Statistica Sinica Preprint No: SS-2017-0073	
Title	Adaptive Estimation in Two-way Sparse Reduced-rank Regression
Manuscript ID	SS-2017-0073
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202017.0073
Complete List of Authors	Zhuang Ma Zongming Ma and Tingni Sun
Corresponding Author	Tingni Sun
E-mail	suntingni@gmail.com
Notice: Accepted version subject to English editing.	

Adaptive Estimation in Two-way Sparse Reduced-rank Regression

Zhuang Ma, Zongming Ma and Tingni Sun

University of Pennsylvania and University of Maryland

Abstract: This paper studies the problem of estimating a large coefficient matrix in a multiple response linear regression model when the coefficient matrix could be both of low rank and sparse in the sense that most nonzero entries concentrate on a few rows and columns. We are especially interested in the high dimensional settings where the number of predictors and/or response variables can be much larger than the number of observations. We propose a new estimation scheme, which achieves competitive numerical performance and at the same time allows fast computation. Moreover, we show that (a slight variant of) the proposed estimator achieves near optimal non-asymptotic minimax rates of estimation under a collection of squared Schatten norm losses simultaneously by providing both the error bounds for the estimator and minimax lower bounds. The effectiveness of the proposed algorithm is also demonstrated on an *in vivo* calcium imaging dataset.

Key words and phrases: Adaptive estimation, dimension reduction, group spar-

An earlier version of the present paper (Ma and Sun, 2014) under the title "Adaptive sparse reduced-rank regression" studied a one-way sparse reduced-rank regression model, which can be viewed as a special case of the model considered in this paper. The earlier version has been uploaded on arXiv, but is not intended for publication.

sity, high dimensionality, low rank matrices, minimax rates, neuroimaging, variable selection.

1. Introduction

High dimensional sparse linear regression has been one of the central topics of high dimensional statistical inference. When the response is univariate, researchers have developed a dazzling collection of tools to take advantage of the potential sparsity of the regression coefficients, e.g., Lasso (Tibshirani, 1996; Chen et al., 1998), SCAD (Fan and Li, 2001), Dantzig selector (Candes and Tao, 2007), MCP (Zhang, 2010), etc. In contemporary applications, we routinely face multivariate or even high dimensional response variables together with a large number of predictors, while the sample size can be much smaller. For example, in a cognitive neuroscience study, Vounou et al. (2012) used around ten thousand voxels from fMRI imaging as the response variables for each subject, and over four hundred thousand SNPs (single-nucleotide polymorphisms) as predictors. In comparison, the sample size was just several hundred.

Let n denote the sample size, m the number of responses, and p the number of predictors. We observe a pair of matrices Y and X from the following linear model

$$Y = XA + Z, \tag{1}$$

where Y is an $n \times m$ response matrix, X is an $n \times p$ design matrix, A is a $p \times m$ coefficient matrix that we are interested in estimating, and Z is an unobserved $n \times m$ matrix with i.i.d. noise entries. Thus, the i^{th} rows of Y and X collect the measurements of the response and the predictor variables on the i^{th} subject, respectively. When either the number of predictors p or the number of response variables m is large, it is hard to estimate the coefficient matrix A accurately unless certain structural assumption is imposed so that its intrinsic dimension is low.

In the literature, researchers have considered several important types of structural assumptions. One is *low-rankness* where the rank of A is assumed to be much smaller than its matrix dimensions p and m. Model (1) with such a structure has been referred to as reduced-rank regression and has been widely used in econometrics. See, for instance, Izenman (1975), Reinsel and Velu (1998) and the references therein. The other is *sparsity* where a large number of entries in the coefficient matrix are zeros. One may consider several different types of sparsity depending on the application problem one has in mind. If only s out of the p rows in A have non-zero entries, it is called *row sparsity*. In other words, only a small subset of size s out of the p predictors contribute to the variation of Y. Structures of this kind arise naturally in the context of multi-task learning (Koltchinskii et al., 2011). It can also be viewed as a leading example of *group sparsity* (Yuan and

Lin, 2006), where the rows of A form natural groups. If only k out of the m columns in A have non-zero entries, it is called *column sparsity*. In this case, only k out of the m response variables are affected by the predictors under consideration.

In this paper, we are interested in the situation where low-rankness, row sparsity and column sparsity could be present in the coefficient matrix simultaneously. In what follows, we refer to model (1) with these structures as the *two-way sparse reduced-rank regression* model. The interest in such a model comes from both applications and theory, and has risen significantly in recent years. In applications such as genomics and neurosciences, researchers can now measure a lot of response and predictor variables and so the size of the coefficient matrix is ever increasing. Thus, imposing both low-rankness and two-way sparsity leads to enhanced interpretability and hence can be more attractive than simply imposing one type of structure. For instance, Ma et al. (2014) conducted a case study of regulatory relationships between different genome-wide measurements, in which the predictors are micro-RNA measurements and the response variables are gene expression levels. The sparsity results from the fact that a relatively small number of micro-RNAs regulated a small collection of genes under the specific experiments of interest, and the low-rankness assumption is reasonable since only a handful of regulatory programs were present. For estimating the

coefficient matrix in this model, several algorithms have been introduced. See, for instance, Chen et al. (2012) and Ma et al. (2014). However, to the best of our limited knowledge, there is no theoretical guarantee on the performance of these procedures in the high dimensional regime where the number of predictors and/or response variables exceeds the sample size.

Main contributions The main contributions of the present paper are two-folded. On one hand, we propose a new computationally efficient estimator for the coefficient matrix in (1) that could take advantage of the potential presence of low-rankness and two-way sparsity adaptively. The new estimator shows competitive numerical performance under a variety of simulation settings when compared with state-of-the-art methods. We also demonstrate how the estimation scheme can play a critical role in analyzing the spatial-temporal structure in calcium imaging data. On the other hand, we obtain new minimax estimation rates of the coefficient matrix with respect to a large class of squared Schatten norm losses and show that (a slight variant of) our estimator can achieve the near optimal rates adaptively for this large collection of loss functions simultaneously when the noise terms are homoscedastic and Gaussian.

Connection to the literature When the coefficient matrix is *either* sparse *or* of low rank, researchers have obtained deep understanding on how the optimal mean squared estimation/prediction error depends on the

model parameters and on how to achieve near optimal error rates without knowing the true rank or sparsity. See, for instance, Bunea et al. (2011) for the low rank case, and Huang and Zhang (2010) and Lounici et al. (2011) for the row sparse case.

In addition, researchers have performed extensive study of the case where both low-rankness and row sparsity are present. Chen and Huang (2012) proposed a weighted rank-constrained group Lasso approach with two heuristic numerical algorithms and studied its fixed dimension large sample asymptotics. Bunea et al. (2012) derived oracle inequalities and studied the minimax rates under squared prediction error loss for this model in the high dimensional setting. See also She (2014) and an earlier version of the present paper (Ma and Sun, 2014).

The line of work that is closest to the present paper includes Chen et al. (2012) and Ma et al. (2014). The main focus of these two papers was on methodology. In comparison, the present paper not only proposes a new method but also justifies its practical effectiveness by both numerical and theoretical studies. From a slightly different perspective, a series of papers have considered the problem of sparse SVD (Lee et al., 2010; Yang et al., 2014, 2016), which can be viewed as a special case of two-way sparse reduced-rank regression with orthogonal design.

Organization The rest of the paper is organized as follows. Section 2

presents our new methodology for obtaining a simultaneously sparse and low rank estimator of the coefficient matrix. Its competitive numerical performance is demonstrated in Section 3 through both simulated and real data examples. In Section 4, we provide finite sample upper bounds for (a slight variant of) the proposed estimator with respect to a collection of squared Schatten norm losses. In addition, we derive minimax lower bounds and hence show that the proposed estimator is simultaneously adaptive and near optimal with respect to all loss functions under consideration. Section 5 discusses interesting related problems for future research. The proofs of the theorems are presented in Section 6 in the supplement.

Notation For an $n \times p$ matrix $X = (x_{ij})$, the i^{th} row of X is denoted by X_{i*} and the j^{th} column X_{*j}. For a positive integer k, $[k]$ denotes the index set $\{1, 2, ..., k\}$. For any set I, $|I|$ denotes its cardinality and I^{c} its complement. For two subsets I and J of indices, we write X_{IJ} for the $|I| \times |J|$ submatrices formed by x_{ij} with $(i, j) \in I \times J$. For conciseness, we let $X_{I*} = X_{I[p]}$ and $X_{*J} = X_{[n]J}$. For any matrix X, $\text{supp}(X)$ stands for the index set of its nonzero rows. We denote the rank of X by $\mathsf{rank}(X)$, and $\sigma_i(X)$ stands for its i^{th} largest singular value. For any $q \in [1, \infty)$, the Schatten-q norm of X is $\|X\|_{\text{S}_q} = \left(\sum_{i=1}^{n \wedge p} \sigma_i^q(X)\right)^{1/q}$, and for $q = \infty$, $\|X\|_{\text{S}_\infty} = \sigma_1(X)$. Note that $\|X\|_{\text{S}_2} = \|X\|_{\text{F}}$ is the Frobenius norm and $\|X\|_{\text{S}_\infty} = \|X\|_{\text{op}}$ is the operator norm of X. For any vector a, $\|a\|$ denotes its ℓ_2 norm. The ℓ_2/ℓ_1

norm of X is defined as the ℓ_1 norm of the vector consisting of its row ℓ_2 norms: $\|X\|_{2,1} = \sum_{j=1}^{n} \|X_{j*}\|$. If $n \geq p$ and X has orthonormal columns, then we say X is an orthonormal matrix, and we write $X \in O(n, p)$. We use $\mathbf{1}_d$ to denote the all-one vector in $\mathbb{R}^d$. For any real number a and b, set $a \vee b = \max\{a, b\}$, $a \wedge b = \min\{a, b\}$ and $a_+ = a \vee 0$.

2. Methodology

2.1 Main Algorithm

The proposed estimation scheme, called *Double Projected Penalization* (DPP), is summarized in Algorithm 1. To initialize the algorithm, we need to specify the rank r of the estimated coefficient matrix and a penalty function $\rho(\cdot\,; \lambda)$ to be used in group penalized regression. In what follows, we explain the main ideas underlying the algorithm, while the choice of penalty and other initialization details are deferred to Sections 2.2 and 2.3.

The algorithm consists of two stages. The first stage involves steps 1–2 and the second stage steps 3–5. In either stage, one first screens the columns of Y, then computes the r leading right singular vectors of the screened response matrix, and finally performs a group penalized regression on the projected data where the projection is onto the subspace spanned by the leading right singular vectors. The purpose of the screening step is to pick those response variables the signals of which stand out of noise. To motivate

the projection step, we observe that if the right singular vector matrix V of XA were known, then one could immediately reduce dimensionality by considering the new regression problem which replaces Y and A in (1) with their projected counterparts YV and AV. Thus, in either stage, we first estimate V by the r leading right singular vectors of the screened response matrix (a further projection is involved in the second stage), and then project the data by post-multiplying the response matrix with the estimated right singular vector matrix. When regressing the projected responses on X, we actually estimate AV. Note that if A has at most s nonzero rows, so does AV. Thus, the rows of AV form natural groups and it makes sense to induce row sparsity in our estimator of AV by performing a group penalized regression.

We now move on to discuss the necessity of the second stage. Comparing the two stages, we note that both the screening step and the estimation of the right singular matrix V are different, but both differences are due to the involvement of the matrix $U_{(1)}$. By definition, $U_{(1)} \in \mathbb{R}^{n\times r}$ consists of the left singular vectors of $XB_{(1)}$. Since $B_{(1)}$ is an estimate of AV, the column subspace of $U_{(1)}$ estimates the left singular subspace of XAV, or equivalently, the left singular subspace of XA. By projecting onto $U_{(1)}$, we increase the signal-to-noise ratio in the screening step. As a result, we would be able to select more columns the signals of which might have been

drowned in noise in the first stage. The inclusion of more signal columns of Y would in turn contribute to the estimation accuracy of the final estimator. Similarly, by pre-multiplying $\widetilde{Y}^{(1)}$ with $U_{(1)}U'_{(1)}$, we further boost the signal-to-noise ratio when estimating the right singular vector matrix V, and thus obtain a better estimator $V_{(1)}$. As to be revealed by later analysis, the second stage is critical for achieving high estimation accuracy for A.

In an earlier version of the present paper (Ma and Sun, 2014), we considered a one-way sparse reduced rank regression model that does not assume column sparsity in A. Compared with the earlier version, the current algorithm takes advantage of the potential column sparsity by column screening in both steps 1 and 3. As we shall show later in Section 4, even when column sparsity is absent, our procedure could still adapt automatically to achieve the best possible accuracy of estimation subject to some multiplicative log factor in the low-rank and row sparse scenario.

2.2 Group Penalized Regression

The penalized regression in steps 2 and 4 of Algorithm 1 can be viewed as a special case of linear regression with group sparsity, where each row of the coefficient matrix is considered as a group and all groups are of the same size r.

Penalized regression with group structure has been extensively studied.

