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

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0073

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 study examines 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 are concentrated in a few rows and columns. We are especially interested in high-dimensional settings in which the numbers of predictors and/or response variables can be much larger than the number of observations. We propose a new estimation scheme, and show that it achieves both competitive numerical performance and fast computation. Moreover, we show that (a slight variant of) the proposed estimator simultaneously achieves near optimal nonasymptotic minimax rates of estimation under a collection of squared Schatten norm losses by providing both the error bounds for the estimator and the minimax lower bounds. The effectiveness of the proposed algorithm is also demonstrated using an *in vivo* calcium imaging data set.

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

An earlier version of this 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 study. The earlier version has been uploaded on arXiv, but is not intended for publication.

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0073

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

1. Introduction

Sparse linear regressions form a central topic in high-dimensional statistical inferences. For univariate responses, many researchers have developed a dazzling collection of tools to take advantage of the potential sparsity of the regression coefficients, including the Lasso (Tibshirani, 1996; Chen et al., 1998), SCAD (Fan and Li, 2001), Dantzig selector (Candes and Tao, 2007) and MCP (Zhang, 2010) among others. However, in contemporary applications, we routinely face multivariate, or even high-dimensional response variables and 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 10,000 voxels from fMRI imaging as the response variables for each subject, and over 400,000 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 $p \times m$

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0073

coefficient matrix that we want to estimate, and Z is an unobserved $n \times m$ matrix with independent and identically distributed (i.i.d.) noise entries. Thus, the ith rows of Y and X collect the measurements of the response and the predictor variables, respectively, on the ith subject. When either the number of predictors p or the number of response variables m is large, it becomes difficult to estimate the coefficient matrix A accurately unless some structural assumptions are imposed so that its intrinsic dimension is low.

Past studies have considered several important types of structural assumptions. For example, *low-rankness* assumes that the rank of A is much smaller than its matrix dimensions p and m. Model (1) with such a structure is referred to a reduced-rank regression model and is widely used in econometrics. See, for instance, Izenman (1975), Reinsel and Velu (1998), and the references therein. Another example of *sparsity* is that many entries in the coefficient matrix are zeros. Several types of sparsity may be considered, depending on the application. If only s of the p rows in A have nonzero entries, we refer to *row sparsity*. In other words, only a small subset (of size s) of the p predictors contribute to the variation of Y. Structures of this kind arise naturally in the context of multitask learning (Koltchinskii et al., 2011). This can also be viewed as an example of *group sparsity* (Yuan and Lin, 2006), where the rows of A form natural groups. If only k of the

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0073

m columns in A have nonzero entries, we refer to *column sparsity*. In this case, only k of the m response variables are affected by the predictors under consideration.

In this study, 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. Interest in such a model stems from both application and theory, and has increased significantly in recent years. In applications in the fields of genomics and neuroscience, researchers can now measure many response and predictor variables, resulting in increasingly large coefficient matrices. Thus, imposing both low-rankness and two-way sparsity leads to enhanced interpretability, which is more appealing 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, where the predictors are micro-RNA measurements and the response variables are gene expression levels. The sparsity occurs because a relatively small number of micro-RNAs regulated a small collection of genes under the specific experiments of interest. Furthermore, the low-rankness assumption is reasonable because only a handful of regulatory programs were present. Several algorithms have been proposed to estimate coefficient matrix in this model. See, for instance, Chen et al.

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0073

(2012) and Ma et al. (2014). However, to the best of our knowledge, there is no theoretical guarantee on the performance of these procedures in a high-dimensional regime, where the number of predictors and/or response variables exceeds the sample size.

