Journal of Machine Learning Research 22 (2021) 1-39 Submitted 11/19; Published 8/21

# Optimal Minimax Variable Selection for Large-Scale Matrix Linear Regression Model

**Meiling Hao*** MEILINGHAO@UIBE.EDU.CN
*School of Statistics*
*University of International Business and Economics*
*Beijing, 100029, China*

**Lianqiang Qu*** QULIANQ@MAIL.CCNU.EDU.CN
*School of Mathematics and Statistics*
*Central China Normal University*
*Wuhan, Hubei, 430079, China*

**Dehan Kong** KONGDEHAN@UTSTAT.TORONTO.EDU
*Department of Statistical Sciences*
*University of Toronto*
*Toronto, Ontario, M5G 1X6, Canada*

**Liuquan Sun†** SLQ@AMT.AC.CN
*Institute of Applied Mathematics, Academy of Mathematics and Systems Science*
*Chinese Academy of Sciences*
*Beijing, 100190, China*

**Hongtu Zhu** HZHU@BIOS.UNC.EDU
*Department of Biostatistics*
*University of North Carolina at Chapel Hill*
*Chapel Hill, North Carolina, 27599, USA*

**Editor:** Genevera Allen

## Abstract

Large-scale matrix linear regression models with high-dimensional responses and high-dimensional variables have been widely employed in various large-scale biomedical studies. In this article, we propose an optimal minimax variable selection approach for the matrix linear regression model when the dimensions of both the response matrix and predictors diverge at the exponential rate of the sample size. We develop an iterative hard-thresholding algorithm for fast computation and establish an optimal minimax theory for the parameter estimates. The finite sample performance of the method is examined via extensive simulation studies and a real data application from the Alzheimer's Disease Neuroimaging Initiative study is provided.

**Keywords:** High dimension, Imaging genetics, Matrix linear regression, Optimal minimax rate, Variable selection

∗. Co-first authors
†. Corresponding author

## 1. Introduction

With the rapid growth of modern technology, many large-scale biomedical studies have been conducted to collect massive datasets with large volumes of multi-modality imaging, genetic, neurocognitive, and clinical information from increasingly large cohorts (Nathoo et al., 2019). To motivate the proposed methodology, we consider a large database with imaging, genetic, and clinical data from 735 subjects collected by the Alzheimer's Disease Neuroimaging Initiative study. Specifically, each subject has a hippocampal surface dataset consisting of the left and right hippocampi, each of which is represented as a $100 \times 150$ matrix, and also has a large genetic dataset with genotyped and imputed genetic data. Our primary problem of interest is to identify novel genetic markers on the local changes of hippocampus structure. However, it is very challenging due to the heterogeneous effects of genetic variants and the high-dimensionality of putative predictors.

Motivated by various data applications, research on matrix-valued data has gained considerable interest in recent years. For example, Li et al. (2010), Leng and Tang (2012), Zhao and Leng (2014), Zhou and Li (2014), Ding and Cook (2014), Fosdick and Hoff (2015) and Hu et al. (2020b) considered the matrix covariates regression. There are also a few works considering matrix response regression (Viroli, 2012; Ding, 2014; Ding and Cook, 2018; Hu et al., 2020a), where all these methods focused on the case that the dimension of the variables is less than the sample size. However, in our real data application, the dimension of the predictors can be much larger than the sample size, and these existing methods suffer from computational expediency, statistical inaccuracy, and algorithm instability (Fan et al., 2009). To address these issues, Kong et al. (2020) developed a low-rank linear regression model to correlate a high-dimensional response matrix with a high dimensional vector of predictors when coefficient matrices have low-rank structures. Their procedure contains two-steps: a first-step sure independence screening procedure based on the spectral norm of each coefficient matrix and a second-step estimation procedure based on the trace norm regularization.

The prominent marginal screening method for variable selection was first proposed by Fan and Lv (2008) for ultrahigh-dimensional linear regression models. The marginal screening method works via ranking the importance of variables according to their marginal correlation with the response. Because of its good numerical performance and novel theoretical properties, the sure independence screening idea has been extensively studied in the last decades. Examples include Fan and Song (2010), Fan et al. (2011), Zhu et al. (2012), He et al. (2013), Liu et al. (2014), Zhao and Li (2012), and so forth. Although the marginal screening method and its variants can reduce the large-scale model size to a moderate one, the performance of the screening methods built upon marginal correlation may be largely discounted because of the complex structure and unprecedented large-scale of the predictors. Since the method in Kong et al. (2020) is based on the sure independence screening, it may overlook some important variables that are marginal uncorrelated with the response variables but are significant. Besides, their method needs a second-step refined estimation to achieve consistent parameter estimates.

To overcome all issues discussed above, we consider a novel variable screening method to handle the matrix response linear model with ultra-high dimensional predictors. Such a model can cover linear regression models with high-dimensional univariate and vector

responses in the literature (Buhlmann and van de Geer, 2011; Yuan et al., 2007). One of the main challenges extending from a univariate or vector response linear model to a matrix response model is to allocate the computer memory for the massive dataset to accommodate all coefficient matrices. The other challenge is to estimate the nonzero coefficient matrices and select the important predictors simultaneously. The proposed method selects the relevant variables via a *Sparsity-Restricted Least Squares* (SRLS) estimator. More precisely, we treat each coefficient matrix as a group, and estimate the coefficient matrices by the least squares method with a group sparse constraint. Then we select the important predictors based on the derived estimates of the coefficient matrices. The proposed method shares the same spirit with Fan et al. (2009), Wang (2009) and Xu and Chen (2014), for considering the joint effects of the variables rather than the marginal correlation. However, our method distinguishes from these methods in the following three perspectives: (i) the new method considers both the structure between different groups inherent in the response matrix and also the feature level sparsity; (ii) the proposed method not only screens out unimportant variables, but also yields a consistent estimate of the coefficient matrix simultaneously; and (iii) our estimate can achieve the optimal minimax rate. Therefore, the proposed method is not an incremental extension of the existing methods.

For implementation, an iterative hard-thresholding (IHT) algorithm is developed for matrix response linear regression models. We prove the convergence of the IHT algorithm, and then show the sure screening property (Fan and Lv, 2008) of our screening procedure. This guarantees that the true model is contained in a set of candidate models selected by our SRLS procedure with overwhelming probability. To further choose the "best" candidate from the candidate models generated by our screening procedure, we employ an extended Bayesian information criterion (EBIC) (Chen and Chen, 2012; Wang et al., 2009), and prove that it enjoys the model selection consistency property.

The rest of the paper is organized as follows. Section 2 introduces the proposed approach and the extended IHT algorithm. Section 3 presents the asymptotic theoretical results of the proposed method. Extensive simulation studies with supportive evidence are reported in Section 4. In Section 5, we perform an imaging genetics analysis using the ADNI dataset. All technical proofs and the detailed information about the data are deferred to the Appendices.

## 2. Methods

In this section, we first describe the matrix linear regression model, and then introduce the methods and an iterative algorithm.

### 2.1 Sparsity-restricted least squares

Let $(Y_i, X_i)$ $(i = 1, \ldots, n)$ be $n$ independent and identically distributed observations, where $Y_i = (Y_{i,sk}) \in \mathbb{R}^{p\times q}$ is a matrix response variable and $X_i = (x_{i1}, \ldots, x_{id_n})^\top \in \mathbb{R}^{d_n}$ is a $d_n$-dimensional predictor variable. We consider the following matrix linear regression model (Kong et al., 2020):

$$Y_i = B_0 + \sum_{j=1}^{d_n} x_{ij} B_j^\star + U_i, \tag{1}$$

where $B_j^\star = (b_{j,sk}^\star) \in \mathbb{R}^{p\times q}$ $(j = 0, 1, \ldots, d_n)$ are unknown true coefficient matrices, and $U_i = (e_{i,sk}) \in \mathbb{R}^{p\times q}$ is a matrix error term satisfying $E(e_{i,sk} \mid X_i) = 0$ and the variance $\mathrm{var}(e_{i,sk}) = \sigma_{sk}^2 < \infty$. Without loss of generality, we assume that the intercept term $B_0$ is zero. If $p = q = 1$, model (1) reduces to the classical linear regression model.

Let $\mathbb{B} = (B_1^\top, \ldots, B_{d_n}^\top)^\top \in \mathbb{R}^{(pd_n)\times q}$, and $\mathbb{B}^\star = (B_1^{\star\top}, \ldots, B_{d_n}^{\star\top})^\top \in \mathbb{R}^{(pd_n)\times q}$. When $d_n = d$ is fixed, the *ordinary least squares* (OLS) estimate of $\mathbb{B}$ is

$$\hat{\mathbb{B}} = \arg\min_{\mathbb{B}} \left\{ \frac{1}{2n} \sum_{i=1}^{n} \|Y_i - \sum_{j=1}^{d} x_{ij} B_j\|_F^2 \right\},$$

where $\|\cdot\|_F$ denotes the Frobenius norm of a matrix. However, when the number of variables $d_n$ is larger than the sample size $n$, the OLS method suffers from multicollinearity. In fact, under this scenario, most of $B_j^\star$ are often assumed to be zero matrices in literatures, see Kong et al. (2020). Here a zero matrix denotes a matrix with every entry zero. We consider the following optimization problem

$$\min_{\mathbb{B}} \left\{ \frac{1}{2n} \sum_{i=1}^{n} \|Y_i - \sum_{j=1}^{d_n} x_{ij} B_j\|_F^2 \right\} \text{ subject to } \sum_{j=1}^{d_n} I(\|B_j\|_F \neq 0) \leq \tau, \tag{2}$$

where $I(\cdot)$ is an indicator function.

We propose an iterative algorithm to obtain an approximate solution to problem (2) in Section 2.2. Since $\tau$ controls the sparse level, if $\tau < n$, there are at least $(d_n - \tau)$ coefficient matrices forced to be zero. To determine the sparse level $\tau$, we derive a solution path for problem (2) motivated by Wang (2009). Specifically, let $\tilde{\tau}$ be a pre-specified integer, we compute problem (2) for each $\tau \in \{1, \ldots, \tilde{\tau}\}$, and denote its corresponding minimizer as $\hat{\mathbb{B}}_\tau = (\hat{B}_{1\tau}^\top, \ldots, \hat{B}_{d_n\tau}^\top)^\top$ and the corresponding selected model as $\hat{\mathcal{M}}_\tau = \{1 \leq j \leq d_n : \|\hat{B}_{j\tau}\|_F \neq 0\}$. Therefore, we get a total of $\tilde{\tau}$ candidate models: $\{\hat{\mathcal{M}}_1, \ldots, \hat{\mathcal{M}}_{\tilde{\tau}}\}$. One practical problem of interest is how to choose $\tilde{\tau}$. First, the true model should be contained in one of the candidate models. Denote the true active set as $\mathcal{M}^\star$, that is, $\mathcal{M}^\star = \{j : \|B_j^\star\|_F \neq 0\}$. Let $\tau^\star = \mathrm{card}(\mathcal{M}^\star)$, where $\mathrm{card}(A)$ denotes the cardinality of a set $A$. As guaranteed by Theorem 2, when $\tilde{\tau} \geq \tau^\star$, $\hat{\mathcal{M}}_{\tilde{\tau}}$ can include the true model with overwhelming probability. In particular, when $\tilde{\tau} = \tau^\star$, $\hat{\mathcal{M}}_{\tilde{\tau}} = \mathcal{M}^\star$ holds with probability approaching one under certain regularity conditions. Thus, if choosing $\tilde{\tau} \geq \tau^\star$, one can always guarantee that $\mathcal{M}^\star$ is contained in one of the candidate models $\{\hat{\mathcal{M}}_1, \ldots, \hat{\mathcal{M}}_{\tilde{\tau}}\}$. On the other side, setting a larger $\tilde{\tau}$ brings heavy computational cost because it needs to search through a larger class of candidate models. In other words, there is a tradeoff between computation and model selection accuracy when choosing $\tilde{\tau}$. In practice, we set $\tilde{\tau} = [n^{1/5}\log(n)]$, where $[a]$ denotes the largest integer part of $a$. This empirical choice is analogous to the recommended $\tilde{\tau}$ values in Fan and Lv (2008) and Xu and Chen (2014), and it works well in both simulation studies and the real data application.

Since the oracle sparse level $\tau^\star$ is unknown in practice, we propose an extended Bayesian information criterion (Chen and Chen, 2012) to find the "best" fitting model among the $\tilde{\tau}$ candidate models. Denote $\mathbb{Y} = (Y_1^\top, \ldots, Y_n^\top)^\top \in \mathbb{R}^{(pn)\times q}$, $\tilde{X} = (X_1, \ldots, X_n)^\top \in \mathbb{R}^{n\times d_n}$ and $\mathbb{X} = \tilde{X} \otimes I_{p\times p}$, where $\otimes$ denotes the Kronecker product. The EBIC is defined as

$$\mathrm{EBIC}(\hat{\mathcal{M}}_\tau) = \log\left\{ \frac{1}{n} \|\mathbb{Y} - \mathbb{X}\hat{\mathbb{B}}_\tau\|_F^2 \right\} + \tau \frac{c_n \log(n)}{n}.$$

We determine the sparse level $\tau$ by $\hat{\tau} = \arg\min_{1\le\tau\le\tilde{\tau}} \text{EBIC}(\hat{\mathcal{M}}_\tau)$, and denote the finally selected model by $\hat{\mathcal{M}}_{\hat{\tau}}$. We will show in Theorem 3 that the EBIC can consistently select the true model.

The proposed procedure employs the joint effects of candidate variables, and is distinguished from the method of Kong et al. (2020), which screens candidate variables based on marginal effects. Wang (2009) proposed a forward regression procedure, which shares the same fashion. However, one main advantage of our method over theirs is that our $\tilde{\tau}$ candidate models can always be guaranteed to include the true model with probability approaching one. If this is not guaranteed, likely, the EBIC can only choose the best model from all wrong candidate models.

### 2.2 Algorithm

In this section, we introduce an iterative algorithm to obtain an approximate solution to problem (2). Note that model (1) can be rewritten as

$$\mathbb{Y} = \mathbb{X}\mathbb{B}^\star + \mathbb{U},$$

where $\mathbb{U} = (U_1^\top, \ldots, U_n^\top)^\top \in \mathbb{R}^{(pn)\times q}$. Then problem (2) is given as follows:

$$\min_{\mathbb{B}} \; g(\mathbb{B}) \;\; \text{subject to} \;\; \sum_{j=1}^{d_n} I(\|B_j\|_F \neq 0) \le \tau, \tag{3}$$

where $g(\mathbb{B}) = \|\mathbb{Y} - \mathbb{X}\mathbb{B}\|_F^2/(2n)$. For problem (3), we borrow ideas from projected gradient descent methods in first-order convex optimization problems (Nesterov, 2004). Specifically, we consider the following quadratic approximation to $g(\mathbb{B})$ at a generic $\mathbb{D}$ :

$$\mathcal{Q}_\lambda(\mathbb{B}|\mathbb{D}) = \frac{1}{2n}\|\mathbb{Y} - \mathbb{X}\mathbb{D}\|_F^2 - \frac{1}{n}\big\langle \mathbb{X}^\top(\mathbb{Y} - \mathbb{X}\mathbb{D}), \mathbb{B} - \mathbb{D}\big\rangle_F + \frac{\lambda}{2}\|\mathbb{B} - \mathbb{D}\|_F^2,$$

where $\lambda > 0$ is a step size and $\langle\cdot,\cdot\rangle_F$ denotes the Frobenius inner product. It can be shown that $g(\mathbb{D}) = \mathcal{Q}_\lambda(\mathbb{D}|\mathbb{D})$ and that $\mathcal{Q}_\lambda(\mathbb{B}|\mathbb{D})$ well approximates $g(\mathbb{D})$ for $\mathbb{B}$ close to $\mathbb{D}$. Based on $\mathcal{Q}_\lambda(\mathbb{B}|\mathbb{D})$, an iterative algorithm (Bertsims et al., 2016; Xu and Chen, 2014) for problem (3) is given by

$$\mathbb{B}^{[l+1]} = \arg\min_{\mathbb{B}} \mathcal{Q}_{\lambda^{[l]}}(\mathbb{B}|\mathbb{B}^{[l]}) \quad \text{subject to} \quad \sum_{j=1}^{d_n} I(\|B_j\|_F \neq 0) \le \tau, \tag{4}$$

where $\mathbb{B}^{[l]}$ is the minimizer obtained at the $l$th iteration. Omitting constant terms, problem (4) is equivalent to

$$\mathbb{B}^{[l+1]} = \arg\min_{\mathbb{B}} \; \big\|\mathbb{B} - \big\{\mathbb{B}^{[l]} + (n\lambda^{[l]})^{-1}\mathbb{X}^\top(\mathbb{Y} - \mathbb{X}\mathbb{B}^{[l]})\big\}\big\|_F^2,$$
$$\text{subject to} \; \sum_{j=1}^{d_n} I(\|B_j\|_F \neq 0) \le \tau.$$

This optimization problem can be solved using the result in the following proposition.

**Proposition 1** Let $\mathbb{D} = (D_1^\top, \ldots, D_{d_n}^\top)^\top \in \mathbb{R}^{(pd_n)\times q}$ be an arbitrary matrix. If $\hat{\mathbb{B}} = (\hat{B}_1^\top, \ldots, \hat{B}_{d_n}^\top)^\top$ is an optimal solution to the following problem:

$$\min_{\mathbb{B}} \|\mathbb{B} - \mathbb{D}\|_F^2 \quad \text{subject to} \quad \sum_{j=1}^{d_n} I(\|B_j\|_F \neq 0) \leq \tau,$$

then $\hat{\mathbb{B}}$ has a closed form with the $j$th block defined as

$$\hat{B}_j = H_\tau(D_j) \equiv \begin{cases} D_j, & \text{if } \ d_j^* \geq d_{(\tau)}^*, \\ 0, & \text{otherwise,} \end{cases} \tag{5}$$

where $d_j^* = \|D_j\|_F$, $d_{(\tau)}^*$ is the $\tau$th largest component of $d_1^*, \ldots, d_{d_n}^*$.

Proposition 1 extends the hard-thresholding operator in Bertsims et al. (2016) to the matrix linear regression, which presents a closed form solution of $\mathbb{B}^{[l+1]}$. The updating rule for $\mathbb{B}^{[l]}$ is

$$\mathbb{B}^{[l+1]} = \mathbb{H}_\tau\big\{\mathbb{B}^{[l]} + (n\lambda^{[l]})^{-1}\mathbb{X}^\top(\mathbb{Y} - \mathbb{X}\mathbb{B}^{[l]})\big\}, \tag{6}$$

where $\mathbb{H}_\tau(\mathbb{D}) = (H_\tau(D_1)^\top, \ldots, H_\tau(D_{d_n})^\top)^\top$.

A good choice of the step size $\lambda^{[l]}$ in the updating rule (6) can greatly reduce the cost of the proposed algorithm, and hence it is critical for the fast convergence of the algorithm. One commonly uses a backtracking method to find $\lambda^{[l]}$ such that the loss function monotonically decreases with steps. Specifically, the selected $\lambda^{[l]}$ satisfies

$$\frac{1}{2n}\|\mathbb{Y} - \mathbb{X}\mathbb{B}^{[l+1]}\|_F^2 \leq \frac{1}{2n}\|\mathbb{Y} - \mathbb{X}\mathbb{B}^{[l]}\|_F^2 - \frac{\varrho\lambda^{[l]}}{2}\|\mathbb{B}^{[l+1]} - \mathbb{B}^{[l]}\|_F^2, \tag{7}$$

where $\varrho \in (0,1)$ is a fixed small constant. Let $L$ be a pre-specified positive integer, and $\epsilon$ be a tolerance parameter. Based on (4), (6) and (7), we summarize the iterative hard-thresholding (IHT) algorithm for problem (3) in Algorithm 1.

**Algorithm 1** (*Iterative Hard-Thresholding Algorithm)*

*Step 1.* Choose an initial value for $\mathbb{B}^{[0]}$, such as $\mathbb{B}^{[0]} = 0$;

*Step 2.* For each $l \in \{0, 1, \ldots, L\}$,

*Step 2.1.* Choose an initial step size $\lambda^{[l]}$;

*Step 2.2.* Compute $\mathbb{B}^{[l+1]}$ by equation (6);

*Step 2.3.* Stop *Step 2* if the linear search criterion (7) is satisfied; otherwise, take the step size to be $2\lambda^{[l]}$, and return to *Step 2.2*;

*Step 3.* Stop the algorithm if $\|\mathbb{B}^{[l+1]} - \mathbb{B}^{[l]}\|_F < \epsilon\|\mathbb{B}^{[l]}\|_F$; otherwise, increase $l$.

We set $\varrho = 10^{-3}$, $\epsilon = 10^{-3}$ and $L = 1000$ in Algorithm 1. The initial selection of $\lambda^{[l]}$ in Step 2.1 is vital to the success of the iterative hard-thresholding algorithm. It balances the non-increasing of $\|\mathbb{Y} - \mathbb{X}\mathbb{B}\|_F^2$ after each iteration and the convergence rate. We choose an initial $\lambda^{[l]}$ in Step 2.1 by adopting the Barzilai-Borwein rule (Barzilai and Borwein, 1988), which uses a diagonal matrix $\mathrm{diag}\{\lambda, \ldots, \lambda\}$ to approximate the Hessian matrix $n^{-1}\mathbb{X}^\top\mathbb{X}$. Specifically, the initial $\lambda^{[l]}$ is chosen as

$$\lambda^{[l]} = \arg\min_{\lambda} \|\lambda w^{[l]} - y^{[l]}\|_F^2 = \frac{\mathrm{trace}({w^{[l]}}^\top \mathbb{X}^\top \mathbb{X} w^{[l]})}{n \times \mathrm{trace}({w^{[l]}}^\top w^{[l]})},$$

where $w^{[l]} = \mathbb{B}^{[l]} - \mathbb{B}^{[l-1]}$ and $y^{[l]} = n^{-1}\mathbb{X}^\top\mathbb{X}w^{[l]}$.

### 2.3 Convergence Analysis

To show the convergence of the proposed algorithm, we introduce a definition of the first-order stationary point for problem (3), which can be viewed as a matrix version of the ordinary first-order stationary point.

**Definition 1** For a step size $\lambda$, $\mathbb{B} = (B_1^\top, \ldots, B_{d_n}^\top)^\top$ is called a *first-order stationary point* of problem (3), if the following holds:

$$(i)\ \ \mathbb{B} \in \mathbb{H}_\tau\Big\{\mathbb{B} + \frac{1}{n\lambda}\mathbb{X}^\top(\mathbb{Y} - \mathbb{X}\mathbb{B})\Big\},\ \ \text{and}\ \ (ii)\ \ \sum_{j=1}^{d_n} I(\|B_j\|_F \neq 0) \leq \tau.$$

The convergence properties of the proposed Algorithm 1 are summarized in the following theorem. Define $\phi$ as the maximum eigenvalue of $\tilde{X}^\top\tilde{X}/n$.

**Theorem 1** Let $\{\mathbb{B}^{[l]}\}$ be the sequence generated by Algorithm 1. If $\lambda^{(l)} \geq \phi/(1-\varrho)$, then we have

(i) All limit points of the sequence $\{\mathbb{B}^{[l]}\}$ are first-order stationary points of problem (3).

(ii) The sequence $\{\|\mathbb{Y} - \mathbb{X}\mathbb{B}^{[l]}\|_F^2\}$ is convergent as $l \to \infty$. In addition, after $L$ iterations, the sequence $\{\mathbb{B}^{[l]}\}$ satisfies that

$$\min_{l=0,1,2,\cdots,L} \|\mathbb{B}^{[l+1]} - \mathbb{B}^{[l]}\|_F^2 \leq \frac{1-\varrho}{\varrho\phi nL}\big\{\|\mathbb{Y} - \mathbb{X}\mathbb{B}^{[0]}\|_F^2 - \|\mathbb{Y} - \mathbb{X}\tilde{\mathbb{B}}\|_F^2\big\},$$

where $\tilde{\mathbb{B}}$ satisfies that $\|\mathbb{Y} - \mathbb{X}\mathbb{B}^{[l]}\|_F^2 \to \|\mathbb{Y} - \mathbb{X}\tilde{\mathbb{B}}\|_F^2$ as $l \to \infty$.

Part (i) of Theorem 1 shows that the limiting point of the sequence $\{\mathbb{B}^{[l]}\}$ is a first-order stationary point. Part $(ii)$ implies that if $L = [1/\varepsilon^2]$, the algorithm stops in a finite number of steps via checking stopping criterion $\|\mathbb{B}^{[l+1]} - \mathbb{B}^{[l]}\|_F \leq O(\varepsilon)$. Furthermore, given the proof of Theorem 1, we know that the monotone linear search criterion (7) is satisfied. Similar to the arguments in Gong et al. (2013), we can show that for each $l \geq 0$, $\lambda^{[l]}$ is bounded above when criterion (7) holds, indicating the existence of $\lambda^{[l]}$ in Step 2.3.

**Remark 1** *Convergence to a first-order stationary point is a necessary but not sufficient condition for convergence to a local optima. This is a typical complex optimization problem; see Bunea et al. (2012) for variable selection under low rank constraint, He et al. (2018) for a dimensionality reduction and variable selection, and Yun et al. (2011) for a coordinate gradient descent method.*

**Remark 2** *In all the simulations and real data application, we have checked that when we adopt the Barzilai-Borwein rule to determine the initial $\lambda^{[l]}$ in Step 2.1 of Algorithm 1, the selected $\lambda^{[l]}$ by Step 2.3 satisfies $\lambda^{[l]} \geq \phi/(1-\varrho)$ ($\varrho = 10^{-3}$) for all $l$. This guarantees the convergence of the algorithm by Theorem 1, though a rigorous theoretical justification of why the $\lambda^{[l]}$ derived by the Barzilai-Borwein rule satisfies $\lambda^{[l]} \geq \phi/(1-\varrho)$ is challenging work and needs further investigation.*

## 3. Theoretical Properties

This section first presents the theoretical variable selection properties, and then provides the minimax rates of the estimate obtained by our algorithm.

### 3.1 Variable Selection Properties

In this section, we study the variable selection properties of our method. Define $\mathcal{M}_\tau^+ = \{\mathcal{M} : \mathcal{M}^\star \subset \mathcal{M},\ \text{card}(\mathcal{M}) \leq \tau\}$ as the collections of the over-fitted models. The following conditions are required for establishing the sure screening property, all of which are regularity conditions and commonly adopted in the analysis of high-dimensional data.

(C1) The coefficient matrix size satisfies $\log(pq) = o(n^{\delta_0})$ for some $0 < \delta_0 < 1$.

(C2) The dimension of predictors satisfies $\log(d_n) = o(n^{\delta_1})$ for some $0 < \delta_1 < 1$.

(C3) $\tau^\star \leq \omega_1 n^{\delta_2}$ for some positive constants $\omega_1$ and $\delta_2$.

(C4) There exist some positive constants $\omega_3$ and $\delta_3$ such that

$$\min_{j \in \mathcal{M}^\star} \left[(pq)^{-1/2} \|B_j^\star\|_F\right] > \omega_2 n^{-\delta_3}.$$

(C5) For sufficiently large $n$, the smallest eigenvalue $\lambda_{\min}\big(n^{-1}\tilde{X}_{\mathcal{M}}^\top \tilde{X}_{\mathcal{M}}\big)$ is bounded away from zero and the largest eigenvalue $\lambda_{\max}\big(n^{-1}\tilde{X}_{\mathcal{M}}^\top \tilde{X}_{\mathcal{M}}\big)$ is bounded away from infinity for any $\mathcal{M} \in \mathcal{M}_+^{2\tau}$ with $\tau^\star \leq \tau < \omega_1 n^{\delta_2}$, where $\tilde{X}_{\mathcal{M}}$ denotes the submatrix of $\tilde{X}$ whose columns are indexed by $\mathcal{M}$.

(C6) There exist positive constants $\eta_1$ and $\eta_2$ such that $E\{\exp(\eta_1 |x_{ij}|)\} < \eta_2$ for each $1 \leq i \leq n$ and $1 \leq j \leq d_n$.

(C7) (i)For each $s = 1, \ldots, p$ and $k = 1, \ldots, q$, the random errors $e_{1,sk}, \ldots, e_{n,sk}$ are independent and identically distributed normal with mean 0 and variance $\sigma_{sk}^2$. (ii) There exists a constant $\sigma$ such that $0 < \sigma^{-2} < \min_{s,k}\{\sigma_{sk}^2\} \leq \max_{s,k}\{\sigma_{sk}^2\} < \sigma^2$. Besides, the covariance matrix $\Sigma_e$ of the errors $\{e_{i,sk} : s = 1, \ldots, p, k = 1, \ldots, q\}$ is nonsingular.

Conditions (C1) and (C2) state that both the coefficient matrix size ($pq$) and the number of predictors $d_n$ are allowed to grow at the exponential rate of the sample size $n$. Condition (C3) assumes that the number of nonzero coefficients is less than the sample size $n$ but can diverge at the rate of $O(n^{\delta_2})$. Condition (C4) indicates at what rate the minimum signal strength can be identified by our procedure. That is, the correlations between the important predictors and the matrix responses can be degenerate but not too fast, so that the signal is detectable (He et al., 2013; Zhu et al., 2012). Condition (C5) requires that the $2\tau$-restricted smallest eigenvalue is bounded away from zero, which is bigger than $\lambda_{\min}\big(n^{-1}\tilde{X}^\top \tilde{X}\big)$; see Xu and Chen (2014). Condition (C6) is much weaker than Condition (T4) in Xu and Chen (2014), which allows that the predictor $x_{ij}$ follows a sub-exponential distribution. A similar condition can be found in Zhu et al. (2012) for example. Conditions (C5) and (C7) imply that our method may break down under strong collinearity of the predictors or the matrix response. The multivariate normal condition is also used in Condition (A10) in Kong et al. (2020).

**Theorem 2** Under Conditions (C1)-(C7), there exist some positive constants $c_1$ and $c_2$ such that

$$pr\big(\mathcal{M}^\star \subseteq \hat{\mathcal{M}}_\tau\big) \geq 1 - c_2 \tau pq d_n^\tau \exp(-c_1 n^{1-\delta_2-2\delta_3-2\upsilon}) - c_2 n d_n \exp(-\eta_1 n^\upsilon),$$

where $\upsilon > 0$ is some constant with $(\delta_2 + \delta_3 + \upsilon) < (1 - \delta_0 - \delta_1)/2$.

Theorem 2 states that our proposed method enjoys the sure screening property (Fan and Lv, 2008) for $\tilde{\tau} \geq \tau^\star$ when we choose $\upsilon > \delta_1$. If $\tilde{\tau} = \tau^\star$, then $\hat{\mathcal{M}}_{\tilde{\tau}} = \mathcal{M}^\star$ holds with probability going to one. Compared to that in Kong et al. (2020), this theorem guarantees that the true model is contained in one of the candidate models $\{\hat{\mathcal{M}}_1, \ldots, \hat{\mathcal{M}}_{\tilde{\tau}}\}$ as long as $\tilde{\tau} \geq \tau^\star$. Theorem 2 also indicates that the probability bound depends on the coefficient matrix size $(pq)$. In particular, the rate of sure screening property becomes slower as $(pq)$ increases. It turns out that the results in Xu and Chen (2014) is a special case of Theorem 2 with $p = q = 1$. Also, Theorem 2 states that with probability approaching one, the true model $\mathcal{M}^\star$ can be included by the solution path within at most $O(n^{\delta_2})$ steps, which is a number much smaller than the sample size $n$ under the condition $(\delta_2+\delta_3+\upsilon) < (1-\delta_0-\delta_1)/2$. Lastly, the success of the sure screening property in Kong et al. (2020) relies on the marginal correlations between the predictors and response, which may miss some important predictors that are marginally uncorrelated but jointly correlated with $Y_i$.

Next, we prove the consistency property for the EBIC procedure.

**Theorem 3** Suppose that Conditions (C1)-(C7) hold with $2(\delta_2 + \delta_3 + \upsilon) < (1 - \delta_0 - \delta_1)$. If $c_n \to \infty$ and $c_n \log(n)/n^{1-\delta_2} \to 0$, then

$$pr\big(\mathcal{M}^\star = \hat{\mathcal{M}}_{\tilde{\tau}}\big) \to 1 \quad \text{as} \quad n \to \infty.$$

**Remark 3** *Theorem 3 shows that the consistency of the* EBIC *relies on the choice of $c_n$. Intuitively, a large $c_n$-value leads to seriously under-fitted models, and vise versa. Note that when $c_n = 1$, the* EBIC *reduces to the classical* BIC *(Schwarz, 1978). Wang et al. (2009) proposed to select $c_n = O(\log\{\log(d_n)\})$ when the number of variables diverges with the sample size. Here we adopt $c_n = \log\{\log(d_n)\}/3$ in the simulation studies.*

### 3.2 Minimax Rates of Estimation

In this subsection, we study the minimax rates of the estimate obtained by solving problem (2). Here let $\tilde{X}$ be column-wisely normalized. Denote $\mathcal{B}_\tau = \{\mathbb{B} : \sum_{j=1}^{d_n} I(\|B_j\|_F \neq 0) \leq \tau\}$. The following technical assumption is needed to establish the lower and upper bounds for estimation.

(A) There exist some constants $0 < \kappa_1 < \kappa_2 < \infty$ such that for any $\theta \in \mathcal{B}_{2\tau}$,

$$\kappa_1 \|\theta\|_F < \frac{1}{\sqrt{n}} \|\mathbb{X}\theta\|_F < \kappa_2 \|\theta\|_F.$$

Assumption (A) is the sparse eigenvalue condition (Raskutti et al., 2011). This assumption depends on the sparse level $\tau$. The $2\tau$-restricted maximal eigenvalue $\kappa_2^2$ can be much smaller than the maximum eigenvalue of $\tilde{X}^\top \tilde{X}/n$.

**Theorem 4** (Upper Bounds) Suppose that Condition (C7) and Assumption (A) hold. Then for $\tau \leq d_n/2$, there exist some positive constants $c_3$, $c_4$ and $c_5$ such that

$$\begin{aligned} pr\left\{ \inf_{\hat{\mathbb{B}}_\tau} \sup_{\mathbb{B}^\star \in \mathcal{B}_\tau} \frac{1}{\sqrt{(pq)}} \|\hat{\mathbb{B}}_\tau - \mathbb{B}^\star\|_F \leq \frac{c_3 \sigma \kappa_2}{\kappa_1} \Big( \frac{\tau \log(d_n/\tau)}{n} \Big)^{1/2} \right\} \\ \geq 1 - c_4 pq \exp\{-c_5 \tau \log(d_n/\tau)\}. \end{aligned}$$

In addition, if $d_n \geq 4\tau$, then there exist some positive constants $c_6,\ c_7$ and $c_8$ such that

$$pr\left\{\inf_{\hat{\mathbb{B}}_\tau}\sup_{\mathbb{B}^\star\in\mathcal{B}_\tau}\frac{1}{\sqrt{(npq)}}\|\mathbb{X}(\hat{\mathbb{B}}_\tau-\mathbb{B}^\star)\|_F\leq\frac{c_6\sigma}{\kappa_2}\Big(\frac{\tau\log(d_n/\tau)}{n}\Big)^{1/2}\right\}$$
$$\geq 1-c_7pq\exp\big[-c_8\tau\log\{d_n/(2\tau)\}\big].$$

Theorem 4 gives nonasymptotic error bounds for any fixed $(n, d_n, p, q)$. If Conditions (C2) and (C3) are satisfied and $pq = o\{n^{1-(\delta_1+\delta_2)}\}$ with $(\delta_1+\delta_2) < 1$, the estimate $\hat{\mathbb{B}}_\tau$ is consistent with probability going to one.

**Theorem 5** (Lower Bounds) Suppose that Condition (C7) and Assumption (A) hold. If $\tau \leq d_n/2$, then there exist some positive constants $c_9$ and $c_{10}$ such that

$$\inf_{\hat{\mathbb{B}}_\tau}\sup_{\mathbb{B}^\star\in\mathcal{B}_\tau}E\Big(\frac{1}{\sqrt{(pq)}}\|\hat{\mathbb{B}}_\tau-\mathbb{B}^\star\|_F\Big)\geq\frac{c_9\sigma\kappa_2}{\kappa_1}\Big\{\frac{\tau\log(d_n/\tau)}{n}\Big\}^{1/2},$$

and

$$\inf_{\hat{\mathbb{B}}_\tau}\sup_{\mathbb{B}^\star\in\mathcal{B}_\tau}E\Big(\frac{1}{\sqrt{(npq)}}\|\mathbb{X}(\hat{\mathbb{B}}_\tau-\mathbb{B}^\star)\|_F\Big)\geq\frac{c_{10}\sigma}{\kappa_2}\Big\{\frac{\tau\log(d_n/\tau)}{n}\Big\}^{1/2}.$$

Theorem 5 establishes the lower error bounds for estimation and prediction, which holds for the minimizer of problem (2). Theorems 4 and 5 identify the optimal minimax rates up to a constant factor. In particular, the minimax Frobenius-norm rate scales as $\{\tau\log(d_n/\tau)/n\}^{1/2}$. When $p = q = 1$ or $(pq)$ is fixed, the minimax lower bounds coincide with Raskutti et al. (2011). Therefore, Theorem 5 generalizes their results to the matrix linear regression model (1).

**Theorem 6** Suppose that Assumption (A) and Conditions (C6)-(C7) hold and $\tau^\star \leq \tau$. If the initial value $\mathbb{B}^{[0]} = 0$, the iterative number $l > [\log_2\{(n\phi)^{1/2}\|\mathbb{B}^\star\|_F/\|\mathbb{X}^\top\mathbb{U}\|_F\}]$, and $\phi < \lambda^{[l]} \leq \kappa_1/(1-1/\sqrt{32})$, then

$$pr\left\{\frac{1}{\sqrt{(pq)}}\|\mathbb{B}^{[l]}-\mathbb{B}^\star\|_F\leq\tilde{c}_3\sigma\Big(\frac{\tau\log(d_n/\tau)}{n}\Big)^{1/2}\right\}$$
$$\geq 1-\tilde{c}_4\big[pq\tau\exp\{-\tilde{c}_5\log(d_n/\tau)n^{2\upsilon}\}+pq\tau\exp\{-\eta_1n^\upsilon\}\big],$$

and

$$pr\left\{\frac{1}{\sqrt{(npq)}}\|\mathbb{X}(\mathbb{B}^{[l]}-\mathbb{B}^\star)\|_F\leq\tilde{c}_6\sigma\Big(\frac{\tau\log(d_n/\tau)}{n}\Big)^{1/2}\right\}$$
$$\geq 1-\tilde{c}_7\big[pq\tau\exp\{-\tilde{c}_8\log(d_n/\tau)n^{1-2\upsilon}\}+n\tau\exp\{-\eta_1n^\upsilon\}\big],$$

where $\tilde{c}_j$ $(j = 3, \ldots, 8)$ are some positive constants.

Theorem 6 states that the estimate generated by Algorithm 1 also enjoys the optimal minimax rate $\{\tau\log(d_n/\tau)/n\}^{1/2}$. Similarly to Theorem 4, the upper bound of $\|\mathbb{B}^{[l]}-\mathbb{B}^\star\|_F$ is negligible under Conditions (C2) and (C3) and $pq = o\{n^{1-(\delta_1+\delta_2)}\}$ with $(\delta_1+\delta_2) < 1$. Hence, Theorem 6 guarantees the consistency of the estimate generated by Algorithm 1. In addition, Theorems 1 and 6 imply that after a finite number of steps the estimator $\mathbb{B}^{[l]}$ would be consistent with probability going to one.

## 4. Simulation Studies

We conduct simulation studies to examine the finite sample performance of the proposed method. The error term $\mathrm{Vec}(U_i)$ is independently generated from the standard multivariate normal distribution, where $\mathrm{Vec}(\cdot)$ denotes the vectorization of a matrix. The total number of predictors is $d_n = 1000$. The sample sizes are $n = 100$ and 200. We consider $(p, q) = (50, 50)$ and $(p, q) = (150, 150)$. The following three models are considered.

**Example 1** *Let $M_1$ and $M_2$ be the subset of $\{1, \ldots, p\}$ and $\{1, \ldots, q\}$, respectively. Define $B[M_1, M_2] = \{b_{sk} : \ s \in M_1, k \in M_2\}$ as the sub-matrix of $B = (b_{sk}) \in \mathbb{R}^{p\times q}$. For the $i$th subject, let $X_i = (x_{ij}) \in \mathbb{R}^{d_n}$ follow a multivariate normal distribution with mean 0 and covariance matrix $(\sigma_{kl})_{d_n\times d_n}$. Let $\sigma_{kl} = \vartheta$ for $k \neq l$ and $k, l \neq 5$, $\sigma_{5l} = \sigma_{l5} = 0$ for $l \neq 5$, and $\sigma_{kl} = 1$ for $k = l$. We set $\vartheta = 0.1, 0.5$ and 0.9. The responses $Y_i$ $(i = 1, 2, \ldots, n)$ are generated by $Y_i = 2x_{i1}B + 2x_{i2}B + 2x_{i3}B - 6\vartheta x_{i4}B + 0.5x_{i5}B + U_i$, and the matrix $B$ is generated as follows. Set $M_1 = \{1, \ldots, 20\}$ and $M_2 = \{1, \ldots, 20\}$. Then, let $B[M_1, M_2] = vv^\top$, where the entries of $v \in \mathbb{R}^{20\times 5}$ follow the standard normal distribution. The other entries of $B$ are set as 0. The regression coefficients $B_j$ are set as 0 for $j \geq 6$. This example is originally considered by Fan and Lv (2008). In this case, the marginal correlation between $x_{i4}$ and $Y_i$ is indeed zero. Moreover, $x_{i5}$ is a relatively weak signal for the response, and does not "borrow" strength from all other variables.*

**Example 2** *We consider a challenging case, similar as the one in Wang (2009). For the $i$th subject, we generate $X_i = (x_{ij}) \in \mathbb{R}^{d_n}$ as follows: independently simulate $Z_i = (z_{ij}) \in \mathbb{R}^{d_n}$ and $W_i = (w_{ij}) \in \mathbb{R}^{d_n}$ from a standard multivariate normal distribution, and obtain $x_{ij}$ by $x_{ij} = (z_{ij} + w_{ij})/\sqrt{2}$ for every $1 \leq j \leq 5$, and $x_{ij} = (z_{ij} + \sum_{j'=1}^{5} z_{ij'})/2$ for every $5 < j \leq d_n$. The responses $Y_i = (Y_{i,sk})$ are generated by $Y_i = \sum_{j=1}^{5}(jx_{ij}B) + U_i$, where $B[M_{1k}, M_{2k}] = v_k u_k^\top$ for $k = 1, \ldots, p/5$, and the entries of $v_k \in \mathbb{R}^{5\times 2}$ and $u_k \in \mathbb{R}^{5\times 2}$ follow from the standard normal distribution. We set $M_{1k} = M_{2k} = \{5k - 4, \ldots, 5k\}$. The other entries of $B$ are set as 0. The coefficient matrices $B_j$ $(j > 5)$ are set to be 0. In this case, it is very difficult to discover (for example) $x_{i1}$ as a significant variable, because that the correlation coefficient of $x_{i1}$ and $Y_i$ is much smaller than that of $x_{ij}$ and $Y_i$ for every $j > 5$.*

**Example 3** *Let $\tau^\star = [\sqrt{n}]$, and $p = q = 4[\sqrt{n}]$. The variable $X_i = (x_{ij}) \in \mathbb{R}^{d_n}$ follows a multivariate normal distribution with mean 0 and covariance matrix $(\sigma_{kl})_{d_n\times d_n}$, where $\sigma_{kl} = 0.5$ for $k \neq l$, and $\sigma_{kl} = 1$ for $k = l$. Set $M_1 = M_2 = \{17, \ldots, 24\}$. The regression coefficients are $B_j[M_1, M_2] = \sum_{k=1}^{3}(a_k v_k u_k^\top)$, where the entries of $v_k \in \mathbb{R}^{8\times 3}$ and $u_k \in \mathbb{R}^{8\times 3}$ follow the standard normal distribution, and $a_k = 2^{2-k}\log(n)n^{-1/2}$. The other entries of $B_j$ are set as 0. The coefficient matrices $B_j$ $(j > \tau^\star)$ are set to be 0.*

All simulation results are based on 100 Monte Carlo repetitions. Let $\hat{\mathbb{B}}^m$ be the estimate realized in the $m$th simulation replication, and $\hat{\mathcal{M}}_{\hat{\tau},m}$ be the corresponding selected model. Let $\mathcal{M}_c^\star = \{1 \leq j \leq d_n : \ j \notin \mathcal{M}^\star\}$. We consider the following four criteria: (I) the proportion of times when $\hat{\mathcal{M}}_{\hat{\tau},m}$ contains $x_j$ (denoted by $p_j$); (II) the average number of true positives (TP) and false positives (FP), defined as $100^{-1}\sum_{m=1}^{100}\mathrm{card}(\hat{\mathcal{M}}_{\hat{\tau},m} \cap \mathcal{M}^\star)$ and $100^{-1}\sum_{m=1}^{100}\mathrm{card}(\hat{\mathcal{M}}_{\hat{\tau},m} \cap \mathcal{M}_c^\star)$, respectively; (III) the proportion of times when $\hat{\mathcal{M}}_{\hat{\tau}} = \mathcal{M}^\star$ (correct fitting rate, CF); and (IV) the average of the scaled mean squared errors, defined

as $100^{-1}\sum_{m=1}^{100}\{\|\mathbb{X}(\hat{\mathbb{B}}^m-\mathbb{B}^\star)\|_F^2/(npq)\}$ for prediction (Pred) and as $100^{-1}\sum_{m=1}^{100}\{\|\hat{\mathbb{B}}^m-\mathbb{B}^\star\|_F^2/(pq)\}$ for estimation (Est). We summarize the results in Tables 1-4.

