} = \beta'_j/||\beta'_{G_{l_j}^{**}}||_2$. In addition, $||\beta'_{G_{k_j}^*}||_2 = 0$ implies that all coefficients with indices in $G_{k_j}^*$ are zero. Therefore $\beta'_j = 0$ and $v_j^{**} = \beta'_j/||\beta'_{G_{l_j}^{**}}||_2 = 0$. In addition, v_j^* is an arbitrary quantity such that $0 \leq (v_j^*)^2 \leq \sum_{j \in G_{k_j}^*}(v_j^*)^2 \leq 1$. Here we let $v_j^* = \tau$. Therefore under case (iii),

$$\begin{aligned}
(\sqrt{q^*}u^* - \sqrt{q^{**}}u^{**})_j &= \frac{1}{\tau\log(1+\tau^{-1})}\left[\frac{(\tau + ||\beta'_{G_{l_j}^{**}}||_2)\sqrt{q_{k_j}^*}v_j^*}{\tau + ||\beta'_{G_{l_j}^{**}}||_2} - \frac{\tau\sqrt{q_{l_j}^{**}}v_j^{**}}{\tau + ||\beta'_{G_{l_j}^{**}}||_2}\right] \\
&= \frac{\sqrt{q_{k_j}^*}}{\log(1+\tau^{-1})}. \qquad \text{(B.8)}
\end{aligned}$$

Finally we consider case (iv). Under case (iv), $v_j^* = \beta'_j/||\beta'_{G_{k_j}^*}||_2$ and $v_j^{**} = \beta'_j/||\beta'_{G_{l_j}^{**}}||_2$. Further by direct calculation, we have

$$\begin{aligned}
(\sqrt{q^*}u^* - \sqrt{q^{**}}u^{**})_j &= \frac{1}{\log(1+\tau^{-1})}\left[\frac{\sqrt{q_{k_j}^*}v_j^*}{(\tau + ||\beta'_{G_{k_j}^*}||_2)} - \frac{\sqrt{q_{l_j}^{**}}v_j^{**}}{(\tau + ||\beta'_{G_{l_j}^{**}}||_2)}\right] \\
&= \frac{1}{\log(1+\tau^{-1})}\left[\frac{\sqrt{q_{k_j}^*}\beta'_j}{(\tau + ||\beta'_{G_{k_j}^*}||_2)||\beta'_{G_{k_j}^*}||_2}\right. \\
&\qquad \left. - \frac{\sqrt{q_{l_j}^{**}}\beta'_j}{(\tau + ||\beta'_{G_{l_j}^{**}}||_2)||\beta'_{G_{l_j}^{**}}||_2}\right]. \qquad \text{(B.9)}
\end{aligned}$$

Now with $\tau = n^{-1}$ and the results from (B.6), (B.7), (B.8), and (B.9), we can see that $|U_\tau(\beta', G^*, G^{**})| = O(\rho_2 \max_k \sqrt{q_k}[\log(n)]^{-1})$. Therefore if $\rho_2 \max_k \sqrt{q_k} = o(\log(n))$, $U_\tau(\beta', G^*, G^{**})$ will approach to zero when $n \to \infty$. It further implies that the right hand side of (B.4) will become zero when $n \to \infty$. On the other hand, the left hand side of (B.4) is just the subgradient vector of the objective function $V_\tau(0, \beta', G^{**})$. Therefore when $\tau \to 0$, (B.4) becomes the subgradient equations of $\lim_{\tau\to 0} V_\tau(0, \beta', G^{**})$. Since $\widehat{\beta}^*_{\text{gvsnss}}$ is the solution of the subgradient equations (B.3) when $\tau \to 0$ and (B.4) is just a rearrangement of (B.3), therefore $\widehat{\beta}^*_{\text{gvsnss}}$ is also the solution to (B.4) when $\tau \to 0$. Since (B.4) becomes the subgraident equations of $\lim_{\tau\to 0} V_\tau(0, \beta', G^{**})$ when $\tau \to 0$, and the solution of (B.4) at $\tau \to 0$ is the minimizer of $\lim_{\tau\to 0} V_\tau(0, \beta', G^{**})$, there we conclude that $\widehat{\beta}^*_{\text{gvsnss}}$ is the minimizer of $\lim_{\tau\to 0} V_\tau(0, \beta', G^{**})$, which completes the proof. □

C Proof of Theorem 5.4

Proof of Theorem 5.4. Define $w = \widehat{\beta}^\tau - \beta$. It can be shown that given β and G are fixed, w is the minimizer of $V_\tau(w^*, \beta, G)$, therefore w is also the solution to the following subgradient equations:

$$2X^T Xw - 2X^T\epsilon + 2\lambda(\beta + w) + \rho_1 g + \rho_2\sqrt{q}u = 0, \tag{C.1}$$

where

$$(g)_j = \frac{h_j}{[\tau \log(1 + \tau^{-1})](1 + \tau^{-1}|w_j + \beta_j|)},$$

with $h_j = \text{sign}(w_j + \beta_j)$ if $w_j + \beta_j \neq 0$ and $-1 \leq h_j \leq 1$ if $w_j + \beta_j = 0$, and

$$(u)_j = \frac{v_j}{[\tau \log(1 + \tau^{-1})](1 + \tau^{-1}||w_{G_{k_j}} + \beta_{G_{k_j}}||_2)}$$

with $v_j = (w_j + \beta_j)/||w_{G_{k_j}} + \beta_{G_{k_j}}||_2$ if $||w_{G_{k_j}} + \beta_{G_{k_j}}||_2 > 0$ and $\sum_{j\in G_{k_j}}(v_j)^2 \leq 1$ if $||w_{G_{k_j}} + \beta_{G_{k_j}}||_2 = 0$, where k_j is the index for the group that j belongs to.

Let $S_1^c = S^c \cap G_R$ and $S_2^c = S^c \cap G_{R^c}$. Here S^c is the set of indices for redundant covariates, i.e. the covariates with zero coefficients. In addition, S_1^c is the set of indices for the redundant covariates covered by G_R, and S_2^c is the set of indices for the redundant covariates covered by G_{R^c}. By definition, $G_{R^c} \subseteq S^c$, therefore we have $S_2^c = G_{R^c}$. In addition, S, S_1^c and G_{R^c} are three disjoint index sets and $S \cup S_1^c \cup G_{R^c} = \{1, 2, \cdots, p\}$. With the results given above, we can re-express (C.1) as

$$\begin{aligned} &2\begin{pmatrix} X_S^T X_S & X_S^T X_{S_1^c} & X_S X_{G_{R^c}} \\ X_{S_1^c}^T X_S & X_{S_1^c}^T X_{S_1^c} & X_{S_1^c}^T X_{G_{R^c}} \\ X_{G_{R^c}}^T X_S & X_{G_{R^c}}^T X_{S_1^c} & X_{G_{R^c}}^T X_{G_{R^c}} \end{pmatrix} \begin{pmatrix} w_S \\ w_{S_1^c} \\ w_{G_{R^c}} \end{pmatrix} - 2\begin{pmatrix} X_S^T\epsilon \\ X_{S_1^c}^T\epsilon \\ X_{G_{R^c}}^T\epsilon \end{pmatrix} \\ &+2\lambda\begin{pmatrix} w_S + \beta_S \\ w_{S_1^c} + \beta_{S_1^c} \\ w_{G_{R^c}} + \beta_{G_{R^c}} \end{pmatrix} + \rho_1\begin{pmatrix} g_S \\ g_{S_1^c} \\ g_{G_{R^c}} \end{pmatrix} + \rho_2\begin{pmatrix} \sqrt{q_S}u_S \\ \sqrt{q_{S_1^c}}u_{S_1^c} \\ \sqrt{q_{G_{R^c}}}u_{G_{R^c}} \end{pmatrix} = 0. \end{aligned} \tag{C.2}$$

For practical purposes, we define ϑ_j^S as the position of index j in the set S. It is equivalent to say that index j is the ϑ_j^Sth element in S. If $j \notin S$, then we just leave ϑ_j^S undefined. Similar definitions are applied to $\vartheta_j^{S_1^c}$ and $\vartheta_j^{G_{R^c}}$.