Main contributions The main contributions of this study are as follows. First, we propose a new computationally efficient estimator for the coefficient matrix in (1) that takes advantage of the potential presence of low-rankness and two-way sparsity adaptively. The proposed estimator shows competitive numerical performance under a variety of simulation settings compared with that of 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. Second, we obtain new minimax estimation rates of the coefficient matrix with respect to a large class of squared Schatten norm losses. Furthermore, we show that (a slight variant of) our estimator simultaneously and adaptively achieves near optimal rates for this large collection of loss functions when the noise terms are homoscedastic and Gaussian.

Connection to the literature Many studies have examined coefficient matrices that are *either* sparse *or* of low rank. As a result, we now have a deep understanding of how the optimal mean squared estimation/prediction error depends on the model parameters, and how to achieve near optimal

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0073

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, extensive research exists for the case in which 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 a squared prediction error loss for this model in a high-dimensional setting. See also She (2014) and an earlier version of the present paper (Ma and Sun, 2014).

The previous studies most closely related to ours are those of Chen et al. (2012) and Ma et al. (2014), both of which focus on methodology. In comparison, in addition to proposing a new method, we justify its practical effectiveness using both numerical and theoretical studies. From a slightly different perspective, a series of studies 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 a two-way sparse reduced-rank regression with an orthogonal design.

Organization The rest of the paper is organized as follows. Section 2 presents our new methodology for obtaining a simultaneously sparse and

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0073

low-rank estimator of the coefficient matrix. Then, we demonstrate its competitive numerical performance in Section 3 using 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 **??** in the online Supplementary Material.

Notation For an $n \times p$ matrix $X = (x_{ij})$, the ith row of X is denoted by X_{i*} and the jth column by 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} denotes 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, $\mathrm{supp}(X)$ represents the index set of its nonzero rows. We denote the rank of X by $\mathsf{rank}(X)$, and $\sigma_i(X)$ represents its ith largest singular value. For any $q \in [1, \infty)$, the Schatten-q norm of X is $\|X\|_{\mathrm{S}_q} = \left(\sum_{i=1}^{n \wedge p} \sigma_i^q(X)\right)^{1/q}$, and for $q = \infty$, $\|X\|_{\mathrm{S}_\infty} = \sigma_1(X)$. Note that $\|X\|_{\mathrm{S}_2} = \|X\|_{\mathrm{F}}$ is the Frobenius norm, and $\|X\|_{\mathrm{S}_\infty} = \|X\|_{\mathrm{op}}$ is the operator norm of X. For any vector a, $\|a\|$ denotes

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0073

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 numbers 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 the group penalized regression. In what follows, we explain the main ideas underlying the algorithm; 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 comprises steps 3–5. In either stage, we first screen the columns of Y, then compute the r leading right singular vectors of the screened response matrix and, finally, perform 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

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0073

to identify those response variables with signals that stand out from the noise. To motivate the projection step, we observe that if the right singular vector matrix V of XA were known, then we could immediately reduce the dimensionality by considering a new regression problem in which we replace Y and A in (1) with their projected counterparts YV and AV, respectively. 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 by 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, then AV does as well. 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.

Next, we 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

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0073

be able to select additional columns, the signals of which might have been drowned in noise in the first stage. The inclusion of more signal columns of Y would then improve the estimation accuracy of the final estimator. Similarly, by pre-multiplying $\widetilde{Y}^{(1)}$ by $U_{(1)}U_{(1)}'$, 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 shown in a later analysis, the second stage is critical to achieve a high estimation accuracy for A.

In an earlier version of this 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 including column screening in both steps 1 and 3. As we show in Section 4, even when column sparsity is absent, our procedure still adapts automatically to achieve the best possible estimation accuracy, 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.

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0073

Algorithm 1: Estimation scheme for A using 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 $\widehat{A} = B_{(2)} V_{(1)}{}'$.

Penalized regressions with a group structure have been studied extensively. One of the most popular procedures is the group Lasso (Bakin,

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0073

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 the group Lasso have been studied in the literature using ideas originating from the original Lasso method. Huang and Zhang (2010) showed the upper bounds for the estimation and prediction errors of the group Lasso with a 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 present a theoretically justified choice of the penalty level λ for the group Lasso penalty function (2) when we have i.i.d. Gaussian noise.