The simulation results show that the SRLS procedure performs well. Specifically, the average of the number of true positives is close to the number of true active variables. Besides, the average of the number of false positives is close to zero. These results are consistent with Theorem 3 that the proposed method enjoys the model selection consistency. The results of Examples 2 and 3 suggest that the proposed method can also handle the case when the coefficient matrix size is larger than the sample size or diverges as the sample size increases. Furthermore, the prediction results indicate that the SRLS method enjoys estimation consistency. Moreover, Tables 1 and 2 show that the proposed method yields robust results when the correlations between $x_j$ $(j=1,\ldots,d_n)$ increase, which suggests that our method can handle the highly correlated variables. We also consider the settings with the error term $\mathrm{Vec}(U_i)$ independently generated from $t$-distribution and chi-square distribution. The results are provided in Tables S1-S4 of Appendix C. We observe that the proposed method performs well for the sub-exponential distributions considered here.

We compare our method with three procedures: the method of Kong et al. (2020) (KAZZ, for short), the SIS* and ISIS*. The SIS* and ISIS* methods for the matrix responses are designed as follows. For the SIS*, we first standardize $\{x_{ij}, i=1,\ldots,n\}$ and $\{Y_{i,lk}, i=1,\ldots,n\}$ for any $j=1,\ldots,d_n$, $l=1,\ldots,p$ and $k=1,2,\ldots,q$, denoted by $\{\tilde{x}_{ij}, i=1,\ldots,n\}$ and $\{\tilde{Y}_{i,lk}, i=1,\ldots,n\}$. Then we calculate $n^{-1}\sum_{i=1}^n \tilde{x}_{ij}\tilde{Y}_i, 1\le j\le d_n$, where $\tilde{Y}_i=(\tilde{Y}_{i,lk})$. Define the selected model as

$$\hat{\mathcal{M}}_\tau^{SIS}=\big\{1\le j\le d_n : \|n^{-1}\sum_{i=1}^n \tilde{x}_{ij}\tilde{Y}_i\|_F \text{ is among the first } \tau \text{ largest of all}\big\}, \tag{8}$$

where $1\le\tau\le\tilde{\tau}$. Based on $\hat{\mathcal{M}}_\tau^{SIS}$, we estimate the coefficient matrices using the OLS method:

$$\min_{B_j}\left\{\frac{1}{2n}\sum_{i=1}^n\|Y_i-\sum_{j\in\hat{\mathcal{M}}_\tau^{SIS}}x_{ij}B_j\|_F^2\right\}.$$

Finally, we carry out this procedure for each $\tau\in\{1,,\ldots,\tilde{\tau}\}$, and adopt the EBIC to select the 'best' fitting model among the $\tilde{\tau}$ submodels. For the ISIS* method, let $\tau_1=[\tau/2]$, and $\tau_2=\tau-\tau_1$. In the first step, we select a submodel via (8), denoted by $\hat{\mathcal{M}}_{1,\tau_1}^{SIS}$. Then we calculate the residuals by regressing the response $\tilde{Y}_i$ on $(\tilde{x}_{ij}: j\in\hat{\mathcal{M}}_{1,\tau_1}^{SIS})$. In the next step, we treat those residuals as the new responses and apply (8) again to the remaining $d_n-\tau_1$ predictors, which results in a submodel $\hat{\mathcal{M}}_{2,\tau_2}^{SIS}$. Let $\hat{\mathcal{M}}_\tau^{ISIS}=\hat{\mathcal{M}}_{1,\tau_1}^{SIS}\cup\hat{\mathcal{M}}_{2,\tau_2}^{SIS}$. Based on $\hat{\mathcal{M}}_\tau^{ISIS}$, we estimate the coefficient matrices by the OLS method. Finally, we carry out the modified ISIS procedure for each $\tau\in\{1,,\ldots,\tilde{\tau}\}$, and select the 'best' fitting model among the $\tilde{\tau}$ submodels by the EBIC. For the KAZZ method, we choose the threshold the same way as the SIS* method for a fair comparison.

Tables 1 and 2 indicate that under the scenario of Example 1, the SIS* fails to capture $x_{i4}$ as the marginal correlation of $x_{i4}$ and $Y_i$ is zero, and it fails to capture $x_{i5}$ because the signal of $x_{i5}$ is weak. Besides, Tables 1-4 demonstrate that although the ISIS* outperforms the SIS* and KAZZ in terms of model selection, it is not stable when the correlations between $x_{ij}$'s

vary and gives large errors for the estimation and prediction than our proposed method. Under the scenario of Example 2, the SIS* and KAZZ procedures overlook some variables such as $x_{i1}$. The phenomenon makes sense because the correlation between $x_{i1}$ and $Y_i$ is much smaller than that of $x_{ij}$ and $Y_i$ for every $j > 5$, and hence $x_{i1}$ is not captured by the marginal screening methods.

| $\vartheta$ | Method | $p_1$ | $p_2$ | $p_3$ | $p_4$ | $p_5$ | TP | FP | CF | Est | Pred |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | $(p,q) = (50,50)$ | | | | | |
| 0.1 | SRLS | 1.00(0.00) | 1.00(0.00) | 1.00(0.00) | 1.00(0.00) | 1.00(0.00) | 5.00(0.00) | 0.00(0.00) | 1.00(0.00) | 0.05(0.00) | 0.05(0.00) |
| | SIS* | 1.00(0.00) | 1.00(0.00) | 1.00(0.00) | 0.00(0.00) | 0.00(0.00) | 3.00(0.00) | 0.07(0.33) | 0.00(0.00) | 0.64(0.16) | 0.58(0.19) |
| | ISIS* | 1.00(0.00) | 1.00(0.00) | 1.00(0.00) | 1.00(0.00) | 0.98(0.14) | 4.98(0.14) | 0.14(0.57) | 0.92(0.27) | 0.06(0.03) | 0.05(0.02) |
| | KAZZ | 1.00(0.00) | 1.00(0.00) | 1.00(0.00) | 0.00(0.00) | 0.03(0.17) | 3.03(0.17) | 0.10(0.48) | 0.00(0.00) | 0.74(0.22) | 0.66(0.22) |
| 0.5 | SRLS | 1.00(0.00) | 1.00(0.00) | 1.00(0.00) | 1.00(0.00) | 1.00(0.00) | 5.00(0.00) | 0.00(0.00) | 1.00(0.00) | 0.08(0.01) | 0.05(0.00) |
| | SIS* | 0.94(0.24) | 0.90(0.30) | 0.96(0.20) | 0.00(0.00) | 0.00(0.00) | 2.80(0.40) | 0.26(0.56) | 0.00(0.00) | 12.3(3.77) | 5.97(1.92) |
| | ISIS* | 0.97(0.17) | 0.95(0.22) | 0.98(0.14) | 1.00(0.00) | 0.44(0.50) | 4.34(0.57) | 1.77(1.90) | 0.11(0.31) | 0.85(1.85) | 0.37(0.57) |
| | KAZZ | 0.95(0.22) | 0.94(0.24) | 0.93(0.26) | 0.00(0.00) | 0.00(0.00) | 2.82(0.44) | 0.24(0.53) | 0.00(0.00) | 12.6(4.46) | 6.07(1.75) |
| 0.9 | SRLS | 1.00(0.00) | 1.00(0.00) | 1.00(0.00) | 1.00(0.00) | 1.00(0.00) | 5.00(0.00) | 0.18(0.36) | 0.85(0.36) | 0.38(0.11) | 0.05(0.00) |
| | SIS* | 0.29(0.46) | 0.23(0.42) | 0.31(0.46) | 0.04(0.20) | 0.10(0.30) | 0.97(0.69) | 0.36(0.99) | 0.00(0.00) | 38.9(13.1) | 3.92(1.50) |
| | ISIS* | 0.70(0.46) | 0.66(0.48) | 0.67(0.47) | 1.00(0.00) | 1.00(0.00) | 4.03(0.73) | 2.68(2.17) | 0.08(0.27) | 6.01(4.92) | 0.43(0.30) |
| | KAZZ | 0.34(0.48) | 0.23(0.42) | 0.24(0.43) | 0.05(0.22) | 0.24(0.43) | 1.10(0.81) | 0.48(1.14) | 0.00(0.00) | 38.9(14.9) | 3.78(1.52) |
| | | | | | | $(p,q) = (150,150)$ | | | | | |
| 0.1 | SRLS | 1.00(0.00) | 1.00(0.00) | 1.00(0.00) | 0.98(0.14) | 0.92(0.27) | 4.90(0.33) | 0.00(0.00) | 0.91(0.29) | 0.06(0.00) | 0.05(0.00) |
| | SIS* | 1.00(0.00) | 1.00(0.00) | 1.00(0.00) | 0.00(0.00) | 0.01(0.10) | 3.01(0.10) | 0.02(0.14) | 0.00(0.00) | 0.11(0.02) | 0.10(0.02) |
| | ISIS* | 1.00(0.00) | 1.00(0.00) | 1.00(0.00) | 0.12(0.33) | 0.13(0.34) | 3.25(0.54) | 0.07(0.26) | 0.05(0.22) | 0.10(0.02) | 0.09(0.02) |
| | KAZZ | 1.00(0.00) | 1.00(0.00) | 1.00(0.00) | 0.00(0.00) | 0.01(0.10) | 3.01(0.10) | 0.01(0.10) | 0.00(0.00) | 0.09(0.03) | 0.08(0.02) |
| 0.5 | SRLS | 1.00(0.00) | 1.00(0.00) | 1.00(0.00) | 1.00(0.00) | 0.85(0.36) | 4.85(0.36) | 0.00(0.00) | 0.85(0.36) | 0.08(0.01) | 0.05(0.00) |
| | SIS* | 0.86(0.35) | 0.89(0.31) | 0.82(0.39) | 0.00(0.00) | 0.00(0.00) | 2.57(0.50) | 0.12(0.33) | 0.00(0.00) | 1.39(0.40) | 0.67(0.20) |
| | ISIS* | 0.94(0.24) | 0.95(0.22) | 0.91(0.29) | 0.96(0.20) | 0.97(0.17) | 4.73(0.53) | 1.11(1.75) | 0.50(0.50) | 0.27(0.37) | 0.12(0.13) |
| | KAZZ | 0.87(0.34) | 0.87(0.34) | 0.92(0.27) | 0.00(0.00) | 0.00(0.00) | 2.66(0.50) | 0.17(0.45) | 0.00(0.00) | 1.40(0.42) | 0.68(0.23) |
| 0.9 | SRLS | 0.98(0.14) | 0.97(0.17) | 0.98(0.14) | 1.00(0.00) | 1.00(0.00) | 4.93(0.43) | 0.34(0.65) | 0.74(0.44) | 0.40(0.28) | 0.05(0.01) |
| | SIS* | 0.26(0.44) | 0.28(0.45) | 0.27(0.45) | 0.04(0.20) | 0.16(0.37) | 1.01(0.67) | 0.35(0.93) | 0.00(0.00) | 4.45(1.43) | 0.46(0.16) |
| | ISIS* | 0.51(0.50) | 0.51(0.50) | 0.52(0.50) | 1.00(0.00) | 1.00(0.00) | 3.54(0.77) | 1.16(1.09) | 0.07(0.26) | 1.36(0.72) | 0.12(0.04) |
| | KAZZ | 0.31(0.46) | 0.19(0.39) | 0.26(0.44) | 0.04(0.20) | 0.21(0.41) | 1.01(0.86) | 0.28(0.57) | 0.00(0.00) | 4.51(1.45) | 0.44(0.15) |

Table 1: The selection results and standard deviation (in parentheses) for Example 1 with $n = 100$.

## 5. Alzheimer's Disease Neuroimaging Initiative

We performed an imaging genetics analysis by exploring the relationship between the 2D hippocampus surface imaging data and the genes on the 19th chromosome. We first preprocessed the imaging and genetics data from the Alzheimer's Disease Neuroimaging Initiative study, and there were 735 subjects and 2000 SNPs retained after preprocessing. Each subject has left and right hippocampus shape representations, each of which is represented as a $100 \times 150$ matrix. We provided the data usage agreement and the detailed data preprocessing steps in Appendix B. We first fitted the matrix linear regression model (1) with either left (or right) hippocampus shape representation from 735 subjects as $100 \times 150$ matrix responses, and age and gender as clinical variables. We also chose the first 5 principal component scores based on the SNP data as variables to correct for population stratification. We calculated the OLS estimates of coefficient matrices and then computed the corresponding residual matrices for the left and right hippocampi after adjusting the effects

| $\vartheta$ | Method | $p_1$ | $p_2$ | $p_3$ | $p_4$ | $p_5$ | TP | FP | CF | Est | Pred |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | $(p,q)=(50,50)$ | | | | | | |
| 0.1 | SRLS | 1.00(0.00) | 1.00(0.00) | 1.00(0.00) | 1.00(0.00) | 1.00(0.00) | 5.00(0.00) | 0.00(0.00) | 1.00(0.00) | 0.03(0.00) | 0.03(0.00) |
| | SIS* | 1.00(0.00) | 1.00(0.00) | 1.00(0.00) | 0.00(0.00) | 0.01(0.10) | 3.01(0.10) | 0.09(0.81) | 0.00(0.00) | 0.71(0.20) | 0.67(0.20) |
| | ISIS* | 1.00(0.00) | 1.00(0.00) | 1.00(0.00) | 1.00(0.00) | 1.00(0.00) | 5.00(0.00) | 0.00(0.00) | 1.00(0.00) | 0.03(0.00) | 0.03(0.00) |
| | KAZZ | 1.00(0.00) | 1.00(0.00) | 1.00(0.00) | 0.00(0.00) | 0.01(0.10) | 3.01(0.10) | 0.05(0.50) | 0.00(0.00) | 0.71(0.23) | 0.65(0.22) |
| 0.5 | SRLS | 1.00(0.00) | 1.00(0.00) | 1.00(0.00) | 1.00(0.00) | 1.00(0.00) | 5.00(0.00) | 0.00(0.00) | 1.00(0.00) | 0.04(0.00) | 0.03(0.00) |
| | SIS* | 1.00(0.00) | 1.00(0.00) | 1.00(0.14) | 0.00(0.00) | 0.00(0.00) | 3.00(0.00) | 0.08(0.39) | 0.00(0.00) | 11.3(3.61) | 5.93(1.98) |
| | ISIS* | 1.00(0.00) | 1.00(0.00) | 1.00(0.00) | 1.00(0.00) | 0.65(0.48) | 4.65(0.48) | 1.97(2.71) | 0.24(0.43) | 0.15(0.12) | 0.11(0.11) |
| | KAZZ | 1.00(0.00) | 1.00(0.00) | 1.00(0.00) | 0.00(0.00) | 0.00(0.00) | 3.00(0.00) | 0.08(0.31) | 0.00(0.00) | 10.9(3.08) | 5.72(1.81) |
| 0.9 | SRLS | 1.00(0.00) | 1.00(0.00) | 1.00(0.00) | 1.00(0.00) | 1.00(0.00) | 5.00(0.00) | 0.01(0.10) | 0.99(0.10) | 0.17(0.02) | 0.03(0.00) |
| | SIS* | 0.43(0.50) | 0.35(0.48) | 0.39(0.49) | 0.03(0.17) | 0.21(0.41) | 1.41(0.89) | 0.10(0.72) | 0.00(0.00) | 38.9(13.7) | 3.99(1.38) |
| | ISIS* | 0.97(0.17) | 0.96(0.20) | 0.95(0.22) | 1.00(0.00) | 1.00(0.00) | 4.88(0.36) | 1.93(2.66) | 0.50(0.50) | 0.89(1.94) | 0.08(0.14) |
| | KAZZ | 0.40(0.49) | 0.42(0.50) | 0.31(0.46) | 0.01(0.10) | 0.22(0.42) | 1.36(0.81) | 0.06(0.37) | 0.01(0.10) | 42.1(13.9) | 4.28(1.42) |
| | | | | | $(p,q)=(150,150)$ | | | | | | |
| 0.1 | SRLS | 1.00(0.00) | 1.00(0.00) | 1.00(0.00) | 1.00(0.00) | 1.00(0.00) | 5.00(0.00) | 0.00(0.00) | 1.00(0.00) | 0.04(0.00) | 0.03(0.00) |
| | SIS* | 1.00(0.00) | 1.00(0.00) | 1.00(0.00) | 0.00(0.00) | 0.00(0.10) | 3.00(0.00) | 0.00(0.00) | 0.00(0.00) | 0.09(0.03) | 0.09(0.03) |
| | ISIS* | 1.00(0.00) | 1.00(0.00) | 1.00(0.00) | 0.72(0.45) | 0.85(0.36) | 4.57(0.67) | 0.23(0.65) | 0.62(0.49) | 0.04(0.02) | 0.04(0.02) |
| | KAZZ | 1.00(0.00) | 1.00(0.00) | 1.00(0.00) | 0.00(0.00) | 0.00(0.00) | 3.00(0.00) | 0.00(0.00) | 0.00(0.00) | 0.08(0.02) | 0.07(0.02) |
| 0.5 | SRLS | 1.00(0.00) | 1.00(0.00) | 1.00(0.00) | 1.00(0.00) | 1.00(0.00) | 5.00(0.00) | 0.00(0.00) | 1.00(0.00) | 0.04(0.00) | 0.03(0.00) |
| | SIS* | 1.00(0.00) | 1.00(0.00) | 0.99(0.10) | 0.00(0.00) | 0.00(0.00) | 2.99(0.10) | 0.00(0.00) | 0.00(0.00) | 1.29(0.34) | 0.68(0.18) |
| | ISIS* | 1.00(0.00) | 1.00(0.00) | 1.00(0.00) | 1.00(0.00) | 1.00(0.00) | 5.00(0.00) | 0.00(0.00) | 1.00(0.00) | 0.04(0.00) | 0.03(0.00) |
| | KAZZ | 1.00(0.00) | 1.00(0.00) | 1.00(0.00) | 0.00(0.00) | 0.00(0.00) | 3.00(0.00) | 0.00(0.00) | 0.00(0.00) | 1.26(0.38) | 0.66(0.21) |
| 0.9 | SRLS | 1.00(0.00) | 1.00(0.00) | 1.00(0.00) | 1.00(0.00) | 1.00(0.00) | 5.00(0.00) | 0.02(0.14) | 0.98(0.14) | 0.16(0.02) | 0.02(0.00) |
| | SIS* | 0.38(0.49) | 0.33(0.47) | 0.30(0.46) | 0.01(0.10) | 0.10(0.30) | 1.12(0.48) | 0.04(0.32) | 0.00(0.00) | 4.36(1.44) | 0.46(0.16) |
| | ISIS* | 0.93(0.26) | 0.92(0.27) | 0.88(0.33) | 1.00(0.00) | 1.00(0.00) | 4.73(0.49) | 0.59(1.29) | 0.56(0.50) | 0.34(0.31) | 0.04(0.03) |
| | KAZZ | 0.30(0.46) | 0.31(0.46) | 0.41(0.49) | 0.00(0.00) | 0.18(0.39) | 1.20(0.45) | 0.00(0.00) | 0.00(0.00) | 4.38(1.07) | 0.46(0.12) |

Table 2: The selection results and standard deviation (in parentheses) for Example 1 with $n = 200$.

| $n$ | Matrix Size | Method | $p_1$ | $p_2$ | $p_3$ | $p_4$ | $p_5$ | TP | FP | CF | Est | Pred |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 100 | $p=q=50$ | SRLS | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 5.00 | 0.00 | 1.00 | 0.05 | 0.05 |
| | | SIS* | 0.00 | 0.00 | 0.00 | 0.04 | 0.56 | 0.60 | 2.23 | 0.00 | 9.83 | 3.88 |
| | | ISIS* | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 5.00 | 3.90 | 0.00 | 0.22 | 0.90 |
| | | KAZZ | 0.00 | 0.00 | 0.00 | 0.02 | 0.57 | 0.59 | 2.32 | 0.00 | 10.2 | 4.17 |
| | $p=q=150$ | SRLS | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 5.00 | 0.00 | 1.00 | 0.05 | 0.05 |
| | | SIS* | 0.00 | 0.00 | 0.00 | 0.02 | 0.57 | 0.59 | 2.46 | 0.00 | 3.34 | 1.32 |
| | | ISIS* | 0.99 | 1.00 | 1.00 | 1.00 | 1.00 | 4.99 | 3.81 | 0.00 | 0.22 | 0.09 |
| | | KAZZ | 0.00 | 0.00 | 0.00 | 0.02 | 0.50 | 0.52 | 2.16 | 0.00 | 3.65 | 1.44 |
| 200 | $p=q=50$ | SRLS | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 5.00 | 0.00 | 1.00 | 0.03 | 0.03 |
| | | SIS* | 0.00 | 0.00 | 0.00 | 0.07 | 0.72 | 0.79 | 2.78 | 0.00 | 8.65 | 3.96 |
| | | ISIS* | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 5.00 | 3.50 | 0.00 | 0.09 | 0.04 |
| | | KAZZ | 0.00 | 0.00 | 0.00 | 0.05 | 0.74 | 0.79 | 2.94 | 0.00 | 9.07 | 4.13 |
| | $p=q=150$ | SRLS | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 5.00 | 0.00 | 1.00 | 0.03 | 0.03 |
| | | SIS* | 0.00 | 0.00 | 0.00 | 0.04 | 0.76 | 0.80 | 2.38 | 0.00 | 2.80 | 1.31 |
| | | ISIS* | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 5.00 | 3.38 | 0.00 | 0.09 | 0.04 |
| | | KAZZ | 0.00 | 0.00 | 0.00 | 0.04 | 0.78 | 0.82 | 2.84 | 0.00 | 2.86 | 1.34 |

Table 3: The selection results for Example 2.

of the clinical variables and the SNP principal component scores. This first step analysis is to exclude the confounding factors age, gender and population stratification.

