



HHS Public Access

Author manuscript

J Am Stat Assoc. Author manuscript; available in PMC 2021 April 30.

Published in final edited form as:

J Am Stat Assoc. 2020 April 30; 115(529): 403–424. doi:10.1080/01621459.2018.1555092.

L2RM: Low-rank Linear Regression Models for High-dimensional Matrix Responses *

Dehan Kong,

Department of Statistical Sciences, University of Toronto

Baiguo An,

School of Statistics, Capital University of Economics and Business

Jingwen Zhang,

Department of Biostatistics, University of North Carolina at Chapel Hill

Hongtu Zhu

Department of Biostatistics, University of North Carolina at Chapel Hill

Abstract

The aim of this paper is to develop a low-rank linear regression model (L2RM) to correlate a high-dimensional response matrix with a high dimensional vector of covariates when coefficient matrices have low-rank structures. We propose a fast and efficient screening procedure based on the spectral norm of each coefficient matrix in order to deal with the case when the number of covariates is extremely large. We develop an efficient estimation procedure based on the trace norm regularization, which explicitly imposes the low rank structure of coefficient matrices. When both the dimension of response matrix and that of covariate vector diverge at the exponential order of the sample size, we investigate the sure independence screening property under some mild conditions. We also systematically investigate some theoretical properties of our estimation procedure including estimation consistency, rank consistency and non-asymptotic error bound under some mild conditions. We further establish a theoretical guarantee for the overall solution of our two-step screening and estimation procedure. We examine the finite-sample performance of our screening and estimation methods using simulations and a large-scale imaging genetic dataset collected by the Philadelphia Neurodevelopmental Cohort (PNC) study.

Keywords

Imaging Genetics; Low Rank; Matrix Linear Regression; Spectral norm; Trace norm

*Address for correspondence and reprints: Hongtu Zhu, Ph.D., htzhu@email.unc.edu; Phone No: 919-929-9010. Dr. Hongtu Zhu is Professor, Department of Biostatistics, The University of North Carolina at Chapel Hill, Chapel Hill, NC, USA. Dr. Dehan Kong is Assistant Professor, Department of Statistical Sciences, University of Toronto, Toronto, ON, Canada. (kongdehan@utstat.toronto.edu).

Supplementary Material

Supplementary Material available online includes modified algorithm for our regularized low rank estimation procedure when response and covariates are not centered and additional simulation results.

1 Introduction

Multivariate regression modeling with a matrix response $\mathbf{Y} \in \mathbb{R}^{p \times q}$ and a multivariate covariate $\mathbf{x} \in \mathbb{R}^s$ is an important statistical tool in modern high-dimensional inference, with wide applications in various large-scale applications, such as imaging genetic studies. Specifically, in imaging genetics, matrix responses (\mathbf{Y}) as phenotypic variables often represent the weighted (or binary) adjacency matrix of a finite graph for characterizing structural (or functional) connectivity pattern, whereas covariates (\mathbf{x}) include genetic markers (e.g., single nucleotide polymorphisms (SNPs)), age, and gender, among others. The joint analysis of imaging and genetic data may ultimately lead to discoveries of genes for many neuropsychiatric and neurological disorders, such as schizophrenia (Scharinger et al., 2010; Peper et al., 2007; Chiang et al., 2011; Thompson et al., 2013; Medlan et al., 2014). This motivates us to systematically investigate a statistical model with a multivariate response \mathbf{Y} and a multivariate covariate \mathbf{x} .

Let $\{(\mathbf{x}_i, \mathbf{Y}_i) : 1 \leq i \leq n\}$ denote independent and identically distributed (i.i.d.) observations, where $\mathbf{x}_i = (x_{i1}, \dots, x_{is})^T$ is a $s \times 1$ vector of scalar covariates (e.g., clinical variables and genetic variants) and \mathbf{Y}_i is a $p \times q$ response matrix. Without loss of generality, we assume that x_{il} has mean 0 and variance 1 for every $1 \leq l \leq s$, and \mathbf{Y}_i has mean $\mathbf{0}$. Throughout the paper, we consider a L2RM, which is given by

$$\mathbf{Y}_i = \sum_{l=1}^s x_{il} * \mathbf{B}_l + \mathbf{E}_i, \quad (1)$$

where \mathbf{B}_l is a $p \times q$ coefficient matrix characterizing the effect of the l th covariate on \mathbf{Y}_i and \mathbf{E}_i is a $p \times q$ matrix of random errors with mean $\mathbf{0}$. The symbol “ $*$ ” denotes the scalar multiplication. Model (1) differs significantly from the existing matrix regression, which was developed for matrix covariates and univariate responses (Leng and Tang, 2012; Zhao and Leng, 2014; Zhou and Li, 2014). Our goal is to discover a small set of important covariates from \mathbf{x} that strongly influence \mathbf{Y} .

We focus on the most challenging setting that both the dimension of \mathbf{Y} (or pq) and that of \mathbf{x} (or s) can diverge with the sample size. Such a setting is general enough to cover high-dimensional univariate and multivariate linear regression models in the literature (Negahban et al., 2012; Fan and Lv, 2010; Buhlmann and van de Geer, 2011; Tibshirani, 1997; Yuan et al., 2007; Candes and Tao, 2007; Breiman and Friedman, 1997; Cook et al., 2013; Park et al., 2017). In the literature, there are two major categories of statistical methods for jointly analyzing high-dimensional matrix \mathbf{Y} and high-dimensional vector \mathbf{x} .

The first category is a set of mass univariate methods. Specifically, it fits a marginal linear regression to correlate each element of \mathbf{Y}_i with each element of \mathbf{x}_i , leading to a total of pqs massive univariate analyses and an expanded search space with pqs elements. It is also called voxel-wise genome-wide association analysis (VGAWS) in the imaging genetics literature (Hibar et al., 2011; Shen et al., 2010; Huang et al., 2015; Zhang et al., 2014; Medland et al., 2014; Zhang et al., 2014; Thompson et al., 2014; Liu and Calhoun, 2014). For instance, Stein et al. (2010) used 300 high performance CPU nodes to run approximately

27 hours to carry out a VGWAS analysis on an imaging genetic dataset with only 448,293 SNPs and 31,622 imaging measures for 740 subjects. Such computational challenges are becoming more severe as the field is rapidly advancing to the most challenging setting with large pq and s . More seriously, for model (1), the massive univariate method can miss some important components of \mathbf{x} that strongly influence \mathbf{Y} due to the interaction among \mathbf{x} .

The second category is to fit a model accommodating all (or part of) covariates and responses (Vounou et al., 2010, 2012; Zhu et al., 2014; Wang et al., 2012a,b; Peng et al., 2010). These methods use regularization methods, such as Lasso or group Lasso, to select a set of covariate-response pairs. However, when the product pqs is extremely large, it is very difficult to allocate computer memory for such an array of size pqs in order to accommodate all coefficient matrices \mathbf{B}_s , rendering all these regularization methods being intractable. Therefore, almost all existing methods in this category have to use some dimension reduction techniques (e.g., screening methods) to reduce both the number of responses and that of covariates. Subsequently, these methods fit a multivariate linear regression model with the selected elements of \mathbf{Y} as new responses and those of \mathbf{x} as new covariates. However, this approach can be unsatisfactory, since it does not incorporate the matrix structural information.

The aim of this paper is to develop a low-rank linear regression model (L2RM) as a novel extension of both VGWAS and regularization methods. Specifically, instead of repeatedly fitting a univariate model to each covariate-response pair, we consider all elements in \mathbf{Y}_j as a high-dimensional matrix response and focus on the coefficient matrix of each covariate, which is approximately low-rank (Candès and Recht, 2009). There is a literature on the development of matrix variate regression (Ding and Cook, 2014; Fosdick and Hoff, 2015; Zhou and Li, 2014), but these papers focus on the case when covariates have a matrix structure. In contrast, there is a large literature on the development of various function-on-scalar regression models that emphasize the inherent functional structure of responses. See Chapter 13 of Ramsay and Silverman (2005) for a comprehensive review on this topic. Variable selection methods have been developed for some function-on-scalar regression models (Wang et al., 2007; Chen et al., 2016), but these methods focus on one-dimensional functional response rather than two-dimensional matrix response. Recently, there has been some literature considering matrix or tensor responses regression (Ding and Cook, 2018; Li and Zhang, 2017; Raskutti and Yuan, 2018; Rabusseau and Kadri, 2016), but they only consider the case when the dimension of the covariates is fixed or slowly diverging with the sample size.

In this paper, we aim at efficiently correlating matrix responses with a high dimensional vector of covariates. Four major methodological contributions of this paper are as follows.

- We introduce a low-rank linear regression model to fit high-dimensional matrix responses with a high dimensional vector of covariates, while explicitly accounting for the low-rank structure of coefficient matrices.
- We introduce a novel rank-one screening procedure based on the spectral norm of the estimated coefficient matrix to eliminate most “noisy” scalar covariates and show that our screening procedure enjoys the sure independence screening

property (Fan and Lv, 2008) with vanishing false selection rate. The use of such spectral norm is critical for dealing with a large number of noisy covariates.

- When the number of covariates is relatively small, we propose a low rank estimation procedure based on trace norm regularization, which explicitly characterizes the low-rank structure of coefficient matrices. An efficient algorithm for solving the optimization problem is developed. We systematically investigate some theoretical properties of our estimation procedure, including estimation and rank consistency when both p and q are fixed and a non-asymptotic error bound when both p and q are allowed to diverge.
- We investigate how incorrectly screening results can affect the low-rank regression model estimation both numerically and theoretically. We establish a theoretical guarantee for the overall solution, while accounting for the randomness of the first-step screening procedure.

The rest of this paper is organized as follows. In Section 2, we introduce a rank-one screening procedure to deal with a high dimensional vector of covariates and describe our estimation procedure when the number of covariates is relatively small. Section 3 investigates the theoretical properties of our method. Simulations are conducted in Section 4 to evaluate the finite-sample performance of the proposed two-step screening and estimation procedure. Section 5 illustrates an application of L2RM in the joint analysis of imaging and genetic data from the Philadelphia Neurodevelopmental Cohort (PNC) study discussed above. We finally conclude with some discussions in Section 6.

2 Methodology

Throughout the paper, we focus on addressing three fundamental issues for L2RM as follows:

- (I) The first one is to eliminate most ‘noisy’ covariates x_{jl} when the number of candidate covariates and that of response matrix are much larger than n , that is $\min(s, pq) \gg n$.
- (II) The second one is to estimate the coefficient matrix \mathbf{B}_J when \mathbf{B}_J does have a low-rank structure.
- (III) The third one is to investigate some theoretical properties of the screening and estimation methods.

2.1 Rank-one Screening Method

We consider the case that both pq and s diverge at an exponential order of n , and we also denote s by s_n . To address (I), it is common to assume that most scalar covariates have no effects on the matrix responses, that is, $\mathbf{B}_{l0} = \mathbf{0}$ for most $1 \leq l \leq s_n$, where \mathbf{B}_{l0} is the true value for \mathbf{B}_J . In this case, we define the true model and its size as

$$\mathcal{M} = \{1 \leq l \leq s_n: \mathbf{B}_{l0} \neq \mathbf{0}\} \text{ and } s_0 = |\mathcal{M}| < n. \quad (2)$$

Our aim is to estimate the set \mathcal{M} and coefficient matrices \mathbf{B}_l . Simultaneously estimating \mathcal{M} and \mathbf{B}_l is difficult since it is computationally infeasible to fit a model when both s_n and pq are quite high. For example, in the PNC data, we have $pq = 69^2 = 4,761$ and $s_n \approx 5 \times 10^6$. Therefore, it may be imperative to employ a screening technique to reduce the model size. However, developing a screening technique for model (1) can be more challenging than many existing screening methods, which focus on univariate responses (Fan and Lv, 2008; Fan and Song, 2010).

Similar to Fan and Lv (2008) and Fan and Song (2010), it is assumed that all covariates have been standardized so that

$$E(x_{il}) = 0 \text{ and } E(x_{il}^2) = 1 \text{ for } l = 1, \dots, s_n.$$

We also assume that every element of $\mathbf{Y}_i = (Y_{i,jk})$ has been standardized, that is,

$$E(Y_{i,jk}) = 0 \text{ and } E(Y_{i,jk}^2) = 1 \text{ for } j = 1, \dots, p \text{ and } k = 1, \dots, q.$$

We propose to screen covariates based on the estimated marginal ordinary least squares (OLS) coefficient matrix $\widehat{\mathbf{B}}_l^M = n^{-1} \sum_{i=1}^n x_{il} * \mathbf{Y}_i$ for $l = 1, \dots, s_n$. Although the interpretations and implications of the marginal models are biased from the joint model, the nonsparse information about the joint model can be passed along to the marginal model under a mild condition. Hence it is suitable for the purpose of variable screening (Fan and Song, 2010). Specifically, we calculate the spectral norm (operator norm or largest singular value) of $\widehat{\mathbf{B}}_l^M$, denoted as $\|\widehat{\mathbf{B}}_l^M\|_{op}$, and define a submodel as

$$\widehat{\mathcal{M}}_{\gamma_n} = \{1 \leq l \leq s_n: \|\widehat{\mathbf{B}}_l^M\|_{op} \geq \gamma_n\}, \quad (3)$$

where γ_n is a prefixed threshold.

The key advantage of using $\|\widehat{\mathbf{B}}_l^M\|_{op}$ is that it explicitly accounts for the low-rank structure of \mathbf{B}_l s for most noisy covariates, while being robust to noise and more sensitive to various signal patterns (e.g., sparsely strong signals and low rank weak signals) in coefficient matrices. In our screening step, we use the marginal OLS estimates of the coefficient matrices, which can be regarded as the true coefficient matrices corrupted with some noise. One may directly use some other summary statistics of $\widehat{\mathbf{B}}_l^M$ based on the component-wise information of $\widehat{\mathbf{B}}_l^M$, such as $\|\widehat{\mathbf{B}}_l^M\|_1$ (sum of the absolute value of all the elements), $\|\widehat{\mathbf{B}}_l^M\|_F$, or the global Wald-type statistic used in Huang et al. (2015). It is well known that those summary statistics are sensitive to noise and suffer from the curse of dimensionality. This is further confirmed in our simulation studies that our rank-one screening based on $\|\widehat{\mathbf{B}}_l^M\|_{op}$ is more robust to noise and sensitive to small signal regions. Moreover, the other advantage of using $\|\widehat{\mathbf{B}}_l^M\|_{op}$ is that it is computationally efficient. In contrast, we may calculate some other

regularized estimates (e.g., Lasso or fused Lasso) for screening, but it is computationally infeasible for L2RM when s_n is much larger than the sample size.

A difficult issue in (3) is how to properly select γ_n . As shown in Section 3.1, when γ_n is chosen properly, our screening procedure enjoys the sure independence property (Fan and Lv, 2008). However, it is difficult to precisely determine γ_n in practice since it involves in two unknown positive constant terms C_1 and α as shown in Theorem 1, which cannot be easily determined for finite sample. We propose to use random decoupling to select γ_n , which is similar to that used in Barut et al. (2016). Let $\{x_i^*, i = 1, \dots, n\}$ be a random permutation of the original data $(\mathbf{x}_j = 1, \dots, n)$. We apply our screening procedure on the random decoupling data $\{x_i^*, \mathbf{Y}_i\}_{i=1}^n$. As the original association between \mathbf{x}_j and \mathbf{Y}_j is destroyed by random decoupling, when we perform screening using $\{x_i^*, \mathbf{Y}_i\}_{i=1}^n$, it mimics the null model, i.e. the model when there is no association. We obtain the estimated marginal coefficient matrix $(\widehat{\mathbf{B}}_l^M)^*$, which is a statistical estimate of zero matrix, and the corresponding operator norm $\|(\widehat{\mathbf{B}}_l^M)^*\|_{op}$ for all $1 \leq l \leq s_n$. Define $v_n = \max_{1 \leq l \leq s_n} \|(\widehat{\mathbf{B}}_l^M)^*\|_{op}$, which is the minimum thresholding parameter that makes no false positives. Since v_n depends on the realization of the permutation, we set the threshold value γ_n as the median of these threshold values $\{v_n^{(k)}, 1 \leq k \leq K\}$ from K different random permutations, where $v_n^{(k)}$ is the threshold value for the k th permutation. We set $K = 10$ in this paper.

2.2 Estimation Method

To address (II), we consider the estimation of \mathbf{B} when the true coefficient matrices \mathbf{B}_l s truly have a low-rank structure. The following refined estimation step can be applied after the screening step when the number of covariates is relatively small. For simplicity, we denote the set selected by the screening step $\widehat{\mathcal{M}}_{\gamma_n}$ by $\widehat{\mathcal{M}}$. Suppose $\widehat{\mathcal{M}} = \{l_1, \dots, l_{|\widehat{\mathcal{M}}|}\}$, where $1 \leq l_1 < \dots < l_{|\widehat{\mathcal{M}}|} \leq s_n$. Define $\mathbf{B} = [\mathbf{B}_l, l \in \widehat{\mathcal{M}}] = [\mathbf{B}_{l_1}, \dots, \mathbf{B}_{l_{|\widehat{\mathcal{M}}|}}] \in \mathbb{R}^{p \times q \times |\widehat{\mathcal{M}}|}$.

Recently, the trace norm regularization $\|\mathbf{B}\|_* = \sum_k \sigma_k(\mathbf{B})$ has been widely used to recover the low-rank structure of \mathbf{B}_l due to its computational efficiency, where $\alpha_k(\mathbf{B}_l)$ is the k th singular value of \mathbf{B}_l . For instance, the trace norm has been used for matrix completion (Candès and Recht, 2009), for matrix regression models with matrix covariates and univariate responses (Zhou and Li, 2014), and for multivariate linear regression with vector responses and scalar covariates (Yuan et al., 2007). Similarly, we propose to calculate the regularized least squares estimator of \mathbf{B} by minimizing

$$Q(\mathbf{B}) = \frac{1}{2n} \sum_{i=1}^n \left\| \mathbf{Y}_i - \sum_{l \in \widehat{\mathcal{M}}} x_{il} * \mathbf{B}_l \right\|_F^2 + \lambda \sum_{l \in \widehat{\mathcal{M}}} \|\mathbf{B}_l\|_*, \tag{4}$$

where $\|\cdot\|_F$ is the Frobenius norm of a matrix and λ is a tuning parameter. The low rank structure can be regarded as a special spatial structure, since it is very similar to functional principal component analysis. We use the five-fold cross validation to select the tuning

parameter λ . Ideally, we may choose different tuning parameters for different \mathbf{B}_l but it can dramatically increase computational complexity.

We apply the Nesterov gradient method to solve problem (4) even though $Q(\mathbf{B})$ is non-smooth (Nesterov, 2004; Beck and Teboulle, 2009). The Nesterov gradient method utilizes the first-order gradient of the objective function to obtain the next iterate based on the current search point. Unlike the standard gradient descent algorithm, the Nesterov gradient algorithm uses two previous iterates to generate the next search point by extrapolating, which can dramatically improve the convergence rate. Before we introduce the Nesterov gradient algorithm, we first state a singular value thresholding formula for the trace norm (Cai et al., 2010).

Proposition 1. For a matrix \mathbf{A} with $\{a_k\}_{k=1}^r$ being its singular values, the solution to

$$\min_{\mathbf{B}} \left\{ \frac{1}{2} \|\mathbf{B} - \mathbf{A}\|_F^2 + \lambda \|\mathbf{B}\|_* \right\} \quad (5)$$

shares the same singular vectors as \mathbf{A} and its singular values are $b_k = (a_k - \lambda)_+$ for $k = 1, \dots, r$.

We present the Nesterov gradient algorithm for problem (4) as follows. Denote

$R(\mathbf{B}) = (2n)^{-1} \sum_{i=1}^n \|\mathbf{Y}_i - \sum_{l \in \widehat{\mathcal{M}}} x_{il} * \mathbf{B}_l\|_F^2$ and $J(\mathbf{B}) = \lambda \sum_{l \in \widehat{\mathcal{M}}} \|\mathbf{B}_l\|_*$. We also define

$$\begin{aligned} g(\mathbf{B} | \mathbf{S}^{(t)}, \delta) &= R(\mathbf{S}^{(t)}) + \langle \nabla R(\mathbf{S}^{(t)}), \mathbf{B} - \mathbf{S}^{(t)} \rangle + (2\delta)^{-1} \|\mathbf{B} - \mathbf{S}^{(t)}\|_F^2 + J(\mathbf{B}) \\ &= (2\delta)^{-1} \|\mathbf{B} - [\mathbf{S}^{(t)} - \delta \nabla R(\mathbf{S}^{(t)})]\|_F^2 + J(\mathbf{B}) + c^{(t)}, \end{aligned}$$

where $\nabla R(\mathbf{S}^{(t)})$ denotes the first-order gradient of $R(\mathbf{S}^{(t)})$ with respect to $\mathbf{S}^{(t)}$, $\mathbf{S}^{(t)}$ is an interpolation between $\mathbf{B}^{(t)}$ and $\mathbf{B}^{(t-1)}$ and will be defined below, $c^{(t)}$ denotes all terms that are irrelevant to \mathbf{B} , and $\delta > 0$ is a suitable step size. Given a previous search point $\mathbf{S}^{(t)}$, the next search point $\mathbf{S}^{(t+1)}$ would be the minimizer of $g(\mathbf{B} | \mathbf{S}^{(t)}, \delta)$. For the search point $\mathbf{S}^{(t)}$, it can be generated by linearly extrapolating two previous algorithmic iterates. A key advantage of using the Nesterov gradient method is that it has an explicit solution at each iteration.

Specifically, let \mathbf{B}_{l_d} , $\mathbf{S}_{l_d}^{(t)}$, and $\nabla R(\mathbf{S}^{(t)})_{l_d}$ be the $(dq - q + 1)$ th to the dq th columns of the corresponding $p \times q$ $|\widehat{\mathcal{M}}|$ matrices \mathbf{B} , $\mathbf{S}^{(t)}$, and $\nabla R(\mathbf{S}^{(t)})$, respectively. Minimizing $(2\delta)^{-1} \|\mathbf{B} - [\mathbf{S}^{(t)} - \delta \nabla R(\mathbf{S}^{(t)})]\|_F^2 + \lambda \sum_{l \in \widehat{\mathcal{M}}} \|\mathbf{B}_l\|_*$ is equivalent to solving $|\widehat{\mathcal{M}}|$ sub-problems, each of which minimizes $(2\delta)^{-1} \|\mathbf{B}_{l_d} - [\mathbf{S}_{l_d}^{(t)} - \delta \nabla R(\mathbf{S}^{(t)})_{l_d}]\|_F^2 + \lambda \|\mathbf{B}_{l_d}\|_*$ for $d = 1, \dots, |\widehat{\mathcal{M}}|$, while each sub-problem can be exactly solved by using the singular value thresholding formula given in Proposition 1.

Define $\mathbf{X}_{\widehat{\mathcal{M}}} = (x_{il})_{1 \leq i \leq n, l \in \widehat{\mathcal{M}}}$ is an $n \times |\widehat{\mathcal{M}}|$ matrix and $\lambda_{\max}(\cdot)$ denotes the largest eigenvalue of a matrix. Our algorithm can be stated as follows:

1. Initialize $\mathbf{B}^{(0)} = \mathbf{B}^{(1)}$, $\alpha^{(0)} = 0$ and $\alpha^{(1)} = 1$, $t = 1$, and $\delta = n / \lambda_{\max}\{(\mathbf{X}_{\widehat{\mathcal{M}}})^T \mathbf{X}_{\widehat{\mathcal{M}}}\}$.