Algorithm 1: Estimation scheme for A via the Double Projected Penalization

Input: Observed response matrix Y, design matrix X, rank r, noise level σ, positive constants α, β and penalty function $\rho(\cdot;\lambda)$ with penalty level λ.

Output: Estimated coefficient matrix $\widehat{A}$.

1 Column screening of Y. Select columns

$$J_{(0)} = \Big\{ j : \|Y_{*j}\|^2 \geq \sigma^2 (n + \alpha\sqrt{n \log(p \vee m)}) \Big\}.$$

Define $\widetilde{Y}^{(0)}$, where $\widetilde{Y}^{(0)}_{*j} = Y_{*j} I\{j \in J_{(0)}\}$.

Compute the right singular vectors of $\widetilde{Y}^{(0)}$, denoted by an $m \times r$ matrix $V_{(0)}$.

2 Group penalized regression

$$B_{(1)} = \underset{B \in \mathbb{R}^{p \times r}}{\arg\min} \Big\{ \|YV_{(0)} - XB\|_F^2/2 + \rho(B;\lambda) \Big\},$$

3 Column screening of Y. Compute the left singular vectors of $XB^{(1)}$, denoted by an $n \times r$ matrix $U_{(1)}$. Select columns

$$J_{(1)} = J_{(0)} \cup \Big\{ j : \|U_{(1)}' Y_{*j}\|^2 \geq \beta\sigma^2 (r + 2\sqrt{3r \log(p \vee m)} + 6 \log(p \vee m)) \Big\}.$$

Define $\widetilde{Y}^{(1)}$, where $\widetilde{Y}^{(1)}_{*j} = Y_{*j} I\{j \in J_{(1)}\}$.

Compute the first r right singular vectors of $U_{(1)} U_{(1)}{}' \widetilde{Y}^{(1)}$, denoted by an $m \times r$ matrix $V_{(1)}$.

4 Group penalized regression

$$B_{(2)} = \underset{B \in \mathbb{R}^{p \times r}}{\arg\min} \Big\{ \|YV_{(1)} - XB\|_F^2/2 + \rho(B;\lambda) \Big\},$$

5 Compute the estimated coefficient matrix by $\widehat{A} = B_{(2)} V_{(1)}{}'$.

One of the most popular procedures is the group Lasso (Bakin, 1999; Yuan and Lin, 2006), where the penalty function is defined by the ℓ_2/ℓ_1 matrix

norm as follows

$$\rho(B;\lambda) = \lambda\|B\|_{2,1} = \lambda\sum_{j=1}^{p}\|B_{j*}\|_2. \tag{2}$$

The theoretical properties of group Lasso have been studied in the literature, using ideas originating from the study of Lasso. Huang and Zhang (2010) showed the upper bounds for the estimation and prediction errors of group Lasso with proper penalty level under strong group sparsity and group sparse eigenvalue conditions. Lounici et al. (2011) provided similar error bounds under a group version of the restricted eigenvalue condition.

In Section 4, we will present a theoretically justified choice of the penalty level λ for the group Lasso penalty function (2) when we have i.i.d. Gaussian noises.

2.3 Initialization

We now discuss the initialization of Algorithm 1. Throughout, we assume the noise standard deviation σ is known. Otherwise, we can estimate it by

$$\widehat{\sigma} = \text{median}(\sigma(Y))/\sqrt{n \vee m}, \tag{3}$$

where $\sigma(Y)$ is the collection of all nonzero singular values of Y. If the true rank of A is not known, we propose to apply the estimator in Bunea et al.

Algorithm 2: Rank Estimation

Input: Response matrix Y, design matrix X, noise level σ and a threshold level η.

Output: Estimated rank $\widehat{r}$, initial matrix $V_{(0)}$.

1 Compute $P = XM^{-}X'$, where $M = X'X$ and M^{-} its Moore–Penrose pseudo-inverse.

2 Compute the singular values of PY and select

$$\widehat{r} = \max\left\{j : \sigma_j(PY) \geq \sigma\eta\right\}.$$

(2011), which is summarized in Algorithm 2. The user specified parameter can be selected as

$$\eta = \sqrt{2m} + \sqrt{2(n \wedge p)}, \tag{4}$$

which was suggested by Bunea et al. (2012) for Gaussian data.

In practice, we may also select the rank based on cross validation. Suppose the data is split into training and test samples. For any given value of $r \in [m \wedge p]$, we may run Algorithm 1 using only the training sample, and the resulting $\widehat{A}$ is then used to calculate the prediction error on the test sample. Thus, we can select the value of r that leads to the smallest prediction error on the test sample, or the smallest average prediction error if k-fold cross validation is used.

3. Numerical Study

3.1 Simulation

In this part, we compare the proposed DPP method, i.e. Algorithm 1, with the thresholding SVD method (TSVD) in Ma et al. (2014) and the exclusive extraction algorithm (EEA) in Chen et al. (2012). For fair comparison, equations (3)–(4) and Algorithm 2 were applied to estimate the noise variance and the rank of the coefficient matrix for all methods in all simulation settings.

Comparison under different model parameters We first compare these methods under different design matrices, ranks and sparsity levels. To this end, we borrow several simulation settings from Bunea et al. (2012), but also add columns of pure noises in the response matrices to induce two-way sparsity. The rows of the design matrix X are i.i.d. random vectors sampled from a multivariate Gaussian distribution with mean zero and covariance matrix Σ, where $\Sigma_{ij} = \rho^{|i-j|}$. The coefficient matrix $A \in \mathbb{R}^{p\times m}$ has the form

$$A = \begin{pmatrix} A_1 & 0 \\ 0 & 0 \end{pmatrix} = \begin{pmatrix} bB_0B_1 & 0 \\ 0 & 0 \end{pmatrix}$$

with $b > 0$, $B_0 \in \mathbb{R}^{s\times r}$ and $B_1 \in \mathbb{R}^{r\times k}$, where all entries in B_0 and B_1 are filled with i.i.d. random numbers from $N(0,1)$. The noise matrix $Z \in \mathbb{R}^{n\times m}$ has i.i.d. $N(0,\sigma^2)$ entries. The following settings are considered with $\sigma = 1$ and $\rho = 0.1$ or 0.9:

- $n = 30$, $m = 50$, $p = 100$, $s = 15$, $k = 10$, $r = 2$, $b = 0.5$ or 1;
- $n = 100$, $m = 50$, $p = 25$, $s = 15$, $k = 25$, $r = 5$, $b = 0.2$ or 0.4.

Large values of b correspond to large signal-to-noise ratios.

We compare the following five estimators derived from the three methods. The first two estimators are computed by Algorithm 1 with $\alpha = 2\sqrt{3}$, $\beta = 1$ and two possible choices of penalty level λ. The one with an estimated universal penalty level $\lambda_{\text{univ}} = \widehat{\sigma}\sqrt{2\log(p)/n}$ is denoted by DPP, while the estimator DPP.cv selects a penalty level λ from the set $\{2^{i/2}\lambda_{\text{univ}} : i = -5, \ldots, 4\}$ via 5-fold cross validation. The third is the TSVD estimator which was implemented by the R package "tsvd" (version 1.3) with the default penalization option "BICtype=2". The last two are EEA and its iterative extension, denoted by iEEA.

Fig. 1 and Fig. 2 show the boxplots of the prediction errors, estimation errors and sizes of selected models based on 50 replications in each setting. The red lines indicate the true model sizes (the numbers of nonzero rows/columns). The estimated ranks for each simulation setting are re-

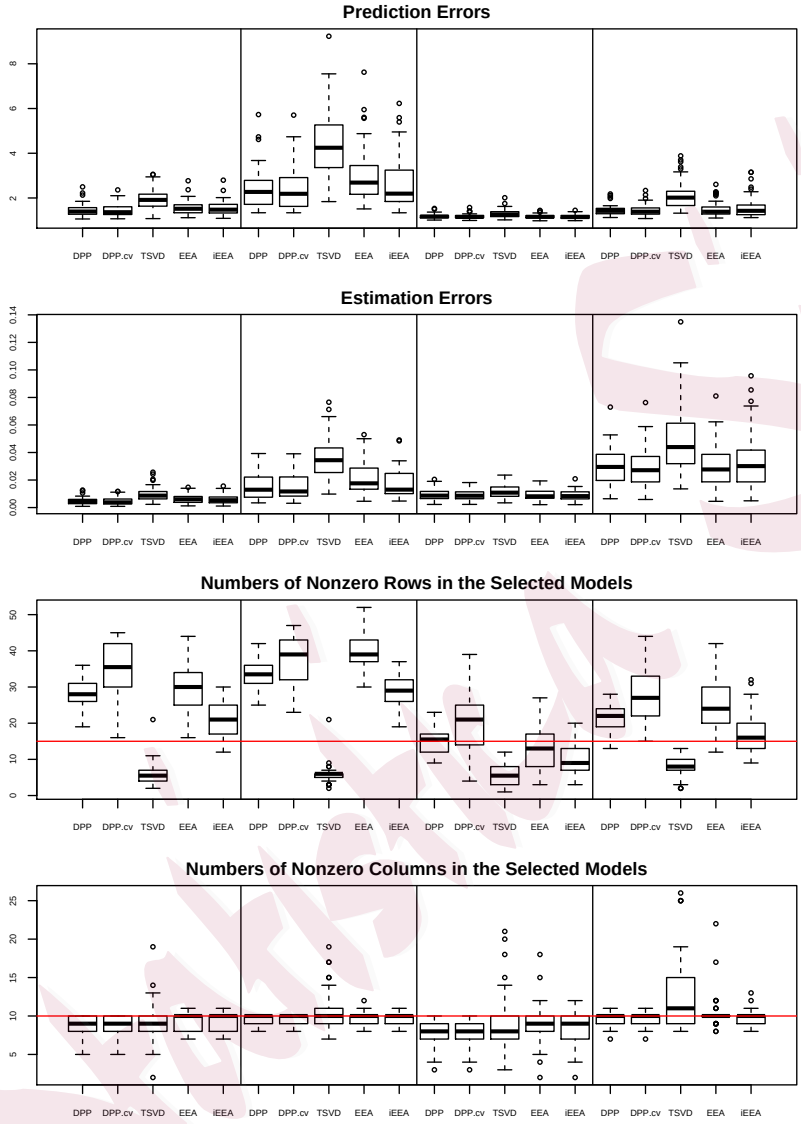

Figure 1: Performance of five methods: prediction errors, estimation errors and sizes of selected models across 50 replications. Sample size $n = 30$, model size $m = 50$, $p = 100$, $s = |\mathrm{supp}(A)| = 15$, $k = |\mathrm{supp}(A')| = 10$ and rank $r = 2$. The four blocks in each plot are for $(\rho, b) = (0.1, 0.5), (0.1, 1), (0.9, 0.5), (0.9, 1)$, respectively.

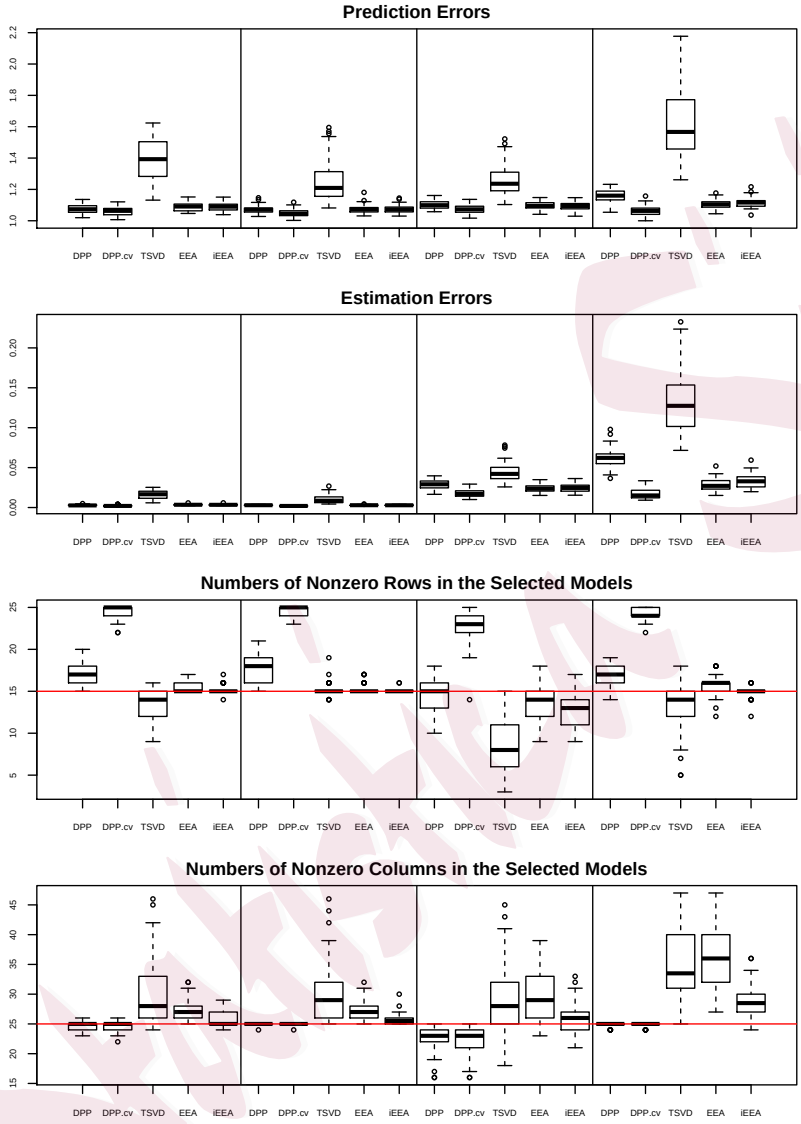

Figure 2: Performance of five methods: prediction errors, estimation errors and sizes of selected models across 50 replications. Sample size $n = 100$, model size $m = 50$, $p = 25$, $s = |\mathrm{supp}(A)| = 15$, $k = |\mathrm{supp}(A')| = 25$ and rank $r = 5$. The four blocks in each plot are for $(\rho, b) = (0.1, 0.2), (0.1, 0.4), (0.9, 0.2), (0.9, 0.4)$, respectively.