We applied the SIS*, ISIS*, and the proposed methods to find the significant SNPs to the left and right hippocampal surfaces, respectively. The selection results are reported in Table 5. Our proposed method selects three SNPs: rs3119815, rs266875 and rs12610273

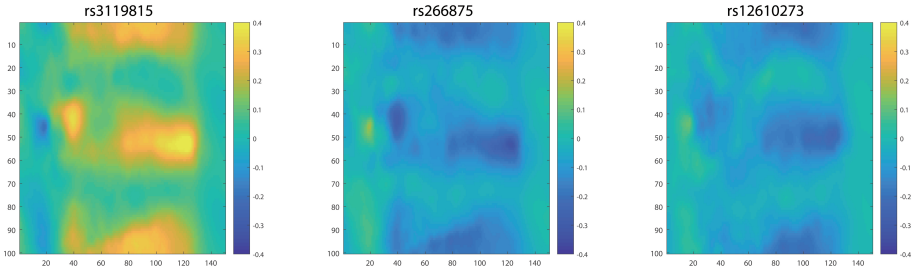
| | $n = 100$ | | | | | $n = 200$ | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Method | TP | FP | CF | Est | Pred | TP | FP | CF | Est | Pred |
| SRLS | 9.99 | 0.00 | 0.99 | 0.20 | 0.10 | 13.6 | 0.00 | 0.70 | 0.14 | 0.07 |
| SIS* | 4.95 | 1.67 | 0.00 | 0.81 | 0.37 | 6.69 | 1.02 | 0.00 | 0.38 | 0.19 |
| ISIS* | 5.81 | 2.06 | 0.00 | 0.69 | 0.30 | 7.04 | 1.21 | 0.00 | 0.37 | 0.17 |
| KAZZ | 5.30 | 1.09 | 0.00 | 0.64 | 0.31 | 7.18 | 0.48 | 0.00 | 0.31 | 0.16 |

Table 4: The selection results for Example 3.

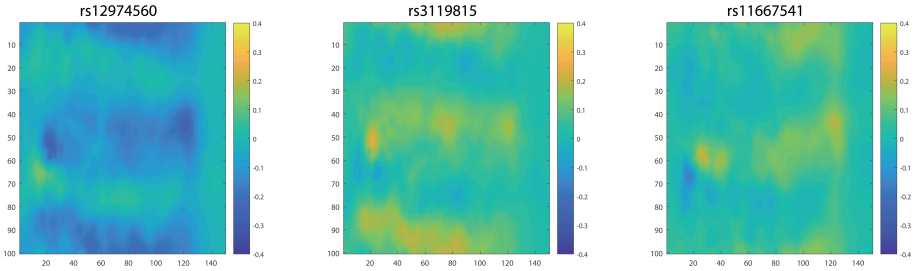
for the left hippocampal surface, and rs3119815, rs12974560 and rs11667541 for the right hippocampal surface. Figure 1 shows the plots for the proposed estimates corresponding to the 6 selected SNPs. The SNP rs3119815 on gene LSM14A is chosen both for the left and right hippocampal surfaces by the proposed procedure. Chowriappa et al. (2013) found that LSM14A is a co-expressed gene in the incipient Alzheimer's disease samples since to some extent, it could be responsible for the phenotypic differences through the progression of Alzheimer's disease. The SNP rs12974560 is on gene LAIR1 that is associated with the left hippocampal surface by the proposed method. Wirz et al. (2013) found that LAIR1 is a significant upregulated gene during the development of $\beta$-amyloid protein pathology. The $\beta$-amyloid protein (A$\beta$) pathology plays one of the most important roles in Alzheimer's disease pathology and prevention (Sadigh-Eteghad et al., 2015). Therefore, LAIR1 may also be a key in the progression of Alzheimer's disease. The SNP rs12974560 on gene ARHGEF18 is selected to be associated with the right hippocampal surface by the proposed procedure. It is identified by Sánchez-Valle (2017) as a significantly differentially expressed gene in lung cancer, malignant glioblastomas and Alzheimer's disease, which are three diseases of the most challenging public health conditions worldwide. However, the genes selected by the SIS* and ISIS* methods have no overlap with those selected by our method, the main reason is that the SIS* and ISIS* have excluded the significant SNPs at the first screening step. This is in accordance with the simulation results, that is, the SIS* and ISIS* may overlook some important predictors.

| Hippocampal surface | SNP | SRLS | SIS* | ISIS* |
|---|---|---|---|---|
| left | rs8105522 | | | ✓ |
| | rs3119815 | ✓ | | |
| | rs8103479 | | ✓ | ✓ |
| | rs266875 | ✓ | | |
| | rs12610273 | ✓ | | |
| right | rs8105522 | | | ✓ |
| | rs12974560 | ✓ | | |
| | rs3119815 | ✓ | | |
| | rs11667541 | ✓ | | |
| | rs10415851 | | ✓ | ✓ |

Table 5: Real data analysis: the selected SNPs associated with the left and right hippocampal surfaces, respectively.




(a)




(b)

Figure 1: Alzheimer's Disease Neuroimaging Initiative data: The panels (a) and (b) show the plots for our proposed estimates corresponding to the 3 selected SNPs associated with the left and right hippocampal surfaces, respectively.

## 6. Smoothing estimator and subregions detection

Potentially there may exist intrinsic information among the entries of the matrix response, e.g., multiple piecewise smooth regions with unknown edges and jumps. In addition, it should be of interest to find out which areas of the matrix response would be mostly affected by the predictors. In this section, we consider a smoothing estimator of $B_j^\star$ and the support recovery procedure for the matrix response. Let $\hat{\mathbb{B}} = (\hat{B}_1^\top, \ldots, \hat{B}_{d_n}^\top)^\top$ be the estimate obtained by the SRLS procedure, and $\hat{B}_j = (\hat{b}_{j,sk})$. Define $\mathcal{N}_{sk} = \{(s',k') : \|(s',k') - (s,k)\|_2 < r\}$ as a neighbourhood of point $(s,k)$, where $r$ denotes a radius. Note that $\hat{B}_j$ is a consistent estimate for $B_j^\star$. Borrowing the idea from Zhu et al. (2014), we obtain an estimator of $b_{j,sk}^\star$ by

$$\bar{b}_{j,sk} = \arg\min_b \sum_{(s',k')\in\mathcal{N}_{sk}} \mathcal{K}_1(\|(s',k') - (s,k)\|_2/r)\mathcal{K}_2(|\hat{b}_{j,s'k'} - \hat{b}_{j,sk}|/h)\big(\hat{b}_{j,s'k'} - b\big)^2, \quad (9)$$

where $\mathcal{K}_1(z)$ and $\mathcal{K}_2(z)$ are two smoothing functions, and $h$ is a bandwidth. The functions $\mathcal{K}_l(z)$ $(l = 1, 2)$ decrease on $[0, \infty)$, and satisfy that $0 \le \mathcal{K}_l(z) \le 1$, $\mathcal{K}_l(0) = 1$, and $\mathcal{K}_l(z) \to 0$ as $z \to \infty$. It is worth noting that the role of $\mathcal{K}_2(z)$ is used to measure the similarity between $b_{j,s'k'}$ and $b_{j,sk}$. Precisely, $\mathcal{K}_2(|b_{j,s'k'} - b_{j,sk}|/h) \approx 1$ if $b_{j,s'k'}$ is close to $b_{j,sk}$; otherwise, $\mathcal{K}_2(|b_{j,s'k'} - b_{j,sk}|/h) \approx 0$.

A direct calculation of (9) yields

$$\bar{b}_{j,sk} = \sum_{(s',k')\in\mathcal{N}_{sk}} w(\hat{b}_{j,s'k'}, \hat{b}_{j,sk}; r, h)\hat{b}_{j,s'k'},$$

where

$$w(v, u; r, h) = \frac{\mathcal{K}_1(\|(s',k') - (s,k)\|_2/r)\mathcal{K}_2(|v-u|/h)}{\sum_{s'k'\in\mathcal{N}_{sk}} \mathcal{K}_1(\|(s',k') - (s,k)\|_2/r)\mathcal{K}_h(|v-u|/h)}.$$

To determine the radius $r$, we propose an adaptive method (Zhu et al., 2014). Consider the following sequence: $r_m = 1.2^m$ $(1 \le m \le M)$, where $M$ is a prespecified positive integer. Let $\mathcal{N}_{sk}^{[m]} = \{(s',k') : \|(s',k') - (s,k)\|_2 < r_m\}$. For each $j \in \hat{\mathcal{M}}_{\hat{\tau}}$, we estimate $B_j^\star$ as follows.

**Algorithm 2** (*Smoothing Estimator)*

*Step 1.* Let $m = 0$, and input $A$, $M_0$, $M$ and the initial value $\bar{B}_j^{[0]} = (\bar{b}_{j,sk}^{[0]})$ with $\bar{b}_{j,sk}^{[0]} = \hat{b}_{j,sk}$, where $A$ is a tolerance parameter and $M_0$ is a positive integer.

*Step 2.* Calculate $\bar{b}_j^{[m+1]}$ by

$$\bar{b}_{j,sk}^{[m+1]} = \sum_{(s',k')\in\mathcal{N}_{sk}^{[m+1]}} w(\bar{b}_{j,s'k'}^{[m]}, \bar{b}_{j,sk}^{[m]}; r_{m+1}, h)\hat{b}_{j,s'k'}.$$

*Step 3.* If $m \le M_0$, increase $m$ to $m+1$, and return to *Step 2*. Otherwise, go to *Step 4*.

*Step 4.* If $m < M$ and $|\bar{b}_{j,sk}^{[m+1]} - \bar{b}_{j,sk}^{[M_0+1]}| < A$, increase $m$ to $m+1$ and return to *Step 2*.

Otherwise, stop the iterative procedure, and set $\bar{b}_{j,sk} = \bar{b}_{j,sk}^{[m]}$.

Here we set $M = 10$, $M_0 = 5$, $A = 10^{-3}$, and $h = \varpi n^{-1/5}$ with $\varpi = 1/2$. More discussions about the settings of these parameters can be found in Zhu et al. (2014). Denote $\bar{B}_j = (\bar{b}_{j,sk})$ for each $j \in \hat{\mathcal{M}}_{\hat{\tau}}$.

In practice, it may be also of interest to find out which areas of the matrix response would be mostly affected by the predictors, that is, the active entries of $Y_i$. Here, we take $Y_{i,sk}$ as an inactive entry of $Y_i$ if $E(Y_{i,sk}|X_i) = 0$ almost surely. Let $\Psi$ be the active set, then the inactive set is $\Psi^c = \{(s,k) : E(Y_{i,sk}|X_i) = 0 \text{ almost surely}\}$. For the analysis, we assume that the matrix response $Y_i$ is sparse, that is, $\mathrm{Card}(\Psi) \ll pq$. Note that under model (1), $\Psi$ can also be denoted by

$$\Psi = \left\{ (s,k) : E\left\{ \left( \sum_{j=1}^{\tau^\star} x_{ij} b^\star_{j,sk} \right)^2 \right\} \neq 0 \right\}.$$

Thus, we can estimate $\Psi$ by

$$\bar{\Psi}_\varsigma = \left\{ (s,k) : \frac{1}{n} \sum_{i=1}^{n} \bar{Y}^2_{i,sk} > \varsigma \right\},$$

where $\varsigma$ is a tuning parameter and $\bar{Y}_{i,sk} = \sum_{j=1}^{\hat{\tau}} x_{ij} \bar{b}_{j,sk}$. Define

$$\textsc{ebic}_\varsigma = \frac{1}{n} \sum_{i=1}^{n} \sum_{(s,k) \in \bar{\Psi}_\varsigma} (\bar{Y}_{i,sk} - Y_{i,sk})^2 + \mathrm{Card}(\bar{\Psi}_\varsigma) \frac{\tilde{c}_n \log(n)}{n}, \tag{10}$$

where $\tilde{c}_n$ is chosen as $\log\log(pq)$. We choose $\hat{\varsigma}$ to minimize $\textsc{ebic}_\varsigma$, and let $\bar{\Psi} = \bar{\Psi}_{\hat{\varsigma}}$. Since this procedure depends on the smoothing estimate $\bar{B}_j$, we refer to it as S-SRLS.

We next consider simulation studies to evaluate the finite performance of the regularized smoothing estimator. The total number of predictors is $d_n = 1000$. We consider $(p,q) = (50,50)$. Let $X_i = (x_{ij}) \in \mathbb{R}^{d_n}$ follow a multivariate normal distribution with mean 0 and covariance matrix $(\sigma_{kl})_{d_n \times d_n}$ with $\sigma_{kl} = 0.5^{|k-l|}$. The errors $\mathrm{Vec}(U_i)$ are independently generated from a multivariate normal distribution with mean 0, and the correlations between $e_{i,sk}$ and $e_{i,s'k'}$ are $0.2^{|s-s'|+|k-k'|}$ for $1 \leq s, s', k, k' \leq 50$. The responses $Y_i$ $(i = 1, 2, \ldots, n)$ are generated by $Y_i = x_{i1} B^\star_1 + x_{i2} B^\star_2 + x_{i3} B^\star_3 + U_i$, where $B^\star_j$ $(1 \leq j \leq 3)$ are depicted in Figure 2 and $B^\star_j$ are set as 0 for $j > 3$. The true values in Figure 2 are displayed with navy blue, blue, green, orange and deep red colors representing $0, 0.2, 0.4, 0.6$ and $0.8$, respectively. The support of $Y_i$ is given in Figure 3. The smoothing functions are set as $\mathcal{K}_1(z) = \max\{0, 1 - z\}$ and $\mathcal{K}_2(z) = \exp(-z^2)$. To evaluate the accuracy of the support recovery, we employ the similarity measure:

$$S(\bar{\Psi}, \Psi) = \frac{\mathrm{Card}(\bar{\Psi} \cap \Psi)}{\sqrt{\mathrm{Card}(\bar{\Psi})\mathrm{Card}(\Psi)}}.$$

For comparison, we also consider a naive method (N-SRLS, for short):

$$\hat{\Psi}_\varsigma = \left\{ (s,k) : \frac{1}{n} \sum_{i=1}^{n} \hat{Y}^2_{i,sk} > \varsigma \right\},$$

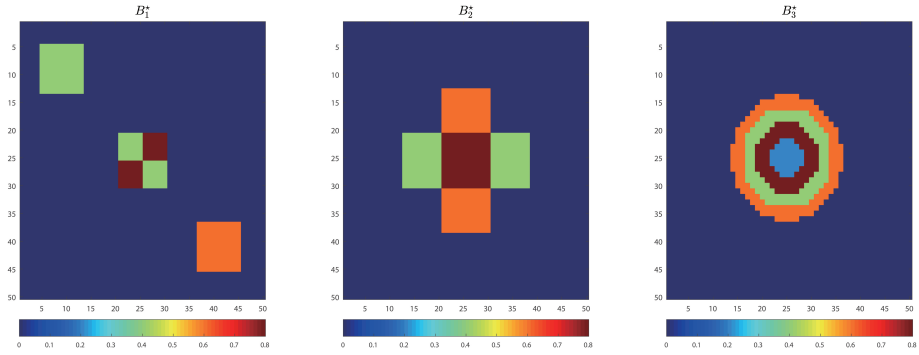



Figure 2: The true coefficient matrices $B_j^\star$. The true values are displayed with navy blue, blue, green, orange and deep red colors representing $0, 0.2, 0.4, 0.6$ and $0.8$, respectively.

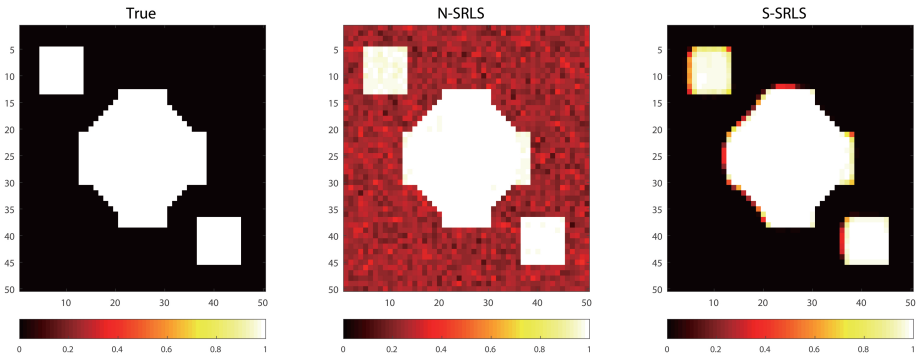



Figure 3: Heat maps of the nonzeros identified out of 100 replications. White indicates 100 1s identified out of 100 replications; black indicates 100 0s identified out of 100 replications.

where $\hat{Y}_{i,sk} = \sum_{j=1}^{\hat{\tau}} x_{ij}\hat{b}_{j,sk}$ and $\varsigma$ is determined by the $\text{EBIC}_\varsigma$ as in (10). The comparison results reported in Table 6 are based on 100 replications with sample sizes $n = 100$ and 200. We observe that the values of $S(\bar{\Psi}, \Psi)$ of the S-SRLS are close to 1 and outperform those of the N-SRLS. To better illustrate the recovery performance, we depict in Figure 3 the heat maps of the nonzeros identified out of 100 replications when $n = 100$. Figure 3 suggests that when $n = 100$, the nonzero regions recovered by the S-SRLS are very close to those of $Y_i$, and the N-SRLS method tends to retain many zero entries of the matrix response in $\hat{\Psi}_{\hat{\varsigma}}$. For theoretical justification of the S-SRLS, it is beyond the scope of the paper and we leave it for future research.

| $n$ | N-SRLS | S-SRLS |
|---|---|---|
| 100 | 0.76(0.01) | 0.98(0.01) |
| 200 | 0.94(0.01) | 0.99(0.00) |

Table 6: The support recovery results and standard deviations (in parentheses).

## Acknowledgments

The authors thank the Action Editor, Professor Genevera Allen, an Associate Editor and two reviewers for their insightful comments and suggestions that greatly improved the quality of the paper. Dr. Hao's research was supported by the National Natural Science Foundation of China (No. 11901087) and the Program for Young Excellent Talents, UIBE (No. 19YQ15). Dr. Kong's work was partially supported by a Discovery Grant and Discovery Accelerator Supplement from Natural Science and Engineering Research Council of Canada (RGPIN-2017-06538 and RGPAS-2017-507944). Dr. Qu's research was partly supported by grants from the National Natural Science Foundation of China (No. 12001219), the Hubei Natural Science Foundation of China (No. 2018CFB256) and the Fundamental Research Funds for the Central Universities in CCNU. Dr. Sun's research was partly supported by grants from the National Natural Science Foundation of China (Nos. 11771431 and 11690015) and Key Laboratory of Random Structures and Data Science, Chinese Academy of Sciences (No. 2008DP173182).

## Appendices

The Appendices include all the technical proofs, the real data usage agreement, the detailed data preprocessing steps and additional simulation studies.

## Appendix A: Proofs of Theorems 1-6

**Proof of Proposition 1.** It suffices to consider $\|D_j\|_F > 0$ for all $j = 1, \cdots, d_n$. Let $\mathcal{M} = \{j : \hat{B}_j \neq 0\}$. The objective function is given by $\sum_{j\notin\mathcal{M}} \|D_j\|_F^2 + \sum_{j\in\mathcal{M}} \|\hat{B}_j - D_j\|_F^2$. Note that by selecting $\hat{B}_j = D_j$ ($j \in \mathcal{M}$), the objective function becomes $\sum_{j\notin\mathcal{M}} \|D_j\|_F^2$. Thus, to minimize the objective function, $\mathcal{M}$ must correspond to the indices of the largest $\tau$ values of $\|D_j\|_F^2$.

**Proof of Theorem 1.** To prove part (i), we first show that the linear search criterion (7) holds. By the definitions of $\mathcal{Q}_\lambda(\mathbb{B}|\mathbb{B}^{[l]})$ and $\mathbb{B}^{[l+1]}$, we have

$$\mathcal{Q}_{\lambda^{[l]}}(\mathbb{B}^{[l]}|\mathbb{B}^{[l]}) = \frac{1}{2n}\|\mathbb{Y} - \mathbb{X}\mathbb{B}^{[l]}\|_F^2 \geq \mathcal{Q}_{\lambda^{[l]}}(\mathbb{B}^{[l+1]}|\mathbb{B}^{[l]}).$$

Thus,

$$\begin{aligned}
&\frac{1}{2n}\|\mathbb{Y}-\mathbb{X}\mathbb{B}^{[l]}\|_F^2\\
\geq&\frac{1}{2n}\|\mathbb{Y}-\mathbb{X}\mathbb{B}^{[l]}\|_F^2-\frac{1}{n}\big\langle\mathbb{X}^\top(\mathbb{Y}-\mathbb{X}\mathbb{B}^{[l]}),\mathbb{B}^{[l+1]}-\mathbb{B}^{[l]}\big\rangle_F+\frac{\lambda^{[l]}}{2}\|\mathbb{B}^{[l+1]}-\mathbb{B}^{[l]}\|_F^2\\
=&\frac{1}{2n}\|\mathbb{Y}-\mathbb{X}\mathbb{B}^{[l]}\|_F^2-\frac{1}{n}\big\langle\mathbb{X}^\top(\mathbb{Y}-\mathbb{X}\mathbb{B}^{[l]}),\mathbb{B}^{[l+1]}-\mathbb{B}^{[l]}\big\rangle_F+\frac{\lambda^{[l]}(1-\varrho)}{2}\|\mathbb{B}^{[l+1]}-\mathbb{B}^{[l]}\|_F^2\\
&+\frac{\lambda^{[l]}\varrho}{2}\|\mathbb{B}^{[l+1]}-\mathbb{B}^{[l]}\|_F^2\\
=&\mathcal{Q}_{(1-\varrho)\lambda^{[l]}}(\mathbb{B}^{[l+1]}|\mathbb{B}^{[l]})+\frac{\varrho\lambda^{[l]}}{2}\|\mathbb{B}^{[l+1]}-\mathbb{B}^{[l]}\|_F^2. \qquad \text{(A.1)}
\end{aligned}$$

It follows from $\lambda^{[l]}\geq\phi/(1-\varrho)$ that

$$\mathcal{Q}_{(1-\varrho)\lambda^{[l]}}(\mathbb{B}^{[l+1]}|\mathbb{B}^{[l]})\geq\frac{1}{2n}\|\mathbb{Y}-\mathbb{X}\mathbb{B}^{[l+1]}\|_F^2. \qquad \text{(A.2)}$$

Using (A.1) and (A.2), we obtain

$$\frac{1}{2n}\|\mathbb{Y}-\mathbb{X}\mathbb{B}^{[l]}\|_F^2-\frac{1}{2n}\|\mathbb{Y}-\mathbb{X}\mathbb{B}^{[l+1]}\|_F^2\geq\frac{\varrho\lambda^{[l]}}{2}\|\mathbb{B}^{[l+1]}-\mathbb{B}^{[l]}\|_F^2\geq 0, \qquad \text{(A.3)}$$

which means that the linear search criterion (7) holds. Hence the non-increasing sequence $n^{-1}\|\mathbb{Y}-\mathbb{X}\mathbb{B}^{[l]}\|_F^2$ is convergent and bounded from below.

We now show that $\mathbb{B}^{[l]}$ is convergent. When $\lambda^{[l]}\to\infty$ as $l\to\infty$, the result is trivial. Next, we assume that $\{\lambda^{[l]}\}$ is bounded. Let $\lambda^*$ be a limit point of $\lambda^{[l]}$, that is, there exists a subsequence $\mathcal{L}$ such that $\lambda^{[l]}\to\lambda^*$ for $l\in\mathcal{L}$. For each $l\in\mathcal{L}$, denote $\mathcal{M}^{[l]}=\{j:\|B_j^{[l]}\|_F\neq 0\}$ as the support of $\mathbb{B}^{[l]}$. By (A.3) and the fact that $n^{-1}\|\mathbb{Y}-\mathbb{X}\mathbb{B}^{[l]}\|_F^2$ is convergent, we know that $\|\mathbb{B}^{[l+1]}-\mathbb{B}^{[l]}\|_F^2\to 0$ as $l\in\mathcal{L}\to\infty$, which implies that $\mathcal{M}^{[l]}$ is convergent. Also, $\mathcal{M}^{[l]}$ is a discrete sequence, and hence there exists an $l^*\in\mathcal{L}$ such that $\mathcal{M}^{[l]}=\mathcal{M}^{[l^*]}$ for all $l\in\mathcal{L}\geq l^*$. Thus, Algorithm 1 becomes a gradient descent algorithm on the space $\mathcal{M}^{[l]}$ for all $l\in\mathcal{L}\geq l^*$. Since a gradient descent algorithm for minimizing $L_2$-loss function over a closed convex set yields a sequence of iterates that converge (Nesterov, 2004), we conclude that the subsequence $\mathbb{B}^{[l]}$ ($l\in\mathcal{L}$) converges to a point $\mathbb{B}^*$, that is, $\mathbb{B}^*\in\mathbb{H}_\tau\big\{\mathbb{B}^*+(n\lambda^*)^{-1}\mathbb{X}^\top(\mathbb{Y}-\mathbb{X}\mathbb{B}^*)\big\}$. This completes the proof.

Next we show part (ii). It follows from (A.3) that

$$\frac{1}{2n}\sum_{l=0}^{L}\Big(\|\mathbb{Y}-\mathbb{X}\mathbb{B}^{[l]}\|_F^2-\|\mathbb{Y}-\mathbb{X}\mathbb{B}^{[l+1]}\|_F^2\Big)\geq\sum_{l=0}^{L}\frac{\varrho\lambda^{[l]}}{2}\|\mathbb{B}^{[l+1]}-\mathbb{B}^{[l]}\|_F^2,$$

which implies

$$\frac{1}{2n}\|\mathbb{Y}-\mathbb{X}\mathbb{B}^{[0]}\|_F^2-\frac{1}{2n}\|\mathbb{Y}-\mathbb{X}\mathbb{B}^{[L+1]}\|_F^2\geq\frac{L\varrho\phi}{2(1-\varrho)}\min_{l=0,1,\cdots,L}\|\mathbb{B}^{[l+1]}-\mathbb{B}^{[l]}\|_F^2.$$

That is,

$$\min_{l=0,1,\cdots,L}\|\mathbb{B}^{[l+1]}-\mathbb{B}^{[l]}\|_F^2 \le \frac{1-\varrho}{\varrho\phi nL}\big(\|\mathbb{Y}-\mathbb{X}\mathbb{B}^{[0]}\|_F^2-\|\mathbb{Y}-\mathbb{X}\mathbb{B}^{[L+1]}\|_F^2\big)$$
$$\le \frac{1-\varrho}{\varrho\phi nL}\big(\|\mathbb{Y}-\mathbb{X}\mathbb{B}^{[0]}\|_F^2-\|\mathbb{Y}-\mathbb{X}\tilde{\mathbb{B}}\|_F^2\big).$$

This completes the proof.

**Proof of Theorem 2.** Let $\mathcal{M}_\tau^- = \{\mathcal{M} : \mathcal{M}^\star \not\subset \mathcal{M},\ \mathrm{card}(\mathcal{M}) \le \tau\}$ be the collection of under-fitted models. Define $Q_n(\mathbb{B}) = n^{-1}\|\mathbb{Y}-\mathbb{X}\mathbb{B}\|_F^2$, and $Q(\mathbb{B}) = E\big(n^{-1}\|\mathbb{Y}-\mathbb{X}\mathbb{B}\|_F^2\big)$. Denote $\|\cdot\|_2$ as the standard Euclidean norm. Let $\mathbb{B}_{sk} = (b_{1,sk},\ldots,b_{d_n,sk})^\top$ and $\mathbb{B}_{\mathcal{M},sk} = (b_{j,sk}, j\in\mathcal{M})^\top$ be the subvector of $\mathbb{B}_{sk}$ with $\mathcal{M}\subset\{1,\ldots,d_n\}$. To prove Theorem 2, we need show that

$$pr\Big(\max_{\mathbb{B}\in\mathcal{M}_\tau^+} Q_n(\mathbb{B}) \ge \min_{\mathbb{B}\in\mathcal{M}_-^\tau} Q_n(\mathbb{B})\Big) \le c_2\tau pq d_n^\tau \exp(-c_1 n^{1-\delta_2-2\delta_3}) + c_2 n d_n \exp(-\eta_1 n^\upsilon).$$

Define $\tilde{\mathcal{M}} = \mathcal{M}\cup\mathcal{M}^\star \in \mathcal{M}_+^{2\tau}$, where $\mathcal{M}\in\mathcal{M}_-^\tau$. For any $\mathbb{B}_{\tilde{\mathcal{M}}}$ such that $\|\mathbb{B}_{\tilde{\mathcal{M}}}-\mathbb{B}^\star_{\tilde{\mathcal{M}}}\|_F = \sqrt{(pq)}\omega_2 n^{-\delta_3}$, we have

$$Q_n(\mathbb{B}_{\tilde{\mathcal{M}}}) - Q_n(\mathbb{B}^\star_{\tilde{\mathcal{M}}}) = \frac{1}{n}\sum_{i=1}^n\sum_{s=1}^p\Big\{\sum_{k=1}^q (Y_{i,sk}-X_{i\tilde{\mathcal{M}}}^\top\mathbb{B}_{\tilde{\mathcal{M}},sk})^2 - (Y_{i,sk}-X_{i\tilde{\mathcal{M}}}^\top\mathbb{B}^\star_{\tilde{\mathcal{M}},sk})^2\Big\}$$
$$=\frac{1}{n}\sum_{i=1}^n\sum_{s=1}^p\sum_{k=1}^q\Big\{(\mathbb{B}_{\tilde{\mathcal{M}},sk}-\mathbb{B}^\star_{\tilde{\mathcal{M}},sk})^\top X_{i\tilde{\mathcal{M}}}X_{i\tilde{\mathcal{M}}}^\top(\mathbb{B}_{\tilde{\mathcal{M}},sk}-\mathbb{B}^\star_{\tilde{\mathcal{M}},sk})$$
$$-2(Y_{i,sk}-X_{i\tilde{\mathcal{M}}}^\top\mathbb{B}^\star_{\tilde{\mathcal{M}},sk})X_{i\tilde{\mathcal{M}}}^\top(\mathbb{B}_{\tilde{\mathcal{M}},sk}-\mathbb{B}^\star_{\tilde{\mathcal{M}},sk})\Big\}$$
$$\ge\lambda_{\min}\omega_1^2 pqn^{-2\delta_3} - 2\sum_{s=1}^p\sum_{k=1}^q\|\frac{1}{n}\sum_{i=1}^n(Y_{i,sk}-X_{i\tilde{\mathcal{M}}}^\top\mathbb{B}^\star_{\tilde{\mathcal{M}},sk})X_{i\tilde{\mathcal{M}}}\|_2\|\mathbb{B}_{\tilde{\mathcal{M}},sk}-\mathbb{B}^\star_{\tilde{\mathcal{M}},sk}\|_2, \quad (A.4)$$

where $\lambda_{\min}$ denotes the smallest eigenvalue of $n^{-1}\sum_{i=1}^n X_{i\tilde{\mathcal{M}}}X_{i\tilde{\mathcal{M}}}^\top$. Since $\sum_{s=1}^p\sum_{k=1}^q\|\mathbb{B}_{\tilde{\mathcal{M}},sk}-\mathbb{B}^\star_{\tilde{\mathcal{M}},sk}\|_2 \le (pq)^{1/2}\|\mathbb{B}_{\tilde{\mathcal{M}}}-\mathbb{B}^\star_{\tilde{\mathcal{M}}}\|_F$, we obtain

$$pr\big(Q_n(\mathbb{B})\le Q_n(\mathbb{B}^\star)\big)$$
$$\le pr\Big(\lambda_{min}\omega_2^2 pqn^{-2\delta_3} - 2\omega_2 pqn^{-\delta_3}\max_{p,q}\|\frac{1}{n}\sum_{i=1}^n(Y_{i,sk}-X_{i\tilde{\mathcal{M}}}^\top\mathbb{B}^\star_{\tilde{\mathcal{M}},sk})X_{i\tilde{\mathcal{M}}}\|_2\le 0\Big)$$
$$=pr\Big(\max_{p,q}\|\frac{1}{n}\sum_{i=1}^n(Y_{i,sk}-X_{i\tilde{\mathcal{M}}}^\top\mathbb{B}^\star_{\tilde{\mathcal{M}},sk})X_{i\tilde{\mathcal{M}}}^\top\|_2\ge\frac{\lambda_{min}}{2}\omega_2 n^{-\delta_3}\Big)$$
$$\le\sum_{s=1}^p\sum_{k=1}^q\sum_{j\in\tilde{\mathcal{M}}}pr\Big(\big|\sum_{i=1}^n(Y_{i,sk}-X_{i\tilde{\mathcal{M}}}^\top\mathbb{B}^\star_{\tilde{\mathcal{M}},sk})x_{ij}\big|\ge\frac{\lambda_{min}\omega_2}{\sqrt{(8\tau)}}n^{1-\delta_3}\tau^{-1/2},|x_{ij}|<n^v\Big)$$
$$+d_n npr\Big(|x_{ij}|\ge n^v\Big). \quad (A.5)$$

Then it follows from Conditions (C4) and (C5) and Lemma A.1 of Xu and Chen (2014) that there exists a constant $c_1 > 0$ such that

$$\begin{aligned} &pr\Big(\big|\sum_{i=1}^{n}(Y_{i,sk} - X_{i\tilde{\mathcal{M}}}^{\top}\mathbb{B}^{\star}_{\tilde{M},sk})x_{ij}\big| \geq \frac{\lambda_{min}\omega_2}{\sqrt{(8\tau)}}n^{1-\delta_3}\tau^{-1/2}\Big) \\ \leq& 2\exp(-c_1 n^{1-\delta_2-2\delta_3-2\upsilon}) \end{aligned}$$

on the event $\{|x_{ij}| < n^{\upsilon}, 1 \leq i \leq n, 1 \leq j \leq d_n\}$. This, together with (A.5) and Condition (C6), implies

$$\begin{aligned} pr\Big(\min_{\mathbb{B}\in\mathcal{M}_{-}^{\tau}} Q_n(\mathbb{B}) \leq Q_n(\mathbb{B}^{\star})\Big) &\leq \sum_{\mathcal{M}\in\mathcal{M}_{-}^{\tau}} pr\Big(Q_n(\mathbb{B}) \leq Q_n(\mathbb{B}^{\star})\Big) \\ &\leq 2\tau d_n^{\tau} pq\exp(-c_1 n^{1-\delta_2-2\delta_3-2\upsilon}) + 2nd_n\eta_2\exp(-\eta_1 n^{\upsilon}). \end{aligned}$$

Since $Q_n(\mathbb{B})$ is convex, the result holds for $\|\mathbb{B}_{\tilde{\mathcal{M}}} - \mathbb{B}^{\star}_{\tilde{\mathcal{M}}}\|_F \geq \sqrt{(pq)}\omega_2 n^{-\delta_3}$. For any $\mathcal{M} \in \mathcal{M}_{-}^{\tau}$, take $\mathbb{B}_{\tilde{\mathcal{M}}}$ as the augmented matrix of $\mathbb{B}_{\mathcal{M}}$ with the component in $\tilde{\mathcal{M}} - \mathcal{M}$ being zero. Then, Condition (C4) implies that $\|\mathbb{B}_{\tilde{\mathcal{M}}} - \mathbb{B}^{\star}_{\tilde{\mathcal{M}}}\|_F \geq \|\mathbb{B}^{\star}_{\mathcal{M}^{\star}-\mathcal{M}}\|_F \geq w_1 n^{-\delta_3}(pq)^{1/2}$. Therefore, we have that there exists $c_2 > 0$ such that

$$\begin{aligned} pr\Big(\min_{\mathbb{B}\in\mathcal{M}_{-}^{\tau}} Q_n(\mathbb{B}) \leq \max_{\mathbb{B}\in\mathcal{M}_{\tau}^{+}} Q_n(\mathbb{B})\Big) &\leq pr\Big(\min_{\mathbb{B}\in\mathcal{M}_{-}^{\tau}} Q_n(\mathbb{B}) \leq Q_n(\mathbb{B}^{\star})\Big) \\ &\leq c_2\tau pqd_n^{\tau}\exp(-c_1 n^{1-\delta_2-2\delta_3-2\upsilon}) + c_2 nd_n\exp(-\eta_1 n^{\upsilon}). \end{aligned}$$

This completes the proof.

**Proof of Theorem 3.** Denote $\gamma = \min_{j\in\mathcal{M}^{\star}}(pq)^{-1/2}\|B_j\|_F > 0$. Let $\mathcal{M} \in \mathcal{M}_{-}^{\tau}$ be a given under-fitted model, and define $\tilde{\mathcal{M}} = \mathcal{M} \cup \mathcal{M}^{\star} \in \mathcal{M}_{+}^{2\tau}$. Without loss of generality, we may assume $\|\mathbb{B}_{\tilde{\mathcal{M}}} - \mathbb{B}^{\star}_{\tilde{\mathcal{M}}}\|_F < \sqrt{(pq)}\gamma$. Thus, $\|\mathbb{B}_{\mathcal{M}} - \mathbb{B}^{\star}_{\mathcal{M}}\|_F > \sqrt{(pq)}\gamma$ and there exists $\tilde{\mathbb{B}} = (\tilde{B}_1^{\top}, \ldots, \tilde{B}_{d_n}^{\top})^{\top}$ with $\|\tilde{\mathbb{B}}_{\tilde{\mathcal{M}}} - \mathbb{B}^{\star}_{\tilde{\mathcal{M}}}\|_F = \sqrt{(pq)}\gamma$ such that

$$\sum_{i=1}^{n}\Big(\|Y_i - \sum_{j\in\mathcal{M}} x_{ij}B_j\|_F^2 - \|Y_i - \sum_{j\in\tilde{\mathcal{M}}} x_{ij}\tilde{B}_j\|_F^2\Big) \geq 0$$

because of the convexity of the norm $\|\mathbb{B}\|_F^2$. This gives

$$\begin{aligned} &\sum_{i=1}^{n}\Big(\|Y_i - \sum_{j\in\mathcal{M}} x_{ij}B_j\|_F^2 - \|Y_i - \sum_{j\in\tilde{\mathcal{M}}} x_{ij}B_j\|_F^2\Big) \\ \geq& \sum_{i=1}^{n}\Big(\|U_i - \sum_{j\in\tilde{\mathcal{M}}} x_{ij}(\tilde{B}_j - B_j^{\star})\|_F^2 - \|U_i - \sum_{j\in\tilde{\mathcal{M}}} x_{ij}(B_j - B_j^{\star})\|_F^2\Big) \\ \geq& \lambda_{\min} n\|\tilde{\mathbb{B}}_{\tilde{\mathcal{M}}} - \mathbb{B}^{\star}_{\tilde{\mathcal{M}}}\|_F^2 - \lambda_{\max} n\|\mathbb{B}_{\tilde{\mathcal{M}}} - \mathbb{B}^{\star}_{\tilde{\mathcal{M}}}\|_F^2 - 2\sum_{i=1}^{n}\sum_{s=1}^{p}\sum_{k=1}^{q}\sum_{j\in\tilde{\mathcal{M}}} x_{ij}(\tilde{b}_{j,sk} - b^{\star}_{j,sk})e_{i,sk} \\ &+ 2\sum_{i=1}^{n}\sum_{s=1}^{p}\sum_{k=1}^{q}\sum_{j\in\tilde{\mathcal{M}}} x_{ij}(b_{j,sk} - b^{\star}_{j,sk})e_{i,sk}, \end{aligned} \tag{A.6}$$

where $\lambda_{\max}$ represents the largest eigenvalue of $n^{-1}\sum_{i=1}^{n} X_{i\tilde{\mathcal{M}}}X_{i\tilde{\mathcal{M}}}^{\top}$. Note that $\|\tilde{\mathbb{B}}_{\tilde{\mathcal{M}}} - \mathbb{B}^{\star}_{\tilde{\mathcal{M}}}\|_F^2 > pq\gamma^2/2 > 0$. Therefore, the first term on the right hand side of (A.6) is positive and bounded away from zero uniformly over $\mathcal{M} \in \mathcal{M}_{-}^{\tau}$. By arguments similar to those in the proofs of (A.4) and (A.5), we can show that uniformly for all $\tilde{\mathcal{M}} \in \mathcal{M}_{+}^{2\tau}$, $(pq)^{-1/2}\|\mathbb{B}_{\tilde{\mathcal{M}}} - \mathbb{B}^{\star}_{\tilde{\mathcal{M}}}\|_F = O_p(n^{-\delta_3})$. In addition, using the Cauchy-Schwarz and Chebyshev inequalities, we have that for any positive $\epsilon$ and $\upsilon$,

$$
\begin{aligned}
&pr\Big(\Big|\frac{1}{npq}\sum_{i=1}^{n}\sum_{s=1}^{p}\sum_{k=1}^{q}\sum_{j\in\tilde{\mathcal{M}}} x_{ij}(\tilde{b}_{j,sk} - b^{\star}_{j,sk})e_{i,sk}\Big| > \epsilon\Big)\\
\leq& pr\Big(\frac{1}{n^{1-\upsilon}pq}\sum_{i=1}^{n}\sum_{s=1}^{p}\sum_{k=1}^{q}\sum_{j\in\tilde{\mathcal{M}}} |\tilde{b}_{j,sk} - b^{\star}_{j,sk}||e_{i,sk}| > \epsilon\Big) + \eta_2\exp(-\eta_1 n^{\upsilon})\\
\leq& pr\Big(\frac{\sqrt{(2\tau)}}{n^{1-\upsilon}pq}\|\tilde{\mathbb{B}}_{\tilde{\mathcal{M}}} - \mathbb{B}^{\star}_{\tilde{\mathcal{M}}}\|_F\sum_{i=1}^{n}\|U_i\|_F > \epsilon\Big) + \eta_2\exp(-\eta_1 n^{\upsilon})\\
\leq& \frac{2\tau\gamma^2}{n^{1-2\upsilon}\epsilon^2}\Big[\frac{1}{npq}\sum_{i=1}^{n}E\big(\|U_i\|_F^2\big)\Big] + \eta_2\exp(-\eta_1 n^{\upsilon}),
\end{aligned}
$$

which converges to zero as $n \to \infty$ under Condition (C7) and $\tau = o(n^{\delta_2})$. That is, the third term on the right hand side of (A.6) is $o_p(npq)$. Similarly, we can show that the last term on the right hand side of (A.6) is also $o_p(npq)$. Thus, there exists a constant $\nu_2 > 0$ not depending on $\mathcal{M} \in \mathcal{M}_{-}^{\tau}$ such that with probability tending to 1

$$
\sum_{i=1}^{n}\Big(\|Y_i - \sum_{j\in\mathcal{M}} x_{ij}B_j\|_F^2 - \|Y_i - \sum_{j\in\tilde{\mathcal{M}}} x_{ij}B_j\|_F^2\Big) > 2\nu_2 > 0. \tag{A.7}
$$

Also, it can be checked that

$$
\Big|\frac{1}{npq}\sum_{i=1}^{n}\Big(\|Y_i - \sum_{j\in\tilde{\mathcal{M}}} x_{ij}B_j\|_F^2 - \|U_i\|_F^2\Big)\Big| = o_p(1).
$$

Furthermore, under Condition (C7)(i), we obtain that $n^{-1}\sum_{i=1}^{n} e_{i,sk}^2 \to_p \sigma_{sk}^2$, and $\sigma^{-2} < E(e_{sk}^2) < \sigma^2$ for some constant $0 < \sigma^2 < \infty$. Therefore, we have that with probability tending to one,

$$
\frac{1}{npq}\sum_{i=1}^{n}\Big\|Y_i - \sum_{j\in\tilde{\mathcal{M}}} x_{ij}B_j\Big\|_F^2 < 2\sigma^2. \tag{A.8}
$$

It then follows from (A.7) and (A.8) that with probability approaching one,

$$\begin{aligned}
&\min_{\mathcal{M}\subseteq\mathcal{M}_-} \text{EBIC}(\mathcal{M}) - \text{EBIC}(\tilde{\mathcal{M}}) \\
&\geq \min_{\mathcal{M}\in\mathcal{M}_\tau^-} \log\left(1 + \frac{\sum_{i=1}^n \left\{\|Y_i - \sum_{j\in\mathcal{M}} x_{ij}B_j\|_F^2 - \|Y_i - \sum_{j\in\tilde{\mathcal{M}}} x_{ij}B_j\|_F^2\right\}}{\sum_{i=1}^n \|Y_i - \sum_{j\in\tilde{\mathcal{M}}} x_{ij}B_j\|_F^2}\right) + \frac{c_n\log(n)}{n} \\
&\geq \min_{\mathcal{M}\in\mathcal{M}_\tau^-} \min\left\{\log(2), \frac{\sum_{i=1}^n \left\{\|Y_i - \sum_{j\in\mathcal{M}} x_{ij}B_j\|_F^2 - \|Y_i - \sum_{j\in\tilde{\mathcal{M}}} x_{ij}B_j\|_F^2\right\}}{2\sum_{i=1}^n \|Y_i - \sum_{j\in\tilde{\mathcal{M}}} x_{ij}B_j\|_F^2}\right\} \\
&\quad - \tau\frac{c_n\log(n)}{n} \\
&\geq \min\left\{\log(2), \frac{\nu_2}{2\sigma^2}\right\} - \tau\frac{c_n\log(n)}{n} > 0,
\end{aligned}$$

where the first inequality follows from $\log(1+x) \geq \min\{x/2, \log(2)\}$ for any $x > 0$, and the last inequality follows from the assumption $c_n\log(n)/n^{1-\delta_2} = o(1)$. This implies that for any underfitted model $\mathcal{M}$ with size $\tau$, there exists an overfitted model $\tilde{\mathcal{M}} = \mathcal{M}\cup\mathcal{M}^\star$ such that $\text{EBIC}(\mathcal{M}) > \text{EBIC}(\mathcal{M}')$ with probability going to one as $n\to\infty$. Thus, to prove Theorem 3, it suffices to show

$$pr\Big\{\min_{\mathcal{M}\in\mathcal{M}_\tau^+} \text{EBIC}(\mathcal{M}) > \text{EBIC}(\mathcal{M}^\star)\Big\} \to 1. \tag{A.9}$$

By arguments similar to those in the proofs of (A.4) and (A.5) in Theorem 3, we can also show that there exists a positive constant $\nu_1$ such that with probability tending to one,

$$\frac{\nu}{4}n^{-2\delta_3} < \frac{1}{npq}\sum_{i=1}^n \left\{\|Y_i - \sum_{j\in\tilde{\mathcal{M}}} x_{ij}B_j\|_F^2 - \|Y_i - \sum_{j\in\mathcal{M}^\star} x_{ij}B_j\|_F^2\right\} \leq \nu n^{-2\delta_3}. \tag{A.10}$$

Note that $\mathcal{M}_\tau^+ = \mathcal{M}^\star\cup(\mathcal{M}_\tau^+ - \mathcal{M}^\star) \subseteq \mathcal{M}_+^{2\tau}$. Hence (A.10) implies that

$$\begin{aligned}
&\min_{\mathcal{M}\in\mathcal{M}_\tau^+} \text{EBIC}(\mathcal{M}) - \text{EBIC}(\mathcal{M}^\star) \\
&\geq \min_{\mathcal{M}\in\mathcal{M}_\tau^+} \log\left(1 + \frac{\sum_{i=1}^n \left\{\|Y_i - \sum_{j\in\mathcal{M}} x_{ij}B_j\|_F^2 - \|Y_i - \sum_{j\in\mathcal{M}^\star} x_{ij}B_j\|_F^2\right\}}{\sum_{i=1}^n \|Y_i - \sum_{j\in\mathcal{M}^\star} x_{ij}B_j\|_F^2}\right) + \frac{c_n\log(n)}{n} \\
&\geq \min_{\mathcal{M}\in\mathcal{M}_\tau^+} \min\left\{\log(2), \frac{\sum_{i=1}^n \left\{\|Y_i - \sum_{j\in\mathcal{M}} x_{ij}B_j\|_F^2 - \|Y_i - \sum_{j\in\mathcal{M}^\star} x_{ij}B_j\|_F^2\right\}}{\sum_{i=1}^n \|Y_i - \sum_{j\in\mathcal{M}^\star} x_{ij}B_j\|_F^2}\right\} \\
&\quad + \frac{c_n\log(n)}{n} \\
&\geq \min\{\log(2), \frac{\nu}{4\sigma^2}n^{-2\delta_3}\} + c_n\frac{\log(n)}{n} > 0,
\end{aligned}$$

which implies that (A.9) holds, and hence completes the proof.

We next show the minimax risks of the estimate obtained by solving problem (2). We first introduce some notations. Let $\hat{\mathbb{B}} = (\hat{B}_1^\top, \ldots, \hat{B}_{d_n}^\top)^\top$ with each $\hat{B}_j = (\hat{b}_{j,sk}) \in \mathbb{R}^{p\times q}$. Define $\hat{\beta}_{sk} = (\hat{b}_{1,sk}, \ldots, \hat{b}_{d_n,sk})^\top \in \mathbb{R}^{d_n}$, $\beta_{sk}^\star = (b_{1,sk}^\star \ldots, b_{d_n,sk}^\star)^\top \in \mathbb{R}^{d_n}$, and $\hat{\Delta}_{sk} = \hat{\beta}_{sk} - \beta_{sk}^\star$.

Further let $\hat{\Delta} = \hat{\mathbb{B}} - \mathbb{B}^\star$, $\mathbb{U}_{sk} = (e_{1,sk}, \ldots, e_{n,sk})^\top \in \mathbb{R}^n$, and $\mathbb{Y}_{sk} = (Y_{1,sk}, \ldots, Y_{n,sk})^\top \in \mathbb{R}^n$. Denote $K(P, Q)$ as the Kullback-Leibler distance between the probability measures $P$ and $Q$. Let $\theta = (\theta_1^\top, \ldots, \theta_{d_n}^\top)^\top \in \mathbb{R}^{(pd_n)\times q}$ with each $\theta_j = (\theta_{j,sk}) \in \mathbb{R}^{p\times q}$, and $\theta_{sk} = (\theta_{1,sk}, \ldots, \theta_{d_n,sk})^\top$. Lastly, for an arbitrary vector $\eta = (\eta_1, \ldots, \eta_{d_n})^\top$, and $\mathbb{B} = (B_1^\top, \ldots, B_{d_n}^\top)^\top$, define $M_\tau = \{\eta : \eta_j \neq 0 \text{ if and only if } \|B_j\|_F \neq 0 \text{ for } \mathbb{B} \in \mathcal{B}_\tau\}$. The following lemma is a result of Raskutti et al. (2011), which will be used in the proof of Theorem 4.

**Lemma 4** *Suppose that Assumption (A) and Condition (C7)(i) hold. Then*

*(i) There exist some positive constants $\ell_1, \ell_2$ and $\ell_3$ such that for any $r > 0$,*

$$\sup_{\theta_{sk}\in M_{2\tau}, \|\theta_{sk}\|\leq r} \frac{1}{n}|\mathbb{U}_{sk}^\top \tilde{X}\theta_{sk}| \leq \ell_1 \sigma_{sk} r \kappa_2 \Big(\frac{\tau \log(d_n/\tau)}{n}\Big)^{1/2}$$

*with probability greater than $1 - \ell_2 \exp[-\ell_3 \min\{n, \tau\log(d_n/\tau)\}]$.*

*(ii) In addition, if $d_n > 4\tau$, then there exist some positive constants $\ell_4, \ell_5$ and $\ell_6$ such that*

$$\sup_{\theta_{sk}\in M_{2\tau}, \|\tilde{X}\theta_{sk}\|\leq r\sqrt{n}} \frac{1}{n}|\mathbb{U}_{sk}^\top \tilde{X}\theta_{sk}| \leq \ell_4 \sigma_{sk} r \Big(\frac{\tau \log(d_n/\tau)}{n}\Big)^{1/2}$$

*with probability greater than $1 - \ell_5 \exp[-\ell_6 \tau \log\{d_n/(2\tau)\}]$.*

**Proof of Theorem 4.** Note that $\hat{\mathbb{B}}$ is a minimizer of problem (2) and $\mathbb{B}^\star$ is also a feasible point. Then $\|\mathbb{Y} - \mathbb{X}\hat{\mathbb{B}}\|_F^2 \leq \|\mathbb{Y} - \mathbb{X}\mathbb{B}^\star\|_F^2$, which implies the following inequality:

$$\frac{1}{n}\|\mathbb{X}\hat{\Delta}\|_F^2 = \frac{1}{n}\sum_{s=1}^p\sum_{k=1}^q \|\tilde{X}\hat{\Delta}_{sk}\|_2^2 \leq \frac{2}{n}\sum_{s=1}^p\sum_{k=1}^q |\mathbb{U}_{sk}^\top \tilde{X}\hat{\Delta}_{sk}|. \tag{A.11}$$

Define the event $\Omega_1$ as

$$\Omega_1 = \left\{\exists\ \theta \in \mathcal{B}_{2\tau} : \frac{1}{n}\sum_{s,k} |\mathbb{U}_{sk}^\top \tilde{X}\theta_{sk}| \geq \ell\kappa_2\sigma\sqrt{(pq)}\|\theta\|_F \Big(\frac{\tau\log(d_n/\tau)}{n}\Big)^{1/2}\right\},$$

where $\ell > 0$ is a constant. Since

$$\begin{aligned}
&pr\left(\frac{1}{n}\sum_{s=1}^p\sum_{k=1}^q |\mathbb{U}_{sk}^\top \tilde{X}\theta_{sk}| \geq \ell\kappa_2\sigma\sqrt{(pq)}\|\theta\|_F\Big(\frac{\tau\log(d_n/\tau)}{n}\Big)^{1/2}\right)\\
\leq &pr\left(\frac{1}{n}\sum_{s=1}^p\sum_{k=1}^q |\mathbb{U}_{sk}^\top \tilde{X}\theta_{sk}| \geq \ell\kappa_2\sigma\sum_{s=1}^p\sum_{k=1}^q\|\theta_{sk}\|_2\Big(\frac{\tau\log(d_n/\tau)}{n}\Big)^{1/2}\right)\\
\leq &\sum_{s=1}^p\sum_{k=1}^q pr\left(\frac{1}{n}|\mathbb{U}_{sk}^\top \tilde{X}\theta_{sk}| \geq \ell\kappa_2\sigma_{sk}\|\theta_{sk}\|_2\Big(\frac{\tau\log(d_n/\tau)}{n}\Big)^{1/2}\right),
\end{aligned}$$

the part (i) of Lemma 4 implies that for some positive constants $c_4$ and $c_5$,

$$pr(\Omega_1) \leq pqc_4 \exp\{-c_5\tau\log(d_n/\tau)\}.$$

That is,

$$\frac{1}{n}\sum_{s=1}^{p}\sum_{k=1}^{q}|\mathbb{U}_{sk}^{\top}\tilde{X}\theta_{sk}| \le \ell\sigma\kappa_2\sqrt{(pq)}\|\theta\|_F\Big(\frac{\tau\log(d_n/\tau)}{n}\Big)^{1/2}$$

with probability greater than $1-pqc_4\exp\{-c_5\tau\log(d_n/\tau)\}$. Thus, it follows from (A.11) and Assumption (A) that

$$\frac{1}{\sqrt{(pq)}}\|\hat{\Delta}\|_F \le c_3\frac{\kappa_2\sigma}{\kappa_1}\Big(\frac{\tau\log(d_n/\tau)}{n}\Big)^{1/2}$$

with the probability as claimed.

For the predictor error, we replace $\|\theta\|_F$ with $n^{-1/2}\|\mathbb{X}\theta\|_F$ in the right side of the inequality in the definition of event $\Omega_1$. Then by similar arguments as above and using the part (ii) of Lemma 4, we can show that there exist some positive constants $c_6$, $c_7$ and $c_8$ such that

$$pr\left(\frac{1}{\sqrt{(npq)}}\|\mathbb{X}\hat{\Delta}\|_F \le c_6\frac{\kappa_2\sigma}{\kappa_1}\Big(\frac{\tau\log(d_n/\tau)}{n}\Big)^{1/2}\right) \ge 1-c_7pq\exp\big[-c_8\tau\log\{d_n/(2\tau)\}\big].$$

This completes the proof.

**Proof of Theorem 5.** Let $\mathbb{Z}$ and $\mathbb{I}$ denote two $p\times q$ matrices with all entries being 0 and 1, respectively. Define the set

$$\Omega_2 = \big\{\varpi = (\varpi_1^{\top},\ldots,\varpi_{d_n}^{\top})^{\top}\in\mathbb{R}^{(pd_n)\times q} : \varpi_j\in\{\mathbb{Z},\mathbb{I}\},\quad \varpi\in\mathcal{B}_\tau\big\},$$

and its dilation

$$\mathcal{D}(\Omega_2) = \Big\{\ell\frac{\varpi\sigma}{\kappa_2}\Big(\frac{\log(d_n/\tau)}{n}\Big)^{1/2}\ :\ \varpi\in\Omega_2\Big\}$$

for a constant $\ell>0$. For any $\varpi$ and $\varpi'$ in $\Omega_2$, we have that $(\varpi-\varpi')\in\mathcal{B}_{2\tau}$. Let $\mathbb{B}=\varpi(\ell\sigma/\kappa_2)\{\log(d_n/\tau)/n\}^{1/2}$ and $\mathbb{B}'=\varpi'(\ell\sigma/\kappa_2)\{\log(d_n/\tau)/n\}^{1/2}$. Then Assumption (A) implies

$$\begin{aligned}\frac{1}{n}\|\mathbb{X}(\mathbb{B}-\mathbb{B}')\|_F^2 \ge& \frac{\ell^2\sigma^2\kappa_1^2\log(d_n/\tau)}{n\kappa_2^2}\sum_{j=1}^{d_n}\sum_{s=1}^{p}\sum_{k=1}^{q}I(\varpi_{j,sk}=\varpi'_{j,sk})\\ =&\frac{\ell^2\sigma^2\kappa_1^2\log(d_n/\tau)pq}{n\kappa_2^2}\xi(\varpi,\varpi'),\end{aligned}\tag{A.12}$$

and

$$\frac{1}{n}\|\mathbb{X}(\mathbb{B}-\mathbb{B}')\|_F^2 \le\frac{\ell^2\sigma^2\log(d_n/\tau)pq}{n}\xi(\varpi,\varpi'),\tag{A.13}$$

where $\xi(\varpi,\varpi')=\sum_{j=1}^{d_n}I(\varpi_j\neq\varpi'_j)$ is the Hamming distance. Thus, Lemma 2.9 in Tsybakov (2009) yields that there exists a subset $\bar{\mathcal{B}}=\{\varpi^{(0)},\ldots,\varpi^{(M)}\}$ of $\Omega_2$ such that $\varpi^{(0)}=(\mathbb{Z}^{\top},\ldots,\mathbb{Z}^{\top})^{\top}$, and for a constant $\tilde{\ell}>0$,

$$\begin{aligned}\log(M)\ge&\tilde{\ell}\tau\log(d_n/\tau),\\ \text{and }\xi(\varpi^{(j)},\varpi^{(k)})\ge&\tau/4,\quad\forall\ 0\le j<k\le M.\end{aligned}$$

This, together with (A.12), implies that for any $\mathbb{B}^{(l)}$ and $\mathbb{B}^{(j)} \in \mathcal{D}(\bar{\mathcal{B}})$ with $l \neq j$,

$$\begin{aligned}\frac{1}{n}\|\mathbb{X}(\mathbb{B}^{(l)} - \mathbb{B}^{(j)})\|_F^2 \geq & \frac{\ell^2\sigma^2\kappa_1^2\log(d_n/\tau)pq}{n\kappa_2^2}\xi(\varpi^{(l)}, \varpi^{(j)}) \\ \geq & \frac{\tau\ell^2\sigma^2\kappa_1^2\log(d_n/\tau)pq}{4n\kappa_2^2}.\end{aligned}$$

Let $\varpi^{(l)} = (\varpi_{j,sk}^{(l)}) \in \bar{\mathcal{B}}$, $\varpi_{sk}^{(l)} = (\varpi_{1,sk}^{(l)}, \ldots, \varpi_{d_n,sk}^{(l)})^\top$, and $\mathbb{B}_{sk}^{(l)} = \varpi_{sk}^{(l)}(\ell\sigma/\kappa_2)\{\log(d_n/\tau)/n\}^{1/2}$ for $1 \leq l \leq M$. Denote $P_{sk}^{(l)}$ as the probability function of $\mathbb{Y}_{sk}$ given $\tilde{X}\mathbb{B}_{sk}^{(l)}$. Then we have

$$\begin{aligned}\sum_{s=1}^{p}\sum_{k=1}^{q} K(P_{sk}^{(0)}, P_{sk}^{(l)}) = & \sum_{s=1}^{p}\sum_{k=1}^{q}\int \log\frac{dP_{sk}^{(0)}}{dP_{sk}^{(l)}} dP_{sk}^{(0)} \\ = & \sum_{s=1}^{p}\sum_{k=1}^{q}\frac{1}{2\sigma_{sk}^2}\|\tilde{X}(\varpi_{sk}^{(l)} - \varpi_{sk}^{(0)})\|^2 \\ \leq & O(1)\sum_{s=1}^{p}\sum_{k=1}^{q}\frac{1}{2\sigma^2}\|\tilde{X}(\varpi_{sk}^{(l)} - \varpi_{sk}^{(0)})\|^2.\end{aligned}$$

This, combining (A.13) and the fact $\varpi^{(l)} \in \mathcal{B}_\tau$, yields

$$\begin{aligned}\frac{1}{M}\sum_{l=1}^{M}\sum_{s=1}^{p}\sum_{k=1}^{q} K(P_{sk}^{(0)}, P_{sk}^{(l)}) \leq & \frac{\ell^2\log(d_n/\tau)pq}{4n}\Big[\frac{1}{M}\sum_{l=1}^{M}\sum_{j=1}^{d_n} I(\varpi_j^{(l)} = \varpi_j^{(0)})\Big] \\ \leq & \frac{\tau\ell^2\log(d_n/\tau)pq}{8n}. \qquad \text{(A.14)}\end{aligned}$$

We then take a sufficiently small $\ell$ in (A.14) such that

$$\frac{1}{M}\sum_{l=1}^{M}\sum_{s=1}^{p}\sum_{k=1}^{q} K(P_{sk}^{(0)}, P_{sk}^{(l)}) \leq \log(M)/16 = \log[\mathrm{card}\{\mathcal{D}(\bar{\mathcal{M}})\}]/16. \qquad \text{(A.15)}$$

In view of (A.15), an application of Theorem 2.7 in Tsybakov (2009) yields the lower bound of the prediction error.

As similar arguments to (A.12) and (A.15), we have that for $\mathbb{B}$ and $\mathbb{B}' \in \mathcal{D}(\Omega_2)$, and $\mathbb{B} \neq \mathbb{B}'$,

$$\frac{1}{\sqrt{(pq)}}\|\mathbb{B} - \mathbb{B}'\|_F = \ell\frac{\sigma}{\kappa_2}\Big(\frac{\log(d_n/\tau)}{n}\Big)^{1/2}\xi^{1/2}(\varpi, \varpi'),$$

and that for any $\mathbb{B}^{(l)}$ and $\mathbb{B}^{(j)} \in \mathcal{D}(\bar{\mathcal{B}})$ with $l \neq j$,

$$\frac{1}{\sqrt{(pq)}}\|\mathbb{B}^{(l)} - \mathbb{B}^{(j)}\|_F \geq \ell\frac{\tau\sigma}{4\kappa_2}\Big(\frac{\log(d_n/\tau)}{n}\Big)^{1/2}.$$

Based on these facts and (A.15), Theorem 2.7 in Tsybakov (2009) yields the lower bound of the estimation error.