2. Repeat

$$\mathbf{S}^{(t)} = \mathbf{B}^{(t)} + \left(\frac{\alpha^{(t)} - 1}{\alpha^{(t)}}\right)(\mathbf{B}^{(t)} - \mathbf{B}^{(t-1)});$$

for $d = 1: |\widehat{\mathcal{M}}|$,

i. $(\mathbf{A}_{temp})_{l_d} = \mathbf{S}_{l_d}^{(t)} - \delta \nabla R(\mathbf{S}^{(t)})_{l_d}$;

ii. compute singular value decomposition (SVD) $(\mathbf{A}_{temp})_{l_d} = \mathbf{U}_{l_d} \text{diag}(a_{l_d}) \mathbf{V}_{l_d}^T$;

iii. $\mathbf{b}_{l_d} = a_{l_d} - \lambda \delta * \mathbf{1}$;

iv. $(\mathbf{B}_{temp})_{l_d} = \mathbf{U}_{l_d} \text{diag}(\mathbf{b}_{l_d}) \mathbf{V}_{l_d}^T$;

end

Combine $\{(\mathbf{B}_{temp})_{l_d}, 1 \leq d \leq |\widehat{\mathcal{M}}|\}$ sub-matrices and get the entire matrix \mathbf{B}_{temp} ;

$$\mathbf{B}^{(t+1)} = \mathbf{B}_{temp};$$

$$\alpha^{(t+1)} = \{1 + \sqrt{1 + (2\alpha^{(t)})^2}\} / 2;$$

$$t = t + 1;$$

3. until objective function $Q(\mathbf{B}^{(t)})$ converges.

For the above $p \times q \times |\widehat{\mathcal{M}}|$ matrices \mathbf{A}_{temp} and \mathbf{B}_{temp} , $(\mathbf{A}_{temp})_{l_d}$ and $(\mathbf{B}_{temp})_{l_d}$ denote the $(dq - q + 1)$ th to the (dq) th columns of the corresponding matrices, respectively.

A sufficient condition for the convergence of $\{\mathbf{B}^{(t)}\}_{t=1}$ is that the step size δ should be smaller than or equal to $1/L(R)$, where $L(R)$ is the smallest Lipschitz constant of the function $R(\mathbf{B})$ (Beck and Teboulle, 2009; Facchinei and Pang, 2003). In our case, $L(R)$ is equal to $n^{-1} \lambda_{\max}\{(\mathbf{X}_{\widehat{\mathcal{M}}})^T \mathbf{X}_{\widehat{\mathcal{M}}}\}$.

Remarks: For model (1), it is assumed that x_{ij} has mean 0 and variance 1 for every $1 \leq i \leq s$, and \mathbf{Y}_j has mean $\mathbf{0}$ throughout the paper. If these assumptions are not valid in practice, a simple solution is to carry out a standardization step including standardizing covariates and centering responses. We use this approach in simulations and real data analysis. An alternative approach is to introduce an intercept matrix term \mathbf{B}_0 in model (1). Our screening procedure is invariant to such standardization step if we calculate $B_{i,jk}^M$, the (j, k) -th element of \mathbf{B}_j , as the sample correlation between x_{ij} and $Y_{i,jk}$. In the Supplementary Material, we present a modified algorithm of our estimation procedure and evaluate the effects of the standardization step on estimating \mathbf{B}_1 by using simulations. According to our experience, scaling covariates is necessary in order to ensure that all covariates are at the same scale, whereas centering covariates and responses is not critical.

3 Theoretical Properties

To address (III), we systematically investigate several key theoretical properties of the screening procedure and the regularized estimation procedure as well as a theoretical guarantee of our two-step estimator. First, we investigate the sure independence screening property of the rank-one screening procedure when s (also denoted by s_n) diverges at an exponential rate of the sample size. Second, we investigate the estimation and rank

consistency of our regularized estimator when both p and q are fixed. Third, we derive the non-asymptotic error bound for our estimator when both p and q are diverging. Finally, we establish an overall theoretical guarantee for our two-step estimator. We state the following theorems, whose detailed proofs can be found in the Appendix B.

3.1 Sure Screening Property

The following assumptions are used to facilitate the technical details, even though they may not be the weakest conditions but help to simplify the proof.

(A0) The covariates \mathbf{X}_j are i.i.d from a distribution with mean $\mathbf{0}$ and covariance matrix Σ_x . Define $\sigma_l^2 = (\Sigma_x)_{ll}$. The vectorized error matrices $\text{vec}(\mathbf{E}_j)$ are i.i.d from a distribution with mean $\mathbf{0}$ and covariance matrix Σ_e , where $\text{vec}(\cdot)$ denotes the vectorization of a matrix. Moreover, \mathbf{x}_j and $\mathbf{E}_j = (E_{i,jk})$ are independent.

(A1) There exist some constants $C_1 > 0$, $b > 0$, and $0 < \kappa < 1/2$ such that

$$\min_{l \in \mathcal{M}} \|\text{cov}(\sum_{l' \in \mathcal{M}} x_{il'} * \mathbf{B}_{l'0}, x_{il})\|_{op} \geq C_1(pq)^{1/2} n^{-\kappa} \text{ and } \max_{l \in \mathcal{M}} \|\mathbf{B}_{l0}\|_{\infty} < b,$$

where $\text{cov}(\sum_{l' \in \mathcal{M}} x_{il'} * \mathbf{B}_{l'0}, x_{il})$ is a $p \times q$ matrix with the (j, k) th element being $\text{cov}(\sum_{l' \in \mathcal{M}} x_{il'} * \mathbf{B}_{l'0, jk}, x_{il})$, and $\|\mathbf{B}_{l0}\|_{\infty} = \max_{1 \leq j \leq p, 1 \leq k \leq q} |B_{l0, jk}|$.

(A2) There exist positive constants C_2 and C_3 such that

$$\max(\mathbb{E}\{\exp(C_2 x_{ij}^2)\}, \mathbb{E}\{\exp(C_2 E_{i,jk}^2)\}) \leq C_3$$

for every $1 \leq l \leq s_n$, $1 \leq j \leq p$ and $1 \leq k \leq q$.

(A3) There exists a constant $C_4 > 0$ such that $\log(s_n) = C_4 n^{\xi}$ for $\xi \in (0, 1 - 2\kappa)$.

(A4) There exist constants $C_5 > 0$ and $\tau > 0$ such that $\lambda_{\max}(\Sigma_x) \leq C_5 n^{\tau}$.

(A5) We assume $\log(pq) = \alpha n^{1-2\kappa}$.

Remarks: Assumptions (A0)-(A5) are used to establish the theory of our screening procedure when s_n diverges to infinity. Assumption (A1) is analogous to Condition 3 in Fan and Lv (2008) and equation (4) in Fan and Song (2010), in which κ controls the rate of probability error in recovering the true sparse model. Assumption (A2) is analogous to Condition (D) in Fan and Song (2010) and Condition (E) in Fan et al. (2011). Assumption (A2) requires that x_{il} and $E_{i,jk}$ are sub-gaussian, which ensures the tail probability to be exponentially light. Assumption (A3) allows the dimension s_n to diverge at an exponential rate of the sample size n , which is analogous to Condition 1 in Fan and Lv (2008). Assumption (A4) is analogous to Condition 4 in Fan and Lv (2008), which rules out the case of strong collinearity. Assumption (A5) allows the product of the row and column dimensions of the matrix pq to diverge at an exponential rate of the sample size n .

The following theorems show the sure screening property of the screening procedure. We allow p and q to be either fixed or diverging with sample size n .

Theorem 1. *Under Assumptions (A0)-(A3) and (A5), let $\gamma_n = \alpha C_1(pq)^{1/2} n^{-\kappa}$ with $0 < \alpha < 1$, then we have $P(\mathcal{M} \subseteq \widehat{\mathcal{M}}_{\gamma_n}) \rightarrow 1$ as $n \rightarrow \infty$.*

Theorem 1 shows that if γ_n is chosen properly, then our rank-one screening procedure will not miss any significant variables with an overwhelming probability. Since the screening procedure automatically includes all the significant covariates for small values of γ_n , it is necessary to consider the size of $\widehat{\mathcal{M}}_{\gamma_n}$ when $\gamma_n = \alpha C_1(pq)^{1/2} n^{-\kappa}$ holds.

Theorem 2. *Under Assumptions (A0)-(A5), we have $P(|\widehat{\mathcal{M}}_{\gamma_n}| = O(n^{2\kappa + \tau})) \rightarrow 1$ for $\gamma_n = \alpha C_1(pq)^{1/2} n^{-\kappa}$ with $0 < \alpha < 1$ as $n \rightarrow \infty$.*

Theorem 2 indicates that the selected model size with the sure screening property is only at a polynomial order of n , even though the original model size is at an exponential order of n . Therefore, the false selection rate of our screening procedure vanishes as $n \rightarrow \infty$.

3.2 Theory for Estimation Procedure

From this subsection, we will denote $\widehat{\mathcal{M}}_{\gamma_n}$ by $\widehat{\mathcal{M}}$ for notation simplicity. We first provide some theoretical results for our estimation procedure. We assume that we can exactly select all the important variables in \mathcal{M} , i.e. $\widehat{\mathcal{M}} = \mathcal{M}$, and $s_0 = |\mathcal{M}|$ is fixed. The results are also applicable if our original s is fixed, in which we only need to apply our estimation procedure.

We need more notations before we introduce more assumptions. Suppose the rank of \mathbf{B}_J is r_J . For every $J = 1, \dots, s_p$, we denote $\mathbf{U}_{J0} \Theta_{J0} \mathbf{V}_{J0}^T$ as the singular value decomposition of B_J and use \mathbf{U}_{J0}^\perp and \mathbf{V}_{J0}^\perp to denote the orthogonal complements of \mathbf{U}_{J0} and \mathbf{V}_{J0} , respectively.

Define $\Sigma_{\mathcal{M}}$ as the covariance matrix for $x_{i, \mathcal{M}}$, where $x_{i, \mathcal{M}} = (x_{il})_{l \in \mathcal{M}} \in \mathbb{R}^{|\mathcal{M}|}$. We further define $\mathbf{A} = \Sigma_{\mathcal{M}} \otimes \mathbf{I}_{pq \times pq}$, $\mathbf{K}_J = \mathbf{V}_{J0}^\perp \otimes \mathbf{U}_{J0}^\perp$ and $\mathbf{d}_J = -\text{vec}(\mathbf{U}_{J0} \mathbf{V}_{J0}^T)$ for $J \in \mathcal{M}$, where \otimes denotes the Kronecker product. Let $\mathbf{d} = (\mathbf{d}_1^T, \dots, \mathbf{d}_{|\mathcal{M}|}^T)^T$ and $\mathbf{K} = \text{diag}\{\mathbf{K}_1, \dots, \mathbf{K}_{|\mathcal{M}|}\}$. We define $\Lambda_J \in \mathbb{R}^{(p-r_J) \times (q-r_J)}$ for $J \in \mathcal{M}$ such that

$\text{vec}(\Lambda) = (\text{vec}(\Lambda_1)^T, \dots, \text{vec}(\Lambda_{|\mathcal{M}|})^T)^T = (\mathbf{K}^T \mathbf{A}^{-1} \mathbf{K})^{-1} \mathbf{K}^T \mathbf{A}^{-1} \mathbf{d}$. The Λ_J has some interesting interpretation. For instance, it can be shown that it is the Lagrange multiplier of an optimization problem. We include more interpretation of Λ_J in the Appendix C.

We then state additional assumptions that are needed to establish the theory of our estimation procedure when both p and q are assumed to be fixed.

The following assumptions (A6)-(A8) are needed.

(A6) The $\Sigma_{\mathcal{M}}$ is nonsingular.

(A7) The $\max_l \{\text{rank}(\mathbf{B}_{l0}) : l \in \mathcal{M}\} < \min(p, q)$ holds.

(A8) For every $l \in \mathcal{M}$, we assume $\|\mathbf{\Lambda}_l\|_{op} < 1$.

Remarks: Assumption (A6) is a regularity condition in the low dimensional context, which rules out the scenario when one covariate is exactly a linear combination of other covariates. Assumption (A7) is used for rank consistency. Assumption (A8) can be regarded as the irrepresentable condition of Zhao and Yu (2006) in the rank consistency context. A similar condition can be found in Bach (2008).

Define $\widehat{\mathbf{B}}_l$ the regularized low rank estimator of \mathbf{B}_l for $l \in \mathcal{M}$. We have the following consistent results when the tuning parameter converges in different rates when both p and q are fixed.

Theorem 3. (Estimation Consistency) Under Assumptions (A0) and (A6), we have

- (i) if $n^{1/2} \lambda \rightarrow \infty$ and $\lambda \rightarrow 0$, then $\lambda^{-1}(\widehat{\mathbf{B}}_l - \mathbf{B}_{l0}) = O_p(1)$ for all $l \in \mathcal{M}$;
- (ii) if $n^{1/2} \lambda \rightarrow \rho \in [0, \infty)$ and $n \rightarrow \infty$, then $n^{1/2}(\widehat{\mathbf{B}}_l - \mathbf{B}_{l0}) = O_p(1)$ for all $l \in \mathcal{M}$.

Theorem 3 reveals an interesting phase-transition phenomenon. When λ is relatively small or moderate, the convergence rate of $\widehat{\mathbf{B}}_l - \mathbf{B}_{l0}$ is of order $n^{-1/2}$, whereas as λ gets large, the convergence rate of $\widehat{\mathbf{B}}_l - \mathbf{B}_{l0}$ can be approximated as the order of λ . Although we have established the consistency of $\widehat{\mathbf{B}}_l$ as $\lambda \rightarrow 0$, the next question is whether the rank of $\widehat{\mathbf{B}}_l$ is consistent under the same set of conditions. It turns out that such rank consistency only holds for relatively large λ , whose convergence rate is slower than $n^{-1/2}$.

Theorem 4. (Rank Consistency) Under Assumptions (A0) and (A6)-(A8), if $\lambda \rightarrow 0$ and $n^{1/2}\lambda \rightarrow \infty$ hold, we have that $P(\text{rank}(\widehat{\mathbf{B}}_l) = \text{rank}(\mathbf{B}_l)) \rightarrow 1$ for all $l \in \mathcal{M}$.

Theorem 4 establishes the rank consistency of our regularized estimates. Theorems 3 and 4 reveal that both of the element consistency and the rank consistency hold only for $\lambda \rightarrow 0$ and $n^{1/2}\lambda \rightarrow \infty$. This phenomenon is similar to that for the Lasso estimator. Specifically, although the Lasso estimator can achieve model selection consistency, the convergence rate of the Lasso estimator cannot achieve the rate of $n^{-1/2}$ when selection consistency is satisfied (Zou, 2006).

We then consider the case when p and q are assumed to be diverging. The following assumptions (A9)-(A12) are needed.

(A9) There exist positive constants C_L and C_M such that $0 < C_L \leq \lambda_{\min}(\Sigma_{\mathcal{M}}) \leq \lambda_{\max}(\Sigma_{\mathcal{M}}) \leq C_M < \infty$.

(A10) We assume that $x_{i, \mathcal{M}}$ are i.i.d multivariate normal with mean $\mathbf{0}$ and covariance matrix $\Sigma_{\mathcal{M}}$.

(A11) The vectorized error matrices $\text{vec}(\mathbf{E}_i)$ are i.i.d $\mathcal{N}(\mathbf{0}, \Sigma_{\mathcal{E}})$, where $\lambda_{\max}(\Sigma_{\mathcal{E}}) \leq C_U^2 < \infty$.

(A12) We assume $\max(p, q) \rightarrow \infty$ and $\max(p, q) = o(n)$ as $n \rightarrow \infty$.

Remarks: Assumptions (A9)-(A12) are needed for our estimation procedure when both p and q are diverging with the sample size n . Assumption (A9) assumes the largest eigenvalue of $\Sigma_{\mathcal{M}}$ is bounded and the smallest eigenvalue of $\Sigma_{\mathcal{M}}$ is greater than 0. Assumption (A10) assumes that the covariates x_{il} are gaussian. Assumption (A11) assumes that the largest eigenvalue of $\Sigma_{\mathcal{M}}$ is bounded. Assumption (A12) allows p and q to diverge slower than n , but it does allow that $pq > n$.

We then show the following non-asymptotic bound for our estimation procedure when both p and q are diverging.

Theorem 5. (Nonasymptotic bound when both p and q diverge) *Under Assumptions (A9)-(A12), when $\lambda \geq 4C_U C_M^{1/2} n^{-1/2} (p^{1/2} + q^{1/2})$, there exist some positive constants c_1, c_2 and c_3 such that with probability at least $1 - c_1 \exp\{-c_2(p + q)\} - c_3 \exp(-n)$, we have*

$$\|\widehat{\mathbf{B}} - \mathbf{B}_0\|_F^2 \leq C \left(\sum_{l \in \mathcal{M}} r_l \right) \lambda^2 C_L^{-2}$$

for some constant $C > 0$.

Theorem 5 implies that when $\{r_l, l \in \mathcal{M}\}$ and $|\mathcal{M}|$ are fixed and $\lambda \asymp n^{-1/2}(p^{1/2} + q^{1/2})$, the estimator $\widehat{\mathbf{B}}$ would be consistent with probability going to 1. The convergence rate of the estimator in Theorem 5 coincides with that in Corollary 5 of Negahban et al. (2009), where they studied the low-rank matrix learning problem using the trace norm regularization. Although considering different models, they also require the dimension of the matrix $\max(p, q) = o(n)$. It differs significantly from the L1 regularized problem, where the dimension of the matrix may diverge at the exponential order of the sample size. The result in this theorem can also be regarded as a special case of the result in Raskutti and Yuan (2018), where they derived non-asymptotic error bound in a class of tensor regression model with sparse or low-rank penalties.

3.3 Theory for Two-step Estimator

In this section, we give a unified theory for our two-step estimator. In particular, we derive the non-asymptotic bound for our final estimate. To begin with, we first introduce some notations. For simplicity, we will use $\widehat{\mathcal{M}}$ to denote $\widehat{\mathcal{M}}_{\gamma_n}$, which is the set selected from the first step. Define $\mathbf{B}^{\widehat{\mathcal{M}}} = [\mathbf{B}_l, l \in \widehat{\mathcal{M}}] \in \mathbb{R}^{p \times q \times |\widehat{\mathcal{M}}|}$ and the true value of $\mathbf{B}^{\widehat{\mathcal{M}}}$ as $\mathbf{B}_0^{\widehat{\mathcal{M}}} = [\mathbf{B}_{l0}, l \in \widehat{\mathcal{M}}] \in \mathbb{R}^{p \times q \times |\widehat{\mathcal{M}}|}$. Define $\widehat{\mathbf{B}}^{\widehat{\mathcal{M}}} = [\widehat{\mathbf{B}}_l, l \in \widehat{\mathcal{M}}] \in \mathbb{R}^{p \times q \times |\widehat{\mathcal{M}}|}$ as the solution of the regularized trace norm penalization problem given by

$$\min_{\mathbf{B}^{\widehat{\mathcal{M}}}} Q(\mathbf{B}^{\widehat{\mathcal{M}}}) = \min_{\mathbf{B}^{\widehat{\mathcal{M}}}} \left\{ \frac{1}{2n} \sum_{i=1}^n \left\| \mathbf{Y}_i - \sum_{l \in \widehat{\mathcal{M}}} x_{il} * \mathbf{B}_l \right\|_F^2 + \lambda \sum_{l \in \widehat{\mathcal{M}}} \left\| \mathbf{B}_l \right\|_* \right\}. \tag{6}$$

We need the following assumptions.

(A13) Assume $2\kappa + \tau < 1$. Define $\nu_L := \min_{\|\delta\|_0 \leq m, \delta \neq 0} \delta^T (n^{-1} \sum_{i=1}^n x_i x_i^T) \delta / \|\delta\|_2^2$ for any $m = O(n^{2\kappa+\tau})$ and $\delta \in \mathbb{R}^s$. We further assume $\nu_L > 0$.

(A14) Assume $\max(p, q) / \log(n) \rightarrow \infty$ and $\max(p, q) = o(n^{1-2\tau})$ as $n \rightarrow \infty$ with $\tau < 1/2$.

Theorem 6. (Nonasymptotic bound for two-step estimator) *Under Assumptions (A0)-(A5), (A10), (A11), (A13), and (A14), when $\lambda \asymp 4C_5 n^{\tau-1/2} (p^{1/2} + q^{1/2})$, there exist some positive constants c_1, c_2, c_3, c_4, c_5 such that with probability at least $1 - c_1 n^{2\kappa+\tau} \exp\{-c_2(p+q)\} - c_3 n^{2\kappa+\tau} \exp(-n) - c_4 \exp(-c_5 n^{1-2\kappa})$, we have*

$$\left\| \widehat{\mathbf{B}}^{\mathcal{M}} - \mathbf{B}_0^{\mathcal{M}} \right\|_F^2 \leq C \left(\sum_{l \in \mathcal{M}} r_l \right) \lambda^2 \nu_L^{-2}$$

for some constant $C > 0$.

Theorem 6 implies that when $\{r_l : l \in \mathcal{M}\}$ and $|\mathcal{M}|$ are fixed and ν_L is fixed, the estimator $\widehat{\mathbf{B}}^{\mathcal{M}}$ is consistent with probability going to 1 when $\lambda \asymp n^{\tau-1/2} (p^{1/2} + q^{1/2})$. Theorem 6 gives an overall theoretical guarantee for our two-step estimator by considering the random selection procedure in the first step. A key fact that we use in the proof of Theorem 6 is that our first-step screening procedure enjoys the sure independence property. In this case, we only need to derive the non-asymptotic bound for the case when we exactly select or over-select the important variables as it holds with overwhelming probability.

4 Simulations

We conduct simulations to examine the finite sample performance of the proposed estimation and screening procedures. For the sake of space, we include additional simulation results in the Supplementary Material.

4.1 Regularized Low-rank Estimate

In the first simulation, we simulate 64×64 matrix responses according to model (1) with $s = 4$ covariates. We set the four true coefficient matrices to be a cross shape (\mathbf{B}_{10}), a square shape (\mathbf{B}_{20}), a triangle shape (\mathbf{B}_{30}), and a butterfly shape (\mathbf{B}_{40}). The images of \mathbf{B}_0 are shown in Figure 1, and each of them consists of a yellow region of interest (ROI) containing ones and a blue ROI containing zeros.

We independently generate all scalar covariates \mathbf{x}_j from $\mathcal{N}(\mathbf{0}, \Sigma_x)$, where $\Sigma_x = (\sigma_{x, ll'})$ is a covariance matrix with an autoregressive structure such that $\sigma_{x, ll'} = \rho_1^{|l-l'|}$ holds for $1 \leq l, l' \leq s$ with $\rho_1 = 0.5$. We independently generate $\text{vec}(\mathbf{E}_j)$ from $\mathcal{N}(\mathbf{0}, \Sigma_e)$. Specifically, we set the variances of all elements in \mathbf{E}_j to be σ_e^2 and the correlation between $\mathbf{E}_{i,jk}$ and $\mathbf{E}_{i,j'k'}$ to be $\rho_2^{|j-j'| + |k-k'|}$ for $1 \leq j, k, j', k' \leq 64$ with $\rho_2 = 0.5$. We consider three different sample sizes including $n = 100, 200, \text{ and } 500$, and set σ_e^2 to be 1 and 25.