Table 1: The estimated ranks for all simulation settings.

Dimensions (n, m, p, s, k, r)	b	$\rho = 0.1$	$\rho = 0.9$
(30,50,100,15,10,2)	0.5	1.92 ± 0.27	1.54 ± 0.5
	1	2 ± 0	2 ± 0
(100,50,25,15,25,5)	0.2	4.74 ± 0.44	3.16 ± 0.55
	0.4	5 ± 0	4.56 ± 0.5

ported in Table 1. It is noticed that DPP.cv has the best performance for almost all cases considered, while DPP with the estimated universal penalty level tends to choose a smaller model with slightly larger estimation errors. In some settings, DPP.cv was able to reduce the estimation errors by up to 40% when compared to TSVD, EEA and iEEA. Note that when comparing prediction errors, the quantity that makes most sense is the excessive error an estimator makes in addition to the oracle error that one would make even when the true coefficient matrix is given. In the current setting, the (normalized) oracle error is 1. In terms of the excessive prediction error, it is observed that the prediction accuracy of DPP.cv outperformed the other methods by a similar percentage. Additionally, it is worth noting that the proposed method tends to choose more rows than the true model, while the column selection, relying on the screening of the columns of $U'Y$, is more accurate. This is somewhat expected as group Lasso tends to over-select variables when cross validation is deployed for choosing tuning parameter values.

Comparison under different noise distributions We now compare the performance of these methods on non-Gaussian data. To this end, we consider three different noise distributions: $\sqrt{3/5}t_5$, $\sqrt{4/5}t_{10}$ and 3 Uniform (the sum of three uniform $[-1, 1]$ random variables). Here, t_ν stands for the t-distribution with ν degrees of freedom. We note that all three distributions have been normalized to have unit variance. Fig. 3 shows the simulation results for the second setting with $\rho = 0.1, b = 0.2$ and for all three noise distributions along with the standard normal error. It shows that our methods, esp. DPP.cv, preserve competitive performance even for non-Gaussian data. Moreover, when compared with the corresponding performance measures on Gaussian data (the first block of boxplots), we see that all the estimators were relatively robust to the noise distributions, though their performance (with the exception of TSVD) did degrade as the tail of the noise distribution gets heavier.

Performance under heteroscedastic noises Although the proposed method is designed under the model that the responses have equal variances, we test the robustness of our method for heteroscedastic cases. In what follows, the noise matrix $Z \in \mathbb{R}^{n\times m}$ has independent normal entries with mean zero and variance σ_j^2 for the j-th column, where σ_j^2 is selected from a uniform distribution $U[\frac{2}{\omega+1}, \frac{2\omega}{\omega+1}]$ with four choices of $\omega = 1, 2, 5, 10$. In

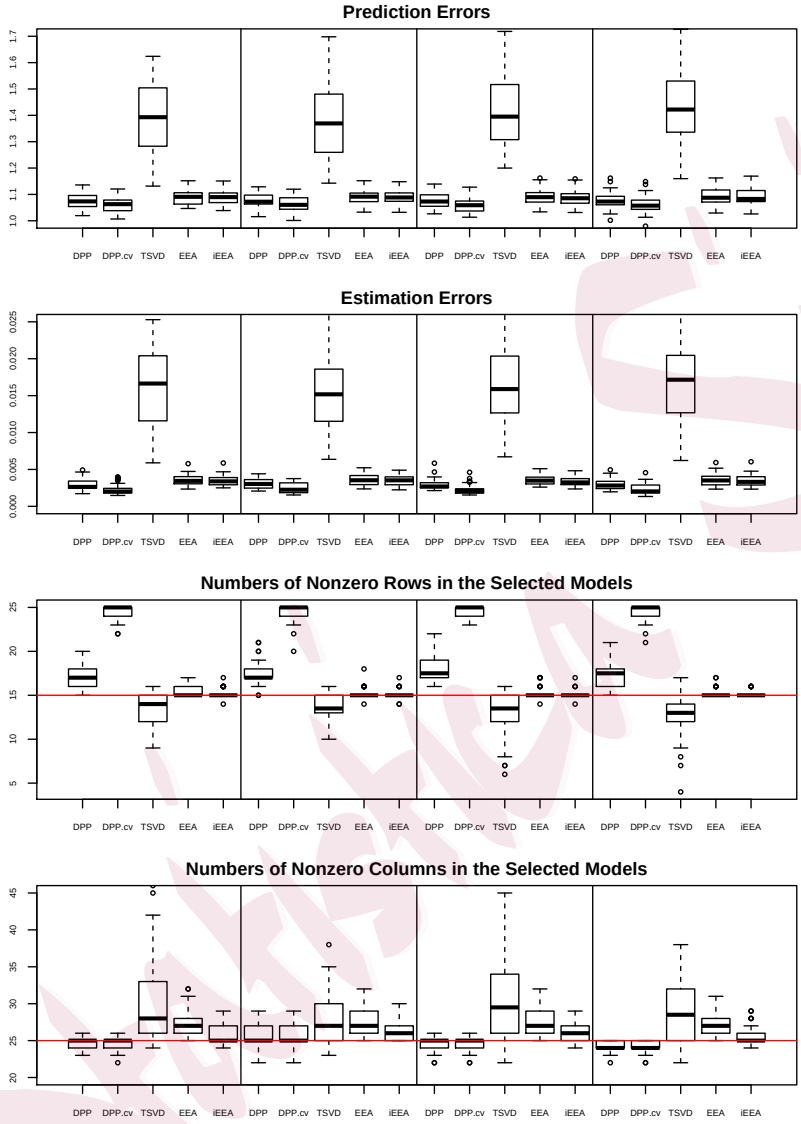

Figure 3: Performance of five methods on non-Gaussian data. Sample size $n = 100$, model size $m = 50$, $p = 25$, $s = |\mathrm{supp}(A)| = 15$, $k = |\mathrm{supp}(A')| = 25$, rank $r = 5$, $\rho = 0.1$ and $b = 0.2$. The four blocks in each plot are for different noise distributions: standard normal, $\sqrt{3/5}t_5$, $\sqrt{4/5}t_{10}$ and 3 Uniform (the sum of three uniform $[-1, 1]$ random variables).

this setting, ω is the ratio of the largest possible variance over the smallest one. When $\omega = 1$, this becomes the case of equal variance as above (the second setting with $\rho = 0.1, b = 0.2$). When ω is getting larger, the noise variance varies among columns, while the average noise variance remains 1. In Fig. 4, we report the prediction, estimation and selection performance of the proposed DPP method for different ω's. When heteroscedasticity occurs, our approach selected more columns than and comparable numbers of rows to the homoscedastic case. The prediction and estimation errors were not significantly affected.

3.2 *In vivo* Calcium Imaging Data

Calcium imaging has become an increasingly important tool in neuroscience to track the activity of neuronal populations by recording the dynamics of the time-varying fluorescence of the neurons (Akerboom et al., 2012; Chen et al., 2013). When a neuron fires an electrical action potential (spike), calcium will enter the cell and change its fluorescent properties by attaching to genetically encoded calcium indicators. By recording the movies of fluorescence activities, researchers hope to identify and demix the regions of interest (ROIs) as well as extract spike traces (Pnevmatikakis et al., 2014; Haeffele et al., 2014).

Following the spatiotemporal model in Pnevmatikakis et al. (2014),

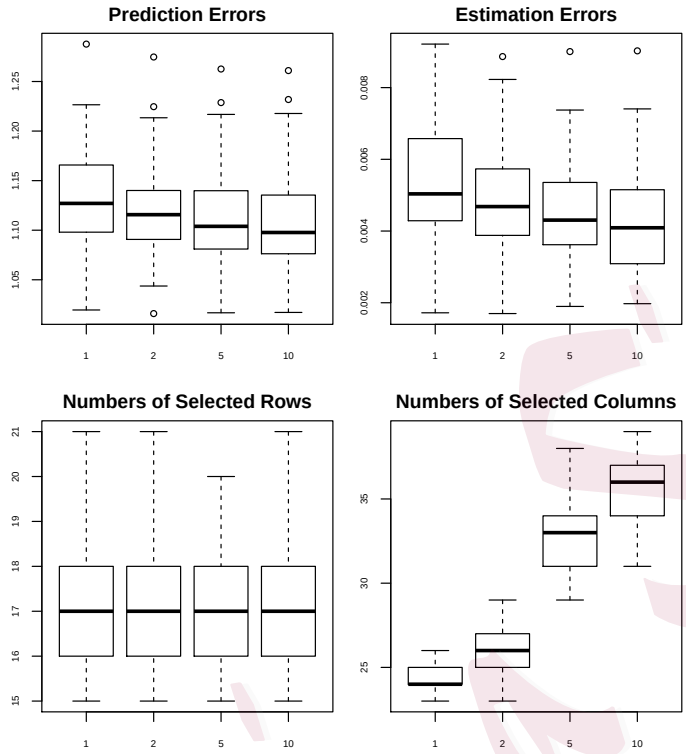

Figure 4: Performance of the proposed DPP method on heteroscedastic data. Sample size $n = 100$, model size $m = 50$, $p = 25$, $s = |\text{supp}(A)| = 15$, $k = |\text{supp}(A')| = 25$, rank $r = 5$, $\rho = 0.1$ and $b = 0.2$. The four boxplots in each plot are for $\omega = 1, 2, 5, 10$, respectively.

suppose an $l_1 \times l_2$ area (2d imaging plane of an original 3d volume) containing K neurons (possibly overlapping) is monitored for T time frames. Here, K is typically much smaller compared to $l_1 \times l_2$ and T. Let $c_i = (c_i(1), \cdots, c_i(T))' \in \mathbb{R}^T$ be the calcium activity and $\omega_i \in \mathbb{R}^m$ $(m = l_1 \times l_2)$ be the spatial footprint (stacked by the monitored area) of the i^{th} neuron.

Then the fluorescence intensity observed at time t can be modeled as

$$y_t = \sum_{i=1}^{K} \omega_i c_i(t) + z_t, \ 1 \le t \le T,$$

where $z_t \overset{iid}{\sim} N(0, \sigma^2 I_m)$ is the noise vector at time t. In matrix notations,

$$Y = C\Omega + Z,$$

where $Y = (y_1, \cdots, y_T)' \in \mathbb{R}^{T\times m}, \Omega = (\omega_1, \cdots, \omega_K)' \in \mathbb{R}^{K\times m}, C = (c_1, \cdots, c_K) \in \mathbb{R}^{T\times K}, Z = (z_1, \cdots, z_T)' \in \mathbb{R}^{T\times m}$. Let $s_i = (s_i(1), \cdots, s_i(T))' \in \mathbb{R}^T$ be the spike trace of the i^{th} neuron. Then the calcium activity can be characterized by a simple first order autoregressive model,

$$c_i(t) = \gamma c_i(t-1) + s_i(t), \ 1 \le t \le T,$$

or equivalently ($c_i(0) = 0$ by convention), $S = GC$, where $S = (s_1, \cdots, s_K) \in \mathbb{R}^{T\times K}$ and

$$G = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ -\gamma & 1 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & -\gamma & 1 \end{pmatrix} \in \mathbb{R}^{T\times T}.$$

In this way,

$$Y = G^{-1}S\Omega + Z = XA + Z \tag{5}$$

where $A = S\Omega$ is the spatiotemporal convolution matrix and $X = G^{-1}$ is a known design matrix. The support of Ω is the location of the neurons and the support of S represents the time frames when the neurons fire. Because the number of neurons in the monitored area is small and the neurons do not fire very frequently, Ω is approximately row sparse and S is approximately column sparse, which together imply that A is two-way sparse (also low-rank since the rank is no greater than the number of neurons K). Therefore, the generative model (5) can be viewed as a special case of model (1) with $n = p = T$ and $m = l_1 \times l_2$. To recover Ω and S, we suggest first estimating A by the proposed algorithm and then running a nonnegative matrix factorization (NMF) on $\widehat{A}$ to obtain $\widehat{\Omega}$ and $\widehat{S}$. Pnevmatikakis et al. (2014) proposed an alternating l_1 minimization strategy to estimate Ω and S but no theoretical guarantee has been established for such heuristic.