To make the sign consistency hold, we must have $w_j = \widehat{\beta}_j^{\tau} - \beta_j = 0$ for all $j \in S_1^c \cup G_{R^c}$, and $\mathrm{sign}(\widehat{\beta}_j) = \mathrm{sign}(\beta_j)$ for all $j \in S$. Given that w is the solution to (C.2), then with the arguments given above, we obtain the following conditions:

$$\big(X_S^T X_S w_S - X_S^T \epsilon + \lambda(w_S + \beta_S) + \frac{\rho_2}{2}\sqrt{q_S} u_S\big)_{\vartheta_j^S} = \Big(- \frac{\rho_1}{2} g_S \Big)_{\vartheta_j^S}, \tag{C.3}$$

for $j \in S$, and

$$\begin{aligned} -\frac{\rho_1}{2\tau\log(1+\tau^{-1})} &< \Big(X_{S_1^c}^T X_S w_S - X_{S_1^c}^T \epsilon + \frac{\rho_2}{2}\sqrt{q_{S_1^c}} u_{S_1^c} \Big)_{\vartheta_j^{S_1^c}} \\ &< \frac{\rho_1}{2\tau\log(1+\tau^{-1})} \end{aligned} \tag{C.4}$$

for $j \in S_1^c$, and

$$||2X_{G_k}^T X_S w_S - 2X_{G_k}^T \epsilon + \rho_1 g_{G_k}||_2 < \frac{\rho_2\sqrt{q_k}}{\tau\log(1+\tau^{-1})} \tag{C.5}$$

for $k \in R^c$.

The subgradient equations (C.3) are a result from the KKT conditions and the inequalities (C.4) and (C.5) are used to ensure that estimated coefficients with indices in S_1^c and G_{R^c} are zero.

Now by solving equations in (C.3) for w_S, we have

$$\begin{aligned} w_S &= (X_S^T X_S + \lambda I_{s\times s})^{-1} X_S^T \epsilon \\ &\quad -(X_S^T X_S + \lambda I_{s\times s})^{-1} \Big(\frac{\rho_1}{2} g_S + \frac{\rho_2}{2}\sqrt{q_S} u_S + \lambda\beta_S \Big). \end{aligned} \tag{C.6}$$

Note that the ϑ_j^Sth element in the last term on the right hand side of (C.6) can be expressed as

$$\begin{aligned} \Big(\frac{\rho_1}{2} g_S + \frac{\rho_2}{2}\sqrt{q_S} u_S + \lambda\beta_S \Big)_{\vartheta_j^S} &= \frac{1}{2[\tau\log(1+\tau^{-1})]} \Big[\frac{\tau\rho_1 h_j}{(\tau + |w_j + \beta_j|)} \\ &\quad + \frac{\tau\rho_2\sqrt{q_{k_j}} v_j}{(\tau + ||w_{G_{k_j}} + \beta_{G_{k_j}}||_2)} + 2\lambda\tau\log(1+\tau^{-1})\beta_j \Big]. \end{aligned} \tag{C.7}$$

Here we define $B_{S,\tau}$ by

$$(B_{S,\tau})_{\vartheta_j^S} = \frac{\tau\rho_1 h_j}{(\tau + |w_j + \beta_j|)} + \frac{\tau\rho_2\sqrt{q_{k_j}} v_j}{(\tau + ||w_{G_{k_j}} + \beta_{G_{k_j}}||_2)} + 2\lambda\tau\log(1+\tau^{-1})\beta_j. \tag{C.8}$$

By Assumption 2, $C_{SS} = n^{-1}(X_S^T X_S + \lambda I)$. Practically we can express w_S as

$$w_S = n^{-1} C_{SS}^{-1} X_S^T \epsilon - \frac{1}{2n\tau\log(1+\tau^{-1})} C_{SS}^{-1} B_{S,\tau}. \tag{C.9}$$

Sign consistency for estimated coefficients with indices in S. Now in order to ensure the sign consistency for estimated coefficients with indices in S, we impose some constraint on each entry of w_S. We focus on the following inequality:

$$|w_j| < |\beta_j|. \tag{C.10}$$

Inequality (C.10) implies that for $j \in S$, $\mathrm{sign}(\widehat{\beta}_j^{\tau}) = \mathrm{sign}(\beta_j)$. To see why it is, let us consider the case when $\beta_j > 0$. If $\beta_j > 0$, then $|w_j| < |\beta_j|$ means that either $-\beta_j < w_j = \widehat{\beta}_j^{\tau} - \beta_j < \beta_j$ or $-\beta_j < -w_j = \beta_j - \widehat{\beta}_j^{\tau} < \beta_j$, which jointly imply that $0 < \widehat{\beta}_j^{\tau} < 2\beta_j$. A similar argument can be applied to the case when $\beta_j < 0$. Therefore given that (C.10) holds, sign consistency holds for estimated coefficients with indices in S.

With representation (C.9), for $j \in S$, we can bound $|w_j|$ in a way such that

$$|w_j| \;\leq\; n^{-1}\Big|\big(C_{SS}^{-1}X_S^T\epsilon\big)_{\vartheta_j^S}\Big| + \frac{1}{2n\tau\log(1+\tau^{-1})}\Big|\big(C_{SS}^{-1}B_{S,\tau}\big)_{\vartheta_j^S}\Big|. \tag{C.11}$$

By plugging the right hand side of (C.11) into the left hand side of (C.10) and doing some rearrangements, we obtain the following inequality:

$$n^{-1}\Big|\big(C_{SS}^{-1}X_S^T\epsilon\big)_{\vartheta_j^S}\Big| \leq |\beta_j| - \frac{1}{2n\tau\log(1+\tau^{-1})}\Big|\big(C_{SS}^{-1}B_{S,\tau}\big)_{\vartheta_j^S}\Big|. \tag{C.12}$$

Further note that for any $j \in S$, $|(C_{SS}^{-1}X_S^T\epsilon)_{\vartheta_j^S}| \leq ||C_{SS}^{-1}X_S^T\epsilon||_\infty \leq \kappa_{\min}^{-1}||X_S^T\epsilon||_\infty$, where $\kappa_{\min}$ is the minimum eigenvalue of C_{SS}. Now with the results given above, we construct the following event:

$$E_1 = \left\{\epsilon : n^{-1}\kappa_{\min}^{-1}||X_S^T\epsilon||_\infty < \min_{j\in S}|\beta_j| - \frac{1}{2n\tau\log(1+\tau^{-1})}||C_{SS}^{-1}B_{S,\tau}||_\infty\right\}. \tag{C.13}$$

Since the left hand side of the inequality stated in E_1 is larger than the left hand side of (C.12), and the right hand side of the inequality stated in E_1 is smaller than the right hand side of (C.12), therefore if the inequality stated in E_1 hold, then (C.12) will hold. In turn, (C.3) and (C.10) will hold, and the sign consistency for estimated coefficients with indices in S can be established.

We go on to derive an estimate for the tail probability of E_1. Define $\psi_{1,n}$ by

$$\psi_{1,n} = \min_{j\in S}|\beta_j| - \frac{1}{2n\tau\log(1+\tau^{-1})}||C_{SS}^{-1}B_{S,\tau}||_\infty. \tag{C.14}$$

Note that E_1 is equivalent to the event $\cap_{j\in S}\{n^{-1}|\sum_{i=1}^n x_{ij}\epsilon_i| < \psi_{1,n}\kappa_{\min}\}$. On the other hand, by the assumptions on ϵ_i's and $\sum_{i=1}^n x_{ij}^2$, one can show that $\sum_{i=1}^n x_{ij}\epsilon_i$ is a normal variable with mean zero and variance $\sigma^2\sum_i x_{ij}^2 = n\sigma^2$. Therefore, we can bound the probability of E_1^c in a way such that

$$\mathbb{P}(E_1^c) \leq \sum_{j\in S}\mathbb{P}\left(\frac{1}{n}\left|\sum_{i=1}^n x_{ij}\epsilon_i\right| \geq \psi_{1,n}\kappa_{\min}\right) \leq s\mathbb{P}\left(|Z| \geq \frac{\sqrt{n}\psi_{1,n}\kappa_{\min}}{\sigma}\right), \tag{C.15}$$

where Z is a standard normal variable. By applying a Chernoff bound argument to the right hand side of (C.15), we further obtain

$$\mathbb{P}(E_1^c) \;\; \le \;\; \exp\bigg(-\frac{n\psi_{1,n}^2\kappa_{\min}^2}{2\sigma^2}+\log s\bigg) = \exp\bigg\{-n\bigg(\frac{\psi_{1,n}^2\kappa_{\min}^2}{2\sigma^2}-\frac{\log s}{n}\bigg)\bigg\}. \tag{C.16}$$