We use 100 replications to evaluate the finite sample performance of our regularized low-rank (RLR) estimates $\widehat{\mathbf{B}}_l$ defined as $\|\widehat{\mathbf{B}}_l - \mathbf{B}_{l0}\|_F^2$. To evaluate the estimation accuracy, we compute the mean squared errors of $\widehat{\mathbf{B}}_l$, denoted by $\text{MSE}(\widehat{\mathbf{B}}_l)$, for all $l = 1, \dots, 4$. We also calculate the prediction errors (PE) by generating $n^{\text{test}} = 500$ independent test observations.

We compare our method with OLS, Lasso, fused Lasso and tensor envelope method (Li and Zhang, 2017). For fair comparison, we also use five-fold cross validation to select regularization parameters of Lasso and fused Lasso and the envelope dimension of the tensor envelope method. The results are shown in Table 1. We also plot the RLR, OLS, Lasso, fused Lasso and tensor envelope estimates of $(\widehat{\mathbf{B}}_l, l = 1, \dots, 4)$ obtained from a randomly selected data set with $n = 500$ and $\sigma_e^2 = 25$ in Figure 2.

Inspecting Figure 2 and Table 1 reveals the following findings. First, our method always outperforms OLS and envelope method. Second, when the images are of low rank (cross and square), our estimation method truly outperforms Lasso. Third, our method outperforms fused Lasso when either the sample size is small or the noise variance is large, whereas fused Lasso outperforms our method in other cases. Fourth, when the images are not of low rank (triangle and butterfly), fused Lasso performs best in most cases, whereas our method outperforms Lasso when either noise level is high or sample size is small.

These findings are not surprising. First, in all settings, since all the true coefficient matrices are piecewise sparse, the fused Lasso method is expected to perform well. Second, Lasso works reasonably well since it still imposes sparse structure. Third, since our method is designed for low rank cases, it performs well for the low rank cross and square cases, whereas it performs relatively worse for the triangle and butterfly cases.

We then conduct the second simulation study when the images only have low rank structure, but no sparse structure. Specifically, we simulate 64×64 matrix responses according to model (1) with $s = 2$ covariates. We set the two true coefficient matrices as $\mathbf{B}_{10} = \sum_{j=1}^{10} \lambda_{1,j} \mathbf{u}_{1,j} \mathbf{v}_{1,j}^T$ and $\mathbf{B}_{20} = \sum_{j=1}^5 \lambda_{2,j} \mathbf{u}_{2,j} \mathbf{v}_{2,j}^T$, where $\boldsymbol{\lambda}_1 = (\lambda_{1,1}, \dots, \lambda_{1,10}^T) = (2, 1.8, 1.6, 1.4, 1.2, 1, 0.8, 0.6, 0.4, 0.2)^T$, $\boldsymbol{\lambda}_2 = (\lambda_{2,1}, \lambda_{2,2}, \lambda_{2,3}, \lambda_{2,4}, \lambda_{2,5})^T = (2, 1.6, 1.2, 0.8, 0.4)^T$, and $\mathbf{u}_{1,j}, \mathbf{u}_{2,j}, \mathbf{v}_{1,j}, \mathbf{v}_{2,j}$ are column vectors of dimension 64. For $\mathbf{U}_1 = (\mathbf{u}_{1,1}, \dots, \mathbf{u}_{1,10})$ and $\mathbf{V}_1 = (\mathbf{v}_{1,1}, \dots, \mathbf{v}_{1,10})$, each of them is generated by orthogonalizing a 64×10 matrix with all elements being i.i.d standard normal. For $\mathbf{U}_2 = (\mathbf{u}_{2,1}, \dots, \mathbf{u}_{2,5})$ and $\mathbf{V}_2 = (\mathbf{v}_{2,1}, \dots, \mathbf{v}_{2,5})$, each of them is generated by orthogonalizing a 64×5 matrix with all elements being i.i.d standard normal. For all other settings, they are the same as those in Section 4.1. Table 2 summarizes the obtained results. Our method outperforms all the comparison methods when the true coefficient matrices are of low rank structure, but of no sparse structure.

4.2 Rank-one Screening using SNP Covariates

We generate 64×64 matrix responses according to model (1). We use the same method as Section 4.1 to generate \mathbf{E}_j with $\rho_2 = 0.5$ and $\sigma_e^2 = 1$ or 25. We generate genetic covariates by mimicking the SNP data used in Section 5. Specifically, we use Linkage Disequilibrium (LD) blocks defined by the default method (Gabriel et al., 2002) of Haploview (Barrett et al.,

2005) and PLINK (Purcell et al., 2007) to form SNP-sets. To calculate LD blocks, n subjects are simulated by randomly combining haplotypes of HapMap CEU subjects. We use PLINK to determine the LD blocks based on these subjects. We randomly select $s_n/10$ blocks, and combine haplotypes of HapMap CEU subjects in each block to form genotype variables for these subjects. We randomly select 10 SNPs in each block, and thus we have s_n SNPs for each subject. We set $s_n = 2,000$ and $5,000$ and choose the first 20 SNPs as the significant SNPs. That is, we set the first 20 true coefficient matrices as nonzero matrices $\mathbf{B}_{1,0} = \dots = \mathbf{B}_{20,0} = \mathbf{B}_{true}$, and the remaining coefficient matrices as zero. We consider three types of coefficient matrices \mathbf{B}_{true} with different significant regions, i.e. $(p_s, q_s) = (4, 4)$, $(8, 8)$, and $(16, 16)$, where p_s and q_s denote the true size of the significant regions of interest. Figure presents the true images \mathbf{B}_{true} and each of them contains a yellow ROI containing ones and a blue ROI containing zeros.

In this subsection, we evaluate the effect of using different γ_n on the finite sample performance of the screening procedure. We will investigate the proposed random decoupling in the next subsection. Specifically, by sorting the magnitude of $\|\widehat{\mathbf{B}}_l^M\|_{op}$ in descending order, we define $\widehat{\mathcal{M}}^k$ as

$$\widehat{\mathcal{M}}^k = \{1 \leq l \leq s_n: \|\widehat{\mathbf{B}}_l^M\|_{op} \text{ is among the first } k \text{ largest of all covariates}\}. \quad (7)$$

We apply our screening procedure to each simulated data set and then report the average true nonzero coverage proportion as k varies from 1 to 200. In this case, $\mathcal{M} = \{1, 2, \dots, 20\}$ is the set of all true nonzero indices, and $\widehat{\mathcal{M}}^k$ is the selected index set by using our screening method. The true nonzero coverage proportion is defined as $|\widehat{\mathcal{M}}^k \cap \mathcal{M}| / |\mathcal{M}|$. We consider three different sample sizes including $n = 100, 200$, and 500 . We run 100 Monte Carlo replications for each scenario.

We consider four screening methods including the rank-one screening method, the L1 entrywise norm screening, the Frobenius norm screening, and the global Wald test screening proposed in Huang et al. (2015). The curves of percentage of the average true nonzero coverage proportion for different threshold values are presented for the case $(\sigma_e^2, s_n) = (1, 2000)$ in Figure 4 and for the case $(\sigma_e^2, s_n) = (25, 2000)$ in Figure 5. Inspecting Figures 4 and 5 reveals that the rank-one screening significantly outperforms all other three methods, followed by the Frobenius norm screening. As expected, increasing the sample size n and/or k increases the true nonzero coverage proportion of all four methods. We also include additional simulation results for the cases $(\sigma_e^2, s_n) = (1, 5000)$ in Figure S1 and $(\sigma_e^2, s_n) = (25, 5000)$ in Figure S2 in the Supplementary Material. The findings are similar. Overall, the rank-one screening method is more robust to noise and signal region size.

4.3 Simulation study for two-step procedure

In this subsection, we perform a simulation study to evaluate our two-step screening and estimation procedure. We simulate 64×64 matrix responses according to model (1) with s_n covariates. We set the first four true coefficient matrices to be a cross shape (\mathbf{B}_{10}), a square

shape (\mathbf{B}_{20}), a triangle shape (\mathbf{B}_{30}), and a butterfly shape (\mathbf{B}_{40}) shown in Figure 1. For the remaining coefficient matrices $\{\mathbf{B}_0, \mathbf{B}_1, \dots, \mathbf{B}_n\}$, we set them as zero matrices. We consider $s_n = 2000$ and 5000.

We independently generate all scalar covariates \mathbf{x}_j from $\mathcal{N}(\mathbf{0}, \Sigma_x)$, where $\Sigma_x = (\sigma_{x, ll'})$ is a covariance matrix with an autoregressive structure such that $\sigma_{x, ll'} = \rho_1^{|l-l'|}$ holds for $1 \leq l, l' \leq s$ with $\rho_1 = 0.5$. We independently generate $\text{vec}(\mathbf{E}_j)$ from $\mathcal{N}(\mathbf{0}, \Sigma_e)$. Specifically, we set the variances of all elements in \mathbf{E}_j to be σ_e^2 and the correlation between $\mathbf{E}_{i,jk}$ and $\mathbf{E}_{i,j'k'}$ to be $\rho_2^{|j-j'| + |k-k'|}$ for $1 \leq j, k, j', k' \leq 64$ with $\rho_2 = 0.5$. We consider three different sample sizes including $n = 100, 200, \text{ and } 500$, and set σ_e^2 to be 1 and 25.

First, we evaluate the finite sample performance of the random decoupling. We perform our screening procedure based on the random decoupling and then apply our regularized low rank estimation procedure. We report the MSEs of $\hat{\mathbf{B}}_l$ ($l = 1, 2, 3, 4$), model size, and prediction error based on 100 replications in Table 3. We report the proportion of times that we exactly select the true model $\mathcal{M} = \{1, 2, 3, 4\}$, the proportion that we over-select some variables, but include all the true ones, and the proportion that we miss some of the important covariates in Table 4. The proposed random decoupling works pretty well in choosing γ_n , since the selected covariate set based on γ_n includes the true covariates with high probabilities in all scenarios.

Second, we consider over-selecting and/or missing some covariates. For each of the three above cases, we report the MSEs of $\hat{\mathbf{B}}_l$ ($l = 1, 2, 3, 4$) and the prediction error in Table 4. When the screening procedure over-selects more irrelevant variables, the MSEs of the true non-zero coefficient matrices and prediction error of the fitted model are similar to those obtained from the model with the correct set of covariates. In contrast, if the screening procedure misses several important variables, then the estimates corresponding to these missed variables completely fail since the corresponding coefficient matrices are estimated zero. However, according to the simulation results, the MSEs corresponding to those important variables that have been selected, are similar to those obtained from the model with the correct set of covariates. The prediction error increases due to missing some important variables.

5 The Philadelphia Neurodevelopmental Cohort

5.1 Data Description and Preprocessing Pipeline

To motivate the proposed methodology, we consider a large database with imaging, genetic, and clinical data collected by the Philadelphia Neurodevelopmental Cohort (PNC) study. This study was a collaboration between the Center for Applied Genomics (CAG) at Children’s Hospital of Philadelphia (CHOP) and the Brain Behavior Laboratory at the University of Pennsylvania (Penn). The PNC cohort consists of youths aged 8-21 years in the CHOP network and volunteered to participate in genomic studies of complex pediatric disorders. All participants underwent clinical assessment and a neuroscience based computerized neurocognitive battery (CNB) and a subsample underwent neuroimaging. We

consider 814 subjects with 429 females and 385 males. The age range of the 814 participants is 8-21 (years) with mean value 14.36 (years) and standard deviation 3.48 (years). Specifically, each subject has a resting state functional magnetic resonance imaging (rs-fMRI) connectivity matrix, which is represented as a 69×69 matrix, and a large genetic data set with around 5,400,000 genotyped and imputed single-nucleotide polymorphisms (SNPs) on all of the 22 chromosomes. Other clinical variables of interest include age and gender, among others. Our primary question of interest is to identify novel genetic effects on the local rs-fMRI connectivity changes.

We preprocess the resting state fMRI data using C-PAC pipeline. First, we register the fMRI data to the standard MNI 2mm resolution level and did segmentation using the C-PAC default setting. Next, we do motion correction using the Friston 24-parameter method. We also perform nuisance signal correction by regressing out the following variables: top 5 principle components in the noise regions of interest (ROIs), Cerebrospinal fluid (CSF), motion parameters, and the linear trends in time series. Finally, we extract the ROI time series by taking the average of voxel-wise time series in each ROI. The atlases that we use are HarvardOxford Cortical Atlas (48 regions) and HarvardOxford Subcortical Atlas (21 regions), which could be found in FSL. In total, we extract time series for each of the 69 regions and each time series has 120 observations after deleting the first and last 3 scans.

5.2 Analysis and Results

We first fit model (1) with the rs-fMRI connectivity matrices from 814 subjects as 69×69 matrix responses and age and gender as clinical covariates. We also include the first 5 principal component scores based on the SNP data as covariates to correct for population stratification. We first calculate the ordinary least squares estimates of coefficient matrices and then compute the corresponding residual matrices for the brain connectivity response matrix after adjusting the effects of the clinical covariates and the SNP principal component scores.

Second, we apply the rank-one screening procedure by using the residual matrices as responses to select important SNPs from the whole set of 5,354,265 SNPs that are highly associated with the residual matrices. We use the random decoupling method described in Section 2.1 to choose the thresholding value γ_n and select all those indices whose $\|\widehat{\mathbf{B}}_l\|_{op}$ is the larger than γ_n . Finally, seven covariates are selected, where the names are shown in Table 5. Among these seven SNPs, the first three ones on Chromosome 5 have exactly the same genotypes for all the subjects and the next four ones on Chromosome 10 have exactly the same genotypes for all the subjects.

Finally, we examine the effects of these selected SNPs on our matrix response. We first fit the OLS to these 7 SNPs. Since the first three ones have exactly the same genotypes and the next four ones have exactly the same genotypes, we regress our matrix response on the first selected SNP and the fourth selected SNP, yielding two coefficient matrix estimates $\widehat{\mathbf{B}}_{(1)}^{ols}$ and $\widehat{\mathbf{B}}_{(2)}^{ols}$. The OLS estimates for the 7 SNPs are defined as $\widehat{\mathbf{B}}_{(1)}^{ols} = \widehat{\mathbf{B}}_{(2)}^{ols} = \widehat{\mathbf{B}}_{(3)}^{ols} = \widehat{\mathbf{B}}_{(1)}^{ols} / 3$ and $\widehat{\mathbf{B}}_{(4)}^{ols} = \widehat{\mathbf{B}}_{(5)}^{ols} = \widehat{\mathbf{B}}_{(6)}^{ols} = \widehat{\mathbf{B}}_{(7)}^{ols} = \widehat{\mathbf{B}}_{(2)}^{ols} / 4$. We then calculate the singular values of these 7 OLS

estimates and plot these singular values in decreasing order in Figure 6. Inspecting Figure 6 reveals that these estimated coefficient matrices have a clear low rank pattern since the first few singular values dominate the remaining ones. This motivates us to apply our RLR estimation procedure to estimate the coefficient matrices corresponding to these 7 SNP covariates. Figure 7(a)-(g) presents the coefficient matrix estimates associated with these SNPs. The coefficient matrices corresponding to the first three selected SNPs are the same and the coefficient matrices corresponding to next four selected SNPs are the same. The estimated ranks of these seven coefficient matrices are given by 11, 11, 11, 8, 8, 8, and 8, respectively.

6 Discussion

Motivated from the analysis of imaging genetic data, we have proposed a low-rank linear regression model to correlate high-dimensional matrix responses with a high dimensional vector of covariates when coefficient matrices are approximately low-rank. We have developed a fast and efficient rank-one screening procedure, which enjoys the sure independence screening property as well as vanishing false selection rate, to reduce the covariate space. We have developed a regularized estimate of coefficient matrices based on the trace norm regularization, which explicitly incorporates the low-rank structure of coefficient matrices, and established its estimation consistency. We have further established a theoretical guarantee for the overall solution obtained from our two-step screening and estimation procedure. We have demonstrated the efficiency of our methods by using simulations and the analysis of PNC dataset.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgement

The authors would like to thank the Editor, Associate Editor and two reviewers for their constructive comments, which have substantially improved the paper.

Dr. Kong's work was partially supported by National Science Foundation, National Institute of Health, and Natural Science and Engineering Research Council of Canada. Dr. Baiguo An is work was partially supported by NSF of China (No. 11601349). Dr. Zhu's work was partially supported by NIH grants R01MH086633 and R01MH116527, and NSF grants SES-1357666 and DMS-1407655. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH or any other funding agency.

The PNC data were obtained through dbGaP (phs000607.v1.p1). Support for the collection of the data sets was provided by grant RC2MH089983 awarded to Raquel Gur and RC2MH089924 awarded to Hakon Hakonarson. All subjects were recruited through the Center for Applied Genomics at The Children's Hospital in Philadelphia.

A: Auxiliary Lemmas

In this section, we include the auxiliary lemmas needed for the theorems and their proofs.

Lemma 1. (Bernstein's inequality) Let Z_1, \dots, Z_n be independent random variables with zero mean such that $E|Z_i|^m \leq m!M^{m-2}v_i/2$ for every $m \geq 2$ (and all i) and some positive constants M and v_i . Then $P(|Z_1 + \dots + Z_n| > x) \leq 2 \exp[-x^2 / \{2(v + Mx)\}]$ for $v = v_1 + \dots + v_n$.

This lemma is Lemma 2.2.11 of van der Vaart and Wellner (2000), and we omit the proof.

Lemma 2. Under Assumptions A0-A2, for arbitrary $t > 0$ and every l, l', j, k , we have that

$$P\left(\left|\sum_{i=1}^n \{x_{il}x_{il'} - E(x_{il}x_{il'})\}\right| \geq t\right) \leq 2 \exp\left\{-\frac{t^2}{2(2nC_2^2e^{C_2C_3+t/C_2})}\right\},$$

and

$$P\left(\left|\sum_{i=1}^n (x_{il}E_{i,jk})\right| \geq t\right) \leq 2 \exp\left\{-\frac{t^2}{2(2nC_2^2e^{C_2C_3+t/C_2})}\right\}.$$

Proofs of Lemma 2: By Assumptions A1 and A2, we have

$$\begin{aligned} E\{\exp\{C_2 | x_{il}x_{il'} - E(x_{il}x_{il'}) | \}\} &\leq e^{C_2} E\{x_{il}x_{il'}\} E\{e^{C_2 | x_{il}x_{il'} |}\} \\ &\leq e^{C_2} E\{e^{C_2x_{il}^2/2} e^{C_2x_{il'}^2/2}\} \leq e^{C_2} \left[E\{e^{C_2x_{il}^2}\} E\{e^{C_2x_{il'}^2}\}\right]^{1/2} \leq e^{C_2C_3}. \end{aligned}$$

For every $m \geq 2$, one has

$$E\{|x_{il}x_{il'} - E(x_{il}x_{il'})|^m\} \leq m!C_2^{-m} E\{\exp(C_2 | x_{il}x_{il'} - E(x_{il}x_{il'}) |)\} \leq m!C_2^{-m} e^{C_2C_3}.$$

It follows from Lemma 1 that we have

$$P\left(\left|\sum_{i=1}^n \{x_{il}x_{il'} - E(x_{il}x_{il'})\}\right| \geq t\right) \leq 2 \exp\left\{-\frac{t^2}{2(2nC_2^2e^{C_2C_3+t/C_2})}\right\}.$$

Similarly, we obtain

$$\begin{aligned} E\{\exp\{C_2 | x_{il}E_{i,jk} | \}\} &\leq e^{C_2} E\{e^{C_2x_{il}^2/2} e^{C_2E_{i,jk}^2/2}\} \\ &\leq e^{C_2} \left[E\{e^{C_2x_{il}^2}\} E\{e^{C_2E_{i,jk}^2}\}\right]^{1/2} \leq e^{C_2C_3}. \end{aligned}$$

For every $m \geq 2$, we have $E\{|x_{il}E_{i,jk}|^m\} \leq m!C_2^{-m} E\{\exp(C_2 | x_{il}E_{i,jk} |)\} \leq m!C_2^{-m} e^{C_2C_3}$.

Therefore, it follows from Lemma 1 that we have

$$P\left(\left|\sum_{i=1}^n (x_{il}E_{i,jk})\right| \geq t\right) \leq 2 \exp\left\{-\frac{t^2}{2(2nC_2^2e^{C_2C_3+t/C_2})}\right\}.$$

This completes the proof of Lemma 2.

The next lemma is about the subdifferential and directional derivatives of the trace norm. For more details about this lemma and its proof, please refer to Recht et al. (2010) and Borwein and Lewis (2010).

Lemma 3. For an arbitrary matrix \mathbf{W} , its singular value decomposition is denoted by $\mathbf{W} = \mathbf{U}\mathbf{D}\mathbf{V}^T$, where $\mathbf{U} \in \mathbb{R}^{p \times m}$ and $\mathbf{V} \in \mathbb{R}^{q \times m}$ have orthonormal columns, $\mathbf{D} = \text{diag}(d_1, \dots, d_m)$, and $d_1 \dots d_m > 0$ are the singular values of \mathbf{W} . Then the trace norm of \mathbf{W} is $\|\mathbf{W}\|_* = \sum_{i=1}^m d_i$ and its subdifferential is equal to

$$\partial\|\mathbf{W}\|_* = \{\mathbf{U}\mathbf{D}\mathbf{V}^T + \mathbf{N}, \text{ such that } \|\mathbf{N}\|_{op} \leq 1, \mathbf{U}^T\mathbf{N} = 0, \mathbf{N}\mathbf{V} = 0\}.$$

The directional derivative at \mathbf{W} is

$$\lim_{\epsilon \rightarrow 0^+} \frac{\|\mathbf{W} + \epsilon\mathbf{Y}\|_* - \|\mathbf{W}\|_*}{\epsilon} = \text{tr}(\mathbf{U}^T\mathbf{Y}\mathbf{V}) + \left\| (\mathbf{U}^\perp)^T\mathbf{Y}\mathbf{V}^\perp \right\|_*,$$

where $\mathbf{U}^\perp, \mathbf{V}^\perp$ are the orthonormal complements of \mathbf{U} and \mathbf{V} .

The following lemma is a standard result called Gaussian comparison inequality (Anderson, 1955).

Lemma 4. Let X and Y be zero-mean vector Gaussian random vectors with covariance matrix Σ_X and Σ_Y respectively. If $\Sigma_X - \Sigma_Y$ is positive semi-definite, then for any convex symmetric set \mathcal{C} , $P(X \in \mathcal{C}) \leq P(Y \in \mathcal{C})$.