The calcium imaging data ($n = p = T = 559$, $m = 135 \times 131$) we use here is taken *in vivo* from the primary auditory cortex of a mouse with genetically encoded calcium indicator GCaMP5 (Akerboom et al., 2012).

Following Vogelstein et al. (2010), γ is set at $\gamma = 1 - 1/(\text{frame rate})$.

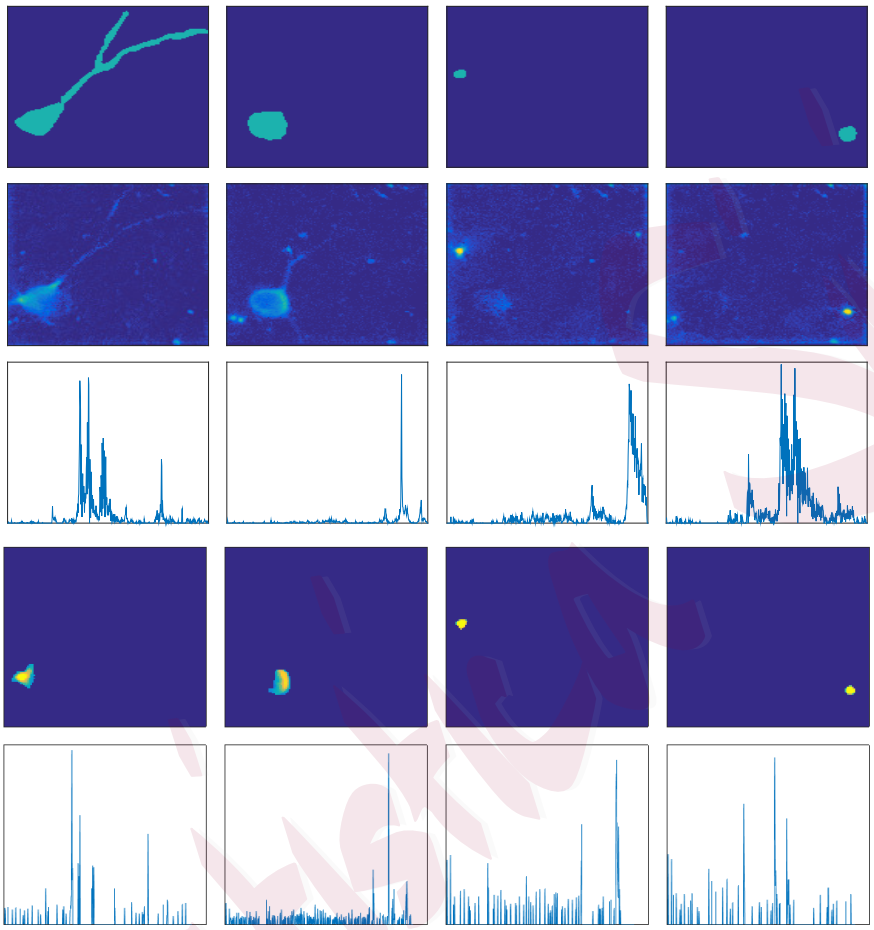

Figure 5: Application to *in vivo* calcium imaging data. First row: manually segmented regions of neurons. Second row: heat maps of the recovered spatial components by Algorithm 1. Third row: estimated spike trace by Algorithm 1. Fourth row: heat maps of the corresponding spatial components recovered by the method in Pnevmatikakis et al. (2014). Fifth row: estimated spike trace by the method in Pnevmatikakis et al. (2014). In the third row and the fifth row, the spatial components have been rescaled to have the same ℓ_2 norms.

We report here four most significant neurons to demonstrate the effectiveness of the proposed method as illustrated in Figure 5. For comparison, we have also included the best matching findings by the method in Pnevmatikakis et al. (2014) and its matlab implementation Giovannucci et al. (2017). In Figure 5, the first row shows the manually segmented regions of the neurons from the raw dataset, which can be approximately regarded as the true support of the spatial component Ω. The first neuron consists of a cell body with a dendritic branch and it heavily overlaps with the second neuron, making manual segmentation very challenging. The second row displays the heat maps of the recovered neurons by the proposed approach and they match the manual segmentation very well. The third row of Figure 5 shows the estimated spike traces by our method. The fourth and the fifth rows show the corresponding components found by the method proposed in Pnevmatikakis et al. (2014). These estimates are in general sparser than those obtained by Algorithm 1. However, they fail to recover the dendritic branch in the top-left subplot. Indeed none of the spatial components extracted by the method in Pnevmatikakis et al. (2014) captured this important structure in our experiment.

4. Theoretical Properties

In this section, we present theoretical results for a slight variant of the proposed estimation scheme when the noise matrix Z in (1) has i.i.d. Gaussian entries. Their proofs are provided in the supplementary materials.

4.1 Minimax Upper Bounds

To facilitate the discussion, we put the estimation problem in a decision–theoretic framework. We are interested in estimating the coefficient matrix A in model (1) where A is both two-way sparse and of low rank, and Z has i.i.d. $N(0, \sigma^2)$ entries. Thus, we assume that A belongs to the following parameter space

$$\begin{aligned}\Theta(s,k,r,d,\gamma) \;=\; \Big\{ & A \in \mathbb{R}^{p\times m} : \mathsf{rank}(A) = r, \gamma d \geq \sigma_1(A) \geq \cdots \geq \sigma_r(A) > d > 0, \\ & |\mathrm{supp}(A)| \leq s, |\mathrm{supp}(A')| \leq k \Big\}, \end{aligned} \tag{6}$$

where $\mathrm{supp}(M)$ is the index set of nonzero rows in matrix M. Here and after, we treat γ as an absolute positive constant. To measure the accuracy of any estimator $\widetilde{A}$, we consider the following class of squared Schatten norm losses:

$$L_q(A, \widetilde{A}) = \|\widetilde{A} - A\|_{\mathrm{s}_q}^2, \qquad q \in [1, 2]. \tag{7}$$

For simplicity, we assume the noise variance σ^2 is known. In addition, we treat the design matrix X as fixed and the only source of randomness is the noise matrix Z. In what follows, we present high probability error bounds for (a slight variant of) the DPP estimator where independent samples are generated and used in steps 1–4. We believe the deviation from Algorithm 1 is an artifact of the proof technique. Numerical studies (not reported) showed that the algorithm produces comparable results whether independent samples are used or a single sample is used repeatedly.

Independent sample generation Note that we can generate the desired independent samples from the observed (X, Y) when the noises are homoscedastic and Gaussian. Indeed, when the entries of the noise matrix Z are i.i.d. $N(0, \sigma^2)$, we can first generate an independent copy $\widetilde{Z}$ so that all entries in $Z + \widetilde{Z}$ and $Z - \widetilde{Z}$ are mutually independent and all follow the same Gaussian distribution $N(0, 2\sigma^2)$. Thus, $Y + \widetilde{Z}$ and $Y - \widetilde{Z}$ are independent, following model (1) with i.i.d. $N(0, 2\sigma^2)$ noises. Employing this trick twice, we can generate four independent copies of responses

$$Y_{(i)} = XA + Z_{(i)}, \quad i = 0, 1, 2, 3,$$

where $Z_{(i)}$ has i.i.d. $N(0, \widetilde{\sigma}^2)$ entries with $\widetilde{\sigma} = 2\sigma$. In the rest of this paper, when we mention Algorithm 1, we refer to the procedure with independent

samples $Y_{(i)}$ used in the $(i+1)^{\text{th}}$ step, $i = 0, 1, 2, 3$, where the noise variance is $\widetilde{\sigma}^2 = 4\sigma^2$.

The design matrix Without of loss of generality, we assume X is of full rank. Otherwise, we can always perform the following operation to reduce to the full rank case. If $\mathsf{rank}(X) = q < n \wedge p$ and let $O \in \mathbb{R}^{n\times q}$ be its left singular vector matrix. Setting $\widetilde{Y} = O'Y$ and $\widetilde{X} = O'X$, we obtain that $\widetilde{Y}$ and $\widetilde{X}$ satisfy model (1) with the same coefficient matrix A, i.i.d. $N(0, \sigma^2)$ noises and a design matrix of full rank.

We write the singular value decomposition of XA as

$$XA = U\Delta V' \tag{8}$$

with $U \in O(n, r)$, $V \in O(m, r)$ and $\Delta = \mathrm{diag}(\delta_1, \ldots, \delta_r)$ collects the non-zero singular values of XA. To introduce appropriate assumptions on X, we first make the following definition.

Definition 1. For any $k \in [p]$, the *ℓ-sparse Riesz constants* $\kappa_{\pm}(\ell)$ of X are defined as

$$\kappa_-^2(\ell; X) = \min_{B\subset[p],|B|=\ell} \sigma_{\min}(X'_{*B}X_{*B}), \qquad \kappa_+^2(\ell; X) = \max_{B\subset[p],|B|=\ell} \sigma_{\max}(X'_{*B}X_{*B}) \tag{9}$$

By definition, if the ℓ-sparse Riesz constants of X are $\kappa_{\pm}(\ell; X)$, then for any $l \in [\ell]$, the l-sparse Riesz constants $\kappa_{\pm}(l; X)$ of X satisfy $\kappa_{-}(\ell; X) \leq \kappa_{-}(l; X) \leq \kappa_{+}(l; X) \leq \kappa_{+}(\ell; X)$.

To establish upper bounds for the proposed estimator, for some integer s_* depending only on s, we require the s_*-sparse Riesz constants of X to satisfy the following condition.

Condition 1 (Sparse eigenvalue condition)**.** There exist positive constants s_* and c_* and $K \geq 1$, such that the s_*-sparse Riesz constants satisfy $K^{-1} \leq \kappa_{-}(s_*; X) \leq \kappa_{+}(s_*; X) \leq K$ and

$$\frac{\kappa_{+}^2(s_*; X) - \kappa_{-}^2(2s_*; X)}{\kappa_{-}^2(s_*; X)} < c_*.$$

We do not put condition on $\kappa_{-}(2s_*; X)$. Following the above definition and discussion, we know that $0 \leq \kappa_{-}(2s_*; X) \leq \kappa_{-}(s_*; X)$ always holds.

The following theorem gives high probability upper bounds, provided that the design matrix satisfies mild regularity conditions and the penalty level is properly chosen.

Theorem 1. *Let $A \in \Theta(s, k, r, d, \gamma)$ where $s \geq r \geq 1$. Set the penalty level*

$$\lambda = 4\sigma \max_{j \leq p} \|X_{*j}\| (\sqrt{r} + \sqrt{4 \log(p \vee m)}) \tag{10}$$

in steps 2 and 4 of Algorithm 1 with the group Lasso penalty (2). Let $\alpha = 2\sqrt{3}$ *and* $\beta = 1.1$ *in Algorithm 1. Suppose that Condition 1 holds with an absolute constant* $K > 1$ *for all* X *and positive constants* s_*, c_* *satisfying*

$$s_* \geq 2s, \quad 6c_* \leq \sqrt{s_*/s - 1}, \tag{11}$$

and that there exist sufficiently small constants $c_0 > 0$ *and* $c_1 > 0$ *such that*

$$\frac{2\sigma}{d}\left\{\sqrt{n} + \sqrt{k} + 2\sqrt{\log(p \vee m)} + \sqrt{k\sqrt{n\log(p \vee m)}}\right\} \leq c_0, \quad \sqrt{s}\lambda/d \leq c_1. \tag{12}$$

Then uniformly over $\Theta(s, k, r, d, \gamma)$ *in (6), with probability at least* $1 - 3(p \vee m)^{-1}$*, the output* $\widehat{A}$ *of Algorithm 1 satisfies*

$$L_q(A, \widehat{A}) \leq C\sigma^2 r^{2/q-1}(k + s)(r + \log(p \vee m)), \qquad \text{for all } q \in [1, 2]$$

where C *is a constant depending only on* $\kappa_\pm(s_*), \gamma, c_*, c_0$ *and* c_1*.*

When we specialize to the case of simultaneously low-rank and row-sparse setting, condition (12) is stronger than some related condition in the literature, e.g. that in Bunea et al. (2012), for establishing minimax rates. However, we deal with the theoretical guarantee of an actual estimator that we compute by Algorithm 1 while Bunea et al. (2012) was concerned with

the global optimum of a non-convex program which is not always attainable by heuristic algorithms. So they are not directly comparable.

4.2 Minimax Lower Bounds

To assess the tightness of the error bounds in Theorem 1, we now provide lower bounds on minimax risk for estimating A under loss functions in (7).

Theorem 2. *Let the observed X, Y be generated by (1) with Z having i.i.d. $N(0, \sigma^2)$ entries. Suppose that the coefficient matrix $A \in \Theta(s, k, r, d, \gamma)$ for some $k \geq 2r$ and $s \geq 2r$ and that the $(2s)$-sparse Riesz constants of the design matrix X satisfy $K^{-1} \leq \kappa_-(2s) \leq \kappa_+(2s) \leq K$ for some absolute constant $K > 1$. Then there exists a positive constant c depending only on γ and $\kappa_+(2s)$ such that the minimax risk for estimating A satisfies*

$$\inf_{\widehat{A}} \sup_{\Theta} \mathbb{E} L_q(A, \widehat{A}) \geq c\sigma^2 \left\{ \left(r^{2/q-1} \frac{d^2}{\sigma^2} \right) \wedge \left[r^{2/q}(s+k) + r^{2/q-1} \left(s \log \frac{ep}{s} + k \log \frac{em}{k} \right) \right] \right\}, \tag{13}$$

for all $q \in [1, 2]$.