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

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

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0073

where $\sigma(Y)$ is the collection of all nonzero singular values of Y. If the true rank of A is not known, we propose applying the estimator of Bunea et al. (2011), which is summarized in Algorithm 2. The user-specified parameter can be selected as

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

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

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^{-} is its Moore–Penrose pseudo-inverse.

2 Compute the singular values of PY and select

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

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

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0073

3. Numerical Study

3.1 Simulation

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

Comparison under different model parameters We first compare the aforementioned methods under different design matrices, ranks, and sparsity levels. To this end, we borrow several simulation settings from Bunea et al. (2012), but add columns of pure noise to 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

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

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0073

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

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

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0073

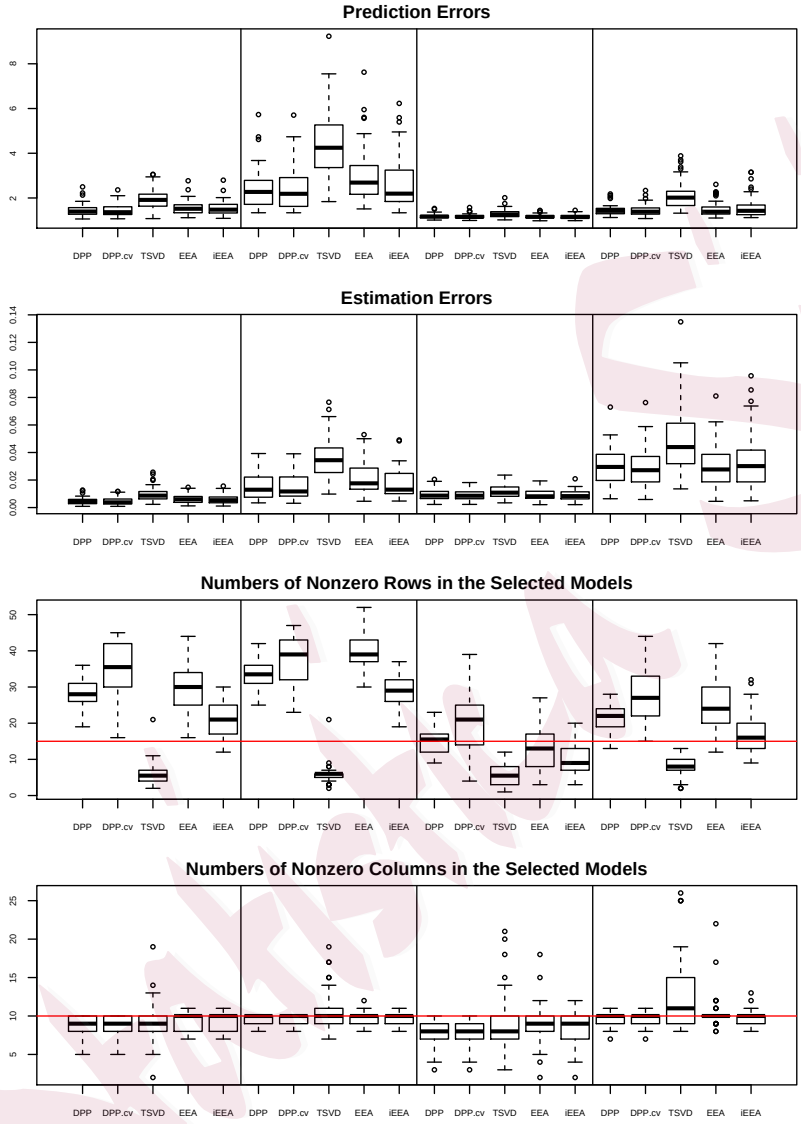

Figure 1: Performance of five methods: prediction errors, estimation errors, and sizes of selected models across 50 replications. Sample size $n = 30$, model size $m = 50$, $p = 100$, $s = |\text{supp}(A)| = 15$, $k = |\text{supp}(A')| = 10$, and rank $r = 2$. The four blocks in each plot are for $(\rho,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

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0073

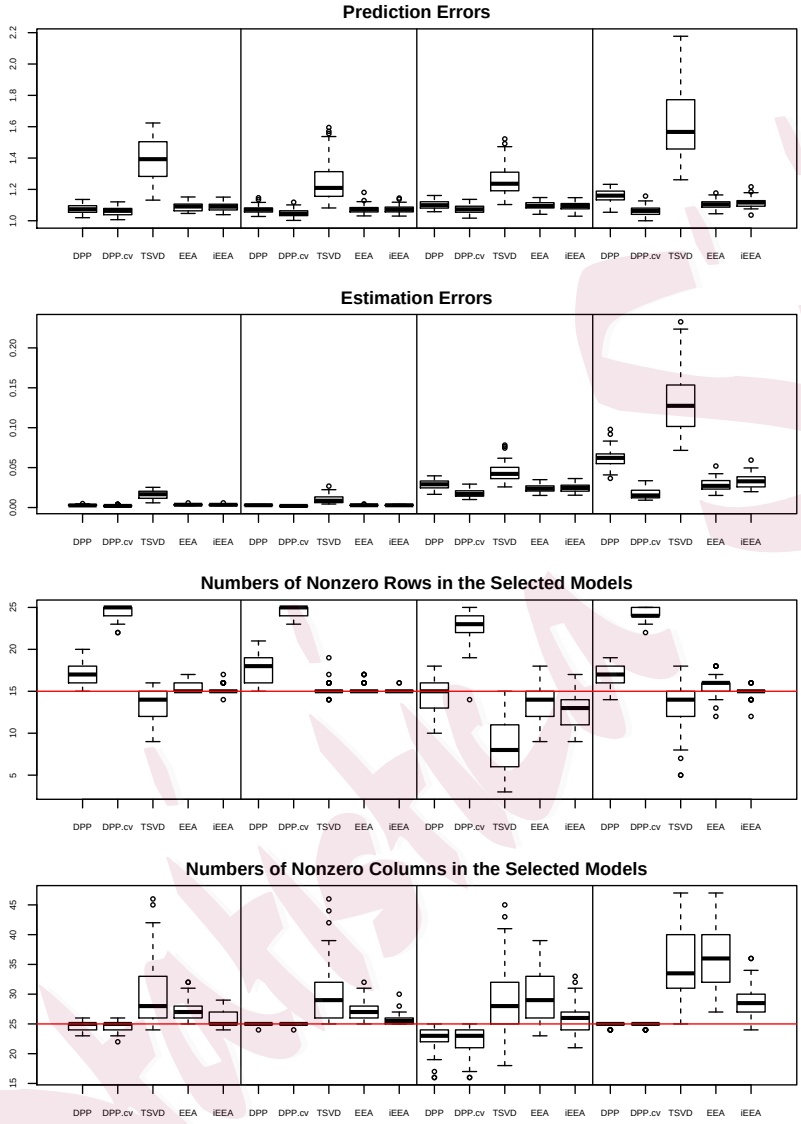

Figure 2: Performance of five methods: prediction errors, estimation errors, and sizes of selected models across 50 replications. Sample size $n = 100$, model size $m = 50$, $p = 25$, $s = |\text{supp}(A)| = 15$, $k = |\text{supp}(A')| = 25$, and rank $r = 5$. The four blocks in each plot are for $(\rho,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

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0073

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

are reported in Table 1. DPP.cv performs best for almost all cases considered, whereas the 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% compared with TSVD, EEA and iEEA. Note that when comparing prediction errors, the quantity that makes the most sense to use is the excessive error an estimator makes in addition to the oracle error one would make even when the true coefficient matrix is given. In the current setting, the (normalized) oracle error is one. In terms of the excessive prediction error, we observe that the prediction accuracy of DPP.cv is better than those of the other methods by a similar percentage. In addition, note that the proposed method tends to choose more rows than the true model, whereas the column selection, relying on the screening of the columns of $U'Y$, is more accurate. This is somewhat expected, because the group Lasso tends to over-select variables when cross-validation is employed to choose the tuning parameter values.

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0073

Comparison under different noise distributions We now compare the performance of the methods using nonGaussian 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 denotes the t-distribution with ν degrees of freedom. Note that all three distributions are normalized to have a 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. The results show that our methods, and particularly DPP.cv, demonstrate competitive performance, even for nonGaussian data. Moreover, compared with the corresponding performance measures on Gaussian data (the first block of box plots), we find that all of the estimators are relatively robust to the noise distributions, although their performance (with the exception of TSVD) does degrade as the tail of the noise distribution becomes heavier.

Performance under heteroscedastic noise Although the proposed method is designed under a model in which 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 jth 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 $\omega = 1, 2, 5, 10$.

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0073

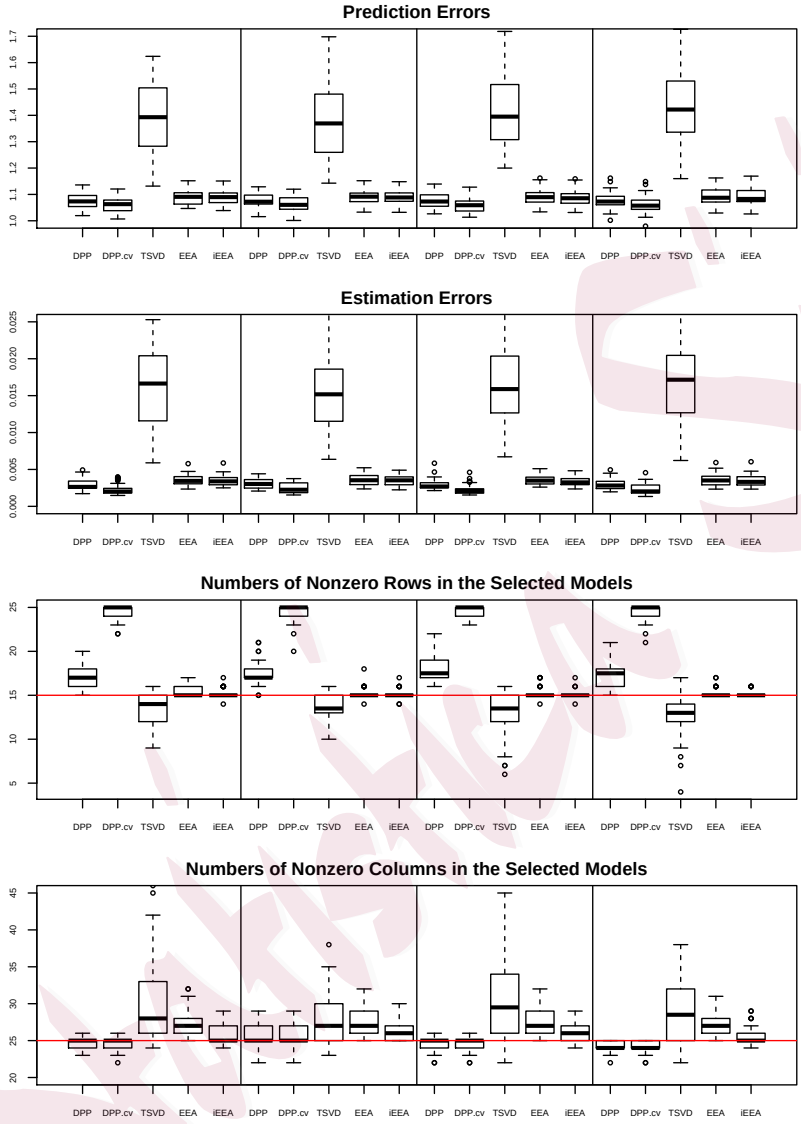

Figure 3: Performance of five methods on nonGaussian data. Sample size $n = 100$, model size $m = 50$, $p = 25$, $s = |\text{supp}(A)| = 15$, $k = |\text{supp}(A')| = 25$, rank $r = 5$, $\rho = 0.1$, and $b = 0.2$. The four blocks in each plot are for different noise distributions: standard normal, $\sqrt{3/5}t_5$,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

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0073

In this setting, ω is the ratio of the largest to the smallest possible variance. When $\omega = 1$, this becomes the case of equal variance, as above (the second setting with $\rho = 0.1, b = 0.2$). When ω increases, the noise variance varies among the columns, whereas the average noise variance remains one. In Fig. 4, we report the prediction, estimation, and selection performance of the proposed DPP method for different ω. When heteroscedasticity occurs, our approach selects 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 for tracking 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 enters the cell and changes its fluorescent properties by attaching to genetically encoded calcium indicators. By recording movies of fluorescence activity, researchers hope to identify and demix the regions of interest (ROIs) and extract spike traces (Pnevmatikakis et al., 2014; Haeffele et al., 2014).

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

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0073

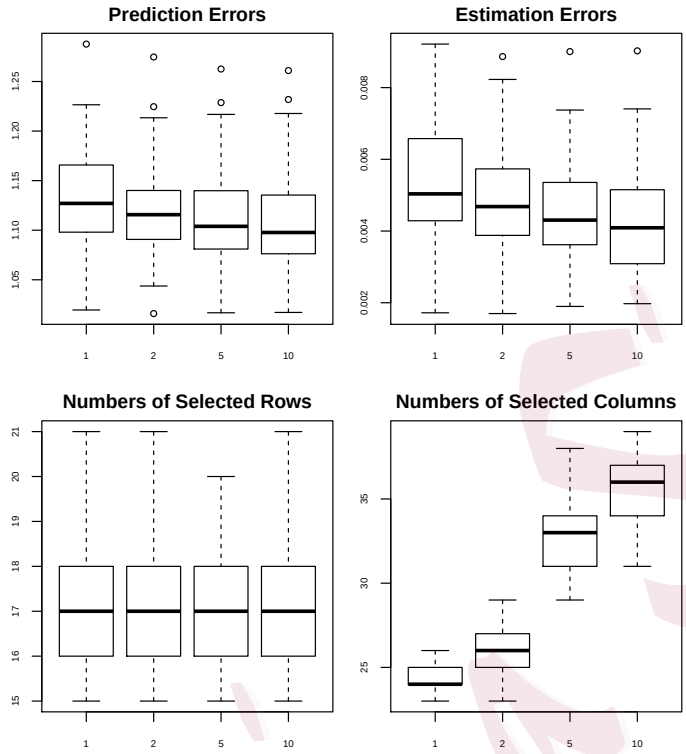

Figure 4: Performance of the proposed DPP method on heteroscedastic data. Sample size $n = 100$, model size $m = 50$, $p = 25$, $s = |\text{supp}(A)| = 15$, $k = |\text{supp}(A')| = 25$, rank $r = 5$, $\rho = 0.1$, and $b = 0.2$. The four box plots in each plot are for $\omega = 1, 2, 5, 10$, respectively.

pose 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 than $l_1 \times l_2$ 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 ith neuron. Then, the fluorescence

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0073

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 notation, we have

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

where $Y = (y_1, \cdots, y_T)' \in \mathbb{R}^{T\times m}, \Omega = (\omega_1, \cdots, \omega_K)' \in \mathbb{R}^{K\times m}, C = (c_1, \cdots, c_K) \in \mathbb{R}^{T\times K}$, and $Z = (z_1, \cdots, z_T)' \in \mathbb{R}^{T\times m}$. Let $s_i = (s_i(1), \cdots, s_i(T))' \in \mathbb{R}^T$ be the spike trace of the ith 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

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

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0073

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. Taken together, these imply that A is two-way sparse (and low-rank, because 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 using 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 this heuristic.

The calcium imaging data ($n = p = T = 559$, $m = 135 \times 131$) we use here are taken *in vivo* from the primary auditory cortex of a mouse with genetically encoded calcium indicator GCaMP5 (Akerboom et al., 2012). We report here the four most significant neurons in order to demonstrate the

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

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0073

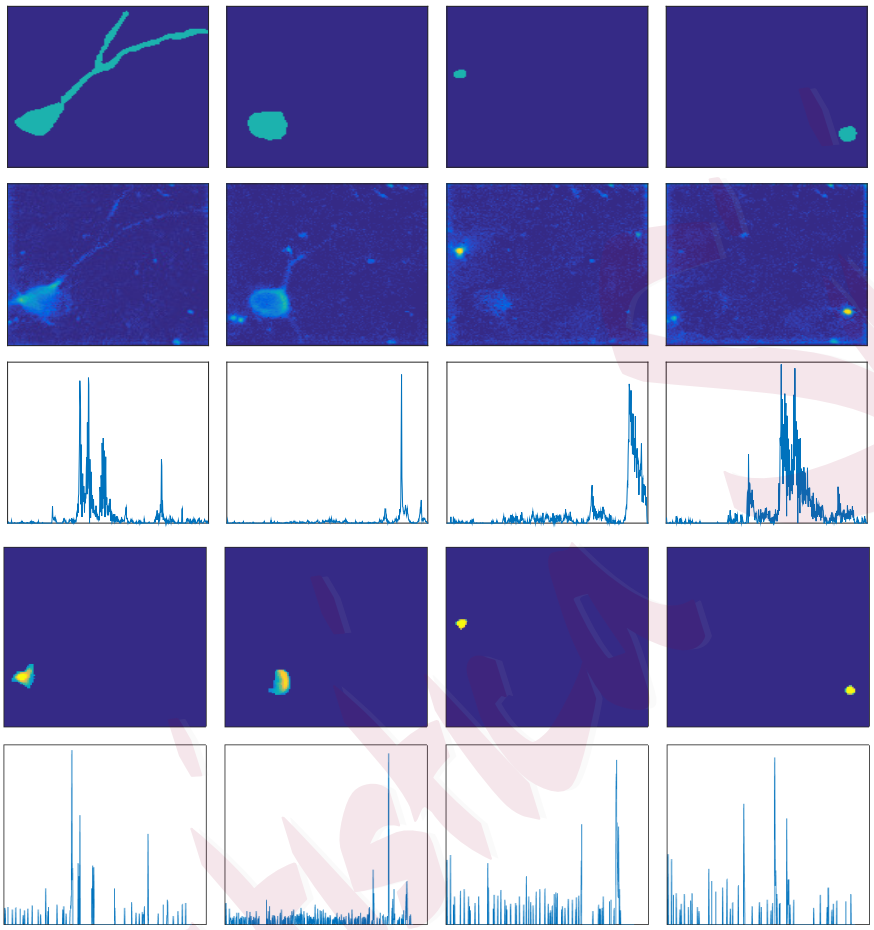

Figure 5: Application to *in vivo* calcium imaging data. First row: manually segmented regions of neurons. Second row: heat maps of the recovered spatial components by Algorithm 1. Third row: estimated spike trace by Algorithm 1. Fourth row: heat maps of the corresponding spatial components recovered 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.

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0073

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 by Giovannucci et al. (2017). In Figure 5, the first row shows the manually segmented regions of the neurons from the raw data set, which can be regarded as the approximately true support of the spatial component Ω. The first neuron consists of a cell body with a dendritic branch, and it overlaps heavily with the second neuron, making manual segmentation very challenging. The second row displays heat maps of the neurons recovered by the proposed approach, showing that they match the manual segmentation very well. The third row of Figure 5 shows the estimated spike traces using our method. The fourth and the fifth rows show the corresponding components found using 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 the theoretical results for a slight variant of the proposed estimation scheme, where the noise matrix Z in (1) has i.i.d. Gaus-

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0073

sian entries. All proofs are provided in the Supplementary Material.

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:

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

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 noise matrix Z as the only source of randomness. In what follows, we present high probability error

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0073

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, regardless of 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 noise is 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 $\widetilde{Z}$, such 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. $N(0, 2\sigma^2)$ noise. Employing this method twice, we 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, Algorithm 1 refers to the procedure with independent samples $Y_{(i)}$ used in the $(i+1)$th step, for $i = 0, 1, 2, 3$, where the noise variance is $\widetilde{\sigma}^2 = 4\sigma^2$.

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0073

The design matrix Without loss of generality, we assume X is of full rank. Otherwise, we can always perform the following operation to reduce it to the full-rank case. If $\mathsf{rank}(X) = q < n \wedge p$, 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. $N(0, \sigma^2)$ noise 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 $\Delta = \mathrm{diag}(\delta_1, \ldots, \delta_r)$ containing the nonzero singular values of XA. To introduce appropriate assumptions on X, we first state the following definition.

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

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

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 $\kappa_-(\ell; X) \le \kappa_-(l; X) \le \kappa_+(l; X) \le \kappa_+(\ell; X)$.

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0073

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 place a 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 the 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 as*

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

in steps 2 and 4 of Algorithm 1 with the group Lasso penalty (2). Let

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0073

$\alpha = 2\sqrt{3}$ 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 a simultaneously low-rank and row-sparse setting, condition (12) is stronger than some related conditions in the literature, such as that in Bunea et al. (2012) for establishing minimax rates. However, we provide the theoretical guarantee of an actual estimator computed using Algorithm 1, whereas Bunea et al. (2012) were concerned with the global optimum of a nonconvex program, which is not always attainable

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0073

by heuristic algorithms. Therefore, the two 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 the minimax risk when estimating A under the 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, when estimating A, the minimax risk 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

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0073

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

As mentioned earlier, the one-way sparse reduced-rank regression model considered in the literature, for example, by 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 to fully exploit any potential column sparsity.

5. Conclusion

We have proposed a new DPP estimator for the coefficient matrix in a two-way sparse reduced-rank regression. The model is well motivated by the massive data sets that are becoming increasingly common in a number of fields, especially genomics and neuroimaging. The proposed estimator is fast to compute and demonstrates competitive performance compared with existing methods in simulation studies. In addition, we have illustrated its potential use in neuroscience by applying it to an analysis of a calcium imaging data set, with the results comparing favorably with some state-of-the-art methods. Lastly, we further justify its nice empirical performance using a decision-theoretic analysis when the data is Gaussian.

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0073

In terms of the DPP estimator, an interesting problem to be studied in future is to establish high-probability error bounds when the data are nonGaussian. 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, note that steps 3–4 of Algorithm 1 can be iterated till a certain convergence criterion is met. Thus, we could also define an iterative projected penalization estimator. However, simulation results not reported here did not find a significant performance gain by employing such an iterative scheme, which is more costly in terms of computation. Furthermore, it is of interest to investigate 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 regressions (Li and Nachtsheim, 2012; Lin et al., 2015). Therefore, it would be of great interest to conduct analogous investigations for multiple responses that include both low-rankness and sparsity.

Supplementary Material

The online Supplementary Material provides all technical proofs.

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0073

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.

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0073

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.

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0073

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.

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0073

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.

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0073

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.

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0073

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.