B: Proof of Theorems

Proof of Theorem 1: Recall that $\mathbf{B}_{l0}^M = \text{cov}(\sum_{l' \in \mathcal{M}} x_{il'} * \mathbf{B}_{l'0}, x_{il})$. For every $1 \leq j \leq p, 1 \leq k \leq q$ and $1 \leq l \leq s_p$, we have

$$\hat{B}_{l,jk}^M - B_{l0,jk}^M = n^{-1} \sum_{i=1}^n \{x_{il}Y_{i,jk} - E(x_{il}Y_{i,jk})\}.$$

It follows from Assumptions (A0) (A1) (A2) and Lemma 2 that for any $t > 0$, we have

$$\begin{aligned} P(|\hat{B}_{l,jk}^M - B_{l0,jk}^M| \geq t) &= P\left(\left|\sum_{i=1}^n \{x_{il}Y_{i,jk} - E(x_{il}Y_{i,jk})\}\right| \geq nt\right) \\ &= P\left(\left|\sum_{l' \in \mathcal{M}} \sum_{i=1}^n \{x_{il}x_{il'} - E(x_{il}x_{il'})\} \mathbf{B}_{l'0,jk} + \sum_{i=1}^n x_{il}E_{i,jk}\right| \geq nt\right) \\ &\leq \sum_{l' \in \mathcal{M}} P\left(\left|\sum_{i=1}^n \{x_{il}x_{il'} - E(x_{il}x_{il'})\}\right| \geq \frac{nt}{b(s_0+1)}\right) + P\left(\sum_{i=1}^n |x_{il}E_{i,jk}| \geq \frac{nt}{(s_0+1)}\right) \\ &\leq 2s_0 \exp\left\{-\frac{nt^2 b^{-2}(s_0+1)^{-2}}{2(2C_2^2 e^2 C_2 C_3 + C_2^{-1} b^{-1}(s_0+1)^{-1} t)}\right\} + 2 \exp\left\{-\frac{nt^2 (s_0+1)^{-2}}{2(2C_2^2 e^2 C_2 C_3 + C_2^{-1} (s_0+1)^{-1} t)}\right\}. \end{aligned}$$

For every $l \in \mathcal{M}$, we have

$$\begin{aligned}
 & P(\|\widehat{\mathbf{B}}_l^M\|_{op} \leq \gamma_n) \leq P(\|\widehat{\mathbf{B}}_l^M - \mathbf{B}_{l0}^M\|_{op} \geq (pq)^{1/2}(1-\alpha)C_1n^{-\kappa}) \\
 & \leq P\left(\left\|\widehat{\mathbf{B}}_l^M - \mathbf{B}_{l0}^M\right\|_F \geq (pq)^{1/2}(1-\alpha)C_1n^{-\kappa}\right) = P\left(\sum_{j,k} |\widehat{B}_{l,jk}^M - B_{l0,jk}^M|^2 \geq pq\{(1-\alpha)C_1n^{-\kappa}\}^2\right) \\
 & \leq \sum_{j,k} P(|\widehat{B}_{l,jk}^M - B_{l0,jk}^M| \geq (1-\alpha)C_1n^{-\kappa}) \\
 & \leq 2pq \left(s_0 \exp\left\{-\frac{n^1 - 2\kappa[(1-\alpha)C_1b^{-1}(s_0+1)^{-1}]^2}{2\{2C_2^2e^{C_2}C_3 + C_2^{-1}b^{-1}(s_0+1)^{-1}(1-\alpha)C_1n^{-\kappa}\}}\right\}\right. \\
 & \quad \left. + \exp\left\{-\frac{n^1 - 2\kappa[(1-\alpha)C_1(s_0+1)^{-1}]^2}{2\{2C_2^2e^{C_2}C_3 + C_2^{-1}(s_0+1)^{-1}(1-\alpha)C_1n^{-\kappa}\}}\right\}\right) \\
 & \leq 2pq \left(s_0 \exp\left\{-\frac{n^1 - 2\kappa[(1-\alpha)C_1b^{-1}(s_0+1)^{-1}]^2}{2\{2C_2^2e^{C_2}C_3 + C_2^{-1}b^{-1}(s_0+1)^{-1}(1-\alpha)C_1\}}\right\}\right. \\
 & \quad \left. + \exp\left\{-\frac{n^1 - 2\kappa[(1-\alpha)C_1(s_0+1)^{-1}]^2}{2\{2C_2^2e^{C_2}C_3 + C_2^{-1}(s_0+1)^{-1}(1-\alpha)C_1\}}\right\}\right).
 \end{aligned}$$

Let $c_1 = 2pq(s_0 + 1)$,

$$\begin{aligned}
 c_2 &= \frac{[(1-\alpha)C_1b^{-1}(s_0+1)^{-1}]^2}{2\{2C_2^2e^{C_2}C_3 + C_2^{-1}b^{-1}(s_0+1)^{-1}(1-\alpha)C_1\}}, \text{ and} \\
 c_3 &= \frac{[(1-\alpha)C_1(s_0+1)^{-1}]^2}{2\{2C_2^2e^{C_2}C_3 + C_2^{-1}(s_0+1)^{-1}(1-\alpha)C_1\}}.
 \end{aligned}$$

We have $P(|\widehat{\mathbf{B}}_l^M| \leq \gamma_n) \leq 2pq(s_0 + 1)\exp(-c_0n^{1-2\kappa})$, where $c_0 = \min\{c_2, c_3\}$. By Assumption (A5), one has

$$\begin{aligned}
 P(\mathcal{M} \subseteq \widehat{\mathcal{M}}_{\gamma_n}) &= P\left(\bigcap_{l \in \mathcal{M}} \{|\widehat{\mathbf{B}}_l^M| > \gamma_n\}\right) \\
 &= 1 - P\left(\bigcup_{l \in \mathcal{M}} \{|\widehat{\mathbf{B}}_l^M| \leq \gamma_n\}\right) \geq 1 - \sum_{l \in \mathcal{M}} P(|\widehat{\mathbf{B}}_l^M| \leq \gamma_n) \\
 &\geq 1 - s_0c_1 \exp(-c_0n^{1-2\kappa}) = 1 - 2pq(s_0 + 1)s_0 \exp(-c_0n^{1-2\kappa}) \rightarrow 1.
 \end{aligned}$$

This completes the proof of Theorem 1.

Proof of Theorem 2: The proof consists of two steps. In Step 1, we will show that $P(\widehat{\mathcal{M}}_{\gamma_n} \subseteq \mathcal{M}^0) \rightarrow 1$, where $\mathcal{M}^0 = \{1 \leq l \leq s_n: \|\mathbf{B}_{l0}^M\|_{op} \geq \gamma_n/2\}$. It follows from the definition of $\widehat{\mathcal{M}}_{\gamma_n}$ that we have

$$P(\widehat{\mathcal{M}}_{\gamma_n} \subseteq \mathcal{M}^0) \geq P\left(\bigcap_{1 \leq l \leq s_n} \left\{\left\|\widehat{\mathbf{B}}_l^M - \mathbf{B}_{l0}^M\right\|_{op} \leq \gamma_n/2\right\}\right),$$

Moreover, we have

$$\begin{aligned}
 & P\left(\bigcap_{1 \leq l \leq s_n} \left\{ \left\| \widehat{\mathbf{B}}_l^M - \mathbf{B}_{l0}^M \right\|_{op} \leq \gamma_n / 2 \right\}\right) = 1 - P\left(\bigcup_{1 \leq l \leq s_n} \left\{ \left\| \widehat{\mathbf{B}}_l^M - \mathbf{B}_{l0}^M \right\|_{op} \geq \gamma_n / 2 \right\}\right) \\
 & \geq 1 - \sum_{1 \leq l \leq s_n} P\left(\left\| \widehat{\mathbf{B}}_l^M - \mathbf{B}_{l0}^M \right\|_{op} \geq \gamma_n / 2\right) \geq 1 - \sum_{1 \leq l \leq s_n} P\left(\left\| \widehat{\mathbf{B}}_l^M - \mathbf{B}_{l0}^M \right\|_F \geq \gamma_n / 2\right) \\
 & \geq 1 - \sum_{1 \leq l \leq s_n} \sum_{j,k} P(|\widehat{B}_{l,jk}^M - B_{l0,jk}^M| \geq C_1 n^{-\kappa} / 2) \\
 & \geq 1 - 2s_n pq \left[s_0 \exp\left\{-\frac{\alpha^2 C_1^2 b^{-2}(s_0+1)^{-2} 2^{-2} n^{1-2\kappa}}{2(2C_2^2 e^{C_2 C_3} + C_2^{-1} b^{-1}(s_0+1)^{-1} \alpha C_1 2^{-1} n^{-\kappa})}\right\} \right. \\
 & \quad \left. + \exp\left\{-\frac{\alpha^2 C_1^2 (s_0+1)^{-2} 2^{-2} n^{1-2\kappa}}{2(2C_2^2 e^{C_2 C_3} + C_2^{-1} (s_0+1)^{-1} \alpha C_1 2^{-1} n^{-\kappa})}\right\} \right] \\
 & \geq 1 - 2s_n pq \left[s_0 \exp\left\{-\frac{\alpha^2 C_1^2 b^{-2}(s_0+1)^{-2} 2^{-2} n^{1-2\kappa}}{2(2C_2^2 e^{C_2 C_3} + C_2^{-1} b^{-1}(s_0+1)^{-1} \alpha C_1 2^{-1})}\right\} \right. \\
 & \quad \left. + \exp\left\{-\frac{\alpha^2 C_1^2 (s_0+1)^{-2} 2^{-2} n^{1-2\kappa}}{2(2C_2^2 e^{C_2 C_3} + C_2^{-1} (s_0+1)^{-1} \alpha C_1 2^{-1})}\right\} \right] \\
 & = 1 - 2pq \exp(C_4 n^\xi) \left[s_0 \exp\left\{-\frac{\alpha^2 C_1^2 b^{-2}(s_0+1)^{-2} 2^{-2} n^{1-2\kappa}}{2(2C_2^2 e^{C_2 C_3} + C_2^{-1} b^{-1}(s_0+1)^{-1} \alpha C_1 2^{-1})}\right\} \right. \\
 & \quad \left. + \exp\left\{-\frac{\alpha^2 C_1^2 (s_0+1)^{-2} 2^{-2} n^{1-2\kappa}}{2(2C_2^2 e^{C_2 C_3} + C_2^{-1} (s_0+1)^{-1} \alpha C_1 2^{-1})}\right\} \right].
 \end{aligned}$$

By Assumptions (A3) and (A5), one has

$P(\bigcap_{1 \leq l \leq s_n} \{ \left\| \widehat{\mathbf{B}}_l^M - \mathbf{B}_{l0}^M \right\|_{op} \leq \gamma_n / 2 \}) \geq 1 - c_4 \exp(-c_5 n^{1-2\kappa})$ for some constants $c_4 > 0$ and $c_5 > 0$. Therefore, we have $P(\widehat{\mathcal{M}}_{\gamma_n} \subseteq \mathcal{M}^0) \rightarrow 1$ by Assumption (A1).

In Step 2, we will show that $|\mathcal{M}^0| = O(n^{2\kappa + \tau})$. Define $\mathcal{M}^1 = \{1 \leq l \leq s_n : \left\| \mathbf{B}_{l0}^M \right\|_F^2 \geq \gamma_n^2 / 4\}$.

As $\left\| \mathbf{B}_{l0}^M \right\|_{op} \leq \left\| \mathbf{B}_{l0}^M \right\|_F$, we have $\mathcal{M}^0 \subseteq \mathcal{M}^1$. By the definition of \mathcal{M}^1 , we have

$$\begin{aligned}
 |\mathcal{M}^1| \gamma_n^2 / 4 & \leq \sum_{l=1}^{s_n} \left\| \mathbf{B}_{l0}^M \right\|_F^2 \\
 & = \sum_{j,k} \sum_{l=1}^{s_n} (B_{l0,jk}^M)^2 = \sum_{j,k} \sum_{l=1}^{s_n} \{E(x_{il} Y_{i,jk})\}^2 = \sum_{j,k} \left\| E(x_i * Y_{i,jk}) \right\|^2.
 \end{aligned}$$

Define $\mathbf{B}_{0,jk} = (B_{10,jk}, \dots, B_{s_0,jk})^T$, we can write $Y_{i,jk} = \mathbf{x}_i^T \mathbf{B}_{0,jk} + E_{i,jk}$. Multiplying \mathbf{X}_j on both sides and taking expectations yield $\mathbf{x}^T \mathbf{B}_{0,jk} = E(\mathbf{x}_j * Y_{i,jk})$. Therefore, we have

$$\begin{aligned} |\mathcal{M}^1| \gamma_n^2 / 4 &\leq \sum_{j,k} \left\| \Sigma_X \mathbf{B}_{0,jk} \right\|^2 \leq \lambda_{\max}(\Sigma_X) \sum_{j,k} \mathbf{B}_{0,jk}^T \mathbf{B}_{0,jk} \\ &= \lambda_{\max}(\Sigma_X) \sum_{j,k} \{\text{var}(Y_{i,jk}) - \text{var}(Y_{i,jk} | x_i)\} \leq pq \lambda_{\max}(\Sigma_X). \end{aligned}$$

By Assumption A4, we have $|\mathcal{M}^1| \leq 4pq \lambda_{\max}(\Sigma_X) \gamma_n^{-2} = O(n^{2\kappa + \tau})$, which implies that $|\mathcal{M}^0| \leq |\mathcal{M}^1| = O(n^{2\kappa + \tau})$.

Combining the results of above two steps leads to

$$P(|\widehat{\mathcal{M}}_{\gamma_n}| = O(n^{2\kappa + \tau})) \geq P(\widehat{\mathcal{M}}_{\gamma_n} \subseteq \mathcal{M}^0) \rightarrow 1.$$

This completes the proof of Theorem 2.

Theorems 3, 4 and 5 are theoretical results for our estimation procedure, and we assume $\widehat{\mathcal{M}} = \mathcal{M}$ and $\widehat{\mathcal{M}}$ is fixed.

Proof of Theorem 3: Without loss of generality, for the proof of Theorem 3, we assume $\widehat{\mathcal{M}} = \mathcal{M} = \{1, \dots, s\}$ with s fixed for notation simplicity. We first prove Theorem 3 (i). We define

$$\begin{aligned} L(\mathbf{\Delta}_1, \dots, \mathbf{\Delta}_s) &= \lambda^{-2} \{Q(\lambda \mathbf{\Delta}_1 + \mathbf{B}_{10}, \dots, \lambda \mathbf{\Delta}_s + \mathbf{B}_{s0}) - Q(\mathbf{B}_{10}, \dots, \mathbf{B}_{s0})\} \\ &= 2^{-1} \sum_{l=1}^s \sum_{l'=1}^s n^{-1} \left(\sum_{i=1}^n x_{il} x_{il'} \right) \text{tr}(\mathbf{\Delta}_l^T \mathbf{\Delta}_{l'}) - \lambda^{-1} \sum_l \text{tr}(\mathbf{\Delta}_l^T n^{-1} \sum_{i=1}^n x_{il} \mathbf{E}_i) \\ &\quad + \lambda^{-1} \sum_l \left\{ \left\| \mathbf{B}_{l0} + \lambda \mathbf{\Delta}_l \right\|_* - \left\| \mathbf{B}_{l0} \right\|_* \right\}, \end{aligned}$$

where $\mathbf{\Delta}_l = \lambda^{-1}(\mathbf{B}_l - \mathbf{B}_{l0})$ for $l = 1, \dots, s$. Therefore, we have

$$(\widehat{\mathbf{\Delta}}_1, \dots, \widehat{\mathbf{\Delta}}_s) = \arg \min \{L(\mathbf{\Delta}_1, \dots, \mathbf{\Delta}_s)\},$$

where $\widehat{\mathbf{\Delta}}_l = \lambda^{-1}(\widehat{\mathbf{B}}_l - \mathbf{B}_{l0})$ for $l = 1, \dots, s$.

When $\lambda \rightarrow 0$, $n^{1/2} \lambda \rightarrow \infty$, we have

$$n^{-1} \sum_{i=1}^n x_{il} x_{il'} \rightarrow_p \Sigma_{\mathcal{M}, ll'}, \quad \text{for every } 1 \leq l, l' \leq s,$$

where $\Sigma_{\mathcal{M}, ll'}$ is the (l, l') -th element of $\Sigma_{\mathcal{M}}$ for $1 \leq l, l' \leq s$. By the Central Limit Theorem, $n^{-1/2} \sum_{i=1}^n x_{il} \mathbf{E}_i$ converges in distribution to a normally distributed matrix \mathbf{D}_l with mean $\mathbf{0}$ and $\text{var}(\text{vec}(\mathbf{D}_l)) = m_{ll}$ for every $1 \leq l \leq s$. Hence

$$\lambda^{-1}n^{-1} \sum_{i=1}^n x_{il} \mathbf{E}_i = \lambda^{-1}n^{-1} / 2 O_p(1) \rightarrow_p 0, \quad \text{for every } 1 \leq l \leq s.$$

For every $l = 1, \dots, s$, recall that the singular value decomposition of \mathbf{B}_{l0} is $\mathbf{U}_{l0} \Theta_{l0} \mathbf{V}_{l0}^T$, and \mathbf{U}_{l0}^\perp , and \mathbf{V}_{l0}^\perp denote orthogonal complements of \mathbf{U}_{l0} and \mathbf{V}_{l0} , respectively. By Lemma 3, we have

$$\lambda^{-1} \sum_j \left\{ \left\| \mathbf{B}_{l0} + \lambda \Delta_l \right\|_* - \left\| \mathbf{B}_{l0} \right\|_* \right\} \rightarrow \sum_{l=1}^s \text{tr}(\mathbf{U}_{l0}^T \Delta_l \mathbf{V}_{l0}) + \sum_{l=1}^s \left\| (\mathbf{U}_{l0}^\perp)^T \Delta_l \mathbf{V}_{l0}^\perp \right\|_*.$$

Consequently, $L(\cdot_1, \dots, \cdot_s) \rightarrow_p L^0(\cdot_1, \dots, \cdot_s)$ for each $\cdot_l \in G_l, l = 1, \dots, s$ with G_l 's compact sets in $\mathbb{R}^{p \times q}$, where

$$\begin{aligned} & L^0(\Delta_1, \dots, \Delta_s) \\ &= 2^{-1} \sum_{l=1}^s \sum_{l'=1}^s \sum_{\mathcal{M}, \Pi} \text{tr}(\Delta_l^T \Delta_{l'}) + \sum_{l=1}^s \text{tr}(\mathbf{U}_{l0}^T \Delta_l \mathbf{V}_{l0}) + \sum_{l=1}^s \left\| (\mathbf{U}_{l0}^\perp)^T \Delta_l \mathbf{V}_{l0}^\perp \right\|_* \end{aligned}$$

One can see that $L^0(\cdot_1, \dots, \cdot_s)$ is convex, hence it has unique minimum value point $(\cdot_{10}, \dots, \cdot_{s0})$. As $L(\cdot_1, \dots, \cdot_s)$ is also convex, by (Knight and Fu, 2000) we have $\widehat{\Delta}_l \rightarrow_p \Delta_{l0}$. This implies that $\lambda^{-1}(\widehat{\mathbf{B}}_l - \mathbf{B}_{l0}) = O_p(1), l = 1, \dots, s$.

We second prove Theorem 3 (ii). We define

$$\begin{aligned} f(\Psi_1, \dots, \Psi_s) &= n(Q(n^{-1} / 2 \Psi_l + \mathbf{B}_{l0}) - Q(\mathbf{B}_{l0})) \\ &= 2^{-1} \sum_{l=1}^s \sum_{l'=1}^s n^{-1} \left(\sum_{i=1}^n x_{il} x_{il'} \right) \text{tr}(\Psi_l^T \Psi_{l'}) - \sum_l \text{tr}(\Psi_l^T n^{-1} / 2 \sum_{i=1}^n x_{il} * \mathbf{E}_i) \\ &\quad + \lambda n \sum_l \left\{ \left\| \mathbf{B}_{l0} + n^{-1} / 2 \Psi_l \right\|_* - \left\| \mathbf{B}_{l0} \right\|_* \right\}, \end{aligned}$$

where $\Psi_l = n^{1/2}(\mathbf{B}_l - \mathbf{B}_{l0})$ for $l = 1, \dots, s$. Let $(\widehat{\Psi}_1, \dots, \widehat{\Psi}_s) = \arg \min \{ f(\Psi_1, \dots, \Psi_s) \}$, then we have that $\widehat{\Psi}_l = n^{1/2}(\widehat{\mathbf{B}}_l - \mathbf{B}_{l0}), l = 1, \dots, s$. Under the Assumption (A6), and $n^{1/2}\lambda \rightarrow \rho$, we have $f(\Psi_1, \dots, \Psi_s) \rightarrow f^\theta(\Psi_1, \dots, \Psi_s)$ and

$$\begin{aligned} & f^\theta(\Psi_1, \dots, \Psi_s) \\ &= 2^{-1} \sum_{l=1}^s \sum_{l'=1}^s \sum_{\mathcal{M}, \Pi} \text{tr}(\Psi_l^T \Psi_{l'}) - \sum_{l=1}^s \text{tr}(\Psi_l^T \mathbf{D}_l) + \rho \left\{ \sum_{l=1}^s \text{tr}(\mathbf{U}_{l0}^T \Delta_l \mathbf{V}_{l0}) + \sum_{l=1}^s \left\| (\mathbf{U}_{l0}^\perp)^T \Delta_l \mathbf{V}_{l0}^\perp \right\|_* \right\}, \end{aligned}$$

where \mathbf{D}_l is a random matrix, and $\text{vec}(\mathbf{D}_l)$ is normally distributed. One can see that $f^\theta(\Psi_1, \dots, \Psi_s)$ is convex, hence it has unique minimum value point $(\Psi_{10}, \dots, \Psi_{s0})$ with $\Psi_{l0} = O_p(1)$ for $l = 1, \dots, s$. Consequently, by (Knight and Fu, 2000), we have $\widehat{\Psi}_l \rightarrow_d \Psi_{l0}$ for $l = 1,$

..., s , which indicates that $n^{1/2}(\widehat{\mathbf{B}}_l - \mathbf{B}_{l0}) = O_p(1)$ for $l = 1, \dots, s$. This completes the proof of Theorem 3.

Proof of Theorem 4: Without loss of generality, for the proof of Theorem 4, we assume $\widehat{\mathcal{M}} = \mathcal{M} = \{1, \dots, s\}$ with s fixed for notation simplicity. It follows from Theorem 3(i) that $\lambda^{-1}(\widehat{\mathbf{B}}_l - \mathbf{B}_{l0}) = O_p(1)$ holds for every $1 \leq l \leq s$. Since the rank function is lower semi-continuous, $P(\text{rank}(\widehat{\mathbf{B}}_l) \geq \text{rank}(\mathbf{B}_{l0})) \rightarrow 1$. We will then prove $\text{rank}(\widehat{\mathbf{B}}) = \text{rank}(\mathbf{B}_{l0})$ for every $1 \leq l \leq s$ with probability tending to one.

Denote the singular value decomposition of $\widehat{\mathbf{B}}_l$ as $\widehat{\mathbf{B}}_l = \widehat{\mathbf{U}}_l \widehat{\boldsymbol{\Theta}}_l \widehat{\mathbf{V}}_l^T$, where $\widehat{\mathbf{U}}_l \in \mathbb{R}^{p \times p}$ and $\widehat{\mathbf{V}}_l \in \mathbb{R}^{q \times q}$. Let $\widehat{\mathbf{U}}_l^\perp$ be the submatrix of $\widehat{\mathbf{U}}_l$ without the first r_l columns, and $\widehat{\mathbf{V}}_l^\perp$ is the submatrix of $\widehat{\mathbf{V}}_l$ without the first r_l columns, where r_l is the rank of \mathbf{B}_{l0} . Denote the rank of $\widehat{\mathbf{B}}_l$ by \widehat{r}_l . We prove the theorem by two steps.

Step 1. In this step, we will show if

$$\left\| (\widehat{\mathbf{U}}_l^\perp)^T \left(n^{-1} \sum_{i=1}^n x_{il} \left[\sum_{l'=1}^s x_{il'} * (\widehat{\mathbf{B}}_{l'} - \mathbf{B}_{l'0}) - \mathbf{E}_i \right] \widehat{\mathbf{V}}_l^\perp \right) \right\|_{op} < \lambda,$$

then $\widehat{r}_l = r_l$. We will prove the statement by contradiction.