Remark 1. Comparing Theorem 1 and Theorem 2, we find that they match up to a multiplicative log factor in general and up to a constant multiplier when r is no smaller than $\log(p \vee m)$ in order of magnitude. Moreover, Theorem 1 imposes an additional condition on the minimum singular value

of A in (12). Therefore, under the conditions of Theorem 1, Algorithm 1 attains nearly optimal convergence rates adaptively for all losses in (7).

As we have mentioned earlier, the one-way sparse reduced rank regression model considered in the literature, such as Chen and Huang (2012), Bunea et al. (2012), She (2014) and Ma and Sun (2014), does not consider column sparsity in A and can be viewed as a special case of model (1) with $k = m$. In view of the foregoing discussion, our estimator is also adaptive to this special case while retaining the ability of fully exploiting potential column sparsity.

5. Conclusion and Discussion

In this paper, we have proposed a new Double Projected Penalization (DPP) estimator for the coefficient matrix in two-way sparse reduced-rank regression. The model is well motivated by massive datasets arising in a number of application fields, especially genomics and neuroimaging. The proposed estimator is fast to compute and demonstrates competitive performance when compared with existing methods in simulation studies. In addition, we have illustrated its potential use in neuroscience by applying it to the analysis of a calcium imaging dataset and it compared favorably with some state-of-the-art method. Last but not least, we have further justified its nice empirical performance by a decision-theoretic analysis when

the data is Gaussian.

In terms of the DPP estimator, an interesting problem to be studied in future is to establish high probability error bounds when the data is not Gaussian. Since one cannot easily generate independent samples in such cases, we anticipate that different proof techniques will be needed to achieve this goal. In addition, it is worth noting that steps 3–4 of Algorithm 1 can be iterated till certain convergence criterion is met. Thus, we could also define an iterative projected penalization estimator. However, based on simulation results not reported here, we did not find significant performance gain by employing such an iterative scheme, which is more costly in terms of computation. Furthermore, it is of interest to investigate what one would be able to achieve in the low signal-to-noise ratio scenario when (12) fails to hold.

Another potential direction for future research is to consider certain nonlinear extensions of the model. When the response is univariate, researchers have considered sparse sliced inverse regression (Li and Nachtsheim, 2012; Lin et al., 2015). It would be of great interest to conduct analogous investigations for multiple responses where both low-rankness and sparsity are involved.

Supplementary Document

The supplementary document provides all technical proofs.

References

Akerboom, J., T.-W. Chen, T. J. Wardill, L. Tian, J. S. Marvin, S. Mutlu, N. C. Calderón, F. Esposti, B. G. Borghuis, X. R. Sun, et al. (2012). Optimization of a gcamp calcium indicator for neural activity imaging. *The Journal of Neuroscience 32*(40), 13819–13840.

Bakin, S. (1999). *Adaptive regression and model selection in data mining problems*. Ph. D. thesis, Australian National University, Canberra.

Birgé, L. (1983). Approximation dans les espaces métriques et théorie de l'estimation. *Z. für Wahrscheinlichkeitstheorie und Verw. Geb. 65*(2), 181–237.

Bunea, F., Y. She, and M. Wegkamp (2011). Optimal selection of reduced rank estimators of high-dimensional matrices. *The Annals of Statistics 39*(2), 1282–1309.

Bunea, F., Y. She, and M. Wegkamp (2012). Joint variable and rank selection for parsimonious estimation of high dimensional matrices. *The Annals of Statistics 40*(5), 2359–2763.

Candes, E. and T. Tao (2007). The Dantzig selector: statistical estimation when p is much larger than n. *The Annals of Statistics*, 2313–2351.

Chen, K., K.-S. Chan, and N. C. Stenseth (2012). Reduced rank stochastic regression with a sparse singular value decomposition. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 74*(2), 203–221.

Chen, L. and J. Huang (2012). Sparse reduced-rank regression for simultaneous dimension reduction and variable selection in multivariate regression. *Journal of the American Statistical Association 107*(500), 1533–1545.

Chen, S. S., D. L. Donoho, and M. A. Saunders (1998). Atomic decomposition by basis pursuit. *SIAM journal on scientific computing 20*(1), 33–61.

Chen, T.-W., T. J. Wardill, Y. Sun, S. R. Pulver, S. L. Renninger, A. Baohan, E. R. Schreiter, R. A. Kerr, M. B. Orger, V. Jayaraman, et al. (2013). Ultrasensitive fluorescent proteins for imaging neuronal activity. *Nature 499*(7458), 295–300.

Davidson, K. and S. Szarek (2001). *Handbook on the Geometry of Banach Spaces*, Volume 1, Chapter Local operator theory, random matrices and Banach spaces, pp. 317–366. Elsevier Science.

Fan, J. and R. Li (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Am. Stat. Assoc. 96*, 1348–1360.

Giovannucci, A., J. Friedrich, B. Deverett, V. Staneva, D. Chklovskii, and E. Pnevmatikakis (2017). CaImAn: An open source toolbox for large scale calcium imaging data analysis on standalone machines. *Cosyne Abstracts*.

Haeffele, B., E. Young, and R. Vidal (2014). Structured low-rank matrix factorization: Optimality, algorithm, and applications to image processing. In *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*, pp. 2007–2015.

Huang, J. and T. Zhang (2010). The benefit of group sparsity. *The Annals of Statistics 38*(4), 1978–2004.

Ibragimov, I. and R. Has'minskii (1981). *Statistical Estimation: Asymptotic Theory*. Springer.

Izenman, A. (1975). Reduced-rank regression for the multivariate linear model. *Journal of Multivariate Analysis 5*(2), 248–264.

Koltchinskii, V., K. Lounici, and A. Tsybakov (2011). Nuclear-norm penalization and optimal rates for noisy low-rank matrix completion. *The Annals of Statistics 39*(5), 2302–2329.

Laurent, B. and P. Massart (2000). Adaptive estimation of a quadratic functional by model selection. *The annals of Statistics 28*(5), 1302–1338.

Lee, M., H. Shen, J. Huang, and J. Marron (2010). Biclustering via sparse singular value decomposition. *Biometrics 66*, 1087–1095.

Li, L. and C. J. Nachtsheim (2012). Sparse sliced inverse regression. *Technometrics*.

Lin, Q., Z. Zhao, and J. S. Liu (2015). On consistency and sparsity for sliced inverse regression in high dimensions. *arXiv preprint arXiv:1507.03895*.

Lounici, K., M. Pontil, S. Van De Geer, and A. B. Tsybakov (2011). Oracle inequalities and optimal inference under group sparsity. *The Annals of Statistics 39*(4), 2164–2204.

Ma, X., L. Xiao, and W. H. Wong (2014). Learning regulatory programs by threshold svd regression. *Proceedings of the National Academy of Sciences 111*(44), 15675–15680.

Ma, Z. and T. Sun (2014). Adaptive sparse reduced-rank regression. *arXiv preprint arXiv:1403.1922v1*.

Ma, Z. and Y. Wu (2015). Volume ratio, sparsity, and minimaxity under unitarily invariant norms. *IEEE Transactions on Information Theory 61*(12), 6939–6956.

Muirhead, R. (1982). *Aspects of Multivariate Statistical Theory*. John Wiley and Sons.

Pnevmatikakis, E. A., Y. Gao, D. Soudry, D. Pfau, C. Lacefield, K. Poskanzer, R. Bruno, R. Yuste, and L. Paninski (2014). A structured matrix factorization framework for large scale calcium imaging data analysis. *arXiv preprint arXiv:1409.2903*.

Reinsel, G. and R. Velu (1998). *Multivariate reduced-rank regression: Theory and applications*. New York: Springer.

Rigollet, P. and A. Tsybakov (2011). Exponential Screening and optimal rates of sparse estimation. *The Annals of Statistics 39*(2), 731–771.

She, Y. (2014). Selectable factor extraction in high dimensions. *arXiv preprint arXiv:1403.6212*.

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

Tsybakov, A. (2009). *Introduction to Nonparametric Estimation*. Springer Verlag.

Vogelstein, J. T., A. M. Packer, T. A. Machado, T. Sippy, B. Babadi, R. Yuste, and L. Paninski (2010). Fast nonnegative deconvolution for

spike train inference from population calcium imaging. *Journal of neurophysiology 104*(6), 3691–3704.

Vounou, M., E. Janousova, R. Wolz, J. Stein, P. Thompson, D. Rueckert, and G. Montana (2012). Sparse reduced-rank regression detects genetic associations with voxel-wise longitudinal phenotypes in alzheimer's disease. *Neuroimage 60*(1), 700–716.

Wedin, P.-A. (1972). Perturbation bounds in connection with singular value decomposition. *BIT 12*, 99–111.

Yang, D., Z. Ma, and A. Buja (2014). A sparse Singular Value Decomposition method for high-dimensional data. *Journal of Computational and Graphical Statistics 23*(4), 923–942.

Yang, D., Z. Ma, and A. Buja (2016). Rate optimal denoising of simultaneously sparse and low rank matrices. *Journal of Machine Learning Research 17*(92), 1–27.

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

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

Supplementary Document for Adaptive Estimation in Two-way Sparse Reduced-rank Regression

Zhuang Ma, Zongming Ma and Tingni Sun

University of Pennsylvania and University of Maryland

We provide the technical proofs for all theorems in this supplementary document.

6. Proofs

6.1 Proof of Theorem 1

We analyze each step of Algorithm 1 to prove Theorem 1. Throughout the proof, some useful lemmas on tail probabilities will be stated without proof.

Analysis of $V_{(0)}$. We first study the property of the right singular vector matrix $V^{(0)}$ obtained in the column-thresholding step of Stage I. For $0 < a_- < 1 < a_+$, define

$$J_{(0)}^{\pm} = \Big\{ j : \|XA_{*j}\|^2 \geq \widetilde{\sigma}^2 a_{\mp} \alpha \sqrt{n \log(p \vee m)} \Big\}.$$

More specifically, let $a_- = 0.1$ and $a_+ = 2$ in the proof. Recall that $\alpha = \sqrt{12}$ and $\widetilde{\sigma} = 2\sigma$.

Lemma 1. *[Stage I column selection] With probabbility at least* $1 - 4(p \vee m)^{-2}$,

$$J_{(0)}^- \subset J_{(0)} \subset J_{(0)}^+$$

Proof of Lemma 1. Due to Gaussianity, $\|Y_{*j}^{(0)}\|^2/\widetilde{\sigma}^2$ follows a non-central χ^2 distribution with n degrees of freedom and noncentrality parameter $\|XA_{*j}\|^2/\widetilde{\sigma}^2$. By Lemma 2,

$$\begin{aligned}
P(J_{(0)}^- \not\subset J_{(0)}) &\leq \sum_{j \in J_{(0)}^-} P\Big\{\|Y_{*j}^{(0)}\|^2 < \widetilde{\sigma}^2(n + \alpha\sqrt{n\log(p \vee m)})\Big\} \\
&\leq mP\Big\{\|Y_{*j}^{(0)}\|^2 < \widetilde{\sigma}^2 n + \|XA_{*j}\|^2 - \widetilde{\sigma}^2\alpha(a_+ - 1)\sqrt{n\log(p \vee m)} \;\Big|\; j \in J_{(0)}^-\Big\} \\
&\leq 2m\exp\Big(-\frac{\alpha^2(a_+ - 1)^2 n\log(p \vee m)}{4(\sqrt{n} + (a_+\alpha)^{1/2}(n\log(p \vee m))^{1/4})^2}\Big) \\
&\leq 2(p \vee m)^{-2}.
\end{aligned}$$

Similarly, it is proved that $J_{(0)} \subset J_{(0)}^+$ holds with probability at least $1 - 2(p \vee m)^{-2}$.