**Proof of Theorem 6.** We first show an upper bound for the estimation error $\|\mathbb{B}^{[l]}-\mathbb{B}^{\star}\|_F$. Note that

$$\begin{aligned}
&\big\|\mathbb{H}_\tau\{\mathbb{B}^{[l]}+(n\lambda^{[l]})^{-1}\mathbb{X}^\top(\mathbb{Y}-\mathbb{X}\mathbb{B}^{[l]})\}-\big[\mathbb{B}^{[l]}+(n\lambda^{[l]})^{-1}\mathbb{X}^\top(\mathbb{Y}-\mathbb{X}\mathbb{B}^{[l]})\big]\\
&\quad+\big[\mathbb{B}^{[l]}+(n\lambda^{[l]})^{-1}\mathbb{X}^\top(\mathbb{Y}-\mathbb{X}\mathbb{B}^{[l]})\big]-\mathbb{B}^{\star}\big\|_F\\
\le&\big\|\mathbb{H}_\tau\{\mathbb{B}^{[l]}+(n\lambda^{[l]})^{-1}\mathbb{X}^\top(\mathbb{Y}-\mathbb{X}\mathbb{B}^{[l]})\}-[\mathbb{B}^{[l]}+(n\lambda^{[l]})^{-1}\mathbb{X}^\top(\mathbb{Y}-\mathbb{X}\mathbb{B}^{[l]})]\big\|_F\\
&\quad+\big\|\mathbb{B}^{[l]}+(n\lambda^{[l]})^{-1}\mathbb{X}^\top(\mathbb{Y}-\mathbb{X}\mathbb{B}^{[l]})-\mathbb{B}^{\star}\big\|_F. \qquad \text{(A.16)}
\end{aligned}$$

Then, by Proposition 1 and condition $\tau^{\star}\le\tau$, we have

$$\begin{aligned}
&\big\|\mathbb{H}_\tau\{\mathbb{B}^{[l]}+(n\lambda^{[l]})^{-1}\mathbb{X}^\top(\mathbb{Y}-\mathbb{X}\mathbb{B}^{[l]})\}-[\mathbb{B}^{[l]}+(n\lambda^{[l]})^{-1}\mathbb{X}^\top(\mathbb{Y}-\mathbb{X}\mathbb{B}^{[l]})]\big\|_F\\
\le&\big\|\mathbb{B}^{\star}-[\mathbb{B}^{[l]}+(n\lambda^{[l]})^{-1}\mathbb{X}^\top(\mathbb{Y}-\mathbb{X}\mathbb{B}^{[l]})]\big\|_F.
\end{aligned}$$

This, together with (A.16), yields that

$$\begin{aligned}
\|\mathbb{B}^{[l+1]}-\mathbb{B}^{\star}\|_F=&\|\mathbb{H}_\tau\{\mathbb{B}^{[l]}+(n\lambda^{[l]})^{-1}\mathbb{X}^\top(\mathbb{Y}-\mathbb{X}\mathbb{B}^{[l]})\}-\mathbb{B}^{\star}\|_F\\
\le&2\|(\mathbb{B}^{[l]}-\mathbb{B}^{\star})+(n\lambda^{[l]})^{-1}\mathbb{X}^\top(\mathbb{Y}-\mathbb{X}\mathbb{B}^{[l]})\|_F\\
=&2\big\|[\mathbb{I}-(n\lambda^{[l]})^{-1}\mathbb{X}^\top\mathbb{X}](\mathbb{B}^{[l]}-\mathbb{B}^{\star})+(n\lambda^{[l]})^{-1}\mathbb{X}^\top\mathbb{U}\big\|_F.
\end{aligned}$$

Then the triangle inequality, combining with $\phi<\lambda^{[l]}\le\kappa_1/(1-1/\sqrt{32})$, yields

$$\|\mathbb{B}^{[l+1]}-\mathbb{B}^{\star}\|_F\le 2\|[\mathbb{I}-(n\lambda^{[l]})^{-1}\mathbb{X}^\top\mathbb{X}](\mathbb{B}^{[l]}-\mathbb{B}^{\star})\|_F+\frac{2}{n\lambda^{[l]}}\|\mathbb{X}^\top\mathbb{U}\|_F,$$

which implies

$$\|\mathbb{B}^{[l+1]}-\mathbb{B}^{\star}\|_F\le 2^{-1}\|\mathbb{B}^{[l]}-\mathbb{B}^{\star}\|_F+\frac{2}{n\phi}\|\mathbb{X}^\top\mathbb{U}\|_F.$$

Iterating this relationship, we obtain

$$\|\mathbb{B}^{[l]}-\mathbb{B}^{\star}\|_F\le 2^{-l}\|\mathbb{B}^{[0]}-\mathbb{B}^{\star}\|_F+\frac{4}{n\phi}\|\mathbb{X}^\top\mathbb{U}\|_F. \qquad \text{(A.17)}$$

Denote $\mathcal{M}_\tau^{[l]}=\{j:\|\mathbb{B}^{[l]}\|_F\neq 0\}$ as the support of $\mathbb{B}^{[l]}$ with sparse level $\tau$, and $\mathcal{M}_\tau=\mathcal{M}^{\star}\cup\mathcal{M}_\tau^{[l]}$. Then (A.17), together with the conditions $\mathbb{B}^{[0]}=0$ and $l>[\log_2(n\phi\|\mathbb{B}^{\star}\|_F/\|\mathbb{X}^\top\mathbb{U}\|_F)]$, leads to

$$\|\mathbb{B}^{[l]}_{\mathcal{M}_\tau}-\mathbb{B}^{\star}_{\mathcal{M}_\tau}\|_F\le\frac{4}{n\phi}\|(\mathbb{X}^\top\mathbb{U})_{\mathcal{M}_\tau}\|_F.$$

Thus, it suffices to show

$$\begin{aligned}
&pr\bigg(\frac{1}{\sqrt{(pqn^2)}}\|(\mathbb{X}^\top\mathbb{U})_{\mathcal{M}_\tau}\|_F\ge\tilde{c}_3\sigma\Big(\frac{\tau\log(d_n/\tau)}{n}\Big)^{1/2}\bigg)\\
&\quad\le\tilde{c}_4pq\tau\exp\Big\{-\tilde{c}_5\log(d_n/\tau)n^{2\upsilon}\Big\}+\tilde{c}_4n\tau\exp\{-\eta_1n^{\upsilon}\}.
\end{aligned}$$

Note that

$$pr\left(\frac{\|(\mathbb{X}^\top\mathbb{U})_{\mathcal{M}_\tau}\|_F}{\sqrt{(pqn^2)}} \geq \tilde{c}_3\sigma\Big(\frac{\tau\log(d_n/\tau)}{n}\Big)^{1/2}\right)$$
$$\leq \sum_{s=1}^{p}\sum_{k=1}^{q}\sum_{j\in\mathcal{M}_\tau} pr\left(\Big|\frac{1}{n}\sum_{i=1}^{n}\frac{x_{ij}e_{i,sk}}{\sigma_{sk}}\Big| \geq \frac{\tilde{c}_3}{2\omega_2}\Big(\frac{\log(d_n/\tau)}{n}\Big)^{1/2}\right).$$

It follows Lemma A.1 of Xu and Chen (2014) that on the event $\{|x_{ij}| < n^\upsilon : 1 \leq i \leq n, j \in \mathcal{M}_\tau\}$,

$$pr\left(\Big|\frac{1}{n}\sum_{i=1}^{n}\frac{x_{ij}e_{i,sk}}{\sigma_{sk}}\Big| \geq \frac{\tilde{c}_3}{2\omega_2}\Big(\frac{\log(d_n/\tau)}{n}\Big)^{1/2}\right) \leq pr\left(\Big|\frac{1}{n}\sum_{i=1}^{n}\frac{x_{ij}e_{i,sk}}{\sigma_{sk}}\Big| \geq \frac{\tilde{c}_3}{2\omega_2}\Big(\frac{\log(d_n/\tau)}{n}\Big)^{1/2}\right)$$
$$\leq \tilde{c}_4\exp\{-\tilde{c}_5\log(d_n/\tau)n^{1-2\upsilon}\}.$$

By Assumption (A) and Condition (C6), we have that there exist some positive constants $\tilde{c}_6$, $\tilde{c}_7$ and $\tilde{c}_8$ such that

$$pr\left\{\frac{1}{\sqrt{(n^2pq)}}\|\mathbb{X}(\mathbb{B}^{[l]}-\mathbb{B}^\star)\|_F \geq \tilde{c}_6\sigma\Big(\frac{\tau\log(d_n/\tau)}{n}\Big)^{1/2}\right\}$$
$$\leq pr\left\{\frac{1}{\sqrt{(pq)}}\|\mathbb{B}^{[l]}-\mathbb{B}^\star\|_F \geq \frac{\tilde{c}_6}{\kappa_2}\sigma\Big(\frac{\tau\log(d_n/\tau)}{n}\Big)^{1/2}\right\}$$
$$\leq \tilde{c}_7\tau n\exp\{-\eta_1 n^\upsilon\} + \tilde{c}_7pq\tau\exp\{-\tilde{c}_8\log(d_n/\tau)n^{1-2\upsilon}\}.$$

This completes the proof.

## Appendix B. Additional Real Data Information

### B.1 Data Usage Acknowledgement

Data used in the preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in the analysis or writing of this report. A complete listing of ADNI investigators can be found at:
`http://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf`.

Data used in the preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (`adni.loni.usc.edu`). The ADNI was launched in 2003 as a public-private partnership, led by Principal Investigator Michael W. Weiner, MD. The primary goal of ADNI has been to test whether serial magnetic resonance imaging (MRI), positron emission tomography (PET), other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of mild cognitive impairment (MCI) and early Alzheimer's disease (AD). For up-to-date information, see www.adni-info.org.

Data collection and sharing for this project was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904) and

DOD ADNI (Department of Defense award number W81XWH-12-2-0012). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: AbbVie, Alzheimer's Association; Alzheimer's Drug Discovery Foundation; Araclon Biotech; BioClinica, Inc.; Biogen; Bristol-Myers Squibb Company; CereSpir, Inc.; Cogstate; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; EuroImmun; F. Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc.; Fujirebio; GE Healthcare; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Lumosity; Lundbeck; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; NeuroRx Research; Neurotrack Technologies; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Takeda Pharmaceutical Company; and Transition Therapeutics. The Canadian Institutes of Health Research is providing funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health (`www.fnih.org`). The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer's Therapeutic Research Institute at the University of Southern California. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of Southern California.

### B.2 Imaging and Genetic Preprocessing

The MRI data, collected across a variety of 1.5 Tesla MRI scanners with protocols individualized for each scanner, includes standard T1-weighted images obtained by using volumetric 3-dimensional sagittal MPRAGE or equivalent protocols with varying resolutions. The typical protocol includes: inversion time (TI) = 1000 ms, flip angle = $8^o$, repetition time (TR) = 2400 ms, and field of view (FOV) = 24 cm with a $256 \times 256 \times 170$ acquisition matrix in the $x-$, $y-$, and $z-$dimensions yielding a voxel size of $1.25 \times 1.26 \times 1.2$ mm$^3$. We adopted a surface fluid registration based hippocampal subregional analysis package (Shi et al, 2014), which uses isothermal coordinates and fluid registration to generate one-to-one hippocampal surface registration for surface statistics computation. It introduced two cuts on a hippocampal surface to convert it into a genus zero surface with two open boundaries. The locations of the two cuts were at the front and back of the hippocampal surface. By using conformal parameterization, it essentially converts a 3D surface registration problem into a two-dimensional (2D) image registration problem. The flow induced in the parameter domain establishes high-order correspondences between 3D surfaces. Finally, the radial distance was computed on the registered surface. This software package and associated image processing methods have been adopted and described by various studies. Although there are several competing methods, such as spherical harmonics representation, our fluid registration method has several unique features (Wang, 2011; Shi et al, 2014; Monje et al, 2013; Colom et al, 2013; Luders et al, 2013). After preprocessing, we obtained left and right hippocampus shape representations as $100 \times 150$ matrices.

We applied the following preprocessing technique to the genetic data. The first line quality control steps include (i) call rate check per subject and per SNP marker, (ii) gender check, (iii) sibling pair identification, (iv) the Hardy-Weinberg equilibrium test, (v) marker removal by the minor allele frequency, and (vi) population stratification. The second line

preprocessing steps include removal of SNPs with (i) more than 5% missing values, (ii) minor allele frequency (MAF) smaller than10%, and (iii) Hardy-Weinberg equilibrium $p$-value $< 10^{-6}$. 503,892 SNPs obtained from 22 chromosomes were included in for further processing. MACH-Admix software (http://www.unc.edu/ yunmli/MaCH-Admix/; Liu et al, 2013) is applied on the 756 Caucasian subjects to perform genotype imputation, using 1000G Phase I Integrated Release Version 3 haplotypes (http://www.1000genomes.org). The 1000 Genomes Project Consortium (2015) was used as a reference panel. Quality control was also conducted after imputation, excluding markers with (i) low imputation accuracy (based on imputation output $R^2$) (ii) Hardy-Weinberg equilibrium $p$-value $10^{-6}$ (iii) minor allele frequency (MAF) $< 5\%$. There were $6,087,205$ bi-allelic markers (including SNPs and indels) of 756 subjects retained in the data analysis. Among these 756 subjects, we deleted those subjects that do not have hippocampus shape representations data, and finally, we included 735 subjects in the study. Moreover, for the illustration of our proposed method, we mainly explored the relationship between the AD and the genes on the 19th chromosome. Specifically, there are 134712 SNPs on 735 patients. Motivated by Huang et al (2008), we first selected 10000 SNPs from all the genes with the largest variance in expression value. Then we chose the top 2000 SNPs whose expression value has the largest Frobenius norm of the marginal correlation with the response.

## Appendix C. Simulation studies

In this section, we conduct more simulation studies to examine the finite sample performance of the proposed method. The noises $\mathrm{Vec}(U_i)$ are generated as follows. Let $W_i \in \mathbb{R}^{pq}$ be independently generated from $t(4)/\sqrt{2}$ and $\{\chi^2(5)-5\}/\sqrt{10}$, where $t(4)$ and $\chi^2(5)$ denote the $t$-distribution and the chi-square distribution with degrees of freedom 4 and 5, respectively. Then, let $\mathrm{Vec}(U_i) = \Sigma_e^{1/2} W_i$, where $\Sigma_e = \{0.2^{|[j]_p-[k]_q|+|\tilde{j}-\tilde{k}|} : 1 \le j,k \le (pq)\}$. That is, the correlations between $e_{i,js}$ and $e_{i,j's'}$ are $\rho^{|j-j'|+|s-s'|} (1 \le j, j', s, s' \le pq)$. Here for any positive integer $a$, $[a]_p = [(a-1)/p] + 1$, $\tilde{a} = a - p([a]_p - 1)$, and $[x]$ denotes the integer part of $x > 0$. The total number of predictors is $d_n = 1000$. The sample sizes are $n = 100$ and 200. The matrix size $(p, q)$ is set as $(50, 50)$. Other settings are the same as in Section 4 of the main text. All simulation results reported in Tables S1-S4 are based on 100 Monte Carlo repetitions.

| $\vartheta$ | Method | $p_1$ | $p_2$ | $p_3$ | $p_4$ | $p_5$ | TP | FP | CF | Est | Pred |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | $t(4)/\sqrt{2}$ | | | | | |
| 0.1 | SRLS | 1.00(0.00) | 1.00(0.00) | 1.00(0.00) | 1.00(0.00) | 1.00(0.00) | 5.00(0.00) | 0.00(0.00) | 1.00(0.00) | 0.05(0.00) | 0.05(0.00) |
| | SIS* | 1.00(0.00) | 1.00(0.00) | 1.00(0.00) | 0.00(0.00) | 0.01(0.10) | 3.01(0.10) | 0.21(0.66) | 0.00(0.00) | 0.67(0.19) | 0.60(0.18) |
| | ISIS* | 1.00(0.00) | 1.00(0.00) | 1.00(0.00) | 0.99(0.10) | 0.99(0.10) | 4.99(0.20) | 0.25(0.91) | 0.90(0.30) | 0.07(0.09) | 0.06(0.09) |
| | KAZZ | 1.00(0.00) | 1.00(0.00) | 1.00(0.00) | 0.00(0.00) | 0.02(0.14) | 3.02(0.14) | 0.02(0.14) | 0.00(0.00) | 0.73(0.22) | 0.66(0.21) |
| 0.5 | SRLS | 1.00(0.00) | 1.00(0.00) | 1.00(0.00) | 1.00(0.00) | 1.00(0.00) | 5.00(0.00) | 0.00(0.00) | 1.00(0.00) | 0.08(0.01) | 0.04(0.00) |
| | SIS* | 0.83(0.38) | 0.79(0.41) | 0.86(0.35) | 0.00(0.00) | 0.00(0.00) | 2.48(0.67) | 1.34(1.71) | 0.00(0.00) | 14.0(4.41) | 6.47(2.09) |
| | ISIS* | 0.87(0.34) | 0.84(0.37) | 0.87(0.34) | 1.00(0.00) | 0.11(0.31) | 3.69(0.71) | 2.57(2.56) | 0.00(0.00) | 2.67(3.51) | 0.93(0.95) |
| | KAZZ | 0.91(0.29) | 0.94(0.24) | 0.94(0.24) | 0.00(0.00) | 0.00(0.00) | 2.79(0.43) | 0.23(0.55) | 0.00(0.00) | 12.0(4.27) | 5.77(2.09) |
| 0.9 | SRLS | 1.00(0.00) | 1.00(0.00) | 1.00(0.00) | 1.00(0.00) | 1.00(0.00) | 5.00(0.00) | 0.22(0.50) | 0.82(0.39) | 0.39(0.12) | 0.05(0.01) |
| | SIS* | 0.28(0.45) | 0.28(0.45) | 0.26(0.44) | 0.05(0.22) | 0.23(0.42) | 1.10(0.86) | 0.47(1.22) | 0.00(0.00) | 39.9(16.9) | 4.05(1.78) |
| | ISIS* | 0.65(0.48) | 0.64(0.48) | 0.62(0.49) | 1.00(0.00) | 0.83(0.38) | 3.74(0.86) | 3.01(2.09) | 0.07(0.26) | 7.16(6.94) | 0.54(0.43) |
| | KAZZ | 0.27(0.45) | 0.28(0.45) | 0.31(0.46) | 0.03(0.17) | 0.16(0.37) | 1.05(0.72) | 0.31(0.72) | 0.00(0.00) | 38.0(11.3) | 3.84(1.38) |
| | | | | | | $(\chi^2(5)-5)/\sqrt{10}$ | | | | | |
| 0.1 | SRLS | 1.00(0.00) | 1.00(0.00) | 1.00(0.00) | 1.00(0.00) | 1.00(0.00) | 5.00(0.00) | 0.00(0.00) | 1.00(0.00) | 0.05(0.00) | 0.04(0.00) |
| | SIS* | 1.00(0.00) | 0.99(0.10) | 0.99(0.10) | 0.01(0.10) | 0.02(0.14) | 3.01(0.22) | 0.46(1.34) | 0.00(0.00) | 0.78(0.46) | 0.69(0.26) |
| | ISIS* | 1.00(0.00) | 1.00(0.00) | 1.00(0.00) | 0.98(0.14) | 0.98(0.14) | 4.96(0.28) | 0.26(0.79) | 0.87(0.34) | 0.07(0.06) | 0.06(0.05) |
| | KAZZ | 1.00(0.00) | 1.00(0.00) | 1.00(0.00) | 0.00(0.00) | 0.01(0.10) | 3.01(0.10) | 0.09(0.45) | 0.00(0.00) | 0.75(0.24) | 0.67(0.22) |
| 0.5 | SRLS | 1.00(0.00) | 1.00(0.00) | 1.00(0.00) | 1.00(0.00) | 1.00(0.00) | 5.00(0.00) | 0.00(0.00) | 1.00(0.00) | 0.08(0.01) | 0.05(0.00) |
| | SIS* | 0.96(0.20) | 0.96(0.20) | 0.93(0.26) | 0.00(0.00) | 0.00(0.00) | 2.85(0.36) | 0.25(0.67) | 0.00(0.00) | 11.3(3.36) | 5.47(1.73) |
| | ISIS* | 0.86(0.35) | 0.86(0.35) | 0.89(0.31) | 1.00(0.00) | 0.16(0.37) | 3.77(0.62) | 2.50(2.16) | 0.01(0.10) | 2.59(3.43) | 0.98(1.04) |
| | KAZZ | 0.95(0.22) | 0.96(0.20) | 0.93(0.26) | 0.00(0.00) | 0.00(0.00) | 2.84(0.37) | 0.26(0.76) | 0.00(0.00) | 11.5(3.38) | 5.49(1.73) |
| 0.9 | SRLS | 1.00(0.00) | 1.00(0.00) | 1.00(0.00) | 1.00(0.00) | 1.00(0.00) | 5.00(0.00) | 0.19(0.46) | 0.84(0.37) | 0.38(0.11) | 0.05(0.01) |
| | SIS* | 0.22(0.42) | 0.33(0.47) | 0.22(0.42) | 0.02(0.14) | 0.22(0.42) | 1.01(0.85) | 0.30(0.46) | 0.01(0.10) | 39.0(13.2) | 3.75(1.36) |
| | ISIS* | 0.73(0.45) | 0.64(0.48) | 0.57(0.50) | 1.00(0.00) | 0.78(0.41) | 3.72(0.89) | 3.49(2.22) | 0.02(0.14) | 7.09(6.03) | 0.52(0.35) |
| | KAZZ | 0.23(0.42) | 0.31(0.46) | 0.21(0.40) | 0.02(0.14) | 0.21(0.41) | 0.98(0.79) | 0.31(0.49) | 0.01(0.10) | 39.1(13.2) | 3.76(1.35) |

Table S1: The selection results and standard deviation (in parentheses) for Example 1 with $n = 100$.

| $\vartheta$ | Method | $p_1$ | $p_2$ | $p_3$ | $p_4$ | $p_5$ | TP | FP | CF | Est | Pred |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | $t(4)/\sqrt{2}$ | | | | | |
| 0.1 | SRLS | 1.00(0.00) | 1.00(0.00) | 1.00(0.00) | 1.00(0.00) | 1.00(0.00) | 5.00(0.00) | 0.00(0.00) | 1.00(0.00) | 0.03(0.00) | 0.03(0.00) |
| | SIS* | 1.00(0.00) | 1.00(0.00) | 1.00(0.00) | 0.00(0.00) | 0.01(0.10) | 3.01(0.10) | 0.11(1.10) | 0.00(0.00) | 0.68(0.22) | 0.65(0.21) |
| | ISIS* | 1.00(0.00) | 1.00(0.00) | 1.00(0.00) | 1.00(0.00) | 1.00(0.00) | 5.00(0.00) | 0.00(0.00) | 1.00(0.00) | 0.03(0.00) | 0.03(0.00) |
| | KAZZ | 1.00(0.00) | 1.00(0.00) | 1.00(0.00) | 0.00(0.00) | 0.05(0.22) | 3.05(0.22) | 0.22(1.17) | 0.00(0.00) | 0.66(0.22) | 0.61(0.20) |
| 0.5 | SRLS | 1.00(0.00) | 1.00(0.00) | 1.00(0.00) | 1.00(0.00) | 1.00(0.00) | 5.00(0.00) | 0.00(0.00) | 1.00(0.00) | 0.04(0.00) | 0.03(0.00) |
| | SIS* | 1.00(0.00) | 0.99(0.10) | 1.00(0.00) | 0.00(0.00) | 0.00(0.00) | 2.99(0.10) | 0.42(0.83) | 0.00(0.00) | 11.5(3.56) | 5.98(1.96) |
| | ISIS* | 1.00(0.00) | 1.00(0.00) | 1.00(0.00) | 1.00(0.00) | 0.29(0.46) | 4.29(0.46) | 1.34(2.49) | 0.06(0.24) | 0.24(0.13) | 0.20(0.12) |
| | KAZZ | 1.00(0.00) | 1.00(0.00) | 1.00(0.00) | 0.00(0.10) | 0.00(0.00) | 3.00(0.00) | 0.08(0.37) | 0.00(0.00) | 10.9(3.46) | 5.55(1.90) |
| 0.9 | SRLS | 1.00(0.00) | 1.00(0.00) | 1.00(0.00) | 1.00(0.00) | 1.00(0.00) | 5.00(0.00) | 0.02(0.14) | 0.98(0.14) | 0.17(0.03) | 0.03(0.00) |
| | SIS* | 0.42(0.50) | 0.37(0.49) | 0.31(0.46) | 0.01(0.10) | 0.21(0.41) | 1.32(0.71) | 0.11(0.47) | 0.00(0.00) | 39.6(13.3) | 4.11(1.45) |
| | ISIS* | 0.93(0.26) | 0.93(0.26) | 0.94(0.24) | 1.00(0.00) | 0.96(0.20) | 4.76(0.45) | 1.82(2.08) | 0.39(0.49) | 1.29(2.22) | 0.13(0.18) |
| | KAZZ | 0.39(0.49) | 0.37(0.49) | 0.36(0.48) | 0.00(0.00) | 0.17(0.38) | 1.29(0.77) | 0.05(0.22) | 0.00(0.00) | 43.0(13.3) | 4.38(1.42) |
| | | | | | | $(\chi^2(5)-5)/\sqrt{10}$ | | | | | |
| 0.1 | SRLS | 1.00(0.00) | 1.00(0.00) | 1.00(0.00) | 1.00(0.00) | 1.00(0.00) | 5.00(0.00) | 0.00(0.00) | 1.00(0.00) | 0.03(0.00) | 0.03(0.00) |
| | SIS* | 1.00(0.00) | 1.00(0.00) | 1.00(0.00) | 0.00(0.00) | 0.03(0.17) | 3.03(0.17) | 0.18(1.07) | 0.00(0.00) | 0.69(0.20) | 0.64(0.20) |
| | ISIS* | 1.00(0.00) | 1.00(0.00) | 1.00(0.00) | 1.00(0.00) | 1.00(0.00) | 5.00(0.00) | 0.00(0.00) | 1.00(0.00) | 0.03(0.00) | 0.03(0.00) |
| | KAZZ | 1.00(0.00) | 1.00(0.00) | 1.00(0.00) | 0.00(0.00) | 0.01(0.10) | 3.01(0.10) | 0.03(0.30) | 0.00(0.00) | 0.68(0.22) | 0.63(0.21) |
| 0.5 | SRLS | 1.00(0.00) | 1.00(0.00) | 1.00(0.00) | 1.00(0.00) | 1.00(0.00) | 5.00(0.00) | 0.00(0.00) | 1.00(0.00) | 0.04(0.01) | 0.03(0.00) |
| | SIS* | 0.99(0.10) | 0.99(0.10) | 1.00(0.00) | 0.00(0.00) | 0.00(0.00) | 2.98(0.14) | 0.33(0.87) | 0.00(0.00) | 12.5(3.88) | 6.53(2.05) |
| | ISIS* | 0.99(0.10) | 1.00(0.00) | 1.00(0.00) | 1.00(0.00) | 0.30(0.46) | 4.29(0.46) | 1.62(2.59) | 0.05(0.22) | 0.31(0.52) | 0.24(0.21) |
| | KAZZ | 1.00(0.00) | 1.00(0.00) | 1.00(0.00) | 0.00(0.00) | 0.00(0.00) | 3.00(0.00) | 0.12(0.66) | 0.00(0.00) | 11.7(3.40) | 6.14(2.04) |
| 0.9 | SRLS | 1.00(0.00) | 1.00(0.00) | 1.00(0.00) | 1.00(0.00) | 1.00(0.00) | 5.00(0.00) | 0.00(0.00) | 1.00(0.00) | 0.17(0.02) | 0.03(0.00) |
| | SIS* | 0.31(0.46) | 0.42(0.50) | 0.40(0.49) | 0.00(0.00) | 0.16(0.37) | 1.29(0.76) | 0.02(0.20) | 0.00(0.00) | 39.3(13.1) | 4.04(1.40) |
| | ISIS* | 0.97(0.17) | 0.96(0.20) | 0.90(0.30) | 1.00(0.00) | 0.96(0.20) | 4.79(0.46) | 2.45(3.22) | 0.40(0.49) | 1.17(2.23) | 0.11(0.17) |
| | KAZZ | 0.31(0.46) | 0.43(0.50) | 0.36(0.48) | 0.00(0.00) | 0.16(0.37) | 1.26(0.69) | 0.02(0.20) | 0.00(0.00) | 39.3(13.1) | 4.04(1.41) |

Table S2: The selection results and standard deviation (in parentheses) for Example 1 with $n = 200$.

| $n$ | Method | $p_1$ | $p_2$ | $p_3$ | $p_4$ | $p_5$ | TP | FP | CF | Est | Pred |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | $t(4)/\sqrt{2}$ | | | | | |
| 100 | SRLS | 1.00(0.00) | 1.00(0.00) | 1.00(0.00) | 1.00(0.00) | 1.00(0.00) | 5.00(0.00) | 0.00(0.00) | 1.00(0.00) | 0.05(0.00) | 0.05(0.00) |
| | SIS* | 0.00(0.00) | 0.00(0.00) | 0.00(0.00) | 0.05(0.20) | 0.58(0.50) | 0.63(0.54) | 2.34(1.25) | 0.00(0.00) | 9.70(3.66) | 3.91(1.35) |
| | ISIS* | 0.04(0.20) | 0.16(0.37) | 0.72(0.45) | 0.98(0.14) | 1.00(0.00) | 2.90(0.63) | 3.42(1.47) | 0.00(0.00) | 2.19(1.65) | 0.90(0.50) |
| | KAZZ | 0.00(0.00) | 0.00(0.00) | 0.00(0.00) | 0.06(0.24) | 0.59(0.49) | 0.65(0.54) | 2.31(1.13) | 0.00(0.00) | 10.2(4.06) | 4.11(1.46) |
| 200 | SRLS | 1.00(0.00) | 1.00(0.00) | 1.00(0.00) | 1.00(0.00) | 1.00(0.00) | 5.00(0.00) | 0.00(0.00) | 1.00(0.00) | 0.03(0.00) | 0.03(0.00) |
| | SIS* | 0.00(0.00) | 0.00(0.00) | 0.00(0.00) | 0.05(0.22) | 0.83(0.38) | 0.88(0.46) | 2.84(1.32) | 0.00(0.00) | 7.72(2.75) | 3.65(1.23) |
| | ISIS* | 0.23(0.42) | 0.31(0.46) | 0.95(0.22) | 1.00(0.00) | 1.00(0.00) | 3.49(0.74) | 3.65(1.10) | 0.00(0.00) | 1.11(0.68) | 0.59(0.33) |
| | KAZZ | 0.00(0.00) | 0.00(0.00) | 0.00(0.00) | 0.05(0.22) | 0.84(0.37) | 0.89(0.45) | 3.22(2.02) | 0.00(0.00) | 8.00(2.81) | 3.76(1.25) |
| | | | | | | $(\chi^2(5)-5)/\sqrt{10}$ | | | | | |
| 100 | SRLS | 1.00(0.00) | 1.00(0.00) | 1.00(0.00) | 1.00(0.00) | 1.00(0.00) | 5.00(0.00) | 0.00(0.00) | 1.00(0.00) | 0.05(0.00) | 0.05(0.00) |
| | SIS* | 0.00(0.00) | 0.00(0.00) | 0.00(0.00) | 0.06(0.24) | 0.61(0.49) | 0.67(0.57) | 2.16(1.06) | 0.00(0.00) | 9.48(3.72) | 3.99(1.62) |
| | ISIS* | 0.01(0.10) | 0.12(0.33) | 0.79(0.41) | 0.99(0.10) | 1.00(0.00) | 2.91(0.53) | 3.42(1.53) | 0.00(0.00) | 2.02(1.49) | 0.82(0.43) |
| | KAZZ | 0.00(0.00) | 0.00(0.00) | 0.00(0.00) | 0.05(0.20) | 0.58(0.50) | 0.63(0.56) | 2.29(1.16) | 0.00(0.00) | 10.2(3.91) | 4.26(1.70) |
| 200 | SRLS | 1.00(0.00) | 1.00(0.00) | 1.00(0.00) | 1.00(0.00) | 1.00(0.00) | 5.00(0.00) | 0.00(0.00) | 1.00(0.00) | 0.03(0.00) | 0.02(0.00) |
| | SIS* | 0.00(0.00) | 0.00(0.00) | 0.00(0.00) | 0.03(0.17) | 0.83(0.38) | 0.86(0.40) | 3.11(2.21) | 0.00(0.00) | 7.86(3.15) | 3.60(1.27) |
| | ISIS* | 0.20(0.40) | 0.25(0.44) | 0.96(0.20) | 1.00(0.00) | 1.00(0.00) | 3.41(0.59) | 3.42(0.78) | 0.00(0.00) | 1.10(0.56) | 0.60(0.28) |
| | KAZZ | 0.00(0.00) | 0.00(0.00) | 0.00(0.00) | 0.02(0.14) | 0.80(0.40) | 0.82(0.41) | 2.68(1.04) | 0.00(0.00) | 8.40(3.28) | 3.85(1.35) |

Table S3: The selection results and standard deviation (in parentheses) for Example 2.

| | | $t(4)/\sqrt{2}$ | | | | | $(\chi^2(5)-5)/\sqrt{10}$ | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| $n$ | Method | TP | FP | CF | Est | Pred | TP | FP | CF | Est | Pred |
| 100 | SRLS | 10.0(0.00) | 0.00(0.00) | 1.00(0.00) | 0.20(0.01) | 0.10(0.00) | 9.99(0.10) | 0.01(0.10) | 0.98(0.14) | 0.20(0.01) | 0.10(0.00) |
| | SIS* | 4.75(1.27) | 1.79(1.62) | 0.00(0.00) | 0.84(0.19) | 0.38(0.09) | 4.84(1.39) | 1.92(1.65) | 0.00(0.00) | 0.82(0.20) | 0.37(0.09) |
| | ISIS* | 5.45(1.40) | 1.96(1.62) | 0.00(0.00) | 0.74(0.19) | 0.33(0.08) | 5.31(1.30) | 2.05(1.67) | 0.00(0.00) | 0.74(0.19) | 0.33(0.08) |
| | KAZZ | 5.42(1.48) | 1.18(1.34) | 0.00(0.00) | 0.64(0.24) | 0.30(0.12) | 5.00(1.66) | 1.15(1.37) | 0.00(0.00) | 0.67(0.25) | 0.32(0.12) |
| 200 | SRLS | 13.3(0.85) | 0.00(0.00) | 0.52(0.50) | 0.15(0.01) | 0.07(0.00) | 13.39(0.80) | 0.01(0.10) | 0.56(0.49) | 0.15(0.01) | 0.07(0.00) |
| | SIS* | 6.36(1.73) | 0.99(1.10) | 0.00(0.00) | 0.39(0.07) | 0.19(0.04) | 6.56(1.79) | 0.86(1.15) | 0.00(0.00) | 0.39(0.07) | 0.19(0.04) |
| | ISIS* | 7.29(1.79) | 1.30(1.42) | 0.00(0.00) | 0.36(0.07) | 0.17(0.03) | 7.11(1.86) | 1.37(1.32) | 0.00(0.00) | 0.36(0.07) | 0.17(0.04) |
| | KAZZ | 7.33(2.26) | 0.55(1.01) | 0.00(0.00) | 0.30(0.11) | 0.15(0.05) | 7.13(1.75) | 0.43(0.90) | 0.00(0.00) | 0.30(0.08) | 0.15(0.04) |

Table S4: The selection results and standard deviation (in parentheses) for Example 3.

## References

J. Barzilai and J. Borwein. Two-point step size gradient methods. *IMA Journal of Numerical Analysis*, 8: 141–148, 1988.

D. Bertsimas, A. King, and R. Mazumder. Best subset selection via a modern optimization lens. *The Annals of Statistics*, 44: 813–852, 2016.

P. Buhlmann and S. van de Geer. *Statistics for High-Dimensional Data: Methods, Theory and Applications.* Springer, New York, 2011.

F. Bunea, Y. She, and M. Wegkamp. Joint variable and rank selection for parsimonious estimation of high-dimensional matrices. *The Annals of Statistics*, 40: 2359–2388, 2012.

J. Chen and Z. Chen. Extended BIC for small-$n$-large-$p$ sparse glm. *Statistica Sinica*, 22: 555–574, 2012.

P. Chowriappa, P. Dua, and W. Lukiw. An exploratory analysis of conservation of co-expressed genes across Alzheimer's disease progression. *Journal of Computer Science and Systems Biology*, 6: 215–227, 2013.

R. Colom, J. Stein, P. Rajagopalan, K. Martínez, D. Hermel, Y. Wang, J. Álvarez-Linera, M. Burgaleta, M. Quiroga, P. Shih, and others. Hippocampal structure and human cognition: Key role of spatial processing and evidence supporting the efficiency hypothesis in females. *Intelligence*, 41: 129–140, 2013.

S. Ding. *Sufficient Dimension Reduction for Complex Data Structures (Ph.D. thesis)*. The University of Minnesota Digital Conservancy. Available electronically via http://hdl.handle.net/11299/164799, 2014.

S. Ding and R. Cook. Dimension folding PCA and PFC for matrix-valued predictors. *Statistica Sinica*, 24: 463–492, 2014.

S. Ding and D. Cook. Matrix variate regressions and envelope models. *Journal of the Royal Statistical Society: Series* B, 80: 387–408, 2018.

J. Fan, Y. Feng, and R. Song. Nonparametric independence screening in sparse ultra-high-dimensional additive models. *Journal of the American Statistical Association*, 106: 544–557, 2011.

J. Fan and J. Lv. Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society: Series* B, 70: 849–911, 2008.

J. Fan, R. Samworth, and Y. Wu. Ultrahigh dimensional feature selection: beyond the linear model. *Journal of Machine Learning Research*, 10: 2013–2038, 2009.

J. Fan and R. Song. Sure independence screening in generalized linear models with NP-dimensionality. *The Annals of Statistics*, 38: 3567–3604, 2010.

B. Fosdick and P. Hoff. Testing and modeling dependencies between a network and nodal attributes. *Journal of the American Statistical Association*, 110: 1047–1056, 2015.

P. Gong, C. Zhang, Z. Lu, J. Huang, and J. Ye. A general iterative shrinkage and thresholding algorithm for non-convex regularized optimization problems. *Proceedings of Machine Learning Research*, 28: 37–45, 2013.

K. He, H. Lian, S. Ma, and J. Huang. Dimensionality reduction and variable selection in multivariate varying-coefficient models with a large number of covariates. *Journal of the American Statistical Association*, 113: 746–754, 2018.

X. He, L. Wang, and H. Hong. Quantile-adaptive model-free variable screening for high-dimensional heterogeneous data. *The Annals of Statistics*, 41: 342–369, 2013.

W. Hu, T. Pan, D. Kong, and W. Shen. Nonparametric matrix response regression with application to brain imaging data analysis. *Biometrics*, doi: 10.1111/biom.13362, 2020.

W. Hu, W. Shen, H. Zhou, and D. Kong. Matrix linear discriminant analysis. *Technometrics*, 62: 196–205, 2020.

J. Huang, S. Ma, and C.-H. Zhang. Adaptive lasso for sparse high dimensional regression models. *Statistica Sinica*, 18: 1603–1618, 2008.

D. Kong, B. An, J. Zhang, and H. Zhu. L2RM: Low-rank linear regression models for high-dimensional matrix responses. *Journal of the American Statistical Association*, 115: 403–424, 2020.

C. Leng and C. Tang. Sparse matrix graphical models. *Journal of the American Statistical Association*, 107: 1187–1200, 2012.

B. Li, K. Kim, and N. Altman. On dimension folding of matrix or array-valued statistical objects. *The Annals of Statistics*, 38: 1094–1121, 2010.

E. Liu, M. Li, W. Wang, and Y. Li. MaCH-Admix: Genotype imputation for admixed populations. *Genetic Epidemiology*, 37: 25–37, 2013.

J. Liu, R. Li, and R. Wu. Feature selection for varying coefficient models with ultrahigh-dimensional covariates. *Journal of the American Statistical Association*, 109: 266–274, 2014.

E. Luders, P. Thompson, F. Kurth, J.-Y. Hong, O. Phillips, Y. Wang, B. Gutman, Y.-Y. Chou, K. Narr, and A. Toga. Global and regional alterations of hippocampal anatomy in long-term meditation practitioners. *Human Brain Mapping*, 34: 3369–3375, 2013.

M. Monje, M. Thomason, L. Rigolo, Y. Wang, D. Waber, S. Sallan, and A. Golby. Functional and structural differences in the hippocampus associated with memory deficits in adult survivors of acute lymphoblastic leukemia. *Pediatric Blood and Cancer*, 60: 293–300, 2013.

F. Nathoo, L. Kong, and H. Zhu. A review of statistical methods in imaging genetics. *Canadian Journal of Statistics*, 47: 108-131, 2019.

Y. Nesterov. Introductory lectures on convex optimization: A basic course. Applied optimization 87. Boston MA, Kluwer Academic, 2004.

G. Raskutti, M. Wainwright, and B. Yu. Minimax rates of estimation for high-dimensional linear regression over $\ell_q$-balls. *IEEE Transactions on Information Theory*, 57: 6976–6994, 2011.

S. Sadigh-Eteghad, B. Sabermarouf, A. Majdi, M. Talebi, M. Farhoudi, and J. Mahmoudi. Amyloid-beta: a crucial factor in Alzheimer's disease. *Medical Principles and Practice*, 24: 1–10, 2015.

J. Sánchez-Valle, H. Tejero, K. lbáñez, J.-L. Portero, M. Krallinger, F. Al-Shahrour, R. Tabarés-Seisdedos, A. Baudot, and A. Valencia. A molecular hypothesis to explain direct and inverse co-morbidities between Alzheimer's disease, glioblastoma and lung cancer. *Scientific Reports*, 7: 1–12, 2017.

G. Schwarz. Estimating the dimension of a model. *The Annals of Statistics*, 6: 461–464, 1978.

J. Shi, P. Thompson, B. Gutman, and Y. Wang. Surface fluid registration of conformal representation: application to detect disease burden and genetic influence on hippocampus. *NeuroImage*, 78: 111–134, 2014.

The 1000 Genomes Project Consortium. A global reference for human genetic variation. *Nature*, 526: 68–74, 2015.

A. Tsybakov. *Introduction to Nonparametric Estimation*. Springer, New York, 2009.

C. Viroli. On matrix-variate regression analysis. *Journal of Multivariate Analysis*, 111: 296–309, 2012.

H. Wang. Forward regression for ultra-high dimensional variable screening. *Journal of the American Statistical Association*, 104: 1512–1524, 2009.

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

Y. Wang, Y. Song, P. Rajagopalan, T. An, K. Liu, Y.-Y. Chou, B. Gutman, A. Toga, P. Thompson, ADNI, and others. Surface-based TBM boosts power to detect disease effects on the brain: an N= 804 ADNI study. *NeuroImage*, 56: 1993–2010, 2011.

K. Wirz, K. Bossers, A. Stargardt, W. Kamphuis, D. Swaab, E. Hol, and J. Verhaagen. Cortical beta amyloid protein triggers an immune response, but no synaptic changes in the APPswe/PS1dE9 Alzheimer's disease mouse model. *Neurobiology of Aging*, 34: 1328–1342, 2013.

C. Xu and J. Chen. The sparse MLE for ultrahigh-dimensional feature screening. *Journal of the American Statistical Association*, 109: 1257–1269, 2014.

M. Yuan, A. Ekici, Z. Lu, and R. Monteiro. Dimension reduction and coefficient estimation in multivariate linear regression. *Journal of the Royal Statistical Society: Series* B, 69: 329–346, 2007.

S. Yun, P. Tseng, and K.-C. Toh. A block coordinate gradient descent method for regularized convex separable optimization and covariance selection. *Mathematical programming*, 129: 331–355, 2011.

D. Zhao and Y. Li. Principled sure independence screening for Cox models with ultra-high-dimensional covariates. *Journal of Multivariate Analysis*, 105: 397–411, 2012.

J. Zhao and C. Leng. Sparse matrix graphical models. *Statistica Sinica*, 24: 799–814, 2014.

H. Zhou and L. Li. Regularized matrix regression. *Journal of the Royal Statistical Society: Series* B, 76: 463–483, 2014.

H. Zhu, J. Fan, and L. Kong. Spatially varying coefficient model for neuroimaging data with jump discontinuities. *Journal of the American Statistical Association*, 109: 1084–1098, 2014.

L. Zhu, L. Li, R. Li, and L. Zhu. Model-free feature screening for ultrahigh-dimensional data. *Journal of the American Statistical Association*, 106: 1464–1474, 2012.