Let $\widehat{\mathbf{U}}_{l1}$ be the submatrix of $\widehat{\mathbf{U}}_l$ corresponding to the first \widehat{r}_l columns, and $\widehat{\mathbf{V}}_{l1}$ be the submatrix of $\widehat{\mathbf{V}}_l$ corresponding to the first \widehat{r}_l columns. If $\widehat{r}_l \geq r_l$, we can write $\widehat{\mathbf{U}}_l^\perp, \widehat{\mathbf{V}}_l^\perp$ as $(\widehat{\mathbf{U}}_{l1}^\perp, \widehat{\mathbf{U}}_{l2}^\perp)$, and $(\widehat{\mathbf{V}}_{l1}^\perp, \widehat{\mathbf{V}}_{l2}^\perp)$ respectively, where $\widehat{\mathbf{U}}_{l1}^\perp \in \mathbb{R}^{p \times (\widehat{r}_l - r_l)$, $\widehat{\mathbf{U}}_{l2}^\perp \in \mathbb{R}^{p \times (p - \widehat{r}_l)}$, $\widehat{\mathbf{V}}_{l1}^\perp \in \mathbb{R}^{q \times (\widehat{r}_l - r_l)$, and $\widehat{\mathbf{V}}_{l2}^\perp \in \mathbb{R}^{q \times (q - \widehat{r}_l)}$. By the definition of $\widehat{\mathbf{B}}_l$, we have

$$\widehat{\mathbf{B}}_l = \underset{\mathbf{B}_l}{\text{argmin}} \frac{1}{2n} \sum_{i=1}^n \left\| Y_i - \sum_{l' \neq l} x_{il'} \widehat{\mathbf{B}}_{l'} - x_{il} \mathbf{B}_l \right\|_F^2 + \lambda \left\| \mathbf{B}_l \right\|_*.$$

Hence, by Lemma 3, we have

$$\left\{ n^{-1} \sum_{i=1}^n x_{il} \left[\sum_{l'=1}^s x_{il'} * (\widehat{\mathbf{B}}_{l'} - \mathbf{B}_{l'0}) - \mathbf{E}_i \right] \right\} + \lambda (\widehat{\mathbf{U}}_{l1} \widehat{\mathbf{V}}_{l1}^T + \mathbf{N}_l) = 0,$$

with $\widehat{\mathbf{U}}_{l1}^T \mathbf{N}_l = 0$, $\widehat{\mathbf{N}}_l \widehat{\mathbf{V}}_{l1} = 0$ and $\|\mathbf{N}_l\|_{op} \leq 1$. Furthermore, we have

$$\begin{aligned}
 & (\widehat{\mathbf{U}}_l^\perp)^T \{n^{-1} \sum_{i=1}^n x_{il} [\sum_{l'=1}^s x_{il'} * (\widehat{\mathbf{B}}_{l'} - \mathbf{B}_{l'0}) - \mathbf{E}_i]\} \widehat{\mathbf{V}}_l^\perp \\
 &= -\lambda (\widehat{\mathbf{U}}_l^\perp)^T (\widehat{\mathbf{U}}_{l1} \widehat{\mathbf{V}}_{l1}^T + \mathbf{N}_l) \widehat{\mathbf{V}}_l^\perp \\
 &= -\lambda (\widehat{\mathbf{U}}_{l1}^\perp, \widehat{\mathbf{U}}_{l2}^\perp)^T (\widehat{\mathbf{U}}_{l1}^\perp (\widehat{\mathbf{V}}_{l1}^\perp)^T + \mathbf{N}_l) (\widehat{\mathbf{V}}_{l1}^\perp, \widehat{\mathbf{V}}_{l2}^\perp) \\
 &= -\lambda \begin{pmatrix} \mathbf{1}(\widehat{r}_l - r_l) \times (\widehat{r}_l - r_l) & 0 \\ 0 & (\widehat{\mathbf{U}}_{l2}^\perp)^T \mathbf{N}_l \widehat{\mathbf{V}}_{l2}^\perp \end{pmatrix}.
 \end{aligned}$$

From the above formula, it follows that we have

$$\left\| (\widehat{\mathbf{U}}_l^\perp)^T \{n^{-1} \sum_{i=1}^n x_{il} [\sum_{l'=1}^s x_{il'} * (\widehat{\mathbf{B}}_{l'} - \mathbf{B}_{l'0}) - \mathbf{E}_i]\} \widehat{\mathbf{V}}_l^\perp \right\|_{op} = \lambda \text{ as long as } \widehat{r}_l > r_l.$$

Consequently,

$$\text{if } \left\| (\widehat{\mathbf{U}}_l^\perp)^T \{n^{-1} \sum_{i=1}^n x_{il} [\sum_{l'=1}^s x_{il'} * (\widehat{\mathbf{B}}_{l'} - \mathbf{B}_{l'0}) - \mathbf{E}_i]\} \widehat{\mathbf{V}}_l^\perp \right\|_{op} < \lambda, \text{ we have } \widehat{r}_l = r_l.$$

Step 2. In this step, we will prove that with probability tending to 1, one has

$$\left\| (\widehat{\mathbf{U}}_l^\perp)^T \{n^{-1} \sum_{i=1}^n x_{il} [\sum_{l'=1}^s x_{il'} * (\widehat{\mathbf{B}}_{l'} - \mathbf{B}_{l'0}) - \mathbf{E}_i]\} \widehat{\mathbf{V}}_l^\perp \right\|_{op} < \lambda.$$

We have

$$\begin{aligned}
 & (\widehat{\mathbf{U}}_l^\perp)^T \{n^{-1} \sum_{i=1}^n x_{il} [\sum_{l'=1}^s x_{il'} * (\widehat{\mathbf{B}}_{l'} - \mathbf{B}_{l'0}) - \mathbf{E}_i]\} \widehat{\mathbf{V}}_l^\perp \\
 &= (\widehat{\mathbf{U}}_l^\perp)^T \{ \lambda \sum_{l'=1}^s (\Sigma_{\mathcal{M}, l'} + o_p(1)) \widehat{\Delta}_l - O_p(n^{-1/2}) \} \widehat{\mathbf{V}}_l^\perp = \lambda (\widehat{\mathbf{U}}_l^\perp)^T \sum_{l'=1}^s \Sigma_{\mathcal{M}, l'} \widehat{\Delta}_l \widehat{\mathbf{V}}_l^\perp + o_p(\lambda).
 \end{aligned}$$

Since $\widehat{\mathbf{B}}_l$ is a consistent estimator of \mathbf{B}_{l0} , we have $\widehat{\mathbf{U}}_l^\perp (\widehat{\mathbf{U}}_l^\perp)^T = \widehat{\mathbf{U}}_{l0}^\perp (\widehat{\mathbf{U}}_{l0}^\perp)^T + o_p(1)$ and $\widehat{\mathbf{V}}_l^\perp (\widehat{\mathbf{V}}_l^\perp)^T = \widehat{\mathbf{V}}_{l0}^\perp (\widehat{\mathbf{V}}_{l0}^\perp)^T + o_p(1)$. Consequently, we have

$$\begin{aligned}
 & \left\| (\widehat{\mathbf{U}}_l^\perp)^T \{n^{-1} \sum_{i=1}^n x_{il} [\sum_{l'=1}^s x_{il'} * (\widehat{\mathbf{B}}_{l'} - \mathbf{B}_{l'0}) - \mathbf{E}_i]\} \widehat{\mathbf{V}}_l^\perp \right\|_{op} \\
 &= \left\| \widehat{\mathbf{U}}_l^\perp (\widehat{\mathbf{U}}_l^\perp)^T \{n^{-1} \sum_{i=1}^n x_{il} [\sum_{l'=1}^s x_{il'} * (\widehat{\mathbf{B}}_{l'} - \mathbf{B}_{l'0}) - \mathbf{E}_i]\} \widehat{\mathbf{V}}_l^\perp (\widehat{\mathbf{V}}_l^\perp)^T \right\|_{op} \\
 &= \lambda \left\| \widehat{\mathbf{U}}_{l0}^\perp (\widehat{\mathbf{U}}_{l0}^\perp)^T \left(\sum_{l'=1}^s \Sigma_{\mathcal{M}, l'} \widehat{\Delta}_l \widehat{\mathbf{V}}_{l0}^\perp (\widehat{\mathbf{V}}_{l0}^\perp)^T \right) \right\|_{op} + o_p(\lambda) \\
 &= \lambda \left\| \widehat{\mathbf{U}}_{l0}^\perp (\widehat{\mathbf{U}}_{l0}^\perp)^T \left(\sum_{l'=1}^s \Sigma_{\mathcal{M}, l'} \Delta_{l'0} \widehat{\mathbf{V}}_{l0}^\perp (\widehat{\mathbf{V}}_{l0}^\perp)^T (1 + o_p(1)) \right) \right\|_{op} + o_p(\lambda) \\
 &= \lambda \left\| \mathbf{U}_{l0}^\perp \mathbf{\Lambda}_L (\mathbf{V}_{l0}^\perp)^T \right\|_{op} + o_p(\lambda) = \lambda \{ \|\mathbf{\Lambda}_l\|_{op} + o_p(1) \}.
 \end{aligned}$$

As $\|\mathbf{A}\|_{op} < 1$, we have $\left\| \left(\widehat{\mathbf{U}}_l^\perp \right)^T \left\{ n^{-1} \sum_{i=1}^n x_{il} \left[\sum_{i'=1}^s x_{i'l'} * \left(\widehat{\mathbf{B}}_{l'} - \mathbf{B}_{l'0} \right) - \mathbf{E}_i \widehat{\mathbf{V}}_l^\perp \right] \right\} \right\|_{op} < \lambda$. with probability 1. This completes the proof of Theorem 4.

Proof of Theorem 5. Without loss of generality, for the proof of Theorem 5, we assume $\widehat{\mathcal{M}} = \mathcal{M} = \{1, \dots, s\}$ with s fixed for notation simplicity. To prove Theorem 5, we first introduce some notations and definitions used in Negahban et al. (2012). Given a pair of subspaces $M \subseteq \overline{M}$, a norm based regularizer J is decomposable with respect to $((M, \overline{M}^\perp))$ if

$$J(\theta + \gamma) = J(\theta) + J(\gamma) \text{ for all } \theta \in M \text{ and } \gamma \in \overline{M}^\perp,$$

where \overline{M}^\perp is the orthogonal complement of the space \overline{M} defined as $\overline{M}^\perp = \{v \mid \langle u, v \rangle = 0 \text{ for all } u \in \overline{M}\}$.

We define the projection operator

$$\Pi_M(u) = \operatorname{argmin}_{v \in M} \|u - v\|.$$

Similarly, we can define the projections Π_{M^\perp} , $\Pi_{\overline{M}}$ and $\Pi_{\overline{M}^\perp}$.

We then introduce the definition of the subspace compatibility constant. For the subspace M , the subspace compatibility constant with respect to the pair $(J, \|\cdot\|)$ is given by

$$\psi(M) := \sup_{u \in M \setminus \{0\}} \frac{J(u)}{\|u\|}$$

We introduce the definition of restricted strong convexity. For a loss function $L(\theta)$, define $\delta L(\Delta, \theta) = L(\theta + \Delta) - L(\theta) - \langle \nabla L(\theta), \Delta \rangle$, where $\nabla L(\theta) = \frac{dL(\theta)}{d\theta}$. The loss function satisfies a restricted strong convexity condition with curvature $\kappa_L > 0$ and tolerance function τ_L if

$$\delta L(\Delta, \theta) \geq \kappa_L \|\Delta\|^2 - \tau_L^2(\theta) \text{ for all } \Delta \in \mathcal{C}(M, \overline{M}^\perp, \theta),$$

where $\mathcal{C}(M, \overline{M}^\perp, \theta) = \{\Delta \mid J(\Delta \overline{M}^\perp) \leq 3J(\Delta \overline{M}) + 4J(\theta \overline{M}^\perp)\}$.

Now we begin to prove Theorem 5. We need to use the result in Theorem 1 of Negahban et al. (2012). We first check the conditions of the theorem under our context.

Recall that $\mathbf{B} = [\mathbf{B}_1, \dots, \mathbf{B}_s] \in \mathbb{R}^{p \times qs}$, and $r_l = \operatorname{rank}(\mathbf{B}_{l0})$. Let us consider the class of matrices $\Theta_l \in \mathbb{R}^{p \times q}$ that have $\operatorname{rank} r_l \leq \min\{p, q\}$ and we define $\Theta = [\Theta_1, \dots, \Theta_s] \in \mathbb{R}^{p \times qs}$.

Let $\operatorname{row}(\Theta_l) \subseteq \mathbb{R}^p$ and $\operatorname{col}(\Theta_l) \subseteq \mathbb{R}^q$ denote its row space and column space, respectively.

Let U_l and V_l be a given pair of r_l -dimensional subspaces $U_l \subseteq \mathbb{R}^p$ and $V_l \subseteq \mathbb{R}^q$. Define $U =$

$[U_1, \dots, U_s]$ and $V = [V_1, \dots, V_s]$. For a given pair (U, V) , we can define the subspaces $M(U, V)$, $\bar{M}(U, V)$ and $\bar{M}^\perp(U, V)$ of $\mathbb{R}^{p \times qs}$ given by

$$M(U, V) = \{\Theta \in \mathbb{R}^{p \times qs} \mid \text{row}(\Theta_l) \subseteq V_l \text{ and } \text{col}(\Theta_l) \subseteq U_l \text{ for } 1 \leq l \leq s\},$$

$$\bar{M}(U, V) = \{\Theta \in \mathbb{R}^{p \times qs} \mid \text{row}(\Theta_l) \subseteq V_l \text{ or } \text{col}(\Theta_l) \subseteq U_l \text{ for } 1 \leq l \leq s\},$$

and

$$\bar{M}^\perp(U, V) = \{\Theta \in \mathbb{R}^{p \times qs} \mid \text{row}(\Theta_l) \subseteq V_l^\perp \text{ and } \text{col}(\Theta_l) \subseteq U_l^\perp \text{ for } 1 \leq l \leq s\},$$

where $\bar{M}^\perp(U, V)$ is the orthogonal complement of the space $\bar{M}(U, V)$. For simplicity, we will use M, \bar{M} and \bar{M}^\perp to denote $M(U, V), \bar{M}(U, V)$ and $\bar{M}^\perp(U, V)$ respectively in the following proof.

Define $J(\mathbf{B}) = \sum_{l=1}^s \|\mathbf{B}_l\|_*$, and we can easily see $J(\mathbf{B})$ is a norm. It is easy to see that the norm J is decomposable with respect to the subspace pair $(M, (M, \bar{M}^\perp))$, where $M \subseteq \bar{M}$. Therefore, the regularizer J satisfies Condition (G1) in Negahban et al. (2012).

Under condition (A9), it is easy to see the loss function R is convex and differentiable, and satisfies the restricted strong convexity with curvature $\kappa_L = C_L$ and tolerance $\tau_L = 0$, and therefore the Condition (G2) in Negahban et al. (2012) holds.

After we check the conditions, we need to calculate $\psi(\bar{M})$ and $R(\{\mathbf{B}_0\}_{\bar{M}^\perp})$. It is easy to see $R(\{\mathbf{B}_0\}_{\bar{M}^\perp})$. For $\psi(\bar{M})$, one has

$$\begin{aligned} \psi(\bar{M}) &= \sup_{u \in \bar{M} \setminus \{0\}} \frac{J(u)}{\|u\|} = \sup_{\mathbf{B}_l \in \bar{M} \setminus \{0\}} \frac{\sum_{l=1}^s \|\mathbf{B}_l\|_*}{\|\mathbf{B}\|_F} \leq \frac{\sum_{l=1}^s \sqrt{2r_l} \|\mathbf{B}_l\|_F}{\|\mathbf{B}\|_F} \\ &\leq \frac{\sqrt{\sum_{l=1}^s (\sqrt{2r_l})^2} \sqrt{\sum_{l=1}^s \|\mathbf{B}_l\|_F^2}}{\|\mathbf{B}\|_F} \leq \sqrt{2 \sum_{l=1}^s r_l}. \end{aligned}$$

Therefore, by Theorem 1 in Negahban et al. (2012), when $\lambda = 2J^*(\nabla R(\mathbf{B}_0))$, one has $\|\hat{\mathbf{B}} - \mathbf{B}_0\|_F^2 \leq C(\sum_{l=1}^s r_l)\lambda^2 C_L^{-2}$ for some constant $C > 0$.

The term $J^*(\nabla R(\mathbf{B}_0))$ is actually a random quantity, and our next step is to derive the order of this term.

Define $J^*(\cdot)$ as the dual norm of $J(\cdot)$. For any matrix $\mathbf{A} = [\mathbf{A}_1, \dots, \mathbf{A}_s] \in \mathbb{R}^{p \times qs}$, we will first prove the following result

$$J^*(\mathbf{A}) = \sup_{J(\mathbf{B}) \leq 1} \langle \mathbf{A}, \mathbf{B} \rangle = \max_{1 \leq l \leq s} \|\mathbf{A}_l\|_{op}. \tag{8}$$

To prove (8), we first show that $J^*(\mathbf{A}) = \max_{1 \leq l \leq s} \|\mathbf{A}_l\|_{op}$. Let $\mathbf{B}^{(l)} = [\mathbf{B}_1^{(l)}, \dots, \mathbf{B}_s^{(l)}]$ with $\mathbf{B}_k^{(l)} = \mathbf{0}$ for any $k \neq l$ and $\|\mathbf{B}_l^{(l)}\|_* \leq 1$. One has

$$J^*(\mathbf{A}) \geq \sup_{\|\mathbf{B}^{(l)}\|_* \leq 1} \left\langle \mathbf{A}, \mathbf{B}^{(l)} \right\rangle = \sup_{\|\mathbf{B}_l^{(l)}\|_* \leq 1} \left\langle \mathbf{A}_l, \mathbf{B}_l^{(l)} \right\rangle = \|\mathbf{A}_l\|_{op}.$$

It is easy to see $J^*(\mathbf{A}) \leq \|\mathbf{A}_l\|_{op}$ holds for any $1 \leq l \leq s$. Consequently, one has $J^*(\mathbf{A}) = \max_{1 \leq l \leq s} \|\mathbf{A}_l\|_{op}$.

Our next step is to show that $J^*(\mathbf{A}) = \max_{1 \leq l \leq s} \|\mathbf{A}_l\|_{op}$. Define the singular value decomposition of $\mathbf{B}_l = \mathbf{U}_l \Theta_l \mathbf{V}_l^T$. One has

$$\begin{aligned} J^*(\mathbf{A}) &= \sup_{J(\mathbf{B}) \leq 1} \left\{ \sum_{l=1}^s \left\langle \mathbf{U}_l \Theta_l \mathbf{V}_l^T, \mathbf{A}_l \right\rangle \right\} \\ &= \sup_{J(\mathbf{B}) \leq 1} \left\{ \sum_{l=1}^s \text{Tr}(\mathbf{V}_l \Theta_l \mathbf{U}_l^T \mathbf{A}_l \mathbf{V}_l) \right\} = \sup_{J(\mathbf{B}) \leq 1} \left\{ \sum_{l=1}^s \text{Tr}(\Theta_l \mathbf{U}_l^T \mathbf{A}_l \mathbf{V}_l) \right\} \\ &= \sup_{J(\mathbf{B}) \leq 1} \left\{ \sum_{l=1}^s \left\langle \mathbf{U}_l^T \mathbf{A}_l \mathbf{V}_l \Theta_l \right\rangle \right\} = \sup_{J(\mathbf{B}) \leq 1} \left\{ \sum_{l=1}^s \sum_{k=1}^{\min\{p, q\}} \theta_{lk} (\mathbf{U}_l^T \mathbf{A}_l \mathbf{V}_l)_{kk} \right\} \\ &= \sup_{J(\mathbf{B}) \leq 1} \left\{ \sum_{l=1}^s \sum_{k=1}^{\min\{p, q\}} \theta_{lk} (\mathbf{U}_l)_{(k)}^T \mathbf{A}_l (\mathbf{V}_l)_{(k)} \right\} \\ &\leq \sup_{J(\mathbf{B}) \leq 1} \left\{ \sum_{l=1}^s \sum_{k=1}^{\min\{p, q\}} \theta_{lk} \|\mathbf{A}_l\|_{op} \right\} \leq \sup_{J(\mathbf{B}) \leq 1} \left\{ \sum_{l=1}^s \sum_{k=1}^{\min\{p, q\}} \theta_{lk} \max_{1 \leq l \leq s} \|\mathbf{A}_l\|_{op} \right\} \\ &\leq \max_{1 \leq l \leq s} \|\mathbf{A}_l\|_{op}, \end{aligned}$$

where θ_{lk} is the lk th diagonal element of the diagonal matrix Θ_l , $(\mathbf{U}_l^T \mathbf{A}_l \mathbf{V}_l)_{kk}$ is the kk th element of the matrix $\mathbf{U}_l^T \mathbf{A}_l \mathbf{V}_l$, $(\mathbf{U}_l)_{(k)}$ and $(\mathbf{V}_l)_{(k)}$ are the k th column of the matrices \mathbf{U}_l and \mathbf{V}_l respectively.

Combining the two inequalities, we show that $J^*(\mathbf{A}) = \max_{1 \leq l \leq s} \|\mathbf{A}_l\|_{op}$.

Next we need to calculate $J^*(\nabla R(\mathbf{B}_0))$, where $\nabla R(\mathbf{B}_0) = [\mathbf{D}_1, \dots, \mathbf{D}_s] \in \mathbb{R}^{p \times q \times s}$ with

$\mathbf{D}_l = -2n^{-1} \sum_{i=1}^n x_{il} * \mathbf{E}_i$. We first need to calculate $\|\mathbf{D}_l\|_{op}$. We know that operator norm is the dual norm of the trace norm.

From the definition of $J^*(\cdot)$, one has

$$\|\mathbf{D}_l\|_{op} = 2 \sup_{\|\mathbf{A}\|_* \leq 1} \left\langle \mathbf{A}, n^{-1} \sum_{i=1}^n x_{il} * \mathbf{E}_i \right\rangle$$

To obtain a bound for $\|\mathbf{D}_l\|_{op}$, we use similar technique as the one used in Raskutti and Yuan (2018). Let \mathbf{W}_i be a $p \times q$ random matrix with each entry i.i.d. standard normal. Assuming condition (A11) and by Lemma 4, conditioning on x_{il} , we get

$$P\left\{\sup_{\|\mathbf{A}\|_* \leq 1} \left\langle \mathbf{A}, n^{-1} \sum_{i=1}^n x_{il} * \mathbf{E}_i \right\rangle > t\right\} \leq P\left\{\sup_{\|\mathbf{A}\|_* \leq 1} \left\langle \mathbf{A}, n^{-1} \sum_{i=1}^n x_{il} * \mathbf{W}_i \right\rangle > \frac{t}{CU}\right\},$$

since $\Sigma_e \leq C_U^2 I_{pq \times pq}$.