□

Lemma 2. *Let* X *follow a non-central chi-square distribution* $\chi_n^2(\lambda)$ *with*

n degrees of freedom and non-centrality parameter λ. Then

$$P\Big\{X \geq (n+\lambda) + 2(\sqrt{n}+\sqrt{\lambda})s\Big\} \leq \Big(1+\frac{1}{\sqrt{2}s}\Big)\exp(-s^2), \quad \textit{if } 0 \leq s \leq \frac{1}{2}n^{9/16},$$
$$P\Big\{X \leq (n+\lambda) - 2(\sqrt{n}+\sqrt{\lambda})s\Big\} \leq 2\exp(-s^2), \quad \textit{if } 0 \leq s \leq \frac{1}{2}n^{1/2}.$$

Lemma 3. *Let X be an $n \times m$ matrix with iid standard Gaussian entries. Then for any $t > 0$,*

$$P\Big\{\|X\| > \sqrt{n}+\sqrt{m}+t\Big\} \leq \exp(-t^2/2).$$

Lemma 4. *[Stage I subspace estimation] With probability at least $1-3(p\vee m)^{-2}$,*

$$\begin{aligned}
\|VV' - V_{(0)}V_{(0)}'\| &\leq \frac{C_1\widetilde{\sigma}}{d}\Big\{\sqrt{n}+\sqrt{k}+2\sqrt{\log(p\vee m)}+\sqrt{k\sqrt{n\log(p\vee m)}}\Big\},\\
\|VV' - V_{(0)}V_{(0)}'\|_F &\leq \frac{C_2\widetilde{\sigma}}{d}\Big\{\sqrt{r}(\sqrt{n}+\sqrt{k}+2\sqrt{\log(p\vee m)})+\sqrt{k\sqrt{n\log(p\vee m)}}\Big\}.
\end{aligned}$$

Proof of Lemma 4. We study the upper bounds in the event where $J_{(0)}^- \subset J_{(0)} \subset J_{(0)}^+$ holds. We may reorder the columns of matrices such that $XA - \widetilde{Y}^{(0)}$ is of the following form

$$XA - \widetilde{Y}^{(0)} = \begin{pmatrix} -Z_{*J_{(0)}} & UDV'_{*J_{(0)}^c} \end{pmatrix}$$

Lemma 3 provides an upper bound for $\|Z_{*J_{(0)}}\|$ as follows

$$\|Z_{*J_{(0)}}\| \le \widetilde{\sigma}(\sqrt{n} + \sqrt{J_{(0)}} + 2\sqrt{\log(p \vee m)}) \le \widetilde{\sigma}(\sqrt{n} + \sqrt{k} + 2\sqrt{\log(p \vee m)})$$

with probability at least $1 - (p \vee m)^2$, since $|J_{(0)}| \le |J^+_{(0)}| = k$. Moreover, it holds that, in the event of $J^-_{(0)} \subset J_{(0)}$,

$$\|U\Delta V'_{*J^c_{(0)}}\|^2 \le \|\Delta V'_{*(J^-_{(0)})^c}\|_F^2 \le \widetilde{\sigma}^2 a_- \alpha k \sqrt{n \log(p \vee m)}.$$

Thus, we have

$$\|XA - \widetilde{Y}^{(0)}\| \le \widetilde{\sigma}(\sqrt{n} + \sqrt{k} + 2\sqrt{\log(p \vee m)}) + \widetilde{\sigma}\sqrt{a_- \alpha k \sqrt{n \log(p \vee m)}}$$

and the desired results then follows from the $\sin\theta$ theorem. □

Analysis of $U_{(1)}$.

Lemma 5. *[Stage I Regression] Under the condition of Theorem 1, there exists a constant C depending only on $\kappa_\pm(s_*), c_*$ and c_0, such that with probability at least $1 - (p \vee m)^{-1}$,*

$$\|U_{(1)}U'_{(1)} - UU'\|_F \le C\sqrt{s}\lambda/d.$$

Proof of Lemma 5. Let $U_* \in \mathbb{R}^{n\times r}$ be the left singular vector matrix of $XAV_{(0)} = UDV'V_{(0)}$. Under condition (12), $V'V_{(0)}$ is an $r \times r$ matrix of full rank, and so the column space of U^* is the same as the column space of U; i.e., $U_*U_*' = UU'$. By Wedin's $\sin\theta$ Theorem (Wedin, 1972),

$$\|U_{(1)}U_{(1)}' - UU'\|_F = \|U_{(1)}U_{(1)}' - U_*U_*'\|_F \le \frac{\|XB_{(1)} - XAV_{(0)}\|_F}{\sigma_r(XAV_{(0)})},$$

where $\sigma_r(XAV_{(0)})$ is the r^{th} singular value of $XAV_{(0)}$.

Since for any unit vector x,

$$\begin{aligned}\|V'V_{(0)}x\|^2 &= x'V_{(0)}'VV'V_{(0)}x \\ &= 1 - x'V_{(0)}'(VV' - V_{(0)}V_{(0)}')V_{(0)}x \\ &\ge 1 - \|VV' - V_{(0)}V_{(0)}'\|.\end{aligned}$$

Thus, we have $\sigma_r^2(V'V_{(0)}) = \min_{\|x\|=1}\|V'V_{(0)}x\|^2 \ge 1 - \|VV' - V_{(0)}V_{(0)}'\|$. When c_0 is small enough, $\|VV' - V_{(0)}V_{(0)}'\|$ is sufficiently small by Lemma 4. So there exists a constant c' such that $\sigma_r(V'V_{(0)}) > c'$. Note that $XAV_{(0)} = XAVV'V_{(0)}$, and so

$$\sigma_r(XAV_{(0)}) \ge \sigma_r(XAV)\sigma_r(V'V_{(0)}) \ge \delta_r c',$$

where the last inequality holds under condition (12) since $\sigma_r(XAV) =$

$\sigma_r(XA) = \delta_r$. Further note that

$$\|XB_{(1)} - XAV_{(0)}\|_F \le \kappa_+(2s)\|B_{(1)} - AV_{(0)}\|_F \le \kappa_+(s_*)\|B_{(1)} - AV_{(0)}\|_F$$

and that $\delta_r \ge \kappa_-(s)\sigma_r(A) \ge \kappa_-(s_*)d$, the desired result then follows from Part (ii) of Theorem 3 with $\eta = 1/(p \vee m)$. □

Analysis of $V_{(1)}$. Recall

$$J_{(1)} = J_{(0)} \cup \Big\{ j : \|U_{(1)}{}' Y^{(2)}_{*j}\|^2 \ge \beta\widetilde{\sigma}^2(r + 2\sqrt{3r\log(p \vee m)} + 6\log(p \vee m)) \Big\}.$$

For $b_- < b_+$, define

$$J^{\pm}_{(1)} = \Big\{ j : \|XA_{*j}\|^2 \ge \widetilde{\sigma}^2 b_{\mp}(r + 2\sqrt{3r\log(p \vee m)} + 6\log(p \vee m)) \Big\}.$$

More specifically, let $b_+ = 4.5$ and $b_- = 0.002$ in the proof. Recall that $\beta = 1.1$.

Lemma 6. *Let X follow a chi-square distribution χ^2_n with n degrees of freedom. Then for any $t > 0$*

$$P(X > n + 2\sqrt{n}t + 2t^2) < \exp(-t^2)$$

Lemma 7. *[Stage II column selection] Assume* $\|U_{(1)}U'_{(1)} - UU'\| < c$ *for some small positive constant* $c < 0.05$. *With probabbility at least* $1 - 2(p \vee m)^{-2}$,

$$J^-_{(1)} \subset J_{(1)} \subset J^+_{(1)}$$

Proof of Lemma 7. For $j \in J^-_{(1)} \setminus J_{(0)}$,

$$\begin{aligned} \|U'_{(1)}Y^{(2)}_{*j}\| &= \|U'_{(1)}(UDV'_{*j} + Z^{(2)}_{*j})\| \\ &\geq \|U'_{(1)}UDV'_{*j}\| - \|U'_{(1)}Z^{(2)}_{*j}\| \end{aligned}$$

The first term is

$$\begin{aligned} \|U'_{(1)}UDV'_{*j}\|^2 &\geq \|XA_{*j}\|^2(1 - \|U_{(1)}U'_{(1)} - UU'\|) \geq \|XA_{*j}\|^2(1-c) \\ &\geq \widetilde{\sigma}^2(1-c)b_+(r + 2\sqrt{3r\log(p \vee m)} + 6\log(p \vee m)) \end{aligned}$$

Since $U'_{(1)}Z^{(2)}_{*j} \sim N(0, \widetilde{\sigma}^2 I_r)$, it follows from Lemma 6 that

$$\|U'_{(1)}Z^{(2)}_{*j}\|^2 \leq \widetilde{\sigma}^2(r + 2\sqrt{3r\log(p \vee m)} + 6\log(p \vee m)),$$

with probability at least $1-(p\vee m)^{-3}$. Thus, in the same event, we have

$$\begin{aligned}\|U'_{(1)}Y^{(2)}_{*j}\| &\geq (\sqrt{(1-c)b_+}-1)\widetilde{\sigma}\Big\{r+2\sqrt{3r\log(p\vee m)}+6\log(p\vee m)\Big\}^{1/2}\\ &\geq \beta^{1/2}\widetilde{\sigma}(r+2\sqrt{3r\log(p\vee m)}+6\log(p\vee m))^{1/2},\end{aligned}$$

due to $(\sqrt{(1-c)b_+}-1)^2>(\sqrt{0.95\times 4.5}-1)^2>1.1=\beta$. Hence, we have $j\in J_{(1)}$. So it holds that $J^-_{(1)}\subset J_{(1)}$ with probability at least $1-(p\vee m)^{-2}$. Similarly, we have $J_{(1)}\subset J^+_{(1)}$ with probability at least $1-(p\vee m)^{-2}$, due to $(\sqrt{(1+c)b_-}+1)^2<1.1=\beta$. □

Lemma 8. *[Stage II subspace estimation] Suppose $\|U_{(1)}U'_{(1)}-UU'\|_F<c'_1$ for a sufficiently small positive constant c'_1. Then there exists a constant C depending only on $\kappa_\pm(s_*),\gamma$ and c'_1 such that with probability at least $1-(p\vee m)^{-1}$,*

$$\|V_{(1)}V'_{(1)}-VV'\|_F\leq C\sigma\sqrt{(k+s)(r+\log(p\vee m))}/d$$

Proof of Lemma 8.

$$\|V_{(1)}V'_{(1)}-VV'\|_F\leq\frac{\|U_{(1)}U'_{(1)}\widetilde{Y}^{(1)}-U_{(1)}U'_{(1)}XA\|_F}{\sigma_r(U_{(1)}U'_{(1)}XA)}. \tag{14}$$

We first upper bound the numerator

$$\begin{aligned}
& \|U_{(1)}U'_{(1)}\widetilde{Y}^{(1)} - U_{(1)}U'_{(1)}XA\|_F \\
\leq\ & \|U'_{(1)}(\widetilde{Y}^{(1)}_{*J_{(1)}} - XA_{*J_{(1)}})\|_F + \|U_{(1)}U'_{(1)}XA_{*J^c_{(1)}}\|_F \\
\leq\ & \|U'_{(1)}(\widetilde{Y}^{(1)}_{*J^{(1)}} - XA_{*J_{(1)}})\|_F + \|(U_{(1)}U'_{(1)} - UU')XA_{*J^c_{(1)}}\|_F + \|UU'XA_{*(J^-_{(1)})^c}\|_F \\
\leq\ & \widetilde{\sigma}(\sqrt{rk} + \sqrt{\log(p\vee m)}) + d\|(U_{(1)}U'_{(1)} - UU'\| + \widetilde{\sigma}\sqrt{k}\sqrt{b_+(r + 2\sqrt{3r\log(p\vee m)} + 6\log(p\vee m))} \\
\leq\ & C\sigma\sqrt{(k+s)(r+\log(p\vee m))} \qquad (15)
\end{aligned}$$

To lower bound the denominator, we apply Weyl's theorem to obtain

$$\begin{aligned}
\sigma_r(U_{(1)}U'_{(1)}XA) &\geq \sigma_r(UU'XA) - \|U_{(1)}U'_{(1)}XA - UU'XA\|_{\text{op}} \\
&\geq \delta_r - \|U_{(1)}U'_{(1)} - UU'\|_{\text{op}}\|XA\|_{\text{op}}.
\end{aligned}$$

Note that $\delta_r \geq \kappa_-(s_*)d$, $\|XA\|_{\text{op}} \leq \kappa_+(s_*)\gamma d$ and that $\|U_{(1)}U'_{(1)} - UU'\|_{\text{op}} \leq \|U_{(1)}U'_{(1)} - UU'\|_F \leq c'_1$. Thus, for sufficiently small value of c'_1, we obtain

$$\sigma_r(U_{(1)}U'_{(1)}XA) \geq C^{-1}d, \qquad (16)$$

where $C > 0$ is a constant depending only on $\kappa_\pm(s_*), \gamma$ and c'_1. Combining (14) – (16), we complete the proof. □

Proof of Theorem 1.

Proof. By the definition of $\widehat{A}$, we have

$$\begin{aligned}\|\widehat{A}-A\|_F &= \|B_{(2)}V_{(1)}'-AVV'\|_F \\ &\le \|B_{(2)}V_{(1)}'-AV_{(1)}V_{(1)}'\|_F+\|AV_{(1)}V_{(1)}'-AVV'\|_F \\ &\le \|V_{(1)}\|_{\mathrm{op}}\|B_{(2)}-AV_{(1)}\|_F+\|A\|_{\mathrm{op}}\|V_{(1)}V_{(1)}'-VV'\|_F.\end{aligned}$$

Assembling the bounds in all lemmas,

$$\|\widehat{A}-A\|_F^2 \lesssim \sigma^2(k+s)(r+\log(p\vee m)) \tag{17}$$

The desired upper bound on other Schatten norm losses is a consequence of (17) and the inequality $\|\widehat{A}-A\|_{\mathrm{s}_q}^2 \le (2r)^{2/q-1}\|\widehat{A}-A\|_F^2$ for all $q\in[1,2]$.