Sign consistency for estimated coefficients with indices in S_1^c. Now by plugging (C.9) in the middle term of (C.4) and then taking absolute value on the quantity, for $j \in S_1^c$, we have

$$\begin{aligned}
&\bigg|\bigg(n^{-1}X_{S_1^c}^T X_S C_{SS}^{-1} X_S^T \epsilon - \frac{1}{2n\tau\log(1+\tau^{-1})} X_{S_1^c}^T X_S C_{SS}^{-1} B_{S,\tau} \\
&\quad -X_{S_1^c}^T\epsilon + \frac{\rho_2}{2}\sqrt{q_{S_1^c}}u_{S_1^c}\bigg)_{\vartheta_j^{S_1^c}}\bigg| \\
\le\;\; & n^{-1}\Big|(X_{S_1^c}^T X_S C_{SS}^{-1} X_S^T\epsilon)_{\vartheta_j^{S_1^c}}\Big| + \Big|(X_{S_1^c}^T\epsilon)_{\vartheta_j^{S_1^c}}\Big| \\
& + \frac{1}{2\tau\log(1+\tau^{-1})}\bigg|\frac{\rho_2\sqrt{q_{k_j}}v_j}{(1+\tau^{-1}||w_{G_{k_j}}+\beta_{G_{k_j}}||_2)}\bigg| \\
& + \frac{1}{2n\tau\log(1+\tau^{-1})}\Big|\big(X_{S_1^c}^T X_S C_{SS}^{-1} B_{S,\tau}\big)_{\vartheta_j^{S_1^c}}\Big|
\end{aligned} \tag{C.17}$$

By plugging the right hand side of (C.17) into the left hand side of (C.4) and doing some rearrangements, we obtain the following inequality:

$$\begin{aligned}
& n^{-1}\Big|(X_{S_1^c}^T X_S C_{SS}^{-1} X_S^T\epsilon)_{\vartheta_j^{S_1^c}}\Big| + \Big|(X_{S_1^c}^T\epsilon)_{\vartheta_j^{S_1^c}}\Big| \\
<\;\; & \frac{1}{2\tau\log(1+\tau^{-1})}\bigg(\rho_1 - n^{-1}\Big|\big(X_{S_1^c}^T X_S C_{SS}^{-1} B_{S,\tau}\big)_{\vartheta_j^{S_1^c}}\Big| - \bigg|\frac{\rho_2\sqrt{q_{k_j}}v_j}{(1+\tau^{-1}||w_{G_{k_j}}+\beta_{G_{k_j}}||_2)}\bigg|\bigg)
\end{aligned} \tag{C.18}$$

Note that by Assumption 3, the maximum eigenvalue value of the matrix $X_S X_S^T$ is $n\varsigma_{\max}$. Therefore,

$$|(X_{S_1^c}^T X_S C_{SS}^{-1} X_S^T\epsilon)_{\vartheta_j^{S_1^c}}| \le ||X_{S_1^c}^T X_S C_{SS}^{-1} X_S^T\epsilon||_\infty \le n\varsigma_{\max}\kappa_{\min}^{-1}||X_{S_1^c}^T\epsilon||_\infty. \tag{C.19}$$

Further define

$$(B_{S_1^c,\tau})_{\vartheta_j^{S_1^c}} = \frac{\rho_2\sqrt{q_{k_j}}v_j}{(1+\tau^{-1}||w_{G_{k_j}}+\beta_{G_{k_j}}||_2)}. \tag{C.20}$$

With (C.19) and (C.20), we construct the following event:

$$\begin{aligned}
E_2 = \bigg\{\epsilon \;\; : \;\; & 2\bigg(\frac{\varsigma_{\max}}{\kappa_{\min}}+1\bigg)||X_{S_1^c}^T\epsilon||_\infty \\
< \;\; & \frac{1}{\tau\log(1+\tau^{-1})}\Big(\rho_1 - n^{-1}||X_{S_1^c}^T X_S C_{SS}^{-1} B_{S,\tau}||_\infty - ||B_{S_1^c,\tau}||_\infty\Big)\bigg\}.
\end{aligned} \tag{C.21}$$

Since the left hand side of the inequality stated in E_2 is larger than the left hand side of (C.18), and the right hand side of the inequality stated in E_2 is smaller than the right hand side of (C.18), therefore if the inequality stated in E_2 holds, then (C.18) will hold. In turn both (C.3) and (C.4) will hold, and the sign consistency for estimated coefficients with indices in S_1^c can be established.

Now define $\psi_{2,n}$ by

$$\psi_{2,n} = \frac{1}{\tau\log(1+\tau^{-1})}\Big(\rho_1 - n^{-1}||X_{S_1^c}^T X_S C_{SS}^{-1} B_{S,\tau}||_\infty - ||B_{S_1^c,\tau}||_\infty\Big). \qquad (C.22)$$

Then following the technique similar to the one used in deriving (C.15) and (C.16), We can bound the probability of E_2^c in a way such that

$$\begin{aligned}\mathbb{P}(E_2^c) &\leq \exp\bigg(-\frac{\psi_{2,n}^2\kappa_{\min}^2}{8(\varsigma_{\max}+\kappa_{\min})^2\sigma^2} + \log s_1^c\bigg)\\ &= \exp\bigg\{-n\bigg[\frac{\psi_{2,n}^2\kappa_{\min}^2}{8n(\varsigma_{\max}+\kappa_{\min})^2\sigma^2} - \frac{\log s_1^c}{n}\bigg]\bigg\}. \qquad (C.23)\end{aligned}$$

Sign consistency for estimated coefficients with indices in G_{R^c}. Now by plugging (C.9) into the left hand side of (C.5), we have

$$\begin{aligned}&\bigg|\bigg|2n^{-1}X_{G_k}^T X_S C_{SS}^{-1} X_S^T\epsilon\\ &-\frac{1}{n\tau\log(1+\tau^{-1})}X_{G_k}^T X_S C_{SS}^{-1} B_{S,\tau} - 2X_{G_k}^T\epsilon + \rho_1 g_{G_k}\bigg|\bigg|_2\\ \leq\ & 2\bigg|\bigg|n^{-1}X_{G_k}^T X_S C_{SS}^{-1} X_S^T\epsilon\bigg|\bigg|_2 + 2\big|\big|X_{G_k}^T\epsilon\big|\big|_2\\ &+\frac{\rho_1}{\tau\log(1+\tau^{-1})}\bigg|\bigg|\frac{h_{G_k}}{(1+\tau^{-1}|w_{G_k}+\beta_{G_k}|)}\bigg|\bigg|_2\\ &+\frac{1}{n\tau\log(1+\tau^{-1})}\big|\big|X_{G_k}^T X_S C_{SS}^{-1} B_{S,\tau}\big|\big|_2 \qquad (C.24)\end{aligned}$$

for $k \in R^c$. Further by plugging the right hand side of (C.24) into the left hand side of (C.5) and doing some rearrangements, we can obtain the following inequality:

$$\begin{aligned}&n^{-1}\bigg|\bigg|X_{G_k}^T X_S C_{SS}^{-1} X_S^T\epsilon\bigg|\bigg|_2 + \big|\big|X_{G_k}^T\epsilon\big|\big|_2\\ <\ & \frac{1}{2\tau\log(1+\tau^{-1})}\bigg(\rho_2\sqrt{q_k} - n^{-1}\big|\big|X_{G_k}^T X_S C_{SS}^{-1} B_{S,\tau}\big|\big|_2\\ &-\rho_1\bigg|\bigg|\frac{h_{G_k}}{(1+\tau^{-1}|w_{G_k}+\beta_{G_k}|)}\bigg|\bigg|_2\bigg). \qquad (C.25)\end{aligned}$$