As $\sup_{\|\mathbf{A}\|_* \leq 1} \left\langle \mathbf{A}, n^{-1} \sum_{i=1}^n x_{il} * \mathbf{W}_i \right\rangle = \|n^{-1} \sum_{i=1}^n x_{il} * \mathbf{W}_i\|_{op}$, conditioning on \mathbf{W}_i , each entry of the matrix $n^{-1} \sum_{i=1}^n x_{il} * \mathbf{W}_i$ is i.i.d $N(0, \frac{\|\mathbf{X}_l\|_{op}^2}{n^2})$, where $\mathbf{X}_l = (x_{1l}, \dots, x_{nl})^T$. Since $\frac{\|\mathbf{X}_l\|_{op}^2}{\sigma_l^2}$ is a χ^2 random variable with n degree of freedom, where $\sigma_l^2 = (\Sigma_{\mathcal{M}})_l$, one has

$$P\left\{\frac{\|\mathbf{X}_l\|_{op}^2}{n\sigma_l^2} \geq 4\right\} \leq \exp(-n)$$

using the tail bounds of χ^2 . Then combining with the standard random matrix theory, we know that $\|n^{-1} \sum_{i=1}^n x_{il} * \mathbf{W}_i\|_{op} \leq 2n^{-1/2} \sigma_l (p^{1/2} + q^{1/2})$ with probability at least $1 - c_1^* \exp\{-c_2^*(p+q)\} - \exp(-n)$ where c_1^* and c_2^* are some positive constants. Therefore, under conditions (A10) and (A12), there exist some positive constants c_1, c_2 and c_3 such that $\max_{l \in \mathcal{M}} \|\mathbf{D}_l\|_{op} \leq 4C_U n^{-1/2} (\max_{l \in \mathcal{M}} \sigma_l) (p^{1/2} + q^{1/2})$ holds with probability at least $1 - c_1 \exp\{-c_2(p+q)\} - c_3 \exp(-n)$. Thus, when $\lambda \geq 4C_U C_M^{1/2} n^{-1/2} (p^{1/2} + q^{1/2})$, $\lambda \mathcal{J}^*(\nabla R(\mathbf{B}_0))$ with probability at least $1 - c_1 \exp\{-c_2(p+q)\} - c_3 \exp(-n)$.

Therefore, with probability $1 - c_1 \exp\{-c_2(p+q)\} - c_3 \exp(-n)$, one has $\|\widehat{\mathbf{B}} - \mathbf{B}_0\|_F^2 \leq C(\sum_{l \in \mathcal{M}} r_l) \lambda^2 C_L^{-2}$ for some positive constant C . This completes the proof of Theorem 5.

Proof of Theorem 6:

To prove the theorem, we consider the event $\{\mathcal{M} \subseteq \widehat{\mathcal{M}}\}$ as it holds with probability goes to 1. We will derive the non-asymptotic error bound under the event $\{\mathcal{M} \subseteq \widehat{\mathcal{M}}\}$. Recall that $r_l = \text{rank}(\mathbf{B}_l)$, one has $r_l = 0$ for $l \in \mathcal{M}$. Let us consider the class of matrices $\Theta_l \in \mathbb{R}^{p \times q}$ that have rank $r_l = \min\{p, q\}$ and we define $\Theta = [\Theta_l, l \in \widehat{\mathcal{M}}] \in \mathbb{R}^{p \times q \times |\widehat{\mathcal{M}}|}$. Let $\text{row}(\Theta_l) \subseteq \mathbb{R}^p$ and $\text{col}(\Theta_l) \subseteq \mathbb{R}^q$ denote its row space and column space, respectively. Let U_l and V_l be a given pair of r_l -dimensional subspaces $U_l \subseteq \mathbb{R}^p$ and $V_l \subseteq \mathbb{R}^q$, respectively. Define $U = [U_l, l \in \widehat{\mathcal{M}}] \in \mathbb{R}^{p \times q \times |\widehat{\mathcal{M}}|}$ and $V = [V_l, l \in \widehat{\mathcal{M}}] \in \mathbb{R}^{p \times q \times |\widehat{\mathcal{M}}|}$. For a given pair (U, V) , we can define the subspaces $\widehat{M}(U, V)$, $\overline{M}(U, V)$ and $\overline{M}^\perp(U, V)$ of $\mathbb{R}^{p \times q \times |\widehat{\mathcal{M}}|}$ as follows:

$$\begin{aligned} \widehat{M}(U, V) &= \{\Theta \in \mathbb{R}^{p \times q} \mid \widehat{\mathcal{M}} \mid \mid \text{row}(\Theta_l) \subseteq V_l \text{ and } \text{col}(\Theta_l) \subseteq U_l \text{ for } l \in \widehat{\mathcal{M}}\}, \\ \overline{M}(U, V) &= \{\Theta \in \mathbb{R}^{p \times q} \mid \widehat{\mathcal{M}} \mid \mid \text{row}(\Theta_l) \subseteq V_l \text{ or } \text{col}(\Theta_l) \subseteq U_l \text{ for } l \in \widehat{\mathcal{M}}\}, \\ \overline{M}^\perp(U, V) &= \{\Theta \in \mathbb{R}^{p \times q} \mid \widehat{\mathcal{M}} \mid \mid \text{row}(\Theta_l) \subseteq V_l^\perp \text{ and } \text{col}(\Theta_l) \subseteq U_l^\perp \text{ for } l \in \widehat{\mathcal{M}}\}, \end{aligned}$$

where $\overline{M}^\perp(U, V)$ is the orthogonal complement of the space $\overline{M}(U, V)$. For simplicity, we will use \widehat{M} , \overline{M} and \overline{M}^\perp to denote $\widehat{M}(U, V)$, $\overline{M}(U, V)$ and $\overline{M}^\perp(U, V)$, respectively.

For the norm $J(\mathbf{B}^{\widehat{\mathcal{M}}}) = \sum_{l \in \widehat{\mathcal{M}}} \|\mathbf{B}_l\|_*$, it is easy to see that the norm J is decomposable with respect to the subspace pair $(\widehat{M}, \overline{M}^\perp)$, where $\widehat{M} \subseteq \overline{M}$. Therefore, the regularizer J satisfies Condition (G1) in Negahban et al. (2012).

We need to calculate $\psi_{\overline{M}}$ and $R(\{\mathbf{B}_0^{\widehat{\mathcal{M}}}\}_{\overline{M}^\perp})$. It is easy to see $R(\{\mathbf{B}_0^{\widehat{\mathcal{M}}}\}_{\overline{M}^\perp}) = 0$. For $\psi_{\overline{M}}$, since $r_l = 0$ holds for $l \in \mathcal{M}$, one has

$$\begin{aligned} \psi_{\overline{M}} &= \sup_{u \in \overline{M} \setminus \{0\}} \frac{J(u)}{\|u\|} = \sup_{\mathbf{B}_l \in \overline{M} \setminus \{0\}} \frac{\sum_{l \in \widehat{\mathcal{M}}} \|\mathbf{B}_l\|_*}{\|\mathbf{B}^{\widehat{\mathcal{M}}}\|_F} \leq \frac{\sum_{l \in \mathcal{M}} \sqrt{2r_l} \|\mathbf{B}_l\|_F}{\sqrt{\sum_{l \in \mathcal{M}} \|\mathbf{B}_l\|_F^2}} \\ &\leq \frac{\sqrt{\sum_{l \in \mathcal{M}} (\sqrt{2r_l})^2} \sqrt{\sum_{l \in \mathcal{M}} \|\mathbf{B}_l\|_F^2}}{\sqrt{\sum_{l \in \mathcal{M}} \|\mathbf{B}_l\|_F^2}} \leq \sqrt{2 \sum_{l \in \mathcal{M}} r_l}. \end{aligned}$$

For any $\Delta \in \mathbb{R}^{p \times q} \mid \widehat{\mathcal{M}} \mid$, we define $F: \mathbb{R}^{p \times q} \mid \widehat{\mathcal{M}} \mid \rightarrow \mathbb{R}$ as

$$F(\Delta) := R(\mathbf{B}_0^{\widehat{\mathcal{M}}} + \Delta) - R(\mathbf{B}_0^{\widehat{\mathcal{M}}}) + \lambda \{J(\mathbf{B}_0^{\widehat{\mathcal{M}}} + \Delta) - J(\mathbf{B}_0^{\widehat{\mathcal{M}}})\}.$$

We will derive a lower bound on $F(\Delta)$. In particular, we have

$$\begin{aligned} F(\Delta) &= R(\mathbf{B}_0^{\widehat{\mathcal{M}}} + \Delta) - R(\mathbf{B}_0^{\widehat{\mathcal{M}}}) + \lambda \{J(\mathbf{B}_0^{\widehat{\mathcal{M}}} + \Delta) - J(\mathbf{B}_0^{\widehat{\mathcal{M}}})\} \\ &\geq \left\langle \nabla(R(\mathbf{B}_0^{\widehat{\mathcal{M}}}), \Delta) \right\rangle + \iota_L \|\Delta\|^2 + \lambda \{J(\mathbf{B}_0^{\widehat{\mathcal{M}}} + \Delta) - J(\mathbf{B}_0^{\widehat{\mathcal{M}}})\} \\ &\geq \left\langle \nabla(R(\mathbf{B}_0^{\widehat{\mathcal{M}}}), \Delta) \right\rangle + \iota_L \|\Delta\|^2 + \lambda \{J(\Delta \overline{M}^\perp) - J(\Delta \overline{M}) - 2J(\mathbf{B}_0^{\widehat{\mathcal{M}}})_{\overline{M}^\perp}\}, \end{aligned}$$

where the first inequality follows from condition (A13) and the second inequality follows from Lemma 3 in Negahban et al. (2012) by applying to the pair $(\overline{M}, \overline{M}^\perp)$.

By the Cauchy-Schwarz inequality applied to the regularizer J and its dual J^* , we have $\left| \left\langle \nabla R(\mathbf{B}_0^{\widehat{\mathcal{M}}}), \Delta \right\rangle \right| \leq J^*(\nabla R(\mathbf{B}_0^{\widehat{\mathcal{M}}})) J(\Delta)$. Since $\lambda \geq 2J^*(\nabla R(\mathbf{B}_0^{\widehat{\mathcal{M}}}))$ holds by assumption, one has

$|\langle \nabla R(\widehat{\mathbf{B}}_0^{\widehat{\mathcal{M}}}), \Delta \rangle| \leq 0.5\lambda J(\Delta) \leq 0.5\lambda(J(\Delta_{\widehat{M}^\perp}) + J(\Delta_{\widehat{M}}))$, where the second inequality holds due to the triangle inequality. Therefore, we have

$$\begin{aligned} F(\Delta) &\geq -\frac{\lambda}{2}\{J(\Delta_{\widehat{M}^\perp}) + J(\Delta_{\widehat{M}})\} + \iota_L \|\Delta\|^2 + \lambda\{J(\Delta_{\widehat{M}^\perp}) - J(\Delta_{\widehat{M}}) - 2J((\widehat{\mathbf{B}}_0^{\widehat{\mathcal{M}}})_{\widehat{M}^\perp})\} \\ &= \iota_L \|\Delta\|^2 + \lambda\{\frac{1}{2}J(\Delta_{\widehat{M}^\perp}) - \frac{3}{2}J(\Delta_{\widehat{M}}) - 2J((\widehat{\mathbf{B}}_0^{\widehat{\mathcal{M}}})_{\widehat{M}^\perp})\} \\ &\geq \iota_L \|\Delta\|^2 - \frac{1}{2}\lambda\{3J(\Delta_{\widehat{M}}) + 4J((\widehat{\mathbf{B}}_0^{\widehat{\mathcal{M}}})_{\widehat{M}^\perp})\}. \end{aligned}$$

By the subspace compatibility, we have $J(\Delta_{\widehat{M}}) \leq \psi(\widehat{M})\|\Delta_{\widehat{M}}\|$. As the projection is non-expansive and $0 \in \widehat{M}$, one has $\|\Delta_{\widehat{M}}\| \leq \|\Delta\|$, and thus $J(\Delta_{\widehat{M}}) \leq \psi(\widehat{M})\|\Delta\|$. Substituting it into the previous inequality, and noticing that $J((\widehat{\mathbf{B}}_0^{\widehat{\mathcal{M}}})_{\widehat{M}^\perp}) = 0$, we obtain

$F(\Delta) \geq \iota_L \|\Delta\|^2 - \frac{3}{2}\lambda\psi(\widehat{M})\|\Delta\|$. The righthand side is a quadratic form of $\|\Delta\|$, as long as

$\|\Delta\|^2 > \frac{9\lambda^2}{4\iota_L^2}\psi(\widehat{M})$, one has $F(\Delta) > 0$. By Lemma 4 in Negahban et al. (2012), we have

$$\|\widehat{\mathbf{B}}^{\widehat{\mathcal{M}}} - \mathbf{B}_0^{\widehat{\mathcal{M}}}\|_F^2 \leq C(\sum_{l \in \widehat{\mathcal{M}}} r_l)\lambda^2 \iota_L^{-2} \text{ for some positive constant } C.$$

Next we need to calculate $J^*(\nabla R(\widehat{\mathbf{B}}_0^{\widehat{\mathcal{M}}}))$, where $\nabla R(\widehat{\mathbf{B}}_0^{\widehat{\mathcal{M}}}) = [\mathbf{D}_l, l \in \widehat{\mathcal{M}}] \in \mathbb{R}^{p \times q}^{|\widehat{\mathcal{M}}|}$ with $\mathbf{D}_l = -2n^{-1}\sum_{i=1}^n x_{il} * \mathbf{E}_i$. By similar argument as the one in the proof of Theorem 5, one has $J^*(\nabla R(\widehat{\mathbf{B}}_0^{\widehat{\mathcal{M}}})) = \max_{l \in \widehat{\mathcal{M}}} \|\mathbf{D}_l\|_{op}$. To calculate $\|\mathbf{D}_l\|_{op}$, by the same argument as the one in proof of Theorem 5, one has $\|n^{-1}\sum_{i=1}^n x_{il} * \mathbf{W}_i\|_{op} \leq 2n^{-1/2}\sigma_l(p^{1/2} + q^{1/2})$ with probability at least $1 - c_1^* \exp\{-c_2^*(p+q)\} - \exp(-n)$, where c_1^* and c_2^* are some positive constants. Therefore, one has

$J^*(\nabla R(\widehat{\mathbf{B}}_0^{\widehat{\mathcal{M}}})) = \max_{l \in \widehat{\mathcal{M}}} \|\mathbf{D}_l\|_{op} \leq 4n^{-1/2}(\max_{l \in \widehat{\mathcal{M}}} \sigma_l)(p^{1/2} + q^{1/2})$ with probability at least $1 - |\widehat{\mathcal{M}}|c_1^* \exp\{-c_2^*(p+q)\} - |\widehat{\mathcal{M}}|\exp(-n)$. By condition (A4), one has

$\max_{l \in \widehat{\mathcal{M}}} \sigma_l \leq \lambda_{\max}(\Sigma_X) \leq C_5 n^\tau$. By the proof of Theorem 2, one has $|\widehat{\mathcal{M}}| = O(n^{2\kappa + \tau})$ with probability at least $1 - c_4^* \exp(-c_5^* n^{1-2\kappa})$ for some positive constants c_4^* and c_5^* . Thus, when

$\lambda \geq 4C_5 n^{\tau-1/2}(p^{1/2} + q^{1/2})$, one has $\lambda \geq J^*(\nabla R(\widehat{\mathbf{B}}_0^{\widehat{\mathcal{M}}}))$ with probability at least

$1 - c_1 n^{2\kappa + \tau} \exp\{-c_2(p+q)\} - c_3 n^{2\kappa + \tau} \exp(-n) - c_4^* \exp(-c_5^* n^{1-2\kappa})$ for some positive constants c_1, c_2, c_3, c_4^* and c_5^* .

By the proof of Theorem 1, the event $\{\mathcal{M} \subseteq \widehat{\mathcal{M}}\}$ holds with probability goes to 1. In particular, $P(\{\mathcal{M} \subseteq \widehat{\mathcal{M}}\}) \geq 1 - c_4^* \exp(-c_5^* n^{1-2\kappa})$ for some positive constants c_4^* and c_5^* . Therefore, there exists some positive constants c_1, c_2, c_3, c_4 and c_5 such that with probability $1 - c_1 n^{2\kappa + \tau} \exp\{-c_2(p+q)\} - c_3 n^{2\kappa + \tau} \exp(-n) - c_4 \exp(-c_5 n^{1-2\kappa})$, one has

$$\|\widehat{\mathbf{B}}^{\widehat{\mathcal{M}}} - \mathbf{B}_0^{\widehat{\mathcal{M}}}\|_F^2 \leq C(\sum_{l \in \widehat{\mathcal{M}}} r_l)\lambda^2 \iota_L^{-2} \text{ for some positive constant } C. \text{ When Assumptions (A5)}$$

and (A14) hold, with probability goes to 1, one has $\|\widehat{\mathbf{B}}^{\widehat{\mathcal{M}}} - \mathbf{B}_0^{\widehat{\mathcal{M}}}\|_F^2 \leq C(\sum_{l \in \mathcal{M}} r_l) \lambda^2 \tau_L^{-2}$. This completes the proof of Theorem 6.

C: Interpretations of Λ

In this section, we include some detailed interpretations of the definition Λ_l . Without loss of generality, we assume $\widehat{\mathcal{M}} = \mathcal{M} = \{1, \dots, s\}$ with s fixed for notation simplicity. We first give a necessary condition for rank consistency presented in Theorem 4. By proposition 18 of (Bach, 2008), for any $1 \leq l \leq s$, we have $(\mathbf{U}_{l0}^\perp)^\top \widehat{\Delta}_l \mathbf{V}_{l0}^\perp = o_p(1)$ if $\text{rank}(\widehat{\mathbf{B}}_l) = \text{rank}(\mathbf{B}_{l0}) = r_l$. Since $\widehat{\Delta}_l \rightarrow_p \Delta_{l0}$ and \mathbf{V}_{l0}^\perp is a nonrandom quantity, we have $(\mathbf{U}_{l0}^\perp)^\top \Delta_{l0} \mathbf{V}_{l0}^\perp = 0$. Recall that $\{\Delta_{l0} : 1 \leq l \leq s\}$ is the minimizer of $\ell^0(\Delta_1, \dots, \Delta_s)$, and thus $\{\Delta_{l0} : 1 \leq l \leq s\}$ is the solution of the optimal problem

$$\min \ell^0(\Delta) \quad \text{subject to} \quad (\mathbf{U}_{l0}^\perp)^\top \Delta_l \mathbf{V}_{l0}^\perp = 0 \quad \text{for every } 1 \leq l \leq s. \tag{9}$$

Using Lagrange multiplier method, consider the minimizer of

$$L(\Delta, \Lambda_1, \dots, \Lambda_s) = 2^{-1} \text{vec}(\Delta)^\top \Sigma_{\mathcal{M}} \text{vec}(\Delta) + \sum_{l=1}^s \text{tr}(\mathbf{U}_{l0}^\top \Delta_l \mathbf{V}_{l0}) + \sum_{l=1}^s \text{tr}(\Lambda_l^\top (\mathbf{U}_{l0}^\perp)^\top \Delta_l \mathbf{V}_{l0}^\perp)$$

where $\{\Lambda_l, l = 1, \dots, s\}$ are Lagrange multipliers. Thus, for $l = 1, \dots, s$, $\{\Delta_{l0} : 1 \leq l \leq s\}$ satisfies

$$\begin{aligned} \frac{\partial L}{\partial \Delta_l} &= \sum_{l'=1}^s \Sigma_{\mathcal{M}, l'l'} \Delta_{l0} + \mathbf{U}_{l0} \mathbf{V}_{l0}^\top + \mathbf{U}_{l0}^\perp \Lambda_l (\mathbf{V}_{l0}^\perp)^\top = 0, \\ \frac{\partial L}{\partial \Lambda_l} &= (\mathbf{U}_{l0}^\perp)^\top \Delta_{l0} \mathbf{V}_{l0}^\perp = 0. \end{aligned}$$

Recall that $\mathbf{A} = \Sigma_{\mathcal{M}} \otimes \mathbf{I}_{pq \times pq}$, $\mathbf{K}_l = \mathbf{V}_{l0}^\perp \otimes \mathbf{U}_{l0}^\perp$, and $\mathbf{d}_l = -\text{vec}(\mathbf{U}_{l0} \mathbf{V}_{l0}^\top)$ for $l = 1, \dots, s$, where \otimes denotes the Kronecker product. Let $\mathbf{d} = (\mathbf{d}_1^\top, \dots, \mathbf{d}_s^\top)^\top$, $\mathbf{K} = \text{diag}\{\mathbf{K}_1, \dots, \mathbf{K}_s\}$, $\Lambda_l \in \mathbb{R}^{(p-r_l) \times (q-r_l)}$ for $l = 1, \dots, s$ such that

$$\text{vec}(\Lambda) = (\text{vec}(\Lambda_1)^\top, \dots, \text{vec}(\Lambda_s)^\top)^\top = (\mathbf{K}^\top \mathbf{A}^{-1} \mathbf{K})^{-1} \mathbf{K}^\top \mathbf{A}^{-1} \mathbf{d}.$$

Then the Lagrange equation can be written as

$$\begin{pmatrix} \mathbf{A} & \mathbf{K} \\ \mathbf{K}^\top & \mathbf{0} \end{pmatrix} \begin{pmatrix} \text{vec}(\Delta) \\ \text{vec}(\Lambda) \end{pmatrix} = \begin{pmatrix} \mathbf{d} \\ \mathbf{0} \end{pmatrix}.$$

It is easy to show that

$$\text{vec}(\mathbf{\Lambda}) = \mathbf{A}^{-1}(\mathbf{d} - \mathbf{K}\text{vec}(\mathbf{\Lambda})) \quad \text{and} \quad \text{vec}(\mathbf{\Lambda}) = (\mathbf{K}^T \mathbf{A}^{-1} \mathbf{K})^{-1} \mathbf{K}^T \mathbf{A}^{-1} \mathbf{d}.$$

From the above calculation, we can see that $\text{vec}(\mathbf{\Lambda}) = (\text{vec}(\mathbf{\Lambda}_1)^T, \dots, \text{vec}(\mathbf{\Lambda}_s)^T)^T$ is actually the Lagrange multiplier for the optimization problem (9).