□

6.2 Proof of Theorem 2

For any probability distributions P and Q, let $D(P||Q)$ denote the Kullback–Leibler divergence of Q from P. For any subset K of $\mathbb{R}^{m\times n}$, the volume of K is $\mathrm{vol}(K)=\int_K \mathrm{d}\mu$ where $\mathrm{d}\mu$ is the usual Lebesgue measure on $\mathbb{R}^{m\times n}$ by taking the product measure of the Lebesgue measures of individual entries. With these definitions, we state the following variant of Fano's lemma (Ibragimov and Has'minskii, 1981; Birgé, 1983; Tsybakov, 2009). This version has been established as Proposition 1 in Ma and Wu (2015). It will be

used repeatedly in the proof of the lower bounds. Throughout the proof, we denote $\kappa_+(2s)$ by κ_+.

Proposition 1. *Let (Θ, ρ) be a metric space and $\{P_\theta : \theta \in \Theta\}$ a collection of probability measures. For any totally bounded $T \subset \Theta$, denote by $\mathcal{M}(T, \rho, \epsilon)$ the ϵ-packing number of T with respect to ρ, i.e., the maximal number of points in T whose pairwise minimum distance in ρ is at least ϵ. Define the* Kullback-Leibler diameter *of T by*

$$d_{\mathrm{KL}}(T) \triangleq \sup_{\theta,\theta' \in T} D(P_\theta \,||\, P_{\theta'}). \tag{18}$$

Then

$$\inf_{\hat{\theta}} \sup_{\theta \in \Theta} \mathbb{E}_\theta[\rho^2(\hat{\theta}(X), \theta)] \geq \sup_{T \subset \Theta} \sup_{\epsilon > 0} \frac{\epsilon^2}{4} \left(1 - \frac{d_{\mathrm{KL}}(T) + \log 2}{\log \mathcal{M}(T, \rho, \epsilon)}\right). \tag{19}$$

In particular, if $\Theta \subset \mathbb{R}^d$ and $\|\cdot\|$ is some norm on $\mathbb{R}^d$, then

$$\inf_{\hat{\theta}} \sup_{\theta \in \Theta} \mathbb{E}_\theta[\|\hat{\theta}(X) - \theta\|^2] \geq \sup_{T \subset \Theta} \sup_{\epsilon > 0} \frac{\epsilon^2}{4} \left(1 - \frac{d_{\mathrm{KL}}(T) + \log 2}{\log \frac{\mathrm{vol}(T)}{\mathrm{vol}(B_{\|\cdot\|}(\epsilon))}}\right). \tag{20}$$

We first prove an oracle version of the lower bound. One can think of it as an lower bound for the minimax risk when we know that the nonzero entries of the coefficient matrix $A \in \mathbb{R}^{p \times m}$ are restricted to the top–left $s \times r$ block (or the top left $r \times k$ block).

Lemma 9. *Let $\Theta_0(s, r, r, d, \gamma) \subset \Theta(s, k, r, d, \gamma)$ be the sub-collection of all matrices whose nonzero entries are in the top left $s \times r$ block. Suppose $\sigma = 1$. There exists a positive constant c that depends only on κ_+ and γ, such that for any $q \in [1, 2]$, the minimax risk for estimating A over Θ_0 satisfies*

$$\inf_{\widehat{A}} \sup_{\Theta_0} \mathbb{E} L_q(A, \widehat{A}) \geq c \left[(r^{2/q-1} d^2) \wedge (r^{2/q} s) \right].$$

Similarly, let $\Theta_0'(r, k, r, d, \gamma) \subset \Theta(s, k, r, d, \gamma)$ be the sub-collection of all matrices whose nonzero entries are in the top left $r \times k$ block. Under the same conditions, we have

$$\inf_{\widehat{A}} \sup_{\Theta_0'} \mathbb{E} L_q(A, \widehat{A}) \geq c \left[(r^{2/q-1} d^2) \wedge (r^{2/q} k) \right].$$

Proof. In what follows, we focus on proving the first claim and the second claim follows from essentially the same argument.

By a simple sufficiency argument, we can reduce to model (1) with $p = s$ and $m = r$, which we assume in the rest of this proof without loss of generality.

Let $A_0 = \mathrm{diag}(1, \ldots, 1) \in \mathbb{R}^{s \times r}$. Moreover, for any δ and any $q \in [1, 2]$, let $B_{\mathrm{S}_q}(\delta) = \{A \in \mathbb{R}^{s \times r} : \|A\|_{\mathrm{s}_q} \leq \delta\}$ denote the Schatten-q ball with radius

δ in $\mathbb{R}^{s\times r}$. For some constant $a>0$ to be specified later, define

$$T(a)=\frac{\gamma d}{2}A_0+B_{\mathrm{S}_2}(\sqrt{a})=\left\{\frac{\gamma d}{2}A_0+M: M\in B_{\mathrm{S}_2}(\sqrt{a})\right\}. \tag{21}$$

For any $A_1, A_2\in T(a)$, we have

$$D(P_{A_1}||P_{A_2})=\frac{1}{2}\left\|XA_1-XA_2\right\|_{\mathrm{S}_2}^2\le\frac{1}{2}\|X\|_{\mathrm{op}}^2\left\|A_1-A_2\right\|_{\mathrm{S}_2}^2\le 2\kappa_+^2 a.$$

Here, the last inequality holds since $\|X\|_{\mathrm{op}}\le\kappa_+$ under the assumption that $X\in\mathbb{R}^{s\times r}$ and $\left\|A_1-A_2\right\|_{\mathrm{S}_2}^2\le 4a$ by definition (21). So

$$d_{\mathrm{KL}}(T(a))\le 2\kappa_+^2 a. \tag{22}$$

By the inverse Santalo's inequality (see, e.g., Lemma 3 of Ma and Wu (2015)), for some universal constants c_0,

$$\begin{aligned}\mathrm{vol}(T(a))^{\frac{1}{sr}}&=\mathrm{vol}(B_{\mathrm{S}_2}(\sqrt{a}))^{\frac{1}{sr}}=\sqrt{a}\cdot\mathrm{vol}(B_{\mathrm{S}_2}(1))^{\frac{1}{sr}}\\&\ge\sqrt{a}\cdot\frac{c_0}{\mathbb{E}\left\|Z\right\|_{\mathrm{S}_2}} \qquad (23)\\&\ge\sqrt{a}\cdot\frac{c_0'}{\sqrt{sr}}. \qquad (24)\end{aligned}$$

In (23), Z is a $s\times r$ matrix with i.i.d. $N(0,1)$ entries. The inequality in (24) holds since by Jensen's inequality, $\mathbb{E}\left\|Z\right\|_{\mathrm{S}_2}\le\sqrt{\mathbb{E}\left\|Z\right\|_{\mathrm{S}_2}^2}=\sqrt{sr}$.

On the other hand, by Urysohn's inequality (see, e.g., Eq.(19) of Ma and Wu (2015)), for any $\epsilon > 0$ and $q \in [1, 2]$,

$$\mathrm{vol}(B_{\mathrm{S}_q}(\epsilon))^{\frac{1}{sr}} \le \frac{\epsilon \mathbb{E}\,\|Z\|_{\mathrm{S}_{q'}}}{\sqrt{sr}} \le \frac{\epsilon r^{\frac{1}{q'}} \mathbb{E}\|Z\|_{\mathrm{op}}}{\sqrt{sr}} \le 2\epsilon r^{\frac{1}{2}-\frac{1}{q}}.$$

Here, $\frac{1}{q'} + \frac{1}{q} = 1$ and Z is a $s \times r$ matrix with i.i.d. $N(0, 1)$ entries. The last inequality is due to Gordon's inequality (see, e.g., Davidson and Szarek (2001)): $\mathbb{E}\|Z\|_{\mathrm{op}} \le \sqrt{s} + \sqrt{r} \le 2\sqrt{s}$.

Now let

$$a = \left(\frac{\gamma \wedge 2 - 1}{2}\right)^2 \left(sr \wedge d^2\right), \quad \text{and} \quad \epsilon = \frac{c_0'}{2\kappa_+} \sqrt{a}\, r^{\frac{1}{q}-\frac{1}{2}}. \tag{25}$$

Then for any $A \in T(a)$ and any $i \in [r]$, $|\sigma_i(A) - \frac{\gamma}{2}d| \le \sqrt{a} \le \frac{\gamma \wedge 2 - 1}{2} d$, and so $\sigma_i(A) \in [d, \gamma d]$ and $T(a) \subset \Theta_0(s, r, d, \gamma)$. Applying Proposition 1 with $T(a)$ and ϵ in (21) and (25), we obtain a lower bound on the order of ϵ^2. This completes the proof. □

Lemma 10. *Let $s \ge r$ be positive integers. There exist a matrix $W \in \mathbb{R}^{s \times r}$ and two absolute constants $c_0 \in (\frac{1}{2}, 1)$ and $c_1 > 0$ such that $\|W\|_{\mathrm{F}} \le 1$ and for any subset $B \subset [s]$ such that $|B| \ge c_0 s$, $\|W_{B*}\|_{\mathrm{s}_q} \ge c_1 r^{\frac{1}{q}-\frac{1}{2}}$ for any $q \in [1, 2]$.*

Proof. We divide the proof into two cases, namely when $s \ge 25$ and when

$s < 25$.

1° When $s \geq 25$, let $Z \in \mathbb{R}^{s\times r}$ have i.i.d. $N(0,1)$ entries. Then $\|Z\|_{\mathrm{F}}^2 \sim \chi_{sr}^2$, and Laurent and Massart (2000, Eq.(4.3)) implies that

$$\mathbb{P}\left\{\|Z\|_{\mathrm{F}}^2 \geq sr + 2s\sqrt{r} + 2s\right\} \leq \mathrm{e}^{-s}.$$

Moreover, for any $c_0 > \frac{1}{2}$,

$$\begin{aligned}
&\mathbb{P}\left\{\exists B \subset [s], \text{ s.t. } |B| = c_0 s \text{ and } \sigma_r(Z_{B*}) < \sqrt{c_0 s} - \sqrt{r} - \frac{1}{2}\sqrt{c_0 s}\right\} \\
&\leq \sum_{B\subset[s],|B|=c_0 s} \mathbb{P}\left\{\sigma_r(Z_{B*}) < \sqrt{c_0 s} - \sqrt{r} - \frac{1}{2}\sqrt{c_0 s}\right\} \\
&\leq \binom{s}{(1-c_0)s} \mathrm{e}^{-c_0 s/4} \\
&\leq \exp\left\{-s\left[\frac{c_0}{4} + (1-c_0)\log(1-c_0)\right]\right\}.
\end{aligned}$$

Here, the first inequality is due to the union bound, the second inequality is due to the Davidson-Szarek bound, and the last inequality holds since for any $\alpha \in (\frac{1}{2}, 1)$, $\binom{s}{\alpha s} = \binom{s}{(1-\alpha)s} \leq (\frac{\mathrm{e}}{1-\alpha})^{(1-\alpha)s}$. If we set $c_0 \geq 0.96$, then the multiplier $\frac{c_0}{4} + (1-c_0)\log(1-c_0) \geq 0.1$.

So when $c_0 = 0.96$ and $s \geq 25$, the sum of the right hand sides of the last two displays is less than 1. Thus, there exists a deterministic matrix Z_0 on which both events happen. Now define $W = Z_0/\|Z_0\|_{\mathrm{F}}$. Then $\|W\|_{\mathrm{F}} = 1$

by definition, and for any $B \subset [s]$ with $|B| = c_0 s$,

$$\begin{aligned}
\|W_{B*}\|_{s_q} &\geq r^{1/q} \sigma_r(W_{B*}) \\
&= r^{1/q} \sigma_r((Z_0)_{B*}) / \|Z_0\|_{\mathrm{F}} \\
&\geq r^{1/q} \frac{\frac{1}{2}\sqrt{c_0 s} - \sqrt{r}}{\sqrt{sr + 2s\sqrt{r} + 2r}} \\
&\geq c_1 r^{1/q - 1/2}.
\end{aligned}$$

Note that the last inequality holds with an absolute constant c_1 when $r \leq \frac{1}{8} c_0 s$. When $r > \frac{1}{8} c_0 s$, we can always let $\tilde{r} = \frac{1}{8} c_0 r \leq \frac{1}{8} c_0 s$ and repeat the above arguments on the $s \times \tilde{r}$ submatrix of Z consisting of its first $\tilde{r}$ columns, and the conclusion continues to hold with a modified absolute constant c_1. This completes the proof for all subsets B with $|B| = c_0 s$. The claim continues to hold for all $|B| \geq c_0 s$ since the Schatten-q norm of a submatrix is always no smaller than the the whole matrix.

2° When $s < 25$, we have $r < 25$ since $r \leq s$ always holds. Let $W = \left[\frac{1}{\sqrt{s}} \mathbf{1}_s \quad \mathbf{0} \right] \in \mathbb{R}^{s \times r}$, i.e., the first column of W consists of s entries all equal to $1/\sqrt{s}$ and the rest are all zeros. So W is rank one. It is straightforward to verify the desired conclusion holds since for any $B \subset [s]$, $\|W_{B*}\|_{s_q} = \|W_{B*}\|_{\mathrm{F}} = \sqrt{|B|/s}$. This completes the proof. □

Lemma 11. *Let $a = d^2 \wedge s \log \frac{ep}{s}$. There exist three positive constants c_1, c_2, c_3 that depend only on γ and κ_+, and a subset $\Theta_1 \subset \Theta(s, k, r, d, \gamma)$,*

such that $c_3 \le c_2/3$, $d_{\mathrm{KL}}(\Theta_1) \le c_3 a$ *and that for any* $q \in [1,2]$,

$$\log \mathcal{M}(\Theta_1, \|\cdot\|_{\mathrm{s}_q}, c_1\sqrt{a}\, r^{1/q-1/2}) \ge c_2 s \log \frac{ep}{s},$$

where d_{KL} *is the Kullback–Leibler diameter and* $\mathcal{M}$ *is the packing number defined in Proposition 1.*

Similarly, for $b = d^2 \wedge k \log \frac{em}{k}$, *there is another subset* $\Theta' \subset \Theta(s,k,r,d,\gamma)$ *such that* $d_{\mathrm{KL}}(\Theta_1') \le c_3 b$ *and that for any* $q \in [1,2]$,

$$\log \mathcal{M}(\Theta_1', \|\cdot\|_{\mathrm{s}_q}, c_1\sqrt{b}\, r^{1/q-1/2}) \ge c_2 k \log \frac{em}{k}.$$

Proof. Let us focus on the first claim and we shall remark on how to establish the second claim at the end of this proof.

Let $W \in \mathbb{R}^{(s-r)\times r}$ satisfy the conclusion of Lemma 10 and define $s_0 = (1-c_0)(s-r)$. Let $\mathcal{B} = \{B_1, \ldots, B_N\}$ be a maximal set consisting of subsets of $[p]\backslash[r]$ with cardinality $s-r$ and for any $B_i \ne B_j$, $|B_i \cap B_j| \le s_0$. By Lemma A.3 of Rigollet and Tsybakov (2011) and Lemma 2.9 of Tsybakov (2009), there exists an absolute positive constant c_2' such that

$$\log N \ge c_2'(s-r) \log \frac{\mathrm{e}(p-r)}{s-r}.$$

Now for each $B_i \in \mathcal{B}$, define $W^{(i)} \in \mathbb{R}^{m\times n}$ by setting the submatrix $W^{(i)}_{B_i[r]} =$

W and filling the remaining entries with zeros. Then for any $i \neq j$, $|B_i \cap B_j| \leq s_0$, and so there exists a set $B_{ij} \subset [s]$ with $|B_{ij}| \geq s-r-s_0 = c_0(s-r)$, such that

$$\|W^{(i)} - W^{(j)}\|_{\mathrm{s}_q} \geq \|W_{B_{ij}*}\|_{\mathrm{s}_q} \geq c_1' r^{1/q-1/2},$$

where c_1' is an absolute constant due to Lemma 10.

Define $M_0 = \begin{bmatrix} I_r & 0 \\ 0 & 0 \end{bmatrix} \in \mathbb{R}^{p\times m}$ and for some positive constant $c_1'' \leq \frac{\gamma\wedge 2-1}{2} \wedge \sqrt{\frac{c_2'}{6\kappa_+^2}}$, let

$$\Theta_1 = \left\{ A^{(i)} = \frac{\gamma d}{2} M_0 + c_1'' \sqrt{a}\, W^{(i)} : i = 1, \ldots, N \right\}.$$

Note that each $A^{(i)}$ has s nonzero rows and r nonzero columns. Moreover, for $i \in [N]$, and $j \in [r]$

$$\left| \sigma_j(A^{(i)}) - \sigma_j(\frac{\gamma d}{2} M_0) \right| \leq \|A^{(i)} - \frac{\gamma d}{2} M_0\|_{\mathrm{op}} = c_1'' \sqrt{a} \|W^{(i)}\|_{\mathrm{op}} \leq c_1'' \sqrt{a} \|W^{(i)}\|_{\mathrm{F}} \leq \frac{\gamma \wedge 2 - 1}{2} d.$$

Here, the second last inequality holds since $\|W^{(i)}\|_{\mathrm{op}} \leq \|W^{(i)}\|_{\mathrm{F}} \leq 1$, and the last inequality holds since $c_1'' \leq \frac{\gamma\wedge 2-1}{2}$ and $\sqrt{a} \leq d$. Since $\sigma_j(\frac{\gamma d}{2} M_0) = \frac{\gamma d}{2}$ for all $j \in [r]$, and so $\sigma_j(A^{(i)}) \in [d, \gamma d]$ for all $j \in [r]$ and $i \in [N]$. Thus, $\Theta_1 \subset \Theta(s, r, d, \gamma)$.

For any $i \neq j$, $D(P_{A^{(i)}}||P_{A^{(j)}}) = \frac{1}{2}\|XA^{(i)} - XA^{(j)}\|_{\mathrm{F}}^2 \leq (c_1''\kappa_+)^2 a$, and

$$\|A^{(i)} - A^{(j)}\|_{\mathrm{s}_q} \geq c_1'' c_1' \sqrt{a}\, r^{1/q-1/2}.$$

Hence, for $c_1 = c_1' c_1''$, $c_2 = c_2'/2$ and $c_3 = (c_1''\kappa_+)^2$, $d_{\mathrm{KL}}(\mathcal{F}_0) \leq c_3 a$ and

$$\log \mathcal{M}(\Theta_1, \|\cdot\|_{\mathrm{s}_q}, c_1\sqrt{a}\, r^{1/q-1/2}) \geq c_2'(s-r)\log\frac{\mathrm{e}(p-r)}{s-r} \geq c_2 s \log\frac{\mathrm{e}p}{s}.$$

Here, the second inequality holds since $s \geq 2r$ and $\frac{p-r}{s-r} \geq \frac{p}{s}$. Moreover, by our choice of c_3, it is guaranteed that $c_3 \leq c_2/3$. This completes the proof of the first claim.

To establish the second claim, we note that Lemma 10 continues to hold if we replace s with k and W with W'. Thus, we could essentially repeat the foregoing arguments to obtain the second claim. This completes the proof. □

Proof of Theorem 2. Throughout the proof, let $c > 0$ denote a generic constant that depends only on γ and κ_+, though its actual value might vary at different occurrences. Note that we only need to prove the lower bounds for $\sigma = 1$, and the case of $\sigma \neq 1$ follows directly from standard scaling argument.

First, by restricting the nonzero entries of any matrix in $\Theta(s, k, r, d, \gamma)$

to the top left $s \times r$ (or $r \times k$) corner, we obtain a minimax lower bound by applying Lemma 9, i.e., for $\Theta = \Theta(s, r, d, \gamma)$ and any $q \in [1, 2]$,

$$\inf_{\widehat{A}} \sup_{\Theta} \mathbb{E} \|\widehat{A} - A\|_{\mathrm{s}_q}^2 \geq c(r^{2/q-1}d^2) \wedge (r^{2/q}(s+k)). \tag{26}$$

Here, we have used the fact that for any $a, b, c > 0$,

$$(a \wedge b) \vee (a \wedge c) = a \wedge (b \vee c) \asymp a \wedge (b + c). \tag{27}$$

Next, by Proposition 1, Lemma 11 and (27), we obtain

$$\inf_{\widehat{A}} \sup_{\Theta} \mathbb{E} \|\widehat{A} - A\|_{\mathrm{s}_q}^2 \geq c(\sqrt{a}\, r^{1/q-1/2})^2 = c(r^{2/q-1}d^2) \wedge \left(r^{2/q-1} \left(s \log \frac{\mathrm{e}p}{s} + k \log \frac{\mathrm{e}m}{k} \right) \right). \tag{28}$$

Thus, the minimax risk is lower bounded by the maximum of the lower bounds in (26) and (28). Applying (27) again, we complete the proof. □

6.3 A Theorem on Group Lasso

Theorem 3. *Consider the linear model $W = XB + Z$, where W is an $n \times r$ response matrix, X is an $n \times p$ design matrix, B is a $p \times r$ coefficient matrix with s-sparse row support for some $s \geq 1$, and Z is an $n \times r$ error matrix.*

Let

$$\widehat{B} = \arg\min_{B \in \mathbb{R}^{p \times r}} \|W - XB\|_F^2/2 + \lambda \|B\|_{2,1},$$

with a given penalty level λ. *Let Condition 1 hold with an absolute constant* $K > 1$ *and positive constants* s_*, c_* *satisfying* (11).

(i) If $2\|X_{*j}'(W - XB)\|_F \leq \lambda$ *for all* j, *then it holds that*

$$\|\widehat{B} - B\|_F \leq \frac{3(1 + (4c_*)^{-1})}{\kappa_-^2(s_*)} \sqrt{s}\lambda. \tag{29}$$

(ii) Assume the error matrix Z *has iid* $N(0, \sigma^2)$ *entries. For any given* $\eta \in (0, 1)$, *if we set*

$$\lambda \geq 2\sigma \max_j \|X_{*j}\|(\sqrt{r} + \sqrt{2\log(p/\eta)}),$$

then (29) holds with probability at least $1 - \eta$.

Proof of Theorem 3. We may rewrite the minimization problem in a vectorized version as follows

$$\min_{B \in \mathbb{R}^{p \times r}} \|\mathrm{vec}(W) - (I_r \otimes X)\mathrm{vec}(B)\|_2^2/2 + \lambda \|B\|_{2,1},$$

where vec is usual vectorization operator and $\otimes$ is the Kronecker product

as defined in (Muirhead, 1982, Section 2.2). In this case, the rows of B form natural groups which are all of size r and $\mathrm{vec}(B)$ satisfies the (s, rs) strong group-sparsity as defined in Huang and Zhang (2010).

We are to prove the desired result by invoking Lemma D.4 of Huang and Zhang (2010). To this end, we first verify that the two conditions of the lemma is satisfied. Note that the penalty level in Huang and Zhang (2010) corresponds to $2\lambda/(nr)$ in our notion, X_{G_j} corresponds to X_{*j}, and the sparse eigenvalues $\rho_+(G_j)$ and $\rho_\pm(rs)$ are identified as

$$\rho_+(G_j) = \|X_{*j}\|^2/(nr), \quad \rho_\pm(rs) = \kappa_\pm^2(s)/(nr).$$

Let $\ell = s_* - s - 1$ and $\lambda_-^2 = \min\{k\lambda^2 : kr \geq \ell r + 1, k \in \mathbb{Z}^+\} = (\ell+1)\lambda^2$. The conditions of Huang and Zhang (2010, Lemma D.4) can be rewritten in our notation as

$$2\|X_{*j}'(W - XB)\|_F \leq \lambda \qquad \text{and} \qquad \frac{\widetilde{\kappa}_+^2(s_*, s_* - s)}{\kappa_-^2(s_*)} \leq \sqrt{\frac{\ell+1}{s}}, \tag{30}$$

where $\widetilde{\kappa}_+^2(s_*, s_* - s) = \sqrt{(\kappa_+^2(s_*) - \kappa_-^2(2s_* - s))(\kappa_+^2(s_* - s) - \kappa_-^2(2s_* - s))}$.

Since by Definition 1, $\kappa_-^2(s) \leq \kappa_-^2(t) \leq \kappa_+^2(t) \leq \kappa_+^2(s)$, $\forall t \leq s$, we obtain

$$\widetilde{\kappa}_+^2(s_*, s_* - s) \leq \kappa_+^2(s_*) - \kappa_-^2(2s_*).$$

Thus, the conditions in (30) are satisfied under the assumption of Theorem 3. Then the conclusion of Huang and Zhang (2010, Lemma D.4) leads to

$$\|\widehat{B} - B\|_F \leq \frac{3}{\kappa_-^2(s_*)}(1 + 1.5\sqrt{s/(\ell+1)})\sqrt{s}\lambda \leq \frac{3(1+(4c_*)^{-1})}{\kappa_-^2(s_*)}\sqrt{s}\lambda.$$

This completes the proof of part (i).

Turning to part (ii), we need to upper bound $2\|X'_{*j}(W - XB)\|_F$. Since $X'_{*j}(W - XB)$ is a vector of length r with iid $N(0, \sigma^2\|X_{*j}\|^2)$ entries, it follows from Laurent and Massart (2000, Eq.(4.3)) that with probability $1 - \eta/p$,

$$\begin{aligned}\|X'_{*j}(W - XB)\|_F^2 &\leq \sigma^2\|X_{*j}\|^2(r + 2\sqrt{r\log(p/\eta)} + 2\log(p/\eta)) \\ &\leq \sigma^2\|X_{*j}\|^2(\sqrt{r} + \sqrt{2\log(p/\eta)})^2.\end{aligned}$$

With probability at least $1 - \eta$, we have $2\|X'_{*j}(W - XB)\|_F \leq \lambda$ for all j and thus (29) holds. □