By Assumption 3, the maximum eigenvalue of the matrix $X_{G_k}X_{G_k}^T$ is $n\nu_{k,\max}$. Further note that

$$||X_{G_k}^T X_S C_{SS}^{-1} X_S^T\epsilon||_2 \leq n\varsigma_{\max}\kappa_{\min}^{-1}||X_{G_k}^T\epsilon||_2 \leq n\varsigma_{\max}\kappa_{\min}^{-1}\sqrt{n\nu_{k,\max}}||\epsilon||_2. \qquad (C.26)$$

With (C.26), we construct the following event:

$$
\begin{aligned}
E_3 = \Bigg\{\epsilon \ : \ & 2\bigg(\frac{\varsigma_{\max}}{\kappa_{\min}}+1\bigg)\sqrt{n\nu_{k,\max}}||\epsilon||_2 \\
& < \frac{1}{\tau\log(1+\tau^{-1})}\bigg(\rho_2 - n^{-1}\max_{k\in R^c}\big|\big|X_{G_k}^T X_S C_{SS}^{-1}B_{S,\tau}\big|\big|_2 \\
& - \max_{k\in R^c}\rho_1\bigg|\bigg|\frac{h_{G_k}}{(1+\tau^{-1}|w_{G_k}+\beta_{G_k}|)}\bigg|\bigg|_2\bigg), \text{ for all } k\in R^c\Bigg\}. \qquad \text{(C.27)}
\end{aligned}
$$

Since the left hand side of the inequality stated in E_3 is larger than the left hand side of (C.25), and the right hand side of the inequality stated in E_3 is smaller than the right hand side of (C.25), therefore if the inequality stated in E_3 holds, then (C.25) will also hold. In turn, if (C.25) holds for all $k \in R^c$, then both (C.3) and (C.5) will hold, and the sign consistency for estimated coefficients with indices in G_{R^c} can be established.

We follow a strategy similar to those given above to derive an estimate for the tail probability of E_3. Define $\psi_{3,n}$ by

$$
\begin{aligned}
\psi_{3,n} \ = \ & \frac{1}{\tau\log(1+\tau^{-1})}\bigg(\rho_2 - n^{-1}\max_{k\in R^c}\big|\big|X_{G_k}^T X_S C_{SS}^{-1}B_{S,\tau}\big|\big|_2 \\
& - \max_{k\in R^c}\rho_1\bigg|\bigg|\frac{h_{G_k}}{(1+\tau^{-1}|w_{G_k}+\beta_{G_k}|)}\bigg|\bigg|_2\bigg). \qquad \text{(C.28)}
\end{aligned}
$$

Note that E_3 is equivalent to the event $\cap_{k\in R^c}\{4n\nu_{k,\max}\kappa_{\min}^{-2}(\varsigma_{\max}+\kappa_{\min})^2||\epsilon||_2^2 < \psi_{3,n}^2\}$. Therefore the probability of E_3^c can be bounded in a way such that

$$
\begin{aligned}
\mathbb{P}(E_3^c) \ \le \ & \sum_{k\in R^c}\mathbb{P}\bigg\{\frac{||\epsilon||_2^2}{\sigma^2} \ge \frac{\kappa_{\min}^2\psi_{3,n}^2}{4n\nu_{k,\max}(\varsigma_{\max}+\kappa_{\min})^2\sigma^2}\bigg\} \\
\le \ & r^c\mathbb{P}\bigg\{\frac{||\epsilon||_2^2}{\sigma^2} \ge \frac{\kappa_{\min}^2\psi_{3,n}^2}{4n\nu_{\max}(\varsigma_{\max}+\kappa_{\min})^2\sigma^2}\bigg\}. \qquad \text{(C.29)}
\end{aligned}
$$

In addition, since ϵ's are i.i.d. normal variables with mean zero and variance σ^2, therefore $||\epsilon||_2^2/\sigma^2$ is a Chi-square variable with n degrees of freedom. It can be shown that $\mathbb{E}[\exp(a||\epsilon||_2^2/\sigma^2)] = (1-2a)^{-n/2}$ for $a < 1/2$. We let $a = 1/4$, then $\mathbb{E}[\exp(4^{-1}||\epsilon||_2^2/\sigma^2)] = 2^{n/2}$. Wit the arguments given above, the probability of E_3^c can be further bounded in a way such that

$$
\begin{aligned}
\mathbb{P}(E_3^c) \ \le \ & \exp\bigg(-\frac{\kappa_{\min}^2\psi_{3,n}^2}{16n\nu_{\max}(\varsigma_{\max}+\kappa_{\min})^2\sigma^2} + \frac{n}{2}\log 2 + \log r^c\bigg) \\
\le \ & \exp\bigg\{-n\bigg[\frac{\kappa_{\min}^2\psi_{3,n}^2}{16n^2\nu_{\max}(\varsigma_{\max}+\kappa_{\min})^2\sigma^2} - 0.35 - \frac{\log r^c}{n}\bigg]\bigg\}. \qquad \text{(C.30)}
\end{aligned}
$$

Since E_1, E_2 and E_3 jointly implies conditions (C.3), (C.10), (C.4) and (C.5), which further implies the sign consistency $\text{sign}(\widehat{\beta}^\tau) = \text{sign}(\beta)$, therefore

$$
\mathbb{P}\big\{\text{sign}(\widehat{\beta}^\tau) = \text{sign}(\beta)\big\} \ge \mathbb{P}(E_1\cap E_2\cap E_3) = 1 - \mathbb{P}\{(E_1\cap E_2\cap E_3)^c\}.
$$

Further note that $\mathbb{P}\{(E_1 \cap E_2 \cap E_3)^c\} = \mathbb{P}(E_1^c \cup E_2^c \cup E_3^c) \leq \mathbb{P}(E_1^c) + \mathbb{P}(E_2^c) + \mathbb{P}(E_3^c)$. Therefore we have

$$\mathbb{P}\{\text{sign}(\widehat{\beta}^\tau) = \text{sign}(\beta)\} \geq 1 - \mathbb{P}(E_1^c) - \mathbb{P}(E_2^c) - \mathbb{P}(E_3^c). \tag{C.31}$$

Then by applying the tail probability results (C.16), (C.23) and (C.30) to construct a lower bound for the quantity on the right hand side of (C.31), we recover the inequality (5.10).

Asymptotic behavior of $\psi_{1,n}$, $\psi_{2,n}$ and $\psi_{3,n}$. Now we go on to show that as $n \to \infty$, $\psi_{1,n}$, $\psi_{2,n}$ and $\psi_{3,n}$, defined in (C.14), (C.22) and (C.28), respectively, can satisfy the requirements stated in Theorem 5.4. We first consider the asymptotic behavior of $(B_{S,\tau})_{\vartheta_j^S}$, which is defined in (C.8). Note that by assumptions, if $|w_j + \beta_j| \neq 0$, then $h_j = 1$ or $h_j = -1$. Therefore given that $\tau = n^{-1}$ and $\rho_1 = O(n^{1/2})$, the first term on the right hand side of (C.8) will be $O(n^{-1/2})$. In addition, if $|w_j + \beta_j| = 0$, then h_j is an arbitrary quantity in $[-1, 1]$. In this situation we may let h_j be proportional to n^{-1}, then the first term on the right hand side of (C.8) will be $O(n^{-1/2})$. An argument similar to the one given above can be applied to the second term on the right hand side of (C.8). Further note that given $\lambda = O(n^{1/2})$, the third term on the right hand side of (C.8) will be $O(n^{-1/2}\log(1+n))$. With the arguments given above, we conclude that

$$(B_{S,\tau})_{\vartheta_j^S} = O(n^{-1/2}\log(1+n)) \tag{C.32}$$

for all $j \in S$. An argument similar to the one given above can be applied to $(B_{S_1^c,\tau})_{\vartheta_j^{S_1^c}}$ in (C.20) and the term $\rho_1 ||h_{G_k}(1 + \tau^{-1}|w_{G_k} + \beta_{G_k}|)^{-1}||_2$ in E_3, which leads to

$$(B_{S_1^c,\tau})_{\vartheta_j^{S_1^c}} = O(q_{k_j}^{1/2} n^{-1/2}) \tag{C.33}$$

for all $j \in S_1^c$ and

$$\rho_1 \left|\left| \frac{h_{G_k}}{1 + \tau^{-1}|w_{G_k} + \beta_{G_k}|} \right|\right|_2 = O(q_k^{1/2} n^{-1/2}) \tag{C.34}$$

for all $k \in R^c$.

Next we go on to deal with the l_∞-norm terms involved in $\psi_{1,n}$, $\psi_{2,n}$ and $\psi_{3,n}$. First note that for a p dimensional vector b, we can bound $||b||_\infty$ in a way such that $||b||_\infty = \sqrt{\max_j |b_j|^2} \leq \sqrt{\sum_{j=1}^p b_j^2} = \sqrt{b^T b}$. Therefore for $\psi_{1,n}$ defined in (C.14), we can bound the term $||C_{SS}^{-1} B_{S,\tau}||_\infty$ in a way such that

$$||C_{SS}^{-1} B_{S,\tau}||_\infty \leq \frac{\sqrt{B_{S,\tau}^T B_{S,\tau}}}{\kappa_{\min}} = O\bigg(\frac{s^{1/2}\log(n+1)}{n^{1/2}\kappa_{\min}}\bigg). \tag{C.35}$$

Now consider $\psi_{2,n}$ defined in (C.22). First note that since $S_1^c \subseteq G_{R^c}$, therefore we can bound the term $n^{-1}||X_{S_1^c}^T X_S C_{SS}^{-1} B_{S,\tau}||_\infty$ in a way such that

$$\begin{aligned} n^{-1}||X_{S_1^c}^T X_S C_{SS}^{-1} B_{S,\tau}||_\infty &\leq n^{-1}||X_{G_{R^c}}^T X_S C_{SS}^{-1} B_{S,\tau}||_\infty \\ &\leq \max_{k \in R^c} n^{-1}||X_{G_k}^T X_S C_{SS}^{-1} B_{S,\tau}||_\infty. \end{aligned} \tag{C.36}$$

The right hand side of (C.36) can be further bounded in a way such that

$$\begin{aligned}\max_{k\in R^c} n^{-1}||X_{G_k}^T X_S C_{SS}^{-1} B_{S,\tau}||_\infty &\le \max_{k\in R^c} n^{-1}\frac{\sqrt{n^2\nu_{k,\max}\varsigma_{\max}}}{\kappa_{\min}}\sqrt{B_{S,\tau}^T B_{S,\tau}} \\ &= O\bigg(\frac{s^{1/2}\sqrt{\nu_{\max}\varsigma_{\max}}\log(1+n)}{n^{1/2}\kappa_{\min}}\bigg). \qquad \text{(C.37)}\end{aligned}$$

A similar argument can be applied to the term $\max_{k\in R^c} n^{-1}||X_{G_k}^T X_S C_{SS}^{-1} B_{S,\tau}||_2$ in $\psi_{3,n}$ defined in (C.28), which leads to

$$\max_{k\in R^c} n^{-1}||X_{G_k}^T X_S C_{SS}^{-1} B_{S,\tau}||_2 = O\bigg(\frac{s^{1/2}\sqrt{\nu_{\max}\varsigma_{\max}}\log(1+n)}{n^{1/2}\kappa_{\min}}\bigg). \qquad \text{(C.38)}$$

Note that we have assumed $p = o(n(\log(n+1))^{-2})$ and since $s \le p$ and $q_k \le p$ for $k = 1, 2, \cdots, m$, therefore we have $s^{1/2} = o(n^{1/2}(\log(n+1))^{-1})$ and $q_k^{1/2} = o(n^{1/2}(\log(1+n))^{-1})$ for $k = 1, 2, \cdots, m$. Then with (C.35), the second term on the right hand side of (C.14) will approach to zero as $n \to \infty$, therefore we have $\psi_{1,n} = O(1)$ as $n \to \infty$. In addition, with results in (C.33), (C.36) and (C.37), the second and third terms on the right hand side of (C.22) will approach to zero as $n \to \infty$, therefore we have $\psi_{2,n} = O(n^{3/2}(\log n)^{-1})$ as $n \to \infty$. Moreover, with results in (C.34) and (C.38), the second and third terms on the right hand side of (C.28) will approach to zero as $n \to \infty$, therefore we have $\psi_{3,n} = O(n^{3/2}(\log n)^{-1})$ as $n \to \infty$, which completes the proof.

Table 1: Out-of-sample mean squared error. Each value is an average over the 415 time blocks. The value in the bracket is the standard error.

Method	Model 1	Model 2
gvsnss	16.99 (1.83)	16.67 (1.81)
lasso	21.87 (1.74)	22.48 (1.81)
gvsnss-PC	17.50 (1.90)	17.03 (1.86)
lasso-PC	17.66 (1.83)	18.39 (1.88)
PC	16.75 (1.88)	17.61 (1.92)
AR	-	18.68 (2.03)

Table 2: Estimation results based on 100 sub-sampling simulations. Each value is an average over 100 sub-sampling simulations and the value in the bracket is the standard error. PMSE: Predictive mean squared error; $\widehat{s}$: The number of covariates with non-zero estimated coefficients; $\widehat{s}^{+}_{\mathrm{SSE}}$: The number of covariates with positive estimated coefficients in the Susan Shepard Effect group.

	gvsnss 5CV	gvsnss BF	lasso 10CV
$\mathrm{PMSE}_{\mathrm{test}}$	0.43 (0.01)	0.38 (0.01)	0.41 (0.01)
$\widehat{s}$	2.81 (0.25)	1.03 (0.02)	3.89 (0.22)
$\widehat{s}^{+}_{\mathrm{SSE}}$	0.35 (0.13)	0.00 (0.00)	1.33 (0.09)

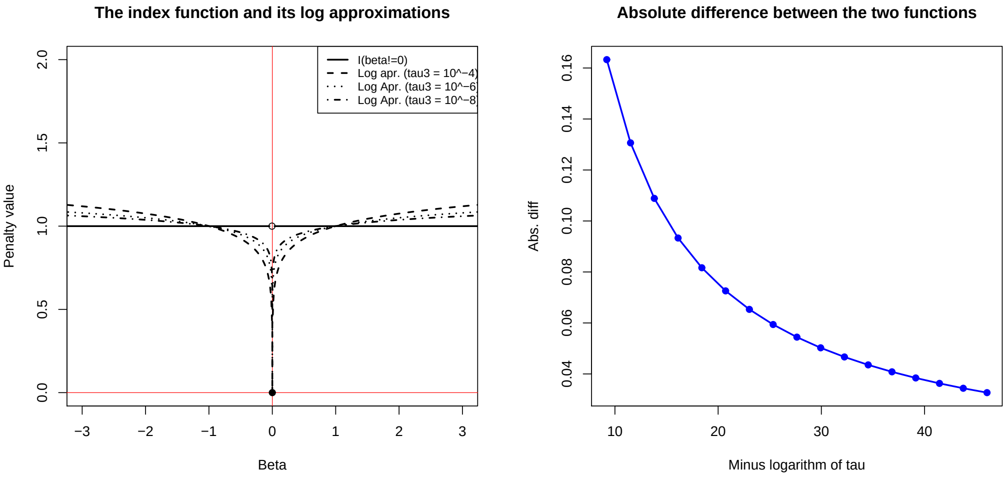

Figure 1: Left: The index function and its log approximations; Right: The mean absolute difference between the index function and its log approximation as a function of $-\log\tau$. Each point is an average over absolute differences with input values from $[-10, 10]$.

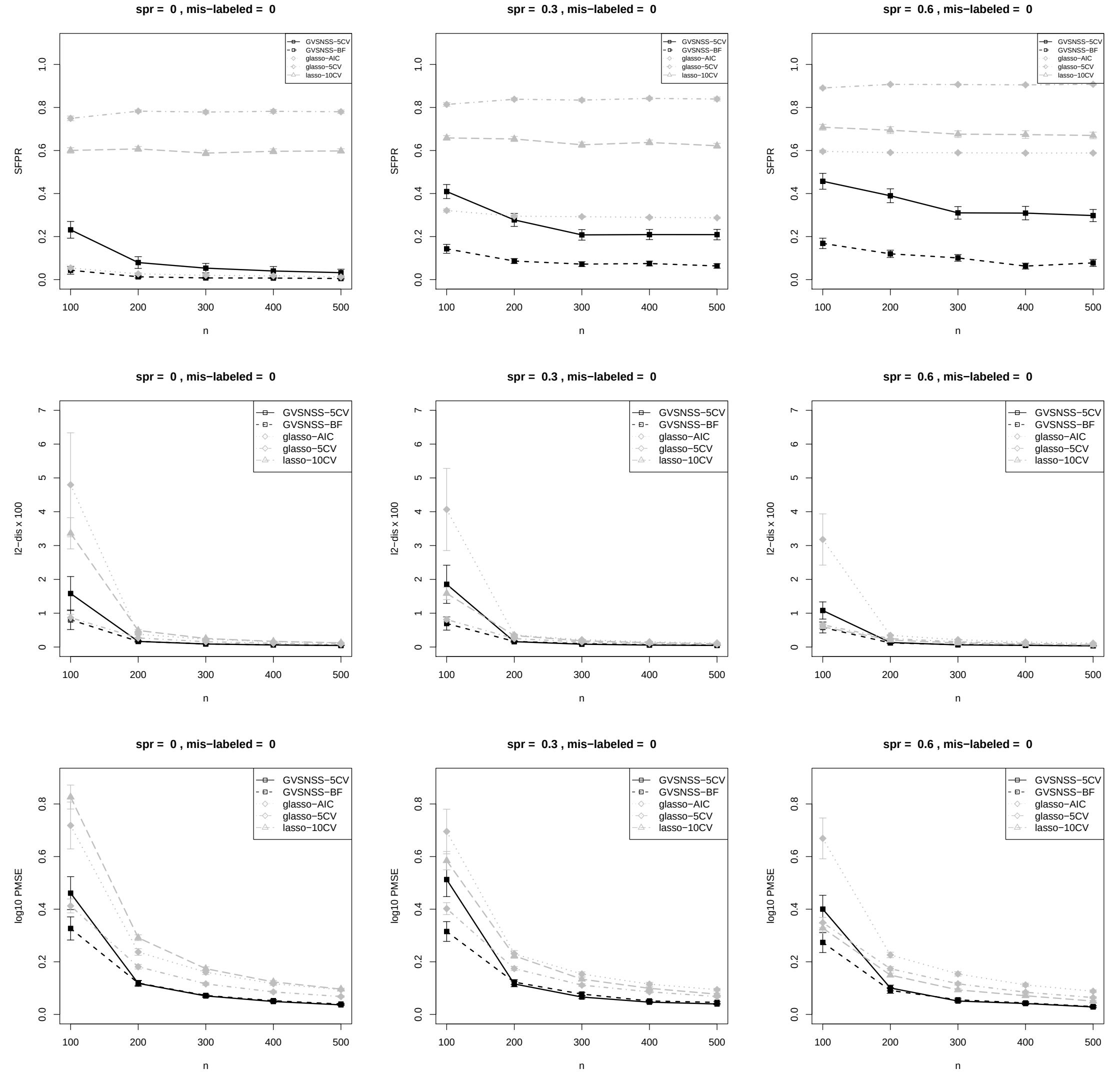

Figure 2: Estimation results from simulated data. Each point is an average over 100 replicates. For all data sets, we set mis-labeled = 0, p = 200, m = 10 and r = 2. Left: spr = 0; Center: spr = 0.3; Right: spr = 0.6. Top: SFPR; Middle: l_2 distance between the estimates and the true values; Bottom: Logarithm of the PMSE with respect to base 10.

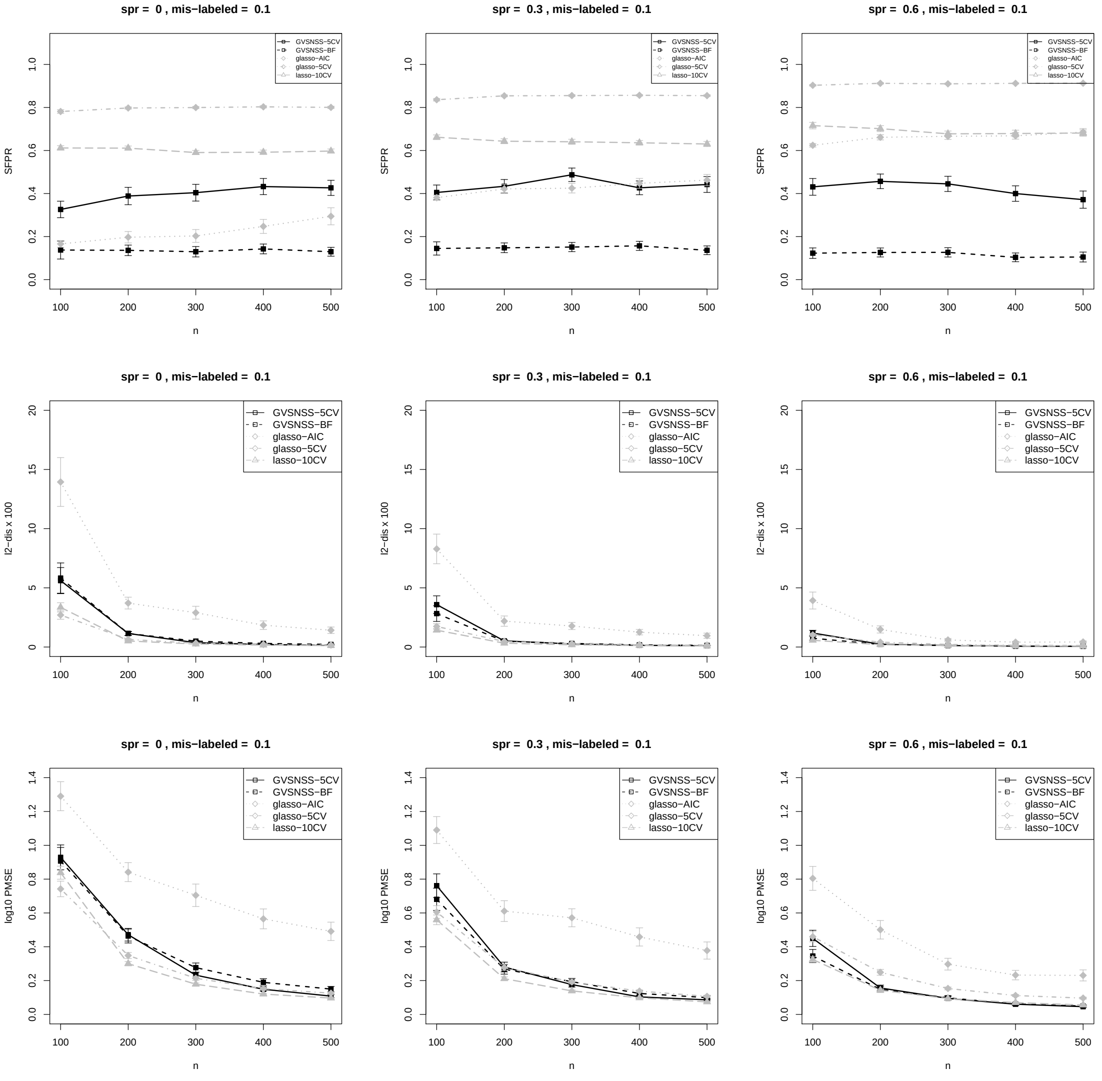

Figure 3: Estimation results from simulated data. Each point is an average over 100 replicates. For all data sets, we set mis-labeled = 0.1, $p = 200$, $m = 10$ and $r = 2$. Left: spr = 0; Center: spr = 0.3; Right: spr = 0.6. Top: SFPR; Middle: l_2 distance between the estimates and the true values; Bottom: Logarithm of the PMSE with respect to base 10.

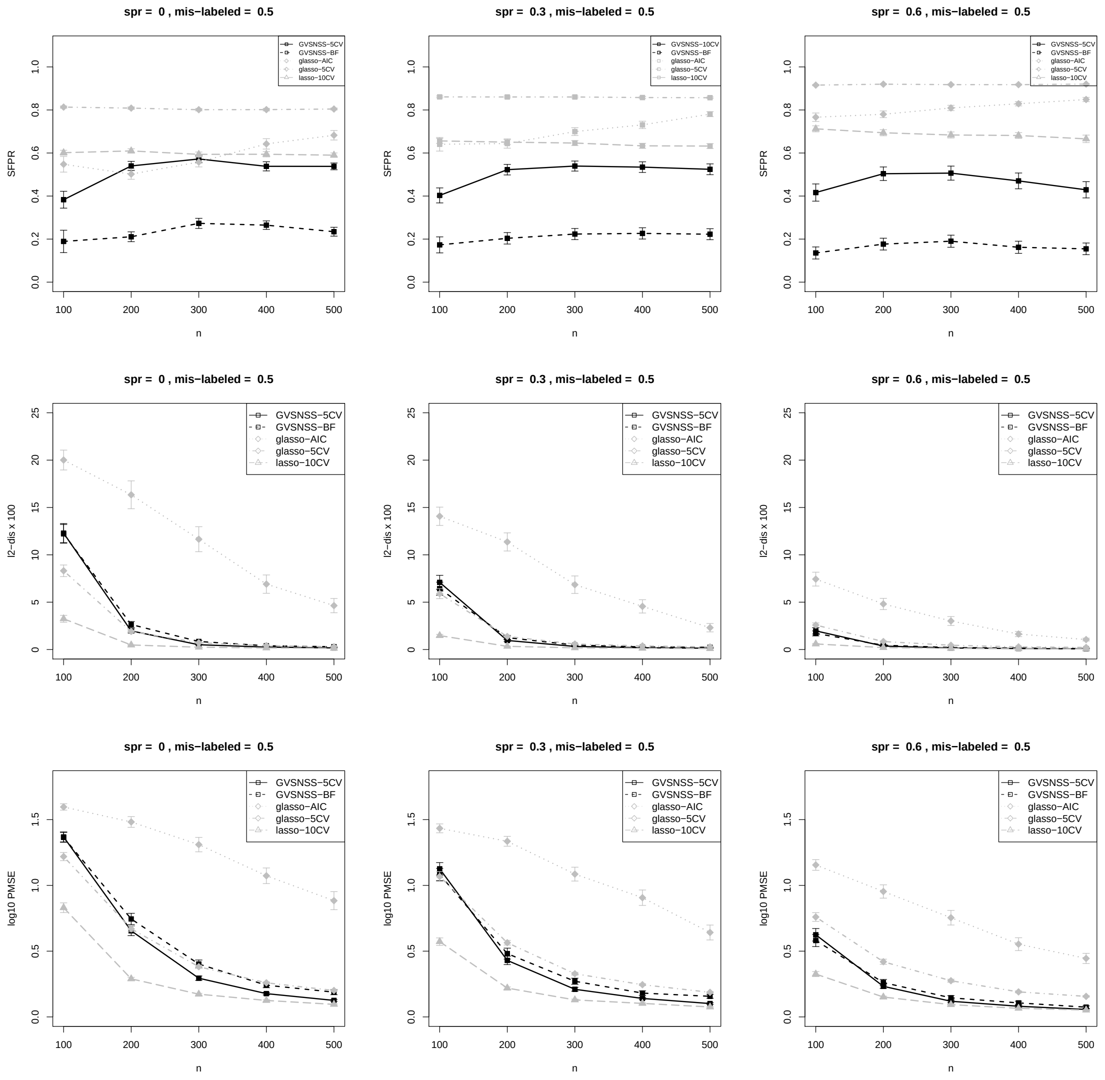

Figure 4: Estimation results from simulated data. Each point is an average over 100 replicates. For all data sets, we set mis-labeled = 0.5, $p = 200$, $m = 10$ and $r = 2$. Left: spr = 0; Center: spr = 0.3; Right: spr = 0.6. Top: SFPR; Middle: l_2 distance between the estimates and the true values; Bottom: Logarithm of the PMSE with respect to base 10.

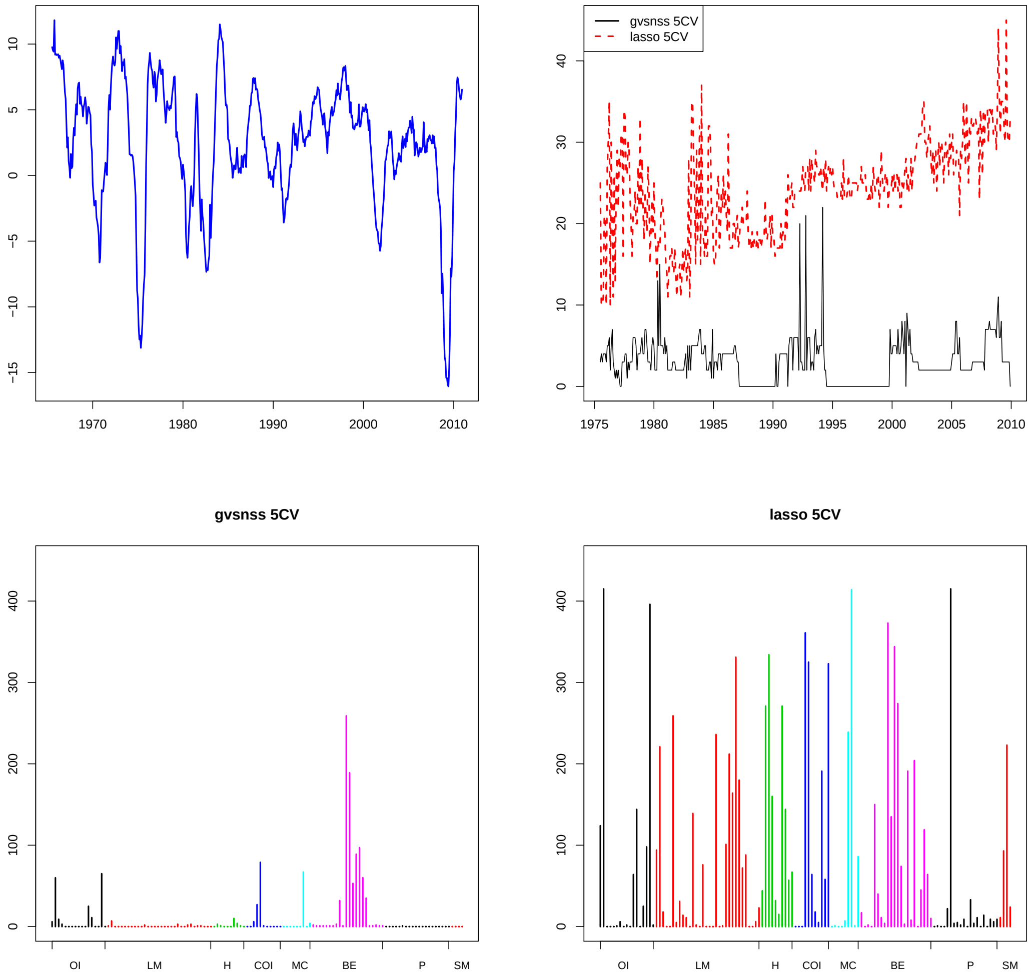

Figure 5: Top Left: Percentage change of the U.S. industrial production index. The change is defined as $100[\log(IP_t) - \log(IP_{t-12})]$. Top Left: The number of selected variables for the 415 time blocks. Bottom Left: Frequencies of variables being selected under the gvsnss. Bottom Right: Frequencies of variables being selected under the lasso. OI: output and income; LM: labor market; H: housing; COI: consumption, orders and inventories; MC: money and credits; BE: bond and exchange rates; P: prices; SM: stock market.

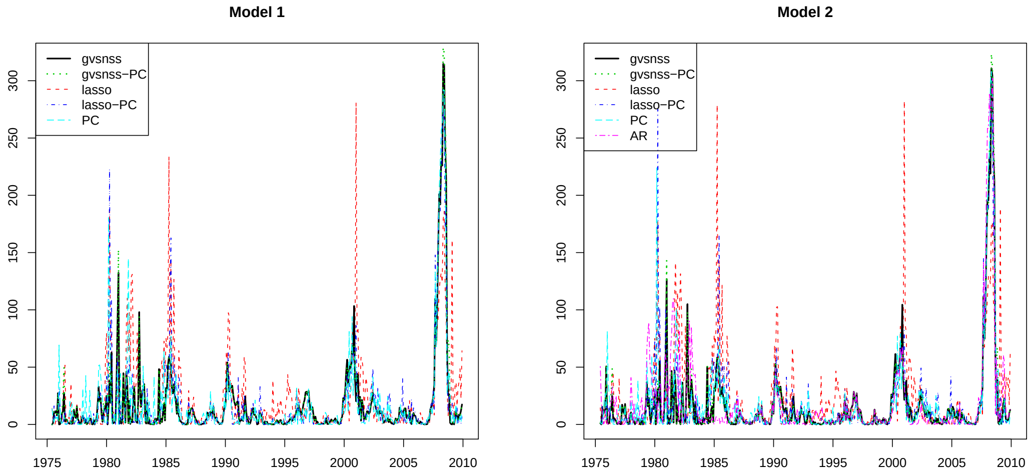

Figure 6: Left: The out-of-sample squared error of Model 1 for the 415 time blocks. Right: The out-of-sample squared error of Model 2 for the 415 time blocks.

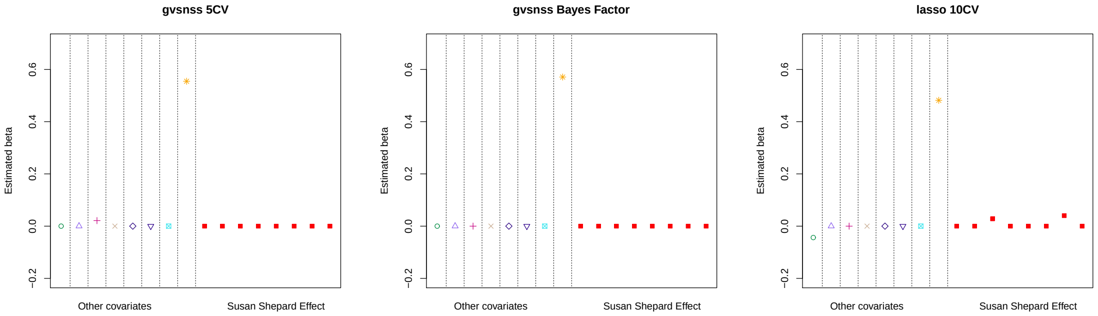

Figure 7: Estimation results from the retirement plan data. Left: The gvsnss estimation with five fold cross validation. Middle: The gvsnss estimation with the Bayes factor. Right: The lasso estimation with ten fold cross validation.

References

[1] A. Armagan, D. Dunson, and J. Lee. Generalized double pareto shrinkage. *http://arxiv.org/abs/arXiv:1104.0861*, 2011.

[2] F. Bach. Consistency of the group lasso and multiple kernel learning. *Journal of Machine Learning Research*, 9:1179–1225, 2008.

[3] J. Bai and S. Ng. Forecasting economic time series using targeted predictors. *Journal of Econometrics*, 146:304–317, 2008.

[4] D. Bertsekas. *Nonlinear Programming.* Athena Scientific, Belmont, 1999.

[5] P. Bickel, Y. Ritov, and A. B. Tsybakov. Simultaneous analysis of lasso and Dantzig selector. *The Annals of Statistics*, 37:1705–1732, 2009.

[6] P. G. Bryant and M. A. Smith. *Practical Data Analysis: Case Studies in Business Statistics.* Irwin, Chicago, 1995.

[7] E. J. Candés, M. B. Wakin, and S. P. Boyd. Enhancing sparsity by reweighted l_1 minimization. *Journal of Fourier Analysis and Applications*, 14:877–905, 2008.

[8] J. Chiquet, Y. Grandvalet, and C. Charbonnier. Sparsity with sign-coherent groups of variables via the cooperative-Lasso. *http://arxiv.org/abs/1103.2697v1*, 2010.

[9] R. Foygel and M. Drton. Exact block-wise optimization in group lasso and sparse group lasso for linear regression. *http://arxiv.org/abs/1010.3320v2*, 2010.

[10] J. Friedman, T. Hastie, and R. Tibshirani. A note on the group lasso and a sparse group lasso. *http://arxiv.org/abs/1001.0736v1*, 2010.

[11] J. Friedman, T. Hastie, and R. Tibshirani. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33:1–22, 2010.

[12] A. Genkin, D. D. Lewis, and D. Madigan. Large scale Bayesian logistic regression for text categorization. *Technometrics*, 49:291–304, 2007.

[13] J. Huang and T. Zhang. The benefit of group sparsity. *The Annals of Statistics*, 38:1978–2004, 2010.

[14] J. J. Huang, J. Horowitz, and F. Wei. Variable selection in nonparametric additive models. *The Annals of Statistics*, 38:2282–2313, 2010.

[15] D. R. Hunter and K. Lange. A tutorial on MM algorithms. *The American Statistician*, 58:30–37, 2004.

[16] K. Lounici, M. Pontil, S. van de Geer, and A. B. Tsybakov. Oracle inequalities and optimal inference under group sparsity. *To be appeared in The Annals of Statistics*, 2011.

[17] S. C. Ludvigson and S. Ng. Macro factors in bond risk premia. *Review of Financial Studies*, 22:5027–5067, 2009.

[18] R. Mazumder, J. Friedman, and T. Hastie. SparseNet: coordinate descent with non-convex penalties. *To be appeared in Journal of the American Statistical Association*, 2011.

[19] L. Meier, S. van de Geer, and P. Bühlmann. The group lasso for logistic regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70:53–71, 2008.

[20] N. Meinshausen and B. Yu. Lasso-type recovery of sparse representations for high-dimensional data. *The Annals of Statistics*, 37:246–270, 2009.

[21] T. J. Mitchell and J. J. Beauchamp. Bayesian variable selection in linear regression. *Journal of the American Statistical Association*, 83:1023–1032, 1988.

[22] G. Obozinski, M. J. Wainright, and M. I. Jordan. Support union recovery in high-dimensional multivariate regression. *The Annals of Statistics*, 39:1–47, 2011.

[23] P. Ravikumar, J. Lafferty, H. Liu, and L. Wasserman. Sparse additive models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71:1009–1030, 2009.

[24] D. Ruppert, M. Wand, and R. J. Carroll. *Semiparametric Regression*. Cambridge University Press, Cambridge, 2003.

[25] F. Scheipl, L. Fahrmeir, and T. Kneib. Spike-and-slab priors for function selection in structured additive regression models. *http://arxiv.org/abs/1105.5250v1*, 2011.

[26] B. K. Sriperumbudur, D. A. Torres, and G. R. G. Lanckriet. A D.C. programming approach to the sparse generalized eigenvalue problem. *http://arxiv.org/abs/0901.1504*, 2009.

[27] J. H. Stock and M. W. Watson. Forecasting using principal components from a large number of predictors. *Journal of the American Statistical Association*, 97:1167–1179, 2002.

[28] M. E. Tipping. Sparse Bayesian learning and the relevance vector machine. *Journal of Machine Learning Research*, 1:211–244, 2001.

[29] T. T. Wu and K. Lange. Coordinate descent algorithms for lasso penalized regression. *The Annals of Applied Statistics*, 2:224–244, 2008.

[30] T. T. Wu and K. Lange. The MM alternative to EM. *Statistical Science*, 25:492–505, 2010.

[31] T. J. Yen. A majorization-minimization approach to variable selection using spike and slab priors. *Accepted by The Annals of Statistics*, 2011.

[32] M. Yuan and Y. Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68:49–67, 2006.

[33] P. Zhao and B. Yu. On model selection consistency of lasso. *Journal of Machine Learning Research*, 7:2541–2564, 2006.

[34] H. Zou and H. H. Zhang. On the adaptive elastic-net with a diverging number of parameters. *The Annals of Statistics*, 37:1733–1751, 2009.