References

- Anderson TW (1955), “The integral of a symmetric unimodal function over a symmetric convex set and some probability inequalities,” *Proceedings of the American Mathematical Society*, 6, 170–176.
- Bach FR (2008), “Consistency of trace norm minimization,” *Journal of Machine Learning Research*, 9, 1019–1048.
- Barrett JC, Fry B, Maller J, and Daly MJ (2005), “Haploview: analysis and visualization of LD and haplotype maps,” *Bioinformatics*, 21, 263–265. [PubMed: 15297300]
- Barut E, Fan J, and Verhasselt A (2016), “Conditional sure independence screening,” *Journal of the American Statistical Association*, 111, 1266–1277. [PubMed: 28360436]
- Beck A and Teboulle M (2009), “A fast iterative shrinkage-thresholding algorithm for linear inverse problems,” *SIAM Journal on Imaging Sciences*, 2, 183–202.
- Borwein JM and Lewis AS (2010), *Convex analysis and nonlinear optimization: theory and examples*, Springer Science & Business Media.
- Breiman L and Friedman J (1997), “Predicting multivariate responses in multiple linear regression,” *Journal of the Royal Statistical Society*, 59, 3–54.
- Buhlmann P and van de Geer S (2011), *Statistics for High-Dimensional Data: Methods, Theory and Applications.*, New York, N. Y.: Springer.
- Cai J-F, Candès EJ, and Shen Z (2010), “A singular value thresholding algorithm for matrix completion,” *SIAM Journal on Optimization*, 20, 1956–1982.
- Candès E and Tao T (2007), “The Dantzig selector: Statistical estimation when p is much larger than n,” *The Annals of Statistics*, 2313–2351.
- Candes EJ and Recht B (2009), “Exact matrix completion via convex optimization,” *Foundations of Computational Mathematics.*, 9, 717–772.
- Chen Y, Goldsmith J, and Ogden T (2016), “Variable Selection in Function-on-Scalar Regression,” *Stat*, 5, 88–101. [PubMed: 27429751]
- Chiang MC, Barysheva M, Toga AW, Medland SE, Hansell NK, James MR, McMahon KL, de Zubicaray GI, Martin NG, Wright MJ, and Thompson PM (2011), “BDNF gene effects on brain circuitry replicated in 455 twins,” *NeuroImage*, 55, 448–454. [PubMed: 21195196]
- Cook R, Helland I, and Su Z (2013), “Envelopes and partial least squares regression,” *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 75, 851–877.
- Ding S and Cook D (2018), “Matrix variate regressions and envelope models,” *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 80, 387–408.
- Ding S and Cook RD (2014), “Dimension folding PCA and PFC for matrix-valued predictors,” *Statistica Sinica*, 24, 463–492.
- Facchinei F and Pang J-S (2003), *Finite-dimensional variational inequalities and complementarity problems Vol. I*, Springer Series in Operations Research, Springer-Verlag, New York.
- Fan J, Feng Y, and Song R (2011), “Nonparametric independence screening in sparse ultra-high-dimensional additive models,” *Journal of the American Statistical Association*, 106, 544–557. [PubMed: 22279246]
- Fan J and Lv J (2008), “Sure independence screening for ultrahigh dimensional feature space,” *Journal of the Royal Statistical Society. Series B.*, 70, 849–911.
- (2010), “A selective overview of variable selection in high dimensional feature space,” *Statistica Sinica*, 20, 101. [PubMed: 21572976]
- Fan J and Song R (2010), “Sure independence screening in generalized linear models with NP-dimensionality,” *The Annals of Statistics*, 38, 3567–3604.

- Fosdick BK and Hoff PD (2015), “Testing and modeling dependencies between a network and nodal attributes,” *Journal of the American Statistical Association*, 110, 1047–1056. [PubMed: 26848204]
- Gabriel SB, Schaffner SF, Nguyen H, Moore JM, Roy J, Blumenstiel B, Higgins J, DeFelice M, Lochner A, Faggart M, et al. (2002), “The structure of haplotype blocks in the human genome,” *Science*, 296, 2225–2229. [PubMed: 12029063]
- Hibar D, Stein JL, Kohannim O, Jahanshad N, Saykin AJ, Shen L, Kim S, Pankratz N, Foroud T, Huentelman MJ, Potkin SG, Jack C, Weiner MW, Toga AW, Thompson P, and ADNI (2011), “Voxelwise gene-wide association study (vGeneWAS): multivariate gene-based association testing in 731 elderly subjects,” *NeuroImage*, 56, 1875–1891. [PubMed: 21497199]
- Huang M, Nichols T, Huang C, Yang Y, Lu Z, Feng Q, Knickmeyer RC, Zhu H, and ADNI (2015), “FVGWAS: Fast Voxelwise Genome Wide Association Analysis of Large-scale Imaging Genetic Data,” *NeuroImage*, 118, 613–627. [PubMed: 26025292]
- Knight K and Fu W (2000), “Asymptotics for lasso-type estimators,” *The Annals of Statistics*, 28, 1356–1378.
- Leng C and Tang CY (2012), “Sparse matrix graphical models,” *Journal of the American Statistical Association*, 107, 1187–1200.
- Li L and Zhang X (2017), “Parsimonious tensor response regression,” *Journal of the American Statistical Association*, 112, 1131–1146.
- Liu J and Calhoun VD (2014), “A review of multivariate analyses in imaging genetics,” *Frontiers in Neuroinformatics*, 8, 29. [PubMed: 24723883]
- Medlan SE, Jahanshad N, Neale BM, and Thompson PM (2014), “Whole-genome analyses of whole-brain data: working within an expanded search space,” *Nature Neuroscience*, 17, 791–800. [PubMed: 24866045]
- Medland SE, Jahanshad N, Neale BM, and Thompson PM (2014), “Whole-genome analyses of whole-brain data: working within an expanded search space,” *Nature Neuroscience*, 17, 791–800. [PubMed: 24866045]
- Negahban S, Yu B, Wainwright MJ, and Ravikumar PK (2009), “A unified framework for high-dimensional analysis of M-estimators with decomposable regularizers,” in *Advances in Neural Information Processing Systems*, pp. 1348–1356.
- Negahban SN, Ravikumar P, Wainwright MJ, and Yu B (2012), “A unified framework for high-dimensional analysis of M-estimators with decomposable regularizers,” *Statistical Science*, 27, 538–557.
- Nesterov Y (2004), *Introductory lectures on convex optimization*, vol. 87 of *Applied Optimization*, Kluwer Academic Publishers, Boston, MA, a basic course.
- Park Y, Su Z, and Zhu H (2017), “Groupwise envelope models for imaging genetic analysis,” *Biometrics*, 73, 1243–1253. [PubMed: 28323341]
- Peng J, Zhu J, Bergamaschi A, Han W, Noh DY, Pollack JR, and Wang P (2010), “Regularized multivariate regression for identifying master predictors with application to integrative genomics study of breast cancer,” *Annals of Applied Statistics*, 4, 53–77.
- Peper JS, Brouwer RM, Boomsma DI, Kahn RS, and Pol HEH (2007), “Genetic Influences on Human Brain Structure: A Review of Brain Imaging Studies in Twins,” *Human Brain Mapping*, 28, 464–473. [PubMed: 17415783]
- Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Maller J, Sklar P, De Bakker PI, Daly MJ, et al. (2007), “PLINK: a tool set for whole-genome association and population-based linkage analyses,” *The American Journal of Human Genetics*, 81, 559–575. [PubMed: 17701901]
- Rabusseau G and Kadri H (2016), “Low-rank regression with tensor responses,” in *Advances in Neural Information Processing Systems*, pp. 1867–1875.
- Ramsay JO and Silverman BW (2005), *Functional Data Analysis*, New York, N. Y.: Springer, 2nd ed.
- Raskutti G and Yuan M (2018), “Convex regularization for high-dimensional tensor regression,” *Annals of Statistics*, to appear.
- Recht B, Fazel M, and Parrilo PA (2010), “Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization,” *SIAM Review*, 52, 471–501.
- Scharinger C, Rabl U, Sitte HH, and Pezawas L (2010), “Imaging genetics of mood disorders,” *NeuroImage*, 53, 810–821. [PubMed: 20156570]

- Shen L, Kim S, Risacher SL, Nho K, Swaminathan S, West JD, Foroud T, Pankratz N, Moore JH, Sloan CD, Huentelman MJ, Craig DW, DeChairo BM, Potkin SG, Jr CRJ, Weiner MW, Saykin AJ, and ADNI (2010), “Whole genome association study of brain-wide imaging phenotypes for identifying quantitative trait loci in MCI and AD: A study of the ADNI cohort,” *NeuroImage*, 53, 1051–1063. [PubMed: 20100581]
- Stein J, Hua X, Lee S, Ho A, Leow A, Toga A, Saykin AJ, Shen L, Foroud T, Pankratz N, Huentelman M, Craig D, Gerber J, Allen A, Corneveaux JJ, Dechairo B, Potkin S, Weiner M, Thompson P, and ADNI (2010), “Voxelwise genome-wide association study (vGWAS),” *NeuroImage*, 53, 1160–1174. [PubMed: 20171287]
- Thompson PM, Ge T, Glahn DC, Jahanshad N, and Nichols TE (2013), “Genetics of the connectome,” *NeuroImage*, 80, 475–488. [PubMed: 23707675]
- Thompson PM, Stein JL, Medland SE, Hibar DP, Vasquez AA, Renteria ME, Toro R, Jahanshad N, Schumann G, Franke B, et al. (2014), “The ENIGMA Consortium: large-scale collaborative analyses of neuroimaging and genetic data,” *Brain Imaging and Behavior*, 8, 153–182. [PubMed: 24399358]
- Tibshirani R (1997), “The lasso method for variable selection in the Cox model,” *Statistics in Medicine*, 16, 385–395. [PubMed: 9044528]
- van der Vaart A and Wellner J (2000), *Weak Convergence and Empirical Processes: With Applications to Statistics* (Springer Series in Statistics), New York, N. Y.: Springer.
- Vounou M, Janousova E, Wolz R, Stein J, Thompson P, Rueckert D, Montana G, and ADNI (2012), “Sparse reduced-rank regression detects genetic associations with voxel-wise longitudinal phenotypes in Alzheimer’s disease,” *NeuroImage*, 60, 700–716. [PubMed: 22209813]
- Vounou M, Nichols TE, Montana G, and ADNI (2010), “Discovering genetic associations with high-dimensional neuroimaging phenotypes: A sparse reduced-rank regression approach,” *NeuroImage*, 53, 1147–1159. [PubMed: 20624472]
- Wang H, Nie F, Huang H, Kim S, Nho K, Risacher S, Saykin A, Shen L, and ADNI (2012a), “Identifying quantitative trait loci via group-sparse multi-task regression and feature selection: An imaging genetics study of the ADNI cohort,” *Bioinformatics*, 28, 229–237. [PubMed: 22155867]
- Wang H, Nie F, Huang H, Risacher S, Saykin A, Shen L, and ADNI (2012b), “Identifying disease sensitive and quantitative trait relevant biomarkers from multi-dimensional heterogeneous imaging genetics data via sparse multi-modal multi-task learning,” *Bioinformatics*, 28, 127–136. [PubMed: 22088842]
- Wang L, Chen G, and Li H (2007), “Group SCAD regression analysis for microarray time course gene expression data,” *Bioinformatics*, 23, 1486–1494. [PubMed: 17463025]
- Yuan M, Ekici A, Lu Z, and Monteiro R (2007), “Dimension reduction and coefficient estimation in multivariate linear regression,” *Journal of the Royal Statistical Society. Series B.*, 69, 329–346.
- Zhang YW, Xu ZY, Shen XT, and Pan W (2014), “Testing for association with multiple traits in generalized estimation equations, with application to neuroimaging data,” *NeuroImage*, 96, 309–325. [PubMed: 24704269]
- Zhao J and Leng C (2014), “Sparse matrix graphical models,” *Statistica Sinica*, 24, 799–814.
- Zhao P and Yu B (2006), “On model selection consistency of Lasso,” *Journal of Machine Learning Research*, 7, 2541–2563.
- Zhou H and Li L (2014), “Regularized matrix regression,” *Journal of the Royal Statistical Society. Series B.*, 76, 463–483.
- Zhu H, Khondker ZS, Lu Z, and Ibrahim JG (2014), “Bayesian generalized low rank regression models for neuroimaging phenotypes and genetic markers,” *Journal of American Statistical Association*, 109, 977–990.
- Zou H (2006), “The adaptive lasso and its oracle properties,” *Journal of the American Statistical Association*, 101, 1418–1429.

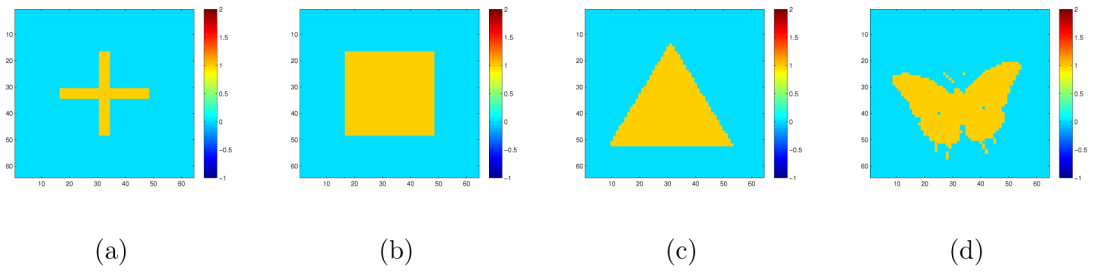


Figure 1:

Simulation I setting: the four 64×64 true coefficient matrices for the first simulation setting: the cross shape for \mathbf{B}_{10} in panel (a), the square shape for \mathbf{B}_{20} in panel (b), the triangle shape of \mathbf{B}_{30} in panel (c), and the butterfly shape for \mathbf{B}_{40} in panel (d). The regression coefficient at each pixel is either 0 (blue) or 1 (yellow).

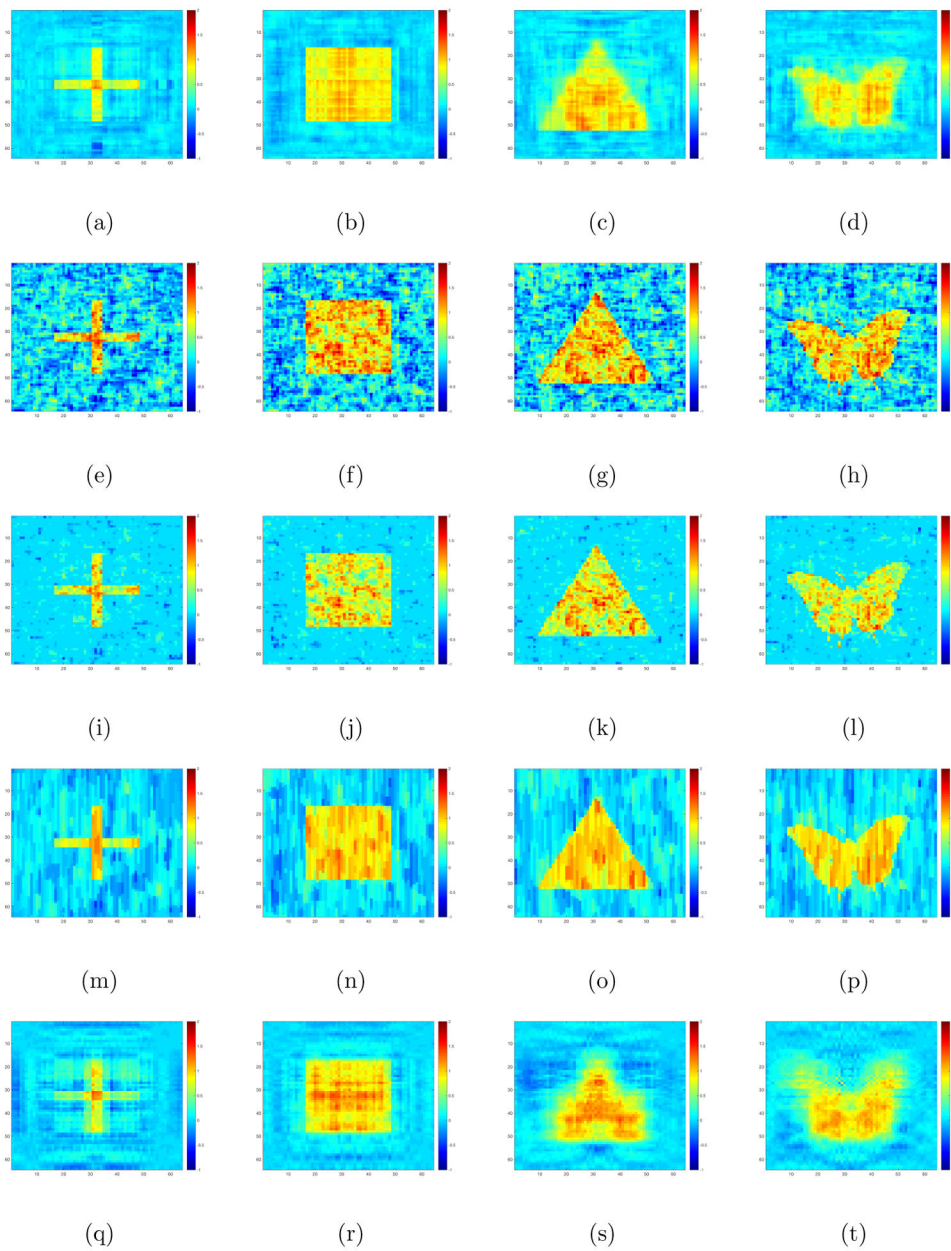


Figure 2: Simulation I results: the RLR (panels (a)-(d)), OLS (panels (e)-(h)), Lasso (panels (i)-(l)), Fused Lasso (panels (m)-(p)) and Envelope (panels (q)-(t)) estimates of coefficient matrices from a randomly selected training dataset with $n = 500$, $\rho_1 = 0.5$, $\rho_2 = 0.5$ and $\sigma^2 = 25$: $\widehat{\mathbf{B}}_1$ (the first column); $\widehat{\mathbf{B}}_2$ (the second column); $\widehat{\mathbf{B}}_3$ (the third column); and $\widehat{\mathbf{B}}_4$ (the fourth column).

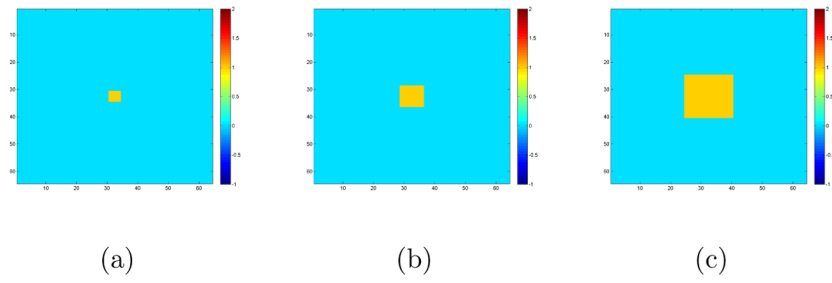


Figure 3: Screening setting: Panel (a)-(c) are the true coefficient images \mathbf{B}_{true} with regions of interest with different sizes: effective regions of interest (yellow ROI) and non-effective regions of interest (blue ROI).

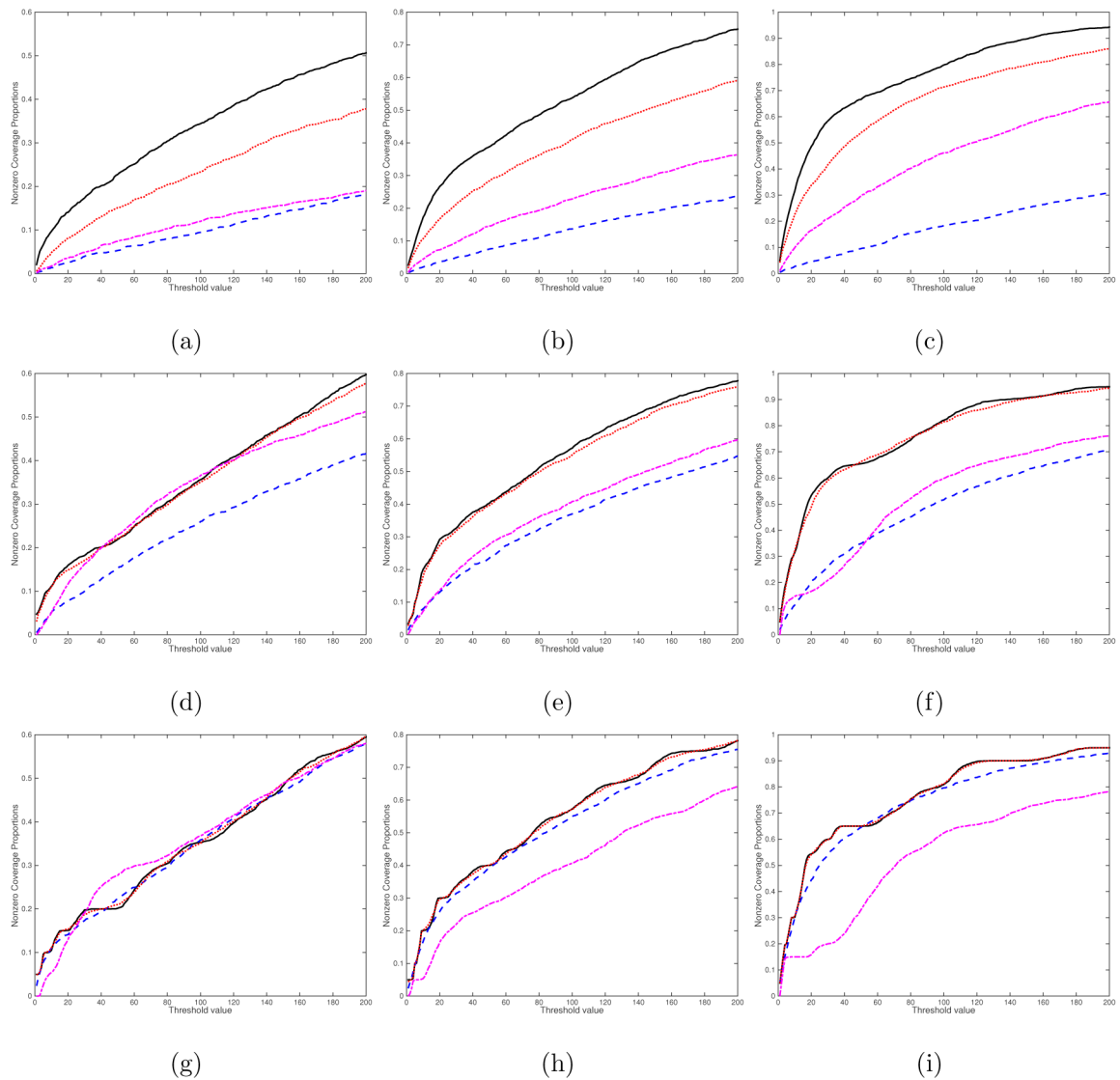


Figure 4: Screening simulation results for the case $(\sigma_e^2, s_n) = (1, 2000)$: the curves of percentage of the average true nonzero coverage proportion. The black solid, blue dashed, red dotted, and purple dashed dotted lines correspond to the rank-one screening, the L1 entrywise norm screening, the Frobenius norm screening, and the global Wald test screening, respectively. Panels (a)-(i) correspond to $(n, p_s, q_s) = (100, 4, 4), (200, 4, 4), (500, 4, 4), (100, 8, 8), (200, 8, 8), (500, 8, 8), (100, 16, 16), (200, 16, 16),$ and $(500, 16, 16)$, respectively.

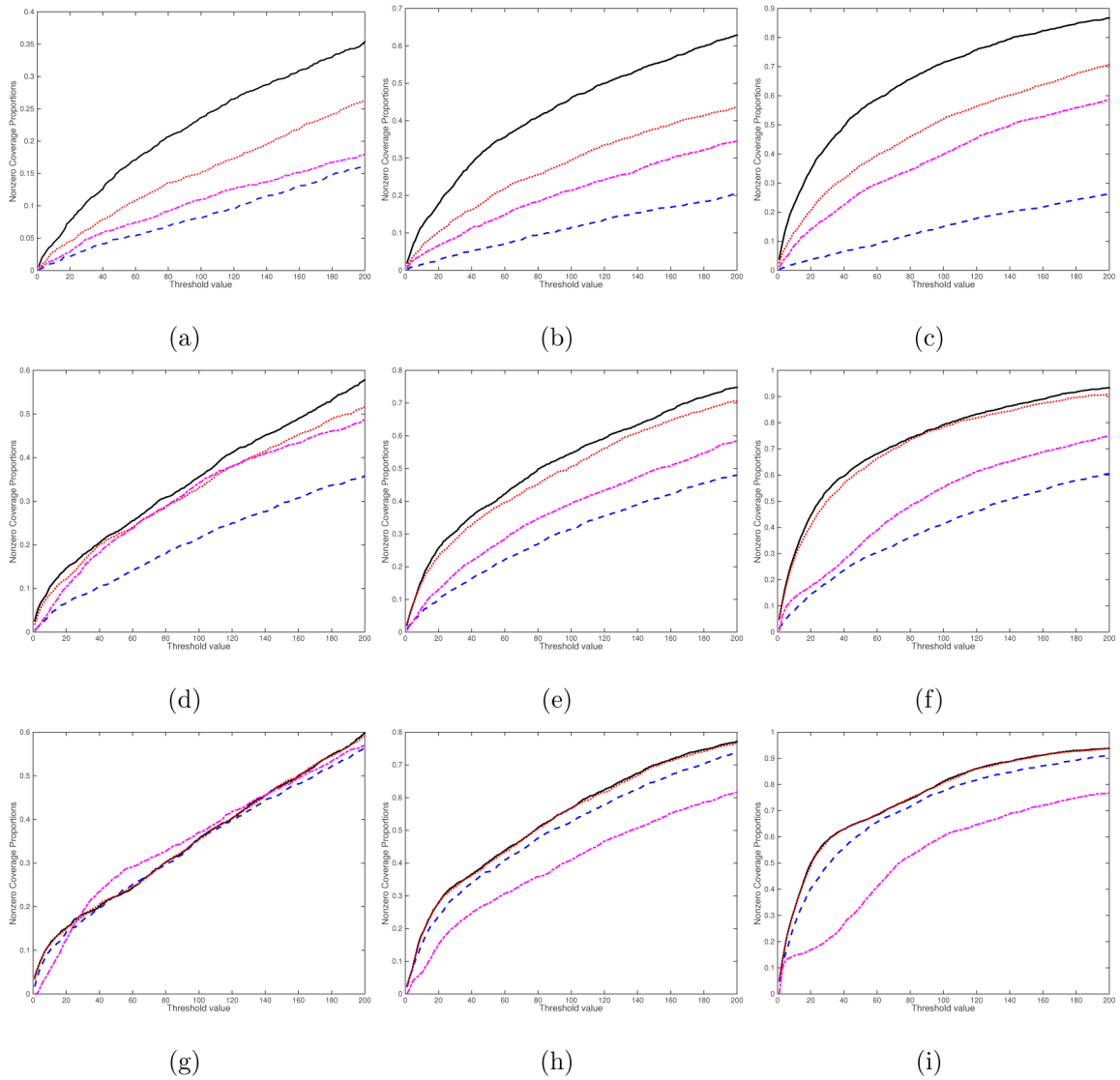


Figure 5:

Screening simulation results for the case $(\sigma_e^2, s_n) = (25, 2000)$: the curves of percentage of the average true nonzero coverage proportion. The black solid, blue dashed, red dotted, and purple dashed dotted lines correspond to the rank-one screening, the L1 entrywise norm screening, the Frobenius norm screening, and the global Wald test screening, respectively. Panels (a)-(i) correspond to $(n, p_s, q_s) = (100, 4, 4), (200, 4, 4), (500, 4, 4), (100, 8, 8), (200, 8, 8), (500, 8, 8), (100, 16, 16), (200, 16, 16),$ and $(500, 16, 16)$, respectively.

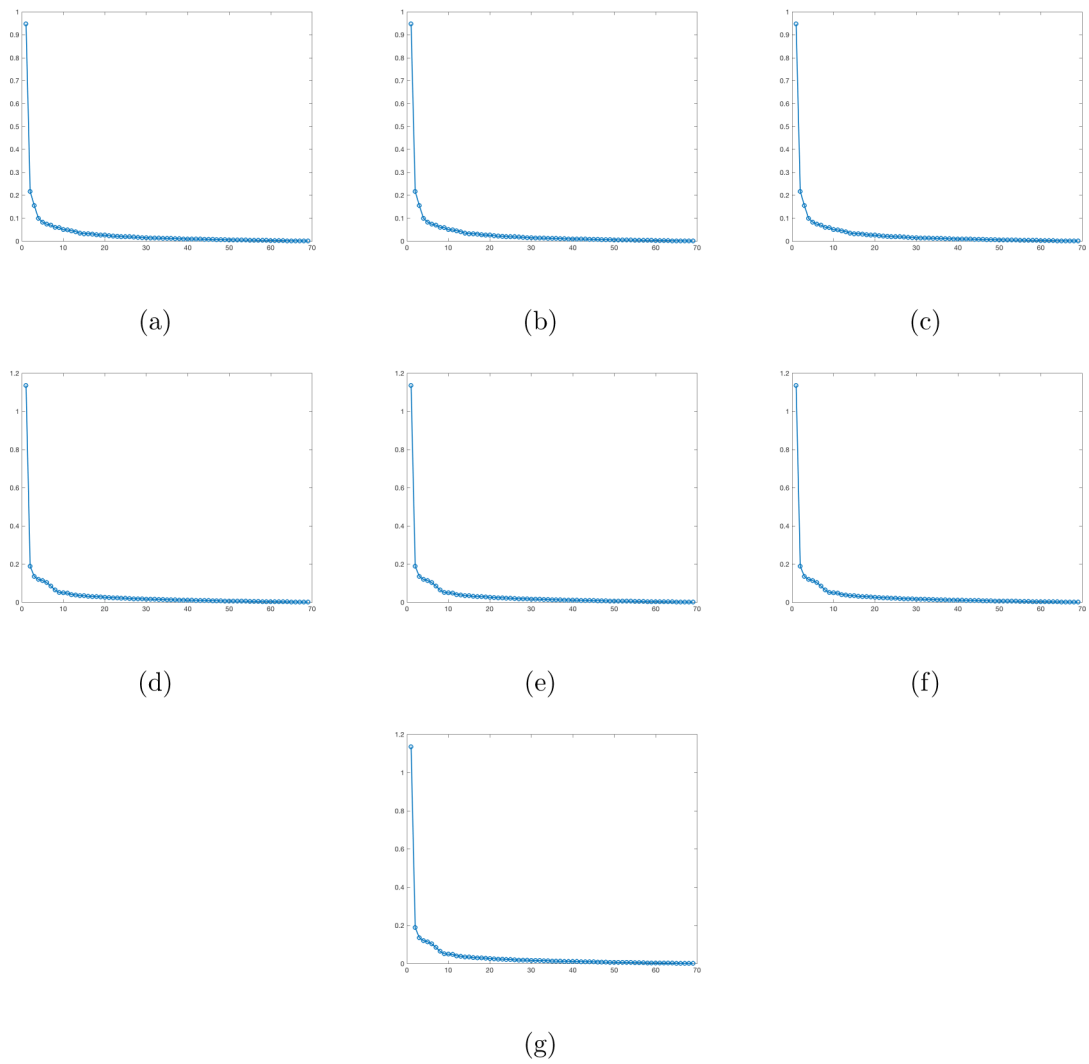


Figure 6: PNC data: Panel (a)-(g) are the plots for the singular values of the OLS estimates corresponding to the 7 SNPs selected by our screening step, with singular values sorted from largest to smallest.

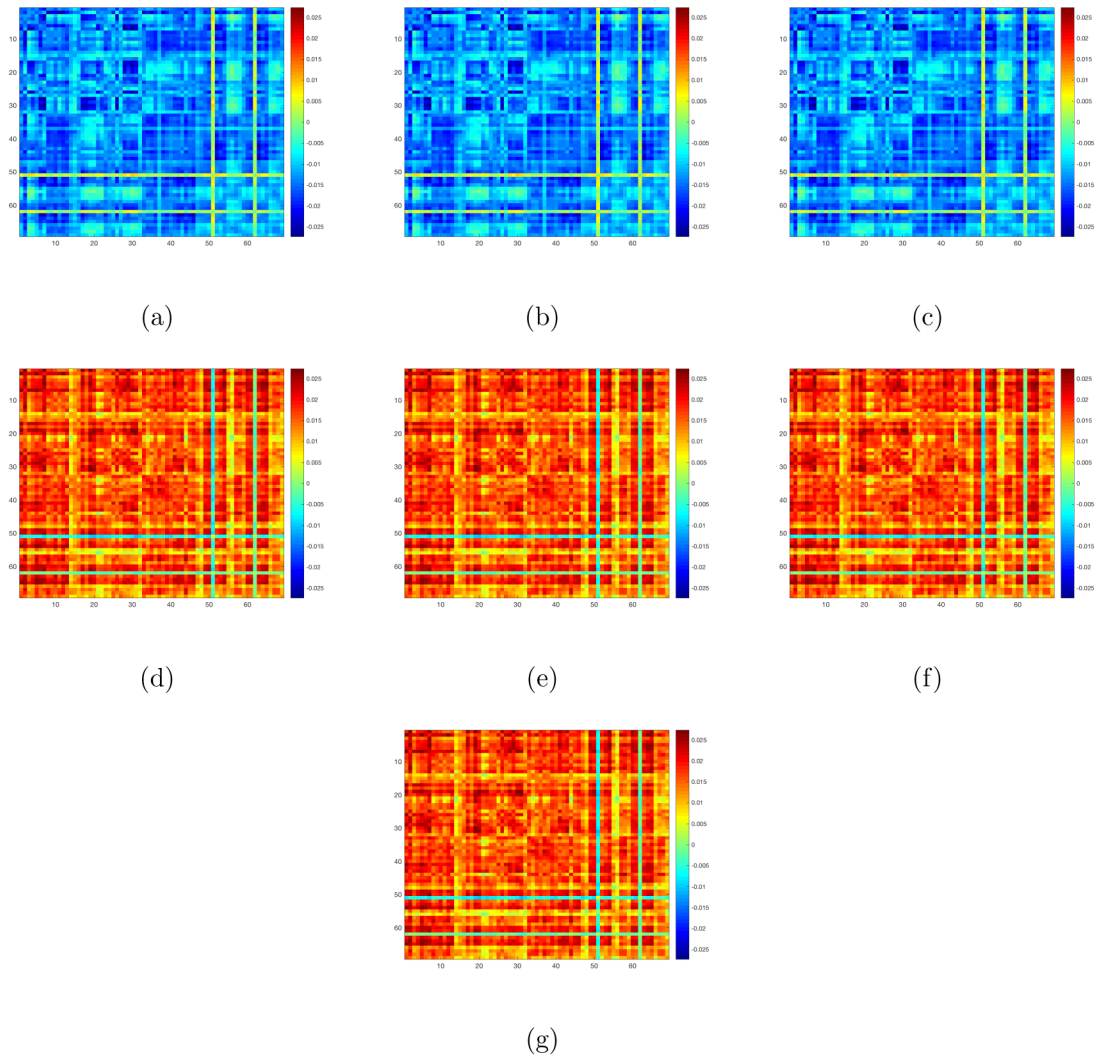


Figure 7: PNC data: Panel (a)-(g) are the plots for our RLR estimates corresponding to the 7 SNPs selected by our screening step.

Table 1:

Simulation I results: the means of PEs and MSEs for regularized low-rank (RLR), OLS, Lasso, fused Lasso (Fused) and tensor envelope (Envelope) estimates and their associated standard errors in the parentheses. For each case, 100 simulated data sets are used.

(n, σ_e^2)	Method	MSE(B ₁)	MSE(B ₂)	MSE(B ₃)	MSE(B ₄)	PE
(100, 1)	RLR	11.67(0.21)	9.96(0.22)	43.21(0.43)	44.88(0.52)	1.03(0.0002)
	OLS	58.08(0.83)	72.38(1.08)	71.66(1.01)	58.00(0.92)	1.05(0.0004)
	Lasso	42.96(0.79)	53.79(1.08)	53.21(0.98)	44.87(0.75)	1.04(0.0004)
	Fused	11.85(0.20)	11.25(0.22)	13.61(0.23)	17.87(0.26)	1.02(0.0002)
	Envelope	21.20(0.34)	24.95(0.36)	51.14(0.49)	55.62(0.67)	1.04(0.0002)
(200, 1)	RLR	7.27(0.09)	6.73(0.10)	23.77(0.20)	23.08(0.23)	1.02(0.0001)
	OLS	28.61(0.30)	34.88(0.36)	34.85(0.38)	28.11(0.32)	1.03(0.0001)
	Lasso	19.29(0.38)	23.86(0.43)	23.73(0.41)	20.30(0.29)	1.02(0.0002)
	Fused	5.93(0.09)	5.62(0.08)	6.59(0.10)	8.63(0.10)	1.01(0.0001)
	Envelope	11.21(0.16)	13.13(0.14)	38.40(0.30)	43.33(0.35)	1.03(0.0001)
(500, 1)	RLR	3.46(0.03)	3.53(0.03)	10.54(0.06)	9.75(0.06)	1.006(0.00003)
	OLS	11.01(0.08)	13.87(0.10)	13.88(0.09)	11.04(0.07)	1.009(0.00003)
	Lasso	5.93(0.17)	7.89(0.17)	7.78(0.18)	6.70(0.13)	1.007(0.00009)
	Fused	2.36(0.03)	2.37(0.02)	2.73(0.03)	3.29(0.03)	1.004(0.00002)
	Envelope	5.44(0.07)	6.60(0.07)	31.72(0.21)	38.29(0.30)	1.02(0.0001)
(100, 25)	RLR	121.61(1.69)	119.58(2.37)	227.58(2.01)	263.90(2.77)	25.37(0.0027)
	OLS	1451.95(20.64)	1809.40(27.01)	1791.53(25.36)	1450.10(23.10)	26.27(0.0099)
	Lasso	1360.23(19.03)	1683.95(24.74)	1669.88(23.97)	1367.71(21.34)	26.22(0.0093)
	Fused	238.87(3.04)	254.78(4.36)	290.50(4.47)	283.07(3.89)	25.42(0.0031)
	Envelope	175.09(1.57)	139.95(2.71)	259.48(2.18)	286.98(1.81)	25.39(0.0023)
(200, 25)	RLR	79.44(1.01)	71.27(1.25)	171.12(1.21)	201.43(1.63)	25.26(0.0013)
	OLS	715.28(7.49)	872.02(8.93)	871.33(9.41)	702.70(8.00)	25.66(0.0037)
	Lasso	657.54(7.19)	798.43(8.58)	798.01(9.04)	652.91(7.14)	25.62(0.0035)
	Fused	156.75(2.00)	162.27(1.95)	174.79(2.34)	175.61(2.20)	25.27(0.0016)
	Envelope	151.68(1.55)	105.18(1.61)	202.96(2.78)	230.24(2.63)	25.29(0.0016)
(500, 25)	RLR	42.17(0.50)	39.70(0.59)	110.16(0.79)	125.08(0.75)	25.10(0.0005)
	OLS	275.31(2.05)	346.69(2.54)	346.93(2.36)	276.05(1.83)	25.24(0.0008)
	Lasso	238.43(2.33)	299.22(2.99)	298.81(2.90)	243.44(1.95)	25.22(0.0011)
	Fused	80.31(0.84)	89.14(0.79)	93.22(0.90)	89.8(0.83)	25.10(0.0006)
	Envelope	95.49(0.99)	75.41(0.99)	142.24(1.51)	171.34(1.31)	25.14(0.0008)

Table 2:

Simulation II results: the means of PEs and MSEs for regularized low-rank (RLR) OLS, Lasso, fused Lasso (Fused) and tensor envelope (Envelope) estimates and their associated standard errors in the parentheses. For each case, 100 simulated data sets are used.

$(n, \sigma_{\varepsilon}^2)$	Method	MSE(B_1)	MSE(B_2)	PE
(100, 1)	RLR	21.86(0.33)	13.91(0.20)	1.02(0.0001)
	OLS	41.85(0.16)	56.82(0.82)	1.03(0.0003)
	Lasso	55.78(0.88)	54.40(0.73)	1.03(0.0002)
	Fused	57.39(0.94)	56.61(0.77)	1.03(0.0002)
	Envelope	41.46(0.15)	33.23(0.54)	1.02(0.0002)
(200, 1)	RLR	10.84(0.12)	6.77(0.07)	1.011(0.00005)
	OLS	20.75(0.07)	27.80(0.30)	1.018(0.0001)
	Lasso	27.90(0.31)	27.03(0.29)	1.018(0.0001)
	Fused	27.69(0.30)	27.29(0.29)	1.018(0.0001)
	Envelope	20.68(0.07)	18.33(0.22)	1.014(0.00008)
(500, 1)	RLR	4.19(0.05)	2.73(0.02)	1.004(0.00002)
	OLS	8.22(0.03)	10.94(0.08)	1.006(0.00003)
	Lasso	12.49(0.18)	12.18(0.17)	1.007(0.00006)
	Fused	10.99(0.08)	10.91(0.09)	1.006(0.00003)
	Envelope	8.26(0.03)	9.37(0.09)	1.006(0.00003)
(100, 25)	RLR	391.95(5.50)	254.67(3.54)	25.37(0.0024)
	OLS	1044(4.10)	1447(23.94)	25.77(0.0060)
	Lasso	1378.32(22.99)	1360.95(18.52)	25.75(0.0059)
	Fused	1232.31(17.74)	1042.72(12.13)	25.68(0.0055)
	Envelope	1033.69(3.66)	626.57(7.71)	25.46(0.0027)
(200, 25)	RLR	219.13(2.14)	136.41(1.33)	25.26(0.0011)
	OLS	518.63(1.83)	694.98(7.47)	25.45(0.0025)
	Lasso	657.39(7.19)	644.78(7.20)	25.44(0.0025)
	Fused	637.01(6.45)	589.9(5.68)	25.43(0.0022)
	Envelope	516.52(1.81)	395.64(3.67)	25.33(0.0015)
(500, 25)	RLR	101.8(0.91)	64.19(0.57)	25.09(0.0005)
	OLS	206.57(0.73)	275.3(2.05)	25.16(0.0008)
	Lasso	259.2(1.95)	254.26(2.17)	25.16(0.0008)
	Fused	265.44(1.89)	255.88(1.99)	25.16(0.0008)
	Envelope	206.41(0.73)	226.94(1.54)	25.14(0.0006)

Table 3:

The means of predictor errors (PEs) and MSEs for our two-step procedure, and the average selected model size for our screening procedure. Their associated standard errors are in the parentheses. For each case, 100 simulated data sets are used.

$(n, s_n, \sigma_{\epsilon}^2)$	MSE(B ₁)	MSE(B ₂)	MSE(B ₃)	MSE(B ₄)	PE	Model Size
(100, 2000, 1)	15.87(3.22)	10.73(0.67)	43.52(0.46)	47.07(0.58)	1.05(0.001)	5.24(0.11)
(200, 2000, 1)	5.92(0.10)	5.10(0.11)	27.61(0.27)	28.23(0.31)	1.03(0.0003)	5.87(0.11)
(100, 5000, 1)	32.28(6.57)	12.60(1.08)	44.28(0.52)	47.14(0.65)	1.07(0.002)	5.03(0.10)
(200, 5000, 1)	5.92(0.09)	4.94(0.09)	28.03(0.29)	28.35(0.29)	1.03(0.0005)	5.83(0.13)
(100, 2000, 25)	126.17(1.84)	119.15(2.56)	227.86(1.96)	279.22(3.17)	25.99(0.027)	5.15(0.11)
(200, 2000, 25)	84.42(1.25)	73.69(1.41)	177.90(1.66)	214.67(2.15)	25.54(0.012)	5.82(0.11)
(100, 5000, 25)	136.27(3.45)	118.73(2.63)	228.65(2.19)	278.43(3.42)	26.04(0.024)	4.96(0.11)
(200, 5000, 25)	82.69(1.12)	73.53(1.37)	177.44(1.57)	211.25(2.12)	25.56(0.012)	5.75(0.13)

Table 4:

The means of PEs and MSEs for our two-step procedure in three scenarios: exact selection (“Exact”), over selection (“Over”) and missing variables (“Miss”). The proportion of times among 100 simulated data sets for each scenario is also reported. The “NA” denotes the values that are not applicable.

$(n, s_n, \sigma_{\epsilon}^2)$	Scenario	MSE(B ₁)	MSE(B ₂)	MSE(B ₃)	MSE(B ₄)	PE	Proportion
(100, 2000, 1)	Exact	11.61(0.18)	9.18(0.16)	43.22(0.38)	45.55(0.47)	1.06(0.001)	0.27
	Over	11.17(0.17)	10.11(0.20)	43.6(0.46)	47.31(0.56)	1.05(0.001)	0.71
	Miss	240(0)	53.49(1.68)	44.58(1.61)	59.11(1.18)	1.1(0.001)	0.02
(200, 2000, 1)	Exact	7.09(0.08)	6.6(0.12)	23.97(0.16)	23.11(0.09)	1.03(0.0005)	0.04
	Over	5.87(0.09)	5.03(0.11)	27.76(0.26)	28.44(0.30)	1.03(0.0003)	0.96
	Miss	NA	NA	NA	NA	NA	0
(100, 5000, 1)	Exact	11.69(0.17)	9.74(0.19)	44.14(0.65)	45.79(0.41)	1.07(0.001)	0.27
	Over	11.76(0.20)	9.75(0.19)	44.24(0.43)	46.60(0.54)	1.06(0.001)	0.64
	Miss	240(0)	41.48(1.93)	44.96(0.65)	55.10(1.23)	1.12(0.001)	0.09
(200, 5000, 1)	Exact	7.52(0.09)	6.69(0.11)	24.91(0.22)	24.24(0.08)	1.04(0.001)	0.05
	Over	5.84(0.08)	4.84(0.08)	28.19(0.28)	28.57(0.28)	1.03(0.0005)	0.95
	Miss	NA	NA	NA	NA	NA	0
(100, 2000, 25)	Exact	121.75(1.27)	114.12(2.19)	229.79(1.95)	270.16(3.47)	26.18(0.024)	0.29
	Over	126.37(1.49)	121.47(2.69)	227.21(1.98)	283.29(3.00)	25.91(0.023)	0.70
	Miss	240(NA)	102.16(NA)	217.49(NA)	256.52(NA)	26.45(NA)	0.01
(200, 2000, 25)	Exact	79.26(0.80)	64.22(0.61)	177.62(0.86)	207.41(1.18)	25.54(0.017)	0.07
	Over	84.81(1.27)	74.4(1.43)	177.92(1.71)	215.21(2.20)	25.54(0.012)	0.93
	Miss	NA	NA	NA	NA	NA	0
(100, 5000, 25)	Exact	122.88(1.74)	114.5(2.49)	217.93(1.87)	260.99(2.64)	26.20(0.019)	0.34
	Over	129.81(1.53)	122.59(2.63)	233.46(2.17)	286.15(3.34)	25.92(0.020)	0.58
	Miss	240(0)	108.61(3.01)	239.31(2.06)	296.51(4.25)	26.23(0.022)	0.08
(200, 5000, 25)	Exact	85.2(0.87)	72.92(1.48)	174.83(1.20)	206.34(2.32)	25.66(0.014)	0.06
	Over	82.53(1.14)	73.57(1.37)	177.61(1.60)	211.56(2.12)	25.55(0.012)	0.94
	Miss	NA	NA	NA	NA	NA	0

Table 5:

PNC data analysis results: the top 7 SNPs selected by our screening procedure.

Ranking	Chromosome	SNP
1	5	rs72775042
2	5	rs6881067
3	5	rs72775059
4	10	rs200328746
5	10	rs75860012
6	10	rs200248696
7	10	rs78309702

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript