Unified Low-Rank Matrix Estimate via Penalized Matrix Least Squares Approximation

Xiangyu Chang, Yan Zhong, Yao Wang, and Shaobo Lin

***Abstract*— Low-rank matrix estimation arises in a number of statistical and machine learning tasks. In particular, the coefficient matrix is considered to have a low-rank structure in multivariate linear regression and multivariate quantile regression. In this paper, we propose a method called penalized matrix least squares approximation (PMLSA) toward a unified yet simple low-rank matrix estimate. Specifically, PMLSA can transform many different types of low-rank matrix estimation problems into their asymptotically equivalent least-squares forms, which can be efficiently solved by a popular matrix fast iterative shrinkage-thresholding algorithm. Furthermore, we derive analytic degrees of freedom for PMLSA, with which a Bayesian information criterion (BIC)-type criterion is developed to select the tuning parameters. The estimated rank based on the BIC-type criterion is verified to be asymptotically consistent with the true rank under mild conditions. Extensive experimental studies are performed to confirm our assertion.**

***Index Terms*— Degrees of freedom, low-rank matrix estimate, multivariate linear regression, multivariate quantile regression (QR).**

I. Introduction

RECOVERING intrinsic data structures from collected data matrices plays an important role in various statistical and machine learning tasks. Low-rank structures, as a data structure of interest, have been surprisingly found in many real-world applications, e.g., stock market data [1], Internet traffic flow data [2], movie ranking data [3] and hyperspectral image data [4], [5]. Based on the low-rank assumption, a number of learning systems have been proposed for handling the aforementioned applications. Prominent examples include multivariate linear regression [1], [6], multitask learning [7], matrix completion [3], [8], robust principle component analysis [9], and relational learning [10]. All these examples provide evidence that exploiting the essential low-rank structure correctly can improve the prediction accuracy. Therefore, estimating the low-rank structure in practical tasks has become a recent focus in statistics and machine learning.

In this paper, we consider the problem of low-rank coefficient matrix estimation for multivariate linear regression. Assume that there are n observations for p explanatory variables $\mathbf{x} = (x_1, \dots, x_p)^\top$, q responses $\mathbf{y} = (y_1, \dots, y_q)^\top$, and

$$\mathbf{Y} = \mathbf{X}B + \mathbf{E} \tag{1}$$

where $\mathbf{Y} = (\mathbf{y}_1, \dots, \mathbf{y}_n)^\top \in \mathbb{R}^{n\times q}$, $\mathbf{X} = (\mathbf{x}_1, \dots, \mathbf{x}_n)^\top \in \mathbb{R}^{n\times p}$, $B \in \mathbb{R}^{p\times q}$ is the coefficient matrix, and $\mathbf{E} = (\mathbf{e}_1, \dots, \mathbf{e}_n)^\top$ is the regression noise, whose components are independently drawn from some unknown distributions. Throughout this paper, we normalize each input variable so that there is no intercept in (1). When the so-called Gram matrix $\mathbf{X}^\top\mathbf{X}$ is invertible, the ordinary least-squares (OLS) approach yields an estimator $\hat{B}_{\text{OLS}} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{Y}$. The OLS approach amounts to performing q separate univariate regressions and completely ignores the multivariate aspect of the problem [11]. As a result, this approach could perform poorly in the presence of highly correlated explanatory variables [1]. In addition, the OLS estimate is unsuitable for the high-dimensional case whereby p and q are both large [1], [12].

Numerous approaches have attempted to overcome the above-mentioned hurdles by imposing specific structure assumptions on the coefficient matrix B [1], [13]–[16]. One of the important research lines focuses on assuming the low-rank structure of the regression coefficient matrix B and penalizing the singular value of B, which leads to the following unified nuclear-norm penalization framework:

$$\min_{B} \left\{ \sum_{i=1}^{n} \sum_{j=1}^{q} L\big(y_{ij} - \mathbf{x}_i^\top B_j\big) + \|B\|_* \right\} \tag{2}$$

where B_j is the jth column of B and $\sigma_i(B)$ is the ith largest singular value of B. Here, we should mention that $\sigma_i(B) \geq 0, i = 1, \dots, \min\{p, q\}$ are obtained by a type of singular value decomposition and that the nuclear norm, $\|B\|_* = \sum_{i=1}^{\min(p,q)} \sigma_i(B)$, can be seen as a measure of the

Manuscript received October 28, 2016; revised September 7, 2017, January 19, 2018, April 27, 2018, and May 30, 2018; accepted May 31, 2018. Date of publication July 2, 2018; date of current version January 21, 2019. This work was supported in part by the National Natural Science Foundation of China under Grant 11771012, Grant 11501440, Grant 61502342, and Grant 91546119, in part by the Major Program of National Natural Science Foundation of China under Project 71731009 and Project 71742005, in part by the China Postdoctoral Science Foundation under Grant 2017M610628, and in part by the Key Research Program of Hunan Province, China under Grant 2017GK2273. *(Corresponding author: Yao Wang.)*

X. Chang is with the Center of Data Science and Information Quality, School of Management, Xi'an Jiaotong University, Xi'an 710049, China.

Y. Zhong is with the Department of Statistics, Texas A&M University, College Station, TX 77843 USA.

Y. Wang is with the School of Management, Xi'an Jiaotong University, Xi'an 710049, China, and also with the State Key Laboratory of Robotics, Shenyang Institute of Automation, Chinese Academy of Sciences, Shenyang 110016, China (e-mail: yao.s.wang@gmail.com).

S. Lin is with the Department of Mathematics, Wenzhou University, Wenzhou 325035, China, and also with the State Key Laboratory of Robotics, Shenyang Institute of Automation, Chinese Academy of Sciences, Shenyang 110016, China.

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TNNLS.2018.2844242

low-rank structure of B. Hence, the nuclear-norm penalization encourages sparsity among the singular values to achieve simultaneous rank reduction and shrinkage coefficient matrix estimation. Here, L is a loss function of B, and different L results in different models. For example, Yuan *et al.* [1] proposed the model

$$\min_B \left\{ \frac{1}{2}\mathrm{tr}[(\mathbf{Y}-\mathbf{X}B)^\top W(\mathbf{Y}-\mathbf{X}B)]+\lambda \sum_{i=1}^{\min(p,q)} \sigma_i(B) \right\} \tag{3}$$

where tr is the trace of a matrix. In (3), the loss function L can be seen as a weighted least squares loss. The flexibility of L makes (2) a unified form of many different nonconvex and nonsmooth optimization problems (a detailed discussion can be found in Section IV-A). The tuning parameter $\lambda > 0$ is also called the regularization parameter, which controls the selected rank of B, that is, larger λ leads to a smaller estimated rank of B. Therefore, we do not distinguish the actual meaning of selecting rank and tuning parameter throughout this paper.

Summarily, there are two key issues concerning any form of (2) before it is applied: a fast algorithm and an efficient tuning parameter selection strategy. The main aim of this paper is to provide a systematic approach to address these two issues.

As a powerful tool for handling complex learning systems, the least squares approximation (LSA) method has been applied to the robust regression [17], [18], robust classification [19], sparse signal recovery [20], [21], and so on. The basic idea of LSA is to use the least squares term to replace the part that is difficult to optimize in a complex model. For example, Wang and Leng [22] employed LSA to obtain a unified and efficient yet simple lasso-type estimate for many different complex models with the ℓ_1-norm penalty. Motivated by their work, we propose in this paper a penalized matrix LSA (PMLSA) method to address the general low-rank matrix estimation problem (2). The basic idea is to use the second-order Taylor expansion of $L(y_{ij} - \mathbf{x}_i^\top B_j)$ in (2) to replace itself, which gives an LSA formulation. The proposed PMLSA method is shown to have the following advantages.

1) It can be quickly solved by an efficient matrix fast iterative shrinkage-thresholding algorithm (FISTA) that has an improved $O(1/\sqrt{\epsilon})$ convergence speed to obtain an ϵ-accurate solution.
2) PMLSA equipped with a Bayesian information criterion (BIC)-type parameter selection strategy that produces a consistent estimate of the true rank of B under mild conditions. Compared with the popular cross-validation (CV) procedure, the BIC-type procedure greatly reduces the computational cost.
3) It can be seen as a unified framework to address numerous different nuclear-norm penalized low-rank models [1], [13]–[16] in (2). For these models, a model selection method with the consistent rank selection property has yet to be developed. Therefore, the main advantage of PMLSA and the proposed BIC procedure is in providing a way to select the true rank consistently for the models included in (2).

The remainder of this paper is organized as follows. In Section II, we introduce PMLSA and present the matrix FISTA algorithm to compute the solution of PMLSA. In Section III, we propose a BIC-type criterion for selecting the tuning parameter of PMLSA. In Section IV, we discuss the related work on PMLSA in detail. In Section V, extensive experimental studies on synthetic data and two real-world applications are performed to further demonstrate the effectiveness of our method. In Section VI, we draw the conclusions for this paper.

II. Penalized Matrix Least Squares Approximation

In this section, we present the details of PMLSA. Moreover, we emphasize that one of the advantages of PMLSA is that it can be quickly solved via an efficient matrix FISTA algorithm.

A. Models

Following the same notation in Section I, a general multivariate linear regression model with n samples on p explanatory variables and q responses has the following matrix form $\mathbf{Y} = \mathbf{X}B + \mathbf{E}$. A classical estimate of B is obtained by minimizing some loss function $L(B)$ such as the least squares loss $L(B) = \sum_{i=1}^n \sum_{j=1}^q (y_{ij} - \mathbf{x}_i^\top B_j)^2$ or least absolute derivation loss $L(B) = \sum_{i=1}^n \sum_{j=1}^q |y_{ij} - \mathbf{x}_i^\top B_j|$. However, this natural estimate $\hat{B} = \arg\min L(B)$ cannot exploit the low-rank structure when p and q are relatively large. Previous works [1], [12], [14], [15] applied the nuclear-norm penalty onto the matrix B to overcome this drawback, leading to the following optimization problem:

$$\min_B \left\{ L(B) + \lambda \sum_{i=1}^{\min(p,q)} \sigma_i(B) \right\} \tag{4}$$

where $L(B) = \sum_{i=1}^n \sum_{q=1}^m L(y_{ij} - \mathbf{x}_i^\top B_j)$. Here, (4) is equivalent to (2) for simplicity.

Now, we focus on establishing a unified method for solving a series of models that have a form of (4). First, we denote $L(B_j) = L(y_{ij} - \mathbf{x}_i^\top B_j)$, $j = 1, \ldots, q$. Then, motivated by the LSA method [17]–[22], we propose a new matrix LSA method. Let us emphasize that $\hat{B} = \arg\min L(B)$. Assume that the second-order derivative of L exists; then, the first derivative of L has the property $\dot{L}(\hat{B}) = 0$. Therefore, the Taylor expansion of $L(B)$ at point $\hat{B}$ can be written as

$$\begin{aligned}
L(B) &= \sum_j L(B_j) \\
&\approx \sum_j \Big[L(\hat{B}_j) + \dot{L}(\hat{B}_j)^\top (B_j - \hat{B}_j) \\
&\qquad + \frac{1}{2}(B_j - \hat{B}_j)^\top \ddot{L}_n(\hat{B}_j)(B_j - \hat{B}_j) \Big] \\
&= \sum_j \left[\frac{n}{2}(B_j - \hat{B}_j)^\top \frac{\ddot{L}_n(\hat{B}_j)}{n}(B_j - \hat{B}_j) + L(\hat{B}_j) \right] \\
&= \sum_j \frac{n}{2}(B_j - \hat{B}_j)^\top \hat{\Sigma}_j^{-1}(B_j - \hat{B}_j) + \sum_j L(\hat{B}_j)
\end{aligned}$$

where $\sum_j L(\hat{B}_j)$ is a constant and $\hat{\Sigma}_j^{-1} = (\ddot{L}_n(\hat{B}_j)/n)$ is an estimate of the covariance of $\hat{B}_j$ [22]. Note that sometimes

$\ddot{L}_n(\hat{B}_j)$ does not exist, but $\hat{\Sigma}_j^{-1}$ can be calculated as shown in (7). Therefore, this approximation can be used without the assumption on the existence of a continuous second-order derivative of L. Finally, ignoring all constants, we obtain a new loss function for any given B

$$L^*(B) = \frac{1}{2}\sum_{j=1}^{q}(B_j - \hat{B}_j)^\top \hat{\Sigma}_j^{-1}(B_j - \hat{B}_j). \tag{5}$$

Before introducing PMLSA, we consider how to obtain $\hat{\Sigma}_j^{-1}$ in (5). For the univariate model $\mathbf{Y}_j = \mathbf{X}B_j + \mathbf{E}_j$, Koenker [23] indicated that the asymptotic covariance of the estimation $\hat{B}_j$ has the common property

$$\Sigma_j^{-1} = f(e_j)\mathrm{COV}(\mathbf{x}) \tag{6}$$

where $f(e_j)$ is a function related to the jth noise and $\mathbf{Y}_j$ and $\mathbf{E}_j$ are the jth column of $\mathbf{Y}$ and $\mathbf{E}$, respectively. For example, the well-known OLS has the form $\Sigma_j^{-1} = (\sigma_j^2)^{-1}\mathrm{COV}(\mathbf{x})$ [24], where $\mathrm{COV}(\mathbf{x})$ is the population covariance matrix of $\mathbf{x}$. Hence, $f_{\mathrm{OLS}}(e_j) = (\sigma_j^2)^{-1}$. For the quantile regression (QR), it has the form $\Sigma_j^{-1} = (\sigma_j(\tau)^2)^{-1}\mathrm{COV}(\mathbf{x})$, where τ is the target quantile, $\sigma_j(\tau)^2 = (\tau(1-\tau)/g^2(e_{j\tau}))$, and $g(e_{j\tau})$ is the probability density of e_j at the τ quantile. Therefore, $f_{\mathrm{QR}}(e_j) = (\sigma_j(\tau)^2)^{-1}$. With these, we obtain the empirical estimate of Σ_j^{-1} as

$$\hat{\Sigma}_j^{-1} = \hat{f}(e_j)\hat{\Sigma}_{\mathbf{X}} \tag{7}$$

where $\hat{\Sigma}_{\mathbf{X}}$ is the sample covariance matrix of $\mathbf{x}$ and $\hat{f}(e_j)$ is a sample estimate of $f(e_j)$.

With (7), (5) can be transformed into the following equivalent form:

$$L^*(B) = \frac{1}{2}\mathrm{tr}[(B^* - \hat{B}^*)^\top \hat{\Sigma}_{\mathbf{X}}(B^* - \hat{B}^*)] \tag{8}$$

where $B^* = (B_1\hat{f}(e_1)^{(1/2)}, B_2\hat{f}(e_2)^{(1/2)}, \ldots, B_q\hat{f}(e_q)^{(1/2)})$ and $\hat{B}^* = (\hat{B}_1\hat{f}(e_1)^{(1/2)}, \hat{B}_2\hat{f}(e_2)^{(1/2)}, \cdots, \hat{B}_q\hat{f}(e_q)^{(1/2)})$. Since the rank of B^* equals B, penalizing B^* produces an equal effect on B. We call (8) as MLSA. Then, by penalizing MLSA with the nuclear norm of B^*, we obtain

$$\min_{B^*}\left\{\frac{1}{2}\mathrm{tr}[(\hat{B}^* - B^*)^\top \hat{\Sigma}_{\mathbf{X}}(\hat{B}^* - B^*)] + \lambda\sum_{i=1}^{\min(p,q)}\sigma_i(B^*)\right\} \tag{9}$$

which is called as PMLSA. We then show in the following how to obtain the valid estimate of (2) via PMLSA.

B. Matrix FISTA for Solving PMLSA

With loss of generality, we still use B to replace B^* in the following. Actually, (9) can be formulated as

$$Q(B) = \frac{1}{2}\mathrm{tr}[(\hat{B}^* - IB)^T \hat{\Sigma}_{\mathbf{X}}(\hat{B}^* - IB)] + \lambda\sum_{i=1}^{\min(p,q)}\sigma_i(B) \tag{10}$$

where I is the identity matrix. Compared with (3), if we set $\mathbf{Y} = \hat{B}^*$ and $\mathbf{X} = I$, then (10) can be seen as a special case of (3). Yuan *et al.* [1] used an algorithm called SDPT3 to solve (3). However, SDPT3 is a second-order cone program solver, which is not computationally efficient for large-scale problems.

The FISTA [25] is an efficient method for solving the penalized linear regression with the ℓ_1-norm penalty. Compared with the traditional algorithms, it significantly improves the convergence speed from $O(1/\epsilon)$ to $O(1/\sqrt{\epsilon})$ to obtain an ϵ-accurate solution. Furthermore, Toh and Yun [26] successfully extended the FISTA to a matrix version. We apply the matrix version of FISTA to solve (10), as shown in Algorithm 1.

Algorithm 1 Matrix FISTA

```
Input: $\hat{\Sigma}_{\mathbf{X}}$, $B^0 := \hat{B}^*$ (the initial value of $B$) and a tolerance $\epsilon$.
Initialization: Choose $B^1 = B^0, t^1 = t^0 = 1$, $l = \sigma(\hat{\Sigma}_{\mathbf{X}})_1$.
Iteration: For $s = 0, 1, \ldots$, generate $B^{s+2}$ from $B^{s+1}$ and $B^s$ by the following steps:
    1) $C^{s+1} \leftarrow B^{s+1} + \frac{t^s - 1}{t^{s+1}}(B^{s+1} - B^s)$.
    2) $G^{s+1} \leftarrow C^{s+1} - l^{-1}\hat{\Sigma}_{\mathbf{X}}(C^{s+1} - \hat{B}^*)$.
    3) SVD on $G^{s+1}$: $G^{s+1} = U^{s+1}diag(\sigma(G^{s+1}))V^{s+1\top}$,
       then $B^{s+2} \leftarrow U^{s+1}diag((\sigma(G^{s+1}) - \lambda/l)_+)V^{s+1\top}$.
    4) $t^{s+2} \leftarrow \frac{1+\sqrt{1+4(t^{s+1})^2}}{2}$.
Stopping condition: Keep the above iteration until $s$ satisfies
    $$\frac{||P^s||_F}{l\max(1, ||B^s||_F)} \le \epsilon$$
  where $P^s \leftarrow l(C^{s-1} - B^s) + \hat{\Sigma}_{\mathbf{X}}(B^s - C^{s-1})$.
Output: $\hat{B} = B^s$.
```

Remark 1: Since the loss function (8) is convex and continuously differentiable, the convergence analysis of Algorithm 1 can be performed using [26, Corollary 3.1], therein showing that Algorithm 1 has a convergence speed of $O(1/\sqrt{\epsilon})$. This implies that the convergence speed of Algorithm 1 is extremely fast compared with the common $O(1/\epsilon)$ convergence speed [1]. Moreover, we should emphasize that implementing Algorithm 1 depends on selecting a proper tuning parameter λ. However, to the best of our knowledge, models included in (2) still do not have a proper model selection procedure with a consistent rank estimate guarantee.

III. Rank Selection of PMLSA

In this section, we show the second advantage of PMLSA, namely, we derive a BIC-type criterion to select λ with the rank selection consistency verification.

In the machine learning community, CV, as a commonly used method, has been widely applied to select the tuning parameters of various models. Recently, the benefits of the information-type criterion have been shown in the model selection of many complex models [27]–[30]. Compared with the CV method, information-type criteria, which do not need to resample the original data set to evaluate the average prediction accuracy, are more time efficient. For example, the classical information-type criterion, BIC [31], for linear models has the form

$$\text{Loss} + \text{df}_\lambda \log(n) \tag{11}$$

where the first part of (11) is usually addressed by recalibrating the goodness of fit based on the whole data set (e.g., likelihood functions and loss functions), the second part of (11) measures the model complexity for different parameters λ that need to be carefully chosen, and df_λ is the degrees of freedom of the model.

It is not difficult to see that to take advantage of the BIC-type criterion for tuning parameter selection, we should first provide a precise estimation of df_λ of PMLSA.

A. BIC-Type Criterion for PMLSA

Indeed, $L^*(B)$ in (8) can be naturally seen as the loss of PMLSA. Then, the BIC-type criterion considered in this paper is defined as

$$\text{BIC}_\lambda(B) = \frac{1}{2}\text{tr}[(B-\hat{B}^*)^\top \hat{\Sigma}_{\mathbf{X}}(B-\hat{B}^*)] + \frac{\log(n)\text{df}_\lambda}{n}. \tag{12}$$

Compared with (11), (12) is rescaled by $1/n$ to obtain $\hat{\Sigma}_{\mathbf{X}}$. Note that the classical BIC generally does not depend on an unregularized estimate such as $\hat{B}^*$. Hence, the formulation (12) is not exactly the traditional BIC criterion. However, motivated by the unified lasso-type estimate proposed in [22] that has a similar formulation as (12), we can still call it a BIC-type criterion. Denote by $\hat{r}_\lambda$ an estimated rank of $\hat{B}^*_\lambda$ with the true rank r. We obtain a best estimate $\hat{B}^*_r$ by restricting the rank to r, which means that

$$\hat{B}^*_r = \arg \min_{\{B:\ \text{rank}(B)=r\}} \text{tr}[(B-\hat{B}^*)^T \hat{\Sigma}_{\mathbf{X}}(B-\hat{B}^*)]. \tag{13}$$

Obviously, for any B^* with $\text{rank}(B^*) = r$, we have

$$\text{tr}[(\hat{B}^* - B^*)^T \hat{\Sigma}_{\mathbf{X}}(\hat{B}^* - B^*)] \geq \text{tr}\big[\big(\hat{B}^* - \hat{B}^*_r\big)^T \hat{\Sigma}_{\mathbf{X}}\big(\hat{B}^* - \hat{B}^*_r\big)\big]. \tag{14}$$

Next, we attempt to present the feasibility of the proposed BIC (12) in Theorem 1, which means that (12) can be used to select the true rank if we can obtain sufficiently many samples, i.e., $n \to \infty$. To this end, we need the following two assumptions.

Assumption 1: $\|\hat{\Sigma}_{\mathbf{X}} - \Sigma_{\mathbf{X}}\|_F^2 = O_p(n)$, where $\|\cdot\|_F$ means the standard matrix Frobenius norm.

Assumption 2: $\max(p,q)/n = o(\lambda^2)$.

Assumption 1 indicates that the empirical covariance matrix $\hat{\Sigma}_{\mathbf{X}}$ is consistent with $\Sigma_{\mathbf{X}}$. In practice, this can be guaranteed by the central limit theory. Assumption 2 is a technical condition, which is provided in [1]. Because any estimation obtained by the penalized method for a fixed tuning parameter λ is biased, Assumption 2 guarantees that $\lambda \to 0$ as $n \to \infty$ and exhibits the so-called $\sqrt{n}$-consistency of $\hat{B}^*_\lambda$, namely, $B^* = \hat{B}^*_\lambda + o_p(1/\sqrt{n})$. We then obtain the following Theorem 1.

Theorem 1: Under Assumptions 1 and 2, there exists λ_0 such that $\mathbb{P}(\inf_{\hat{r}_\lambda \neq r} BIC_\lambda > BIC_{\lambda_0}) \to 1$, as $n \to \infty$.

Proof: We will prove that there exists λ_0 (satisfying Assumption 2) that has the minimal BIC value of all λ leading to $\hat{r}_\lambda \neq r$ asymptotically. To this end, we will discuss the underestimate $\hat{r}_\lambda < r$ and overestimate $\hat{r}_\lambda > r$ cases.

Case 1 ($\hat{r}_\lambda < r$): Based on the proof of [1, Lemma 2, eq. (10), p. 334] and Assumption 2, there exists λ_1 such that $\hat{B}^*_{\lambda_1}$ is $\sqrt{n}$-consistent for B^* and $BIC_{\lambda_1}(\hat{B}^*_{\lambda_1})$ is $o(1)$. In addition, for any $\hat{r}_\lambda < r$, the following holds:

$$\begin{aligned}
&\text{BIC}_\lambda\big(\hat{B}^*_\lambda\big)\\
&\quad= \frac{1}{2}\text{tr}\big[\big(\hat{B}^* - \hat{B}^*_\lambda\big)^\top \hat{\Sigma}_{\mathbf{X}}\big(\hat{B}^* - \hat{B}^*_\lambda\big)\big] + \frac{\text{df}_\lambda \log(n)}{n}\\
&\quad\geq \frac{1}{2}\text{tr}\big[\big(\hat{B}^* - \hat{B}^*_\lambda\big)^\top \hat{\Sigma}_{\mathbf{X}}\big(\hat{B}^* - \hat{B}^*_\lambda\big)\big]\\
&\quad\geq \frac{1}{2}\text{tr}\big[\big(\hat{B}^* - \hat{B}^*_{\hat{r}_\lambda}\big)^\top \hat{\Sigma}_{\mathbf{X}}\big(\hat{B}^* - \hat{B}^*_{\hat{r}_\lambda}\big)\big]\\
&\quad\geq \frac{1}{2}\min_{r'<r}\big\{\text{tr}\big[\big(\hat{B}^* - \hat{B}^*_{r'}\big)^\top \hat{\Sigma}_{\mathbf{X}}\big(\hat{B}^* - \hat{B}^*_{r'}\big)\big]\big\}\\
&\quad\xrightarrow{p} \frac{1}{2}\min_{r'<r}\big\{\text{tr}\big[\big(B^* - B^*_{r'}\big)^\top \Sigma_{\mathbf{X}}\big(B^* - B^*_{r'}\big)\big]\big\}\\
&\quad> 0
\end{aligned} \tag{15}$$

where the last inequality (15) follows from the continuous mapping theorem [32, Th. 5.5, p. 75]. Consequently, $\inf_{\hat{r}_\lambda < r} BIC_\lambda(\hat{B}^*_\lambda) > BIC_{\lambda_1}(\hat{B}^*_{\lambda_1})$ is satisfied with probability approaching one as n increases to ∞ in this case.

Case 2 ($\hat{r}_\lambda > r$): In this case, we obtain an overfitted model with $\text{df}_\lambda - \text{df}_{\lambda_2} \geq 1$, where λ_2 satisfies Assumption 2. The difference between $\text{BIC}_\lambda(\hat{B}^*_\lambda)$ and $\text{BIC}_{\lambda_2}(\hat{B}^*_{\lambda_2})$ is

$$\begin{aligned}
&n\big(\text{BIC}_\lambda\big(\hat{B}^*_\lambda\big) - \text{BIC}_{\lambda_2}\big(\hat{B}^*_{\lambda_2}\big)\big)\\
&\quad= \frac{n}{2}\text{tr}\big[\big(\hat{B}^* - \hat{B}^*_\lambda\big)^\top \hat{\Sigma}_{\mathbf{X}}\big(\hat{B}^* - \hat{B}^*_\lambda\big)\big]\\
&\qquad-\frac{n}{2}\text{tr}\big[\big(\hat{B}^* - \hat{B}^*_{\lambda_2}\big)^\top \hat{\Sigma}_{\mathbf{X}}\big(\hat{B}^* - \hat{B}^*_{\lambda_2}\big)\big]\\
&\qquad+(\text{df}_\lambda - \text{df}_{\lambda_2})\log(n)\\
&\quad\geq \frac{n}{2}\ \text{tr}[\big(\hat{B}^* - \hat{B}^*_{\hat{r}_\lambda}\big)^\top \hat{\Sigma}_{\mathbf{X}}\big(\hat{B}^* - \hat{B}^*_{\hat{r}_\lambda}\big)]\\
&\qquad-\frac{n}{2}\ \text{tr}\big[\big(\hat{B}^* - \hat{B}^*_{\lambda_2}\big)^\top \hat{\Sigma}_{\mathbf{X}}\big(\hat{B}^* - \hat{B}^*_{\lambda_2}\big)\big]\\
&\qquad+(\text{df}_\lambda - \text{df}_{\lambda_2})\log(n)\\
&\quad\geq \frac{n}{2}\ \text{tr}\big[\big(\hat{B}^* - \hat{B}^*_{\hat{r}_\lambda}\big)^\top \hat{\Sigma}_{\mathbf{X}}\big(\hat{B}^* - \hat{B}^*_{\hat{r}_\lambda}\big)\big]\\
&\qquad-\frac{n}{2}\text{tr}\big[\big(\hat{B}^* - \hat{B}^*_{\lambda_2}\big)^\top \hat{\Sigma}_{\mathbf{X}}\big(\hat{B}^* - \hat{B}^*_{\lambda_2}\big)\big] + \log(n)\\
&\quad\geq \inf_{r'>r}\frac{n}{2}\text{tr}\big[\big(\hat{B}^* - \hat{B}^*_{r'}\big)^\top \hat{\Sigma}_{\mathbf{X}}\big(\hat{B}^* - \hat{B}^*_{r'}\big)\big]\\
&\qquad-\frac{n}{2}\text{tr}\big[\big(\hat{B}^* - \hat{B}^*_{\lambda_2}\big)^\top \hat{\Sigma}_{\mathbf{X}}\big(\hat{B}^* - \hat{B}^*_{\lambda_2}\big)\big] + \log(n).
\end{aligned}$$

Note that $\hat{B}^*_{r'}$ is $\sqrt{n}$-consistent for any $r' > r$ and that the first two terms of the last expression are both $O(1)$. However, $\log(n) \to \infty$ when $n \to \infty$. Consequently, for $\hat{r}_\lambda > r$, λ_2 satisfies $\mathbb{P}(\inf_{\hat{r}_\lambda > r} \text{BIC}_\lambda > \text{BIC}_{\lambda_2}) \to 1$. Combining the above-mentioned two cases, we can choose $\lambda_0 = \lambda_1$ for the underestimate case and $\lambda_0 = \lambda_2$ for the overestimate case. The proof of Theorem 1 is completed. □

Theorem 1 implies that as n increases, with probability approaching one, the BIC-type criterion (12) obtains a proper λ_0 satisfying $\hat{r}_{\lambda_0} = r$ if df_λ is known. However, it is not easy to estimate df_λ of (9). Fortunately, using the idea in [34], we can obtain an unbiased estimate of df_λ in Section III-B.

B. Effective Degrees of Freedom

As already mentioned, to apply the proposed BIC (12) to PMLSA, we need to estimate df_λ of PMLSA. The degrees of freedom, which can be used as a powerful tool to quantify the complexity of a modeling procedure, have been extensively studied (see [12], [33], [34]). In the case of the univariate linear regression model, the degree of freedom is the number of estimated parameters p. However, in general, it is difficult to give the exact relationship between the degrees of freedom and the number of free parameters [35]. Therefore, estimating df_λ of the models included in (2) is of great difficulty.

For some simple case, Yuan *et al.* [1] used

$$\hat{\text{df}}_\lambda = q\,\text{tr}(\mathbf{X}(\mathbf{X}^\top\mathbf{X} + 2\lambda R)^{-1}\mathbf{X}^\top) \tag{16}$$

to estimate the degrees of freedom of (3) with $W = I$, where R is a matrix calculated by the nonzero singular values of the minimizer of (3). However, this technique suffers from the problem that the estimated rank is usually larger than the true rank [1], [34]. To remedy this issue, Mukherjee *et al.* [34] considered the problem of estimating df_λ for a broad class of multivariate regression models [includes (3)] with the form

$$\hat{\mathbf{Y}}(\lambda) = \hat{\mathbf{Y}}\sum s_k(\sigma_k, \lambda)v_k v_k^\top \tag{17}$$

where $\hat{\mathbf{Y}}$ and $\hat{B}$ denote the unpenalized prediction of $\mathbf{Y}$ and B, the singular value decomposition (SVD) of B is $B = U\text{diag}(\sigma_k)V^T$, v_k denotes the kth column of V, and $s_k(\sigma_k, \lambda)$ is a function of σ_k and λ. They further proposed an unbiased estimate of df for any multivariate regression model having the property of (17). However, other models, such as multivariate quantile regression [15] and robust multivariate regression [14], are not special cases of (17).

In the following Theorem 2, we derive an unbiased estimate of df_λ of PMLSA by two steps. In the first step, we prove that the solution of PMLSA satisfies (17). Then, in the second step, we use the results provided in [34] to give a precise estimation of df_λ of PMLSA.

Theorem 2: df_λ of PMLSA (9) has an unbiased estimate provided by

$$\begin{aligned}\hat{\text{df}}_\lambda = \max(r(\mathbf{X}), q)\sum_{k=1}^{\hat{r}_\lambda} s_k &+ \sum_{k=1}^{\hat{r}_\lambda}\sum_{l=\hat{r}_\lambda+1}^{\hat{r}} \frac{s_k(\sigma_k^2 + \sigma_l^2)}{\sigma_k^2 - \sigma_l^2}\\ &+ \sum_{k=1}^{\hat{r}_\lambda}\sum_{l\neq k}^{\hat{r}} \frac{\sigma_k^2(s_k - s_l)}{\sigma_k^2 - \sigma_l^2} + \sum_{k=1}^{\hat{r}_\lambda}\sigma_k\dot{s}_k, \quad \hat{r}_\lambda < \hat{r}\end{aligned} \tag{18}$$

and

$$\begin{aligned}\hat{\text{df}}_\lambda = \max(r(\mathbf{X}), q)\sum_{k=1}^{\hat{r}_\lambda} s_k &+ \sum_{k=1}^{\hat{r}_\lambda}\sum_{l\neq k}^{\hat{r}} \frac{\sigma_k^2(s_k - s_l)}{\sigma_k^2 - \sigma_l^2}\\ &+ \sum_{k=1}^{\hat{r}_\lambda}\sigma_k\dot{s}_k, \quad \hat{r}_\lambda = \hat{r}\end{aligned} \tag{19}$$

where $r(\mathbf{X}) = \text{rank}(\mathbf{X}), \hat{r}_\lambda = \text{rank}(\hat{B}_\lambda^*)$, $\hat{r} = \text{rank}(\hat{B}^*)$, $s_k = s_k(\sigma_k, \lambda)$, and $\dot{s}_k = \partial s_k(\sigma_k, \lambda)/\partial\sigma_k$.

Proof: Recall that (9) can be formulated as (10), that is

$$Q(B) = \frac{1}{2}\text{tr}[(\hat{B}^* - IB)^T\hat{\Sigma}_{\mathbf{X}}(\hat{B}^* - IB)] + \lambda\sum_{i=1}^{\min(p,q)}\sigma_i(B)$$

where I is the identity matrix. Compared with (3), if we set $\mathbf{Y} = \hat{B}^*$ and $\mathbf{X} = I$, then (10) can be seen as an orthogonal design of (3). Suppose that $\hat{B}_\lambda^* = \arg\min Q(B)$. Thus, [1, Lemma 2] indicates that for the orthogonal design of (3), $\hat{B}_\lambda^*$ has the analytic form

$$\hat{B}_\lambda^* = \hat{U}^*\hat{D}^*(\lambda)\hat{V}^{*\top} \tag{20}$$

where the SVD of the unpenalized estimate $\hat{B}^* = \hat{U}^*\hat{D}^*\hat{V}^{*\top}$ and $\hat{D}^*(\lambda) = \text{diag}(\max(\sigma_k(\hat{D}^*) - \lambda, 0))$.

Next, we show that $\hat{B}_\lambda^*$ satisfies (17). This is because

$$\begin{aligned}\hat{B}_\lambda^* &= \hat{U}^*\hat{D}^*(\lambda)\hat{V}^{*\top}\\ &= \hat{U}^*\hat{D}^*\text{diag}\left(\frac{\max(\sigma_k(\hat{D}^*) - \lambda, 0)}{\sigma_k(\hat{D}^*)}\right)\hat{V}^{*\top}\\ &= \hat{U}^*\hat{D}^*\hat{V}^{*\top}\hat{V}^*\text{diag}\left(\frac{\max(\sigma_k(\hat{D}^*) - \lambda, 0)}{\sigma_k(\hat{D}^*)}\right)\hat{V}^{*\top}\\ &= \hat{B}^*\hat{V}^*\text{diag}(s_k(\sigma_k(\hat{D}^*), \lambda))\hat{V}^{*\top}\\ &= \hat{B}^*\sum_{k=1}^{\min(p,q)} s_k(\sigma_k(\hat{D}^*), \lambda)v_k^*v_k^{*\top}\end{aligned}$$

with $s_k(\sigma_k(\hat{D}^*), \lambda) = \max(\sigma_k(\hat{D}^*) - \lambda, 0)/\sigma_k(\hat{D}^*)$. Given the above-mentioned derivations, we can directly use the unbiased estimation of df_λ, (18) and (19) derived in [33, Th. 4], to estimate df_λ of PMLSA. □

Putting formulations (18) and (19) into the BIC-type criterion (12), the criterion can be used to select the tuning parameter λ.

IV. Discussion

A. Related Work

This paper is naturally related to a number of learning systems (e.g., penalized multivariate regression models). By penalizing the rank of the regression coefficient matrix B, penalized multivariate regression indeed solves the following optimization problem:

$$\min_{\{B:\ \text{rank}(B)\leq r\}}\left\{\frac{1}{2}\text{tr}[(\mathbf{Y} - \mathbf{X}B)^\top(\mathbf{Y} - \mathbf{X}B)]\right\} \tag{21}$$

which is usually called the reduced-rank regression. Although the rank constraint makes (21) a nonconvex optimization problem, it allows for a closed-form solution, as described in [11]. Therefore, the main challenge in the reduced-rank regression (21) method is to select a proper tuning parameter r. In fact, there are two significant drawbacks to the reduced-rank regression. One drawback is that this type of procedure is unstable due to its discrete constraint [36], [37], namely, small changes in data sets lead to different estimates. The other drawback is that we can use many different types of criteria to select the tuning parameter [12], [14], [38]. However, no method can guarantee that the selected rank is consistent with the true rank theoretically.

To improve the stability of low-rank coefficient matrix estimate, Yuan *et al.* [1] proposed a nuclear-norm penalized low-rank coefficient matrix estimation method in (3). Penalizing the nuclear norm can be seen as a convex relaxation of penalizing the rank, which encourages sparsity among the singular values to achieve simultaneous rank reduction and shrink the coefficient matrix estimate. However, it is computationally intensive and tends to overestimate the rank [12], [34], [39]. Although Toh and Yun [26] proposed an accelerated proximal gradient algorithm to solve (3), which dramatically improves the convergence speed, how to select a proper λ to obtain a more efficient estimation remains an open problem.

Recently, the low-rank coefficient matrix estimate problem was generalized to many other regression models for various purposes. Engle and Manganelli [40] and Chao *et al.* [15] introduced a nuclear-norm penalized multivariate QR for tail event curve estimation. In such models, the loss function is written as $L(B_j) = \rho_\tau(y_{ij} - \mathbf{x}_i^\top B_j), j = 1, \ldots, q$, where $\rho_\tau(u) = u(\tau - \mathbf{1}(u < 0))$ and $\mathbf{1}$ is the commonly accepted indicator function. Furthermore, Zhao *et al.* [38] and She and Chen [14] proposed a robust version of penalized low-rank regression models by employing the least absolute derivation loss $L(B_j) = |y_{ij} - \mathbf{x}_i^\top B_j|$, which attempts to mitigate data corruption. Although all the above-mentioned examples demonstrated the flexibility of the penalized low-rank regression models, this flexibility may lead to difficulties in solving the model [15] and further selecting the true rank of B. How to efficiently solve the models and derive a suitable criterion for the model complexity for selecting the tuning parameter has yet to be completely addressed. Thus, the aforementioned problems have become a hurdle to the widespread application of the models.

To overcome these hurdles, we consider the unified problem (2) including all the aforementioned models [1], [14], [15], [38], [40]. The main contributions of this paper are in the use of PMLSA to solve all such models included in (2) with an efficient tuning parameter selection procedure. We shall also show that PMLSA can improve the estimation accuracy and rank selection consistency of the nuclear-norm penalized multivariate regression [1], multivariate robust regression [14], [38], and multivariate QR [15], [40] methods through an extensive numerical study.

B. Implementation

This section presents some considerations that are useful for implementing the PMLSA and its BIC criterion.

1) For real data analysis, we consider a projected regression problem, i.e., centralization is necessary for both $\mathbf{X}$ and $\mathbf{Y}$.
2) Sometimes, $f(e_j)$ is very small, which leads to the problem of underflows during computation. In practice, a better method is to adjust $f(e_j)$ as $f'(e_j) = f(e_j)/\min_j(f(e_j))$.
3) The proposed BIC-type criterion for PMLSA is proven to consistently select the true rank of B under mild conditions. However, we cannot guarantee that the estimation of B can result in better prediction performance because PMLSA is simply an approximation of the unified framework (2). To this end, we suggest using a two-step method in practice. Specifically, in the first step, we choose the rank $\hat{r}$ by PMLSA with the proposed BIC and obtain $\hat{B}$ by the "no penalty" method [i.e., set $\lambda = 0$ in (2)]. In the second step, $\hat{B}^{\hat{r}}$, as the $\hat{r}$-rank approximation of $\hat{B}$, is constructed by
$$\hat{B}^{\hat{r}} = U\mathrm{diag}(\hat{\lambda}_1, \ldots, \hat{\lambda}_r)V^\top \tag{22}$$
where the SVD decomposition of $\hat{B} = U\Lambda V^\top$ and $\hat{\lambda}_1, \ldots, \hat{\lambda}_r$ are the first $\hat{r}$ singular values in Λ. This two-step procedure has been considered as a standard method for handling this type of problem [41], [42].

Based on the above statements of PMLSA's implementation, we conduct a number of experimental studies to demonstrate the rank selection and prediction performance of PMLSA using the proposed BIC-type criterion.

V. Experimental Study

We present extensive numerical studies to compare PMLSA's finite sample performance with different models included in (2). In the following, we first repeat each simulation 200 times, and the mean and standard deviation of the model error (ME), median relative ME (MRME), coefficient error (CE), and mean squared error (MSE) are recorded to measure their performance. These are denoted as follows:
$$\begin{aligned}\mathrm{ME} &= \mathrm{tr}((\hat{B} - B)^\top \mathrm{COV}(\mathbf{x})(\hat{B} - B))\\ \mathrm{MRME} &= \mathrm{median}(|\hat{B} - B|/B)\\ \mathrm{CE} &= \mathrm{tr}((\hat{B} - B)^\top(\hat{B} - B))\end{aligned}$$
and
$$\mathrm{MSE} = \frac{1}{np}\|\mathbf{Y}_{\mathrm{test}} - \mathbf{X}_{\mathrm{test}}\hat{B}\|_F^2$$
where $\{\mathbf{Y}_{\mathrm{test}}, \mathbf{X}_{\mathrm{test}}\}$ are the generated testing samples. Then, two real-world data examples, i.e., gene expression data and stock market data, are used to illustrate the efficiency and flexibility of PMLSA.

A. Simulation 1 (Least Squares Regression)

In this experiment, we justify the goodness of our model in the multivariate regression with the least squares loss. First, n samples are independently generated from (1). An $p \times q$ matrix is simulated from $N(0, \sigma^2)$, and B is generated by replacing its singular values with $(\lambda_1, \ldots, \lambda_r, 0, \ldots, 0)$, where λ_is are randomly sampled from $\{2, 3, 4, 5, 6, 7, 8, 9, 10\}$. Then, $\mathbf{X}$ is drawn from $N(0, \Sigma)$, where $\Sigma_{ij} = 0.5^{|i-j|}$. We compare the performance of the following methods under different settings of $(n, p, q, r, n_{\mathrm{test}})$, where n_{test} is the number of testing samples.

1) *PMLSA + BIC:* Use Algorithm 1 and the BIC criterion (12) with df_λ estimated by (18) and (19).
2) *PMLSA + CV:* Use Algorithm 1 and the fivefold cross validation (CV) procedure to choose the tuning parameter λ.

TABLE I

SIMULATION RESULTS FOR LEAST SQUARES REGRESSION

(n, p, q, r, n_{test})	Method	PMLSA+BIC	PMLSA + CV	LS+GCV	No-penalty	Two-step
(200,10,10,4,200)	Rank	**4.78**	7.65	7.42	10	4.78
		(0.524)	(1.486)	(2.811)	(0)	(0.524)
$\sigma^2 = 1$	ME	0.707	0.463	0.64	0.523	**0.43**
		(2.507)	(0.787)	(1.101)	(1.096)	(0.813)
	MRME	0.096	0.075	0.09	0.083	**0.072**
		(0.032)	(0.021)	(0.026)	(0.022)	(0.022)
	CE	1.27	0.737	1.111	0.837	**0.66**
		(0.414)	(0.166)	(0.5)	(0.13)	(0.122)
	MSE	0.070	0.085	0.068	0.053	**0.045**
		(0.015)	(0.025)	(0.021)	(0.008)	(0.008)
(400,10,10,4,400)	Rank	**4.51**	7.31	6.34	10	4.51
		(0.522)	(1.862)	(2.836)	(0)	(0.522)
$\sigma^2 = 1$	ME	0.333	0.248	0.343	0.254	**0.206**
		(0.08)	(0.055)	(0.11)	(0.035)	(0.042)
	MRME	0.063	0.053	0.065	0.056	**0.047**
		(0.017)	(0.014)	(0.018)	(0.015)	(0.013)
	CE	0.593	0.403	0.608	0.407	**0.307**
		(0.181)	(0.114)	(0.257)	(0.067)	(0.077)
	MSE	0.034	0.042	0.034	0.026	**0.021**
		(0.007)	(0.014)	(0.010)	(0.004)	(0.004)
(200,100,50,10,200)	Rank	**10**	45.51	19.07	50	10
		(0)	(2.509)	(2.567)	(0)	(0)
$\sigma^2 = 0.25$	ME	27.471	5.558	11.312	12.713	**4.4**
		(4.343)	(0.33)	(1.7)	(0.459)	(0.236)
	MRME	0.344	0.149	0.222	0.239	**0.128**
		(0.04)	(0.017)	(0.029)	(0.031)	(0.018)
	CE	48.321	8.444	20.058	21.102	**6.276**
		(7.771)	(0.599)	(3.2)	(0.8)	(0.372)
	MSE	0.551	0.421	0.235	0.254	**0.088**
		(0.093)	(0.053)	(0.042)	(0.010)	(0.007)
(400,100,50,10,400)	Rank	**10.14**	42.96	10.9	50	10.14
		(0.377)	(1.601)	(1.21)	(0)	(0.377)
$\sigma^2 = 0.25$	ME	7.357	2.002	5.627	4.174	**1.411**
		(1.214)	(0.08)	(0.696)	(0.113)	(0.086)
	MRME	0.194	0.092	0.167	0.141	**0.075**
		(0.047)	(0.016)	(0.028)	(0.025)	(0.014)
	CE	13.815	3.018	10.588	6.936	**2.028**
		(2.081)	(0.145)	(1.421)	(0.219)	(0.173)
	MSE	0.147	0.128	0.117	0.083	**0.028**
		(0.029)	(0.012)	(0.014)	(0.002)	(0.002)

3) *LS + GCV:* Use Toh and Yun's algorithm [26] to solve (3) and the generalized CV (GCV) criterion with df_λ estimated by (16).

4) *No Penalty:* This is the ordinary least squares approach to estimating B.

5) *Two-Step Method:* Use the PMLSA + BIC method to obtain the estimated rank $\hat{r}$ and the no penalty method to estimate B; then, construct $\hat{B}^{\hat{r}}$ via (22).

The average results of the estimated rank, ME, MRME, and CE are summarized in Table I.

It can be easily observed from Table I that PMLSA + BIC demonstrates the best capability of choosing the true rank, which supports the assertion of Theorem 1. On the one hand, the average values of ME, MRME, and CE of PMLSA + BIC are all larger than those of LS + GCV estimation, which is the commonly used approach to solve (3) [1]. This is because the PMLSA is only an approximation of (3), and (3) itself uses the least squares loss in the model. On the other hand, the average values of ME, MRME, CE, and MSE obtained by the two-step method are all smaller than LS + GCV. This indicates that PMLSA + BIC can capture the more accurate low-rank structure of B, which leads to the better estimation of B via the two-step method.

B. Simulation 2 (Least Absolute Deviation Regression)

We consider the following optimization problem:

$$\min_B \left\{ \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{q} \left| y_{ij} - \mathbf{x}_i^\top B_j \right| + \lambda \sum_{i=1}^{\min(p,q)} \sigma_i(B) \right\} \tag{23}$$

which is the penalized least absolute deviation regression (PLADR) with the nuclear-norm penalty. In fact, this regression model is proposed for two fundamental motivations.

TABLE II

SIMULATION RESULTS FOR PLADR

(n, p, q, r, n_{test})	Method	PMLSA+BIC	PMLSA + CV	FST	No-penalty	Two-step
(400,100,50,4,400)	Rank	**5.58**	23.64	8.8	50	5.58
		(1.148)	(8.091)	(0.62)	(0)	(1.148)
	ME	12.592	8.448	7.674	31.775	**7.657**
		(2.025)	(0.992)	(0.918)	(0.885)	(1.626)
	MRME	0.393	0.396	0.387	0.701	**0.309**
		(0.07)	(0.057)	(0.055)	(0.175)	(0.062)
	CE	21.608	13.246	13.266	52.647	**12.335**
		(4.104)	(1.748)	(1.831)	(1.651)	(3.273)
	MSE	0.262	0.155	**0.138**	0.638	0.162
		(0.045)	(0.023)	(0.018)	(0.018)	(0.039)
(1000,100,50,4,1000)	Rank	**5.77**	24.37	8.56	50	5.77
		(1.081)	(10.185)	(0.756)	(0)	(1.081)
	ME	4.381	2.823	2.945	10.056	**2.285**
		(0.471)	(0.311)	(0.159)	(0.247)	(0.491)
	MRME	0.257	0.196	0.175	0.422	**0.181**
		(0.07)	(0.056)	(0.055)	(0.127)	(0.042)
	CE	8.019	4.332	4.517	16.705	**3.769**
		(1.126)	(0.971)	(0.382)	(0.442)	(1.014)
	MSE	0.087	0.056	**0.043**	0.201	0.047
		(0.011)	(0.006)	(0.003)	(0.005)	(0.009)
(400,100,50,10,400)	Rank	**12.73**	35.37	18.75	50	12.73
		(1.563)	(4.419)	(0.796)	(0)	(1.563)
	ME	26.28	**13.447**	15.204	31.783	14.897
		(2.806)	(1.064)	(0.845)	(0.922)	(1.753)
	MRME	0.341	0.253	0.329	0.383	**0.248**
		(0.041)	(0.025)	(0.026)	(0.049)	(0.028)
	CE	45.674	24.511	24.227	52.814	**23.62**
		(5.306)	(2.004)	(1.667)	(1.594)	(3.685)
	MSE	0.533	0.269	**0.250**	0.633	0.295
		(0.065)	(0.021)	(0.020)	(0.019)	(0.036)
(1000,100,50,10,1000)	Rank	**12.52**	34.82	16.57	50	12.52
		(1.251)	(4.229)	(0.64)	(0)	(1.251)
	ME	9.447	4.689	4.525	10.002	**4.332**
		(1.092)	(0.297)	(0.181)	(0.227)	(0.484)
	MRME	0.204	0.134	0.138	0.208	**0.130**
		(0.022)	(0.015)	(0.017)	(0.026)	(0.018)
	CE	17.478	7.259	7.614	16.565	**6.82**
		(2.245)	(0.558)	(0.373)	(0.421)	(0.963)
	MSE	0.187	0.094	**0.085**	0.200	0.086
		(0.022)	(0.006)	(0.004)	(0.005)	(0.011)

First, PLADR is a special case of penalized multivariate QR with $\tau = 0.5$ quantile [15]. Second, PLADR is also a penalized robust regression for addressing corruption data with a heavy tail [14].

In this simulation, we keep the same way of generating B and $\mathbf{X}$ as Simulation 1. Due to the aforementioned two motivations of PLADR, we consider that the noise independently draws from the t-distribution namely $t(5)$. We then compare the performance obtained by the following methods.

1) *PMLAS + BIC:* Use Algorithm 1 and BIC criterion (12) with df_λ (18) and (19).
2) *PMLAS + CV:* Use Algorithm 1 and fivefold CV to choose the tuning parameter.
3) *FST:* It is an algorithm to solve PLADR, uses [15, eq. (3.8)] to choose the tuning parameter.
4) *No Penalty:* We solve the model by LAD regression separately to obtain $\hat{B}^{\mathrm{LAD}}$.

Two-Step

5) *Method:* Use PMLSA + BIC method to obtain the estimated rank $\hat{r}$ and the no penalty method to estimate B and then construct $\hat{B}^{\hat{r}}$ via (22).

We should mention how to obtain the inputs $\hat{B}^{\mathrm{LAD}}$ and $f(e_j) = (\sigma_j(0.5)^2)^{-1}$ of PMLSA. We can estimate $f(e_j)$ by the following function [43]:

$$\hat{\sigma}_j = \frac{n(\hat{\varepsilon}_{j,\kappa_2} - \hat{\varepsilon}_{j,\kappa_1})^2}{4Z^2_{1-\alpha/2}} \tag{24}$$

where $\tau = 0.5$ is used in this case, $\kappa_1 = [n\tau - Z_{1-\alpha/2}(n\lambda(1-\lambda))^{1/2}]$, $\kappa_2 = [n\tau + Z_{1-\alpha/2}(n\lambda(1-\lambda))^{1/2}]$, $\hat{\varepsilon}_{j,\kappa_1}$ and $\hat{\varepsilon}_{j,\kappa_2}$ are the κ_1 and κ_2 quantile of the residues based on $\hat{B}^{LAD}$, and Z_α is the α quantile of the standard normal distribution. Here, we choose $\alpha = 0.05$. The averaged results over 200 trials are reported in Table II.

In Table II, PMLSA + BIC still significantly outperforms other competing methods in terms of the accuracy of the estimated rank. In addition, the average values of ME, MRME, CE, and MSE of the two-step method are almost smaller than the factorisable sparse tail (FST) and no penalty methods, which indicates that this method can be applied in robust multivariate regression [14], [38].

C. Simulation 3 (Quantile Regression)

In this simulation, we consider the model

$$\min_{B}\left\{\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{q}\rho_\tau\left(y_{ij}-\mathbf{x}_i^\top B_j\right)+\lambda\sum_{i=1}^{\min(p,q)}\sigma_i(B)\right\} \tag{25}$$

where $\rho(w) = w[\tau - \mathbf{1}(w < 0)]$. Following the same simulation setting in [15], we set $p = q = 50$ and $n = 400, 1000$. We first build two coefficient matrices B_1 and B_2 with ranks $r_1 = 5$ and $r_2 = 15$, respectively, as follows. We generate $a_g, b_h \in \mathbb{R}^p$, $\alpha_g, \beta_h \in \mathbb{R}^q (g = 1, 2, \ldots, r_1, h = 1, 2, \ldots, r_2)$ with each element of a_g, b_h, α_g, and β_h independently drawn from $U(0, 1)$. Thus

$$B_{1,j} = \sum_{g=1}^{r_1}\alpha_{g,j}a_g \text{ and } B_{2,j} = \sum_{h=1}^{r_2}\beta_{h,j}b_h$$

where $B_{1,j}$ and $B_{2,j}$ are the jth column of B_1 and B_2, respectively. Then, $\mathbf{x}$ is drawn from $N(0, \Sigma)$ with $\Sigma_{ij} = 0.5^{|i-j|}$, and y_{ij} is generated by

$$\begin{aligned} y_{ij} &= \operatorname{sign}(U_{ij} - 0.5)\mathbf{x}_i^T \\ &\quad \times [B_{1,j}1(U_{ij} < 0.5) + B_{2,j}1(U_{ij} \geq 0.5)] \\ &\quad + N(-q_\tau, 0.4^2) \end{aligned} \tag{26}$$

where U_{ij} is simulated by the uniform distribution $U(0, 1)$ and $q_\tau = \text{quantile}(\tau, N(0, 0.4^2))$. Note that we can obtain y_{ij} on $\mathbf{x}_i$, that is

$$y_{ij}(\tau) = -\mathbf{x}_i^\top B_{1,j}, \quad \text{where } \tau < 0.5 \tag{27}$$

$$y_{ij}(\tau) = \mathbf{x}_i^\top B_{2,j}, \quad \text{where } \tau \geq 0.5. \tag{28}$$

We compare the estimations of the 0.25 and 0.75 quantiles by the PMLSA + BIC, PMLSA + CV, FST, no penalty, and two-step methods. The input of PMLSA can be calculated by (24) and unpenalized QR. The numerical performance over 200 trials is averaged and given in Table III.

In Table III, compared with the FST and no penalty methods, PMLSA + BIC still has the best capability in choosing the true rank. Moreover, the average values of the MRME, CE, and MSE of the two-step methods are all smaller than those of the FST and no-penalty methods. This implies that PMLSA+BIC can capture a more accurate low-rank structure in the multivariate QR model (25).

D. Arabidopsis Thaliana Data

In this section, we apply the model (3) to the genetic association study of Wille *et al.* [44]. The goal of this microarray experiment is to understand the regulatory control mechanisms in the isoprenoid gene network of the plant *Arabidopsis thaliana*, more commonly known as thale cress or mouse-ear cress. Isoprenoids have many important biochemical functions in plants. To monitor the gene expression levels, 118 gene chip microarray experiments were performed. The predictors consist of 39 genes from two isoprenoid biosynthesis pathways, i.e., mevalonic acid (MVA) and motor evoked potential (MEP), and the responses consist of the expression levels of 795 genes from 56 metabolic pathways, many of which are downstream of the two pathways considered as predictors. Thus, some of the predictor genes have been shown to present high correlations resulting in the low-rank structure of the coefficient matrix [34], [44].

We select two downstream pathways, i.e., the carotenoid and phytosterol pathways, as our responses. It has already been shown experimentally that the carotenoid pathway is strongly linked to the MEP pathway, whereas the phytosterol pathway is significantly related to the MVA pathway (see [44] and the references therein for a detailed discussion). Finally, we have 118 observations with $p = 39$ predictors and $q = 36$ responses, which are all logarithmically transformed to reduce the skewness. We also standardized the responses to make them comparable. The data set is randomly split into a training set with 50% of the samples and a testing set with the remaining samples. We then fit the model (3) using the PMLSA + BIC, PMLSA + CV, LS + GCV, no penalty, and two-step methods on the training samples and then apply them to measure their prediction performance on the testing set. The performance metric is the MSE. The average estimated rank over 200 independent trials is recorded in Table IV.

It can be observed from Table IV that PMLSA+BIC achieves the best prediction performance in terms of the smallest MSE. Furthermore, we can see that the estimated ranks of PMLAS + BIC, PMLSA + CV, and LS + GCV are all much smaller than 36, which means that they indeed found the low-rank structure underlying the data ensemble. Compared with all other methods, PMLAS + BIC gives the smallest estimated rank while simultaneously achieving the smallest MSE. These observations imply that the PMLSA + BIC method can interpret the data very well. Note that the two-step method cannot give the best performance in terms of MSE. This may be because the sample size is relatively small for this real data example.

E. Stock Market Data

In this section, we evaluate the proposed PMLSA for the penalized multivariate QR (25) on a stock market data set, i.e., a set of stock prices consisting of 119 major U.S. banks. The data set can be downloaded directly from Simone Manganelli's website.[1] The data period that we used is from October 30, 2003 to August 6, 2010 and includes 1767 trading days. Hence, there are 1767 closing prices for each stock.

Denote $\text{lar}_{\text{it}} = \log(\text{Price}_{\text{it}}) - \log(\text{Price}_{\text{it}-1})$ as the log asset return of the ith stock at time point t, where $i = 1, \ldots, 119$, $t = 2, \ldots, 1767$, and Price_{it} as the price of the ith stock at time point t. To predict the lar, the conditional autoregressive value at risk (CAViaR) (MQ-CAViaR) model, which allows a

[1] http://www.simonemanganelli.org/Simone/Research.html

TABLE III
SIMULATION RESULTS FOR QR

(τ, n, r, n_{test})	Method	PMLSA + BIC	PMLSA +CV	FST	No-penalty	Two-step
$(0.25, 400, 5, 400)$	Rank	**6.22**	32.82	21.95	50	6.22
		(0.887)	(5.465)	(1.289)	(0)	(0.887)
	ME	10.508	8.743	34.859	19.864	**7.016**
		(1.233)	(1.772)	(3.381)	(2.927)	(2.991)
	MRME	0.043	0.037	0.067	0.061	**0.031**
		(0.004)	(0.005)	(0.005)	(0.004)	(0.004)
	CE	15.596	11.657	37.207	32.608	**10.66**
		(2.053)	(2.919)	(2.949)	(4.943)	(5.16)
	MSE	0.204	0.172	0.693	0.387	**0.131**
		(0.029)	(0.032)	(0.073)	(0.035)	(0.031)
$(0.75, 400, 15, 400)$	Rank	**16.141**	40.242	30.071	50	16.141
		(0.958)	(3.127)	(1.35)	(0)	(0.958)
	ME	24.477	14.42	71.675	20.036	**13.959**
		(4.956)	(2.968)	(5.645)	(3.171)	(3.19)
	MRME	0.023	0.016	0.031	0.02	**0.015**
		(0.002)	(0.001)	(0.001)	(0.001)	(0.001)
	CE	39.942	21.579	75.634	32.723	**20.649**
		(7.92)	(4.718)	(5.155)	(5.175)	(5.05)
	MSE	0.507	0.285	1.442	0.396	0.279
		(0.122)	(0.046)	(0.115)	(0.049)	(0.051)
$(0.25, 1000, 5, 1000)$	Rank	**6.32**	32.05	16.43	50	6.32
		(0.709)	(5.999)	(1.265)	(0)	(0.709)
	ME	3.095	2.208	14.333	4.535	**1.467**
		(0.324)	(0.372)	(1.22)	(0.148)	(0.175)
	MRME	0.024	0.019	0.043	0.03	**0.015**
		(0.002)	(0.002)	(0.003)	(0.002)	(0.002)
	CE	4.874	2.875	15.482	7.281	**2.104**
		(0.682)	(0.596)	(0.967)	(0.272)	(0.383)
	MSE	0.062	0.045	0.288	0.090	**0.029**
		(0.006)	(0.007)	(0.031)	(0.003)	(0.004)
$(0.75, 1000, 15, 1000)$	Rank	**16.07**	38.2	22.54	50	16.07
		(0.856)	(4.192)	(1.105)	(0)	(0.856)
	ME	6.439	3.298	49.104	4.468	**2.892**
		(0.905)	(0.427)	(3.051)	(0.149)	(0.19)
	MRME	0.012	0.008	0.026	0.01	**0.007**
		(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
	CE	10.935	4.832	52.414	7.174	**4.105**
		(1.715)	(0.858)	(3.179)	(0.291)	(0.387)
	MSE	0.137	0.068	0.989	0.091	**0.058**
		(0.023)	(0.009)	(0.072)	(0.003)	(0.004)

TABLE IV
RESULTS FOR *Arabidopsis Thaliana* DATA

Method	PMLSA+ BIC	PMLSA+ CV	LS + GCV	No-penalty	Two-step
Rank	14.08	16.29	24.21	36	14.08
sd(Rank)	(0.813)	(4.723)	(1.166)	(0)	(0.813)
MSE	0.169	0.171	0.188	0.392	0.363
sd(MSE)	(0.014)	(0.015)	(0.023)	(0.065)	(0.062)

sequence of conditional quantiles of log asset returns to depend on each other, has been proposed in [40], [45]. Formally, it can be formulated as

$$\mathrm{lar}_{\mathrm{it},\tau} = \beta_1 |\mathrm{lar}_{\mathrm{it}-1}| + \beta_2 \mathrm{lar}^{-}_{\mathrm{it}-1} \tag{29}$$

where $\mathrm{lar}_{\mathrm{it},\tau}$ is the τ quantile of the ith lar at time t and $\mathrm{lar}^{-}_{\mathrm{it}} = \min(\mathrm{lar}_t, 0)$ is the negative part of $\mathrm{lar}_{\mathrm{it}}$. If we consider all lar_{t-1} and lar^{-}_{t-1} for each stock as explanatory variables, then we have $p = 119 \times 2 = 238$ explanatory variables. Therefore, based on model (29), we can construct a design matrix $\mathbf{X} \in \mathbb{R}^{n \times p}$ with $p = 238$ and $\mathbf{Y} \in \mathbb{R}^{n \times q}$ with $q = 119$. In fact, the famous three-factor model indicates the existence of an underlying mechanism for the stock market, which imposes the low-rank structure on the coefficient matrix [15], [46].

In this data set, we choose the first $n = 1365$ time points as training samples and the last 400 time points as testing samples. We apply the penalized multivariate regression

TABLE V
RESULTS FOR STOCK MARKET DATA

τ	Method	PMLSA	FST	No-penalty	Two-step
0.9	Rank	2	1	119	2
	PE	**0.00723**	0.0105	0.00836	0.00669
0.1	Rank	5	1	119	5
	PE	**0.00597**	0.00958	0.00826	0.00687

model (25) on the training set. We adopt the PMLSA+BIC, FST, and no penalty methods to solve the penalized multivariate QR (25). Here, $\tau = 0.1$ and 0.9 are considered. Then, we use the prediction error

$$\mathrm{PE} = \frac{\sum_i \sum_j \rho_\tau\left(y_{ij}^{\text{test}} - \mathbf{x}_i^{\text{test}\top} B_j\right)}{nq}$$

to evaluate the performance and summarize the results in Table V.

From Table V, we can find that the estimated rank of the no penalty method is 119, which is very high. This means that it cannot exploit the low-rank structure underlying the data ensemble. Compared with the no penalty method, both PMLSA + BIC and FST can estimate a rank that is extremely small when $\tau = 0.1$ and 0.9. This implies that the low-rank structure of the stock market data set indeed exists. Furthermore, the PE of PMLSA + BIC decreases by approximately 20% compared to that of FST. This may be because the PMLSA does not regressively shrink the rank of the coefficient matrix as FST does, considering that the estimated rank of FST is one on the data set.

VI. CONCLUSION

Low-rank coefficient matrix estimation in multivariate regression has encountered two crucial issues: 1) how to construct a fast algorithm for solving the model and 2) how to derive a corresponding criterion for consistently selecting the tuning parameter or the true rank.

In this paper, we propose a PMLSA method to develop a unified framework (2) that includes a number of multivariate regression models. Using a unified form of the least squares loss to approximate the loss of different models, we employ an efficient algorithm named matrix FISTA to solve the resulting model. We also develop an effective method to calculate an unbiased estimate of df_λ and further propose a BIC-type criterion to choose the proper tuning parameter λ. Moreover, we justify such procedure consistency in finding the true rank. Simulations demonstrate the superiority of PMLAS over other competing methods in choosing the rank, and a real data example further reveals its stability and validity.

Along this line, this method presents several desirable research directions for the future study. First, in statistics, M-estimators are a broad class of estimators, which are obtained as the minima of the sums of several loss functions [47]. Therefore, the unified framework (2) can be seen as a special penalized M-estimator [18], [47], [48], which treats many different models as special cases. How to extend the PMLSA to address more general models is an important problem. Second, the proposed PMLSA is based on the property that $\Sigma_j^{-1} = f(e_j)\mathrm{COV}(\mathbf{x})$, which may not hold in certain regression models such as logistic regression. However, $\Sigma_j^{-1} = f(e_j)\mathrm{COV}(\mathbf{x})$ is related to the Fisher information matrix in the M-estimator [47]. This may represent one way of overcoming this difficulty. Third, the technical Assumption 2 supposes that q and p cannot be set very large beyond $O(n)$. However, this type of application (large p and q) has appeared in ultrahigh-dimensional data analysis [49], [50]. Generalizing the PMLSA for handling large p and q problems is an urgent issue. We will investigate all such issues in the future.

ACKNOWLEDGMENT

The authors would like to thank the associate editor and the three anonymous reviewers for their careful work and constructive comments, which greatly helped them to improve the presentation of this paper.

REFERENCES

[1] M. Yuan, A. Ekici, Z. Lu, and R. Monteiro, “Dimension reduction and coefficient estimation in multivariate linear regression,” *J. Roy. Statist. Soc. B, Stat. Methodol.*, vol. 69, no. 3, pp. 329–346, 2007.

[2] A. Lakhina *et al.*, “Structural analysis of network traffic flows,” *ACM SIGMETRICS*, vol. 32, no. 1, pp. 61–72, 2004.

[3] E. J. Candès and B. Recht, “Exact matrix completion via convex optimization,” *Found. Comput. Math.*, vol. 9, no. 6, pp. 717–772, 2009.

[4] W. Cao, Y. Wang, C. Yang, X. Chang, Z. Han, and Z. Xu, “Folded-concave penalization approaches to tensor completion,” *Neurocomputing*, vol. 152, pp. 261–273, Mar. 2015.

[5] Y. Wang, J. Peng, Q. Zhao, D. Meng, Y. Leung, and X.-L. Zhao, “Hyperspectral image restoration via total variation regularized low-rank tensor decomposition,” *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 11, no. 4, pp. 1227–1243, Apr. 2018.

[6] Y. Su, X. Gao, X. Li, and D. Tao, “Multivariate multilinear regression,” *IEEE Trans. Syst., Man, Cybern. B, Cybern.*, vol. 42, no. 6, pp. 1560–1573, Dec. 2012.

[7] A. Argyriou, A. Evgeniou, and M. Pontil, “Multi-task feature learning,” in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 19, 2007, pp. 41–48.

[8] X. Lu, T. Gong, P. Yan, Y. Yuan, and X. Li, “Robust alternative minimization for matrix completion,” *IEEE Trans. Syst., Man, Cybern. B, Cybern.*, vol. 42, no. 3, pp. 939–949, Jun. 2012.

[9] J. Wright, A. Ganesh, S. Rao, Y. Peng, and Y. Ma, “Robust principal component analysis: Exact recovery of corrupted low-rank matrices via convex optimization,” in *Proc. Adv. Neural Inf. Process. Syst.*, 2009, pp. 2080–2088.

[10] A. P. Singh and G. J. Gordon, “Relational learning via collective matrix factorization,” in *Proc. 14th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2008, pp. 650–658.

[11] T. W. Anderson, *An Introduction to Multivariate Statistical Analysis*, vol. 2. New York, NY, USA: Wiley, 1958.

[12] K. Chen, H. Dong, and K.-S. Chan, “Reduced rank regression via adaptive nuclear norm penalization,” *Biometrika*, vol. 100, no. 4, pp. 901–920, 2012.

[13] D. M. Witten, R. Tibshirani, and T. Hastie, “A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis,” *Biostatistics*, vol. 10, no. 3, pp. 515–534, 2009.

[14] Y. She and K. Chen. (2015). “Robust reduced rank regression.” [Online]. Available: https://arxiv.org/abs/1509.03938

[15] S.-K. Chao, W. K. Härdle, and M. Yuan. (2015). “Factorisable sparse tail event curves.” [Online]. Available: https://arxiv.org/abs/1507.03833

[16] Y. Deng, Q. Dai, R. Liu, Z. Zhang, and S. Hu, “Low-rank structure learning via nonconvex heuristic recovery,” *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 24, no. 3, pp. 383–396, Mar. 2013.

[17] P. W. Holland and R. E. Welsch, “Robust regression using iteratively reweighted least-squares,” *Commun. Statist.-Theory Methods*, vol. 6, no. 9, pp. 813–827, 1977.

[18] V. Roth, “The generalized LASSO,” *IEEE Trans. Neural Netw.*, vol. 15, no. 1, pp. 16–28, Jan. 2004.

[19] J. A. K. Suykens, J. De Brabanter, L. Lukas, and J. Vandewalle, “Weighted least squares support vector machines: Robustness and sparse approximation,” *Neurocomputing*, vol. 48, nos. 1–4, pp. 85–105, 2002.

[20] I. F. Gorodnitsky and B. D. Rao, “Sparse signal reconstruction from limited data using FOCUSS: A re-weighted minimum norm algorithm,” *IEEE Trans. Signal Process.*, vol. 45, no. 3, pp. 600–616, Mar. 1997.

[21] I. Daubechies, R. DeVore, M. Fornasier, and C. S. Güntürk, “Iteratively reweighted least squares minimization for sparse recovery,” *Commun. Pure Appl. Math.*, vol. 63, no. 1, pp. 1–38, 2010.

[22] H. Wang and C. Leng, “Unified lasso estimation by least squares approximation,” *J. Amer. Stat. Assoc.*, vol. 102, no. 479, pp. 1039–1048, 2007.

[23] R. Koenker, *Quantile Regression*, no. 38. Cambridge, U.K.: Cambridge Univ. Press, 2005.

[24] K. P. Murphy, *Machine Learning—A Probabilistic Perspective*. Cambridge, MA, USA: MIT Press, 2012.

[25] A. Beck and M. Teboulle, “A fast iterative shrinkage-thresholding algorithm for linear inverse problems,” *SIAM J. Imag. Sci.*, vol. 2, no. 1, pp. 183–202, 2009.

[26] K.-C. Toh and S. Yun, “An accelerated proximal gradient algorithm for nuclear norm regularized linear least squares problems,” *Pacific J. Optim.*, vol. 6, no. 3, pp. 615–640, Nov. 2010.

[27] A. Mehrjou, R. Hosseini, and B. N. Araabi, “Improved Bayesian information criterion for mixture model selection,” *Pattern Recognit. Lett.*, vol. 69, pp. 22–27, Jan. 2016.

[28] P. Stoica and Y. Selen, “Model-order selection: A review of information criterion rules,” *IEEE Signal Process. Mag.*, vol. 21, no. 4, pp. 36–47, Jul. 2004.

[29] C. Cobos, H. Muñoz-Collazos, R. Urbano-Muñoz, M. Mendoza, E. León, and E. Herrera-Viedma, “Clustering of Web search results based on the cuckoo search algorithm and balanced Bayesian information criterion,” *Inf. Sci.*, vol. 281, pp. 248–264, Oct. 2014.

[30] S. Watanabe, “A widely applicable Bayesian information criterion,” *J. Mach. Learn. Res.*, vol. 14, pp. 867–897, Mar. 2013.

[31] G. Schwarz, “Estimating the dimension of a model,” *Ann. Statist.*, vol. 6, no. 2, pp. 461–464, 1978.

[32] L. Wasserman, *All of Statistics: A Concise Course in Statistical Inference*. New York, NY, USA: Springer-Verlag, 2013.

[33] S.-B. Lin and D.-X. Zhou, “Distributed kernel-based gradient descent algorithms,” *Constructive Approx.*, vol. 47, pp. 249–276, 2018.

[34] A. Mukherjee, K. Chen, N. Wang, and J. Zhu, “On the degrees of freedom of reduced-rank estimators in multivariate regression,” *Biometrika*, vol. 102, no. 2, pp. 457–477, 2015.

[35] J. Ye, “On measuring and correcting the effects of data mining and model selection,” *J. Amer. Stat. Assoc.*, vol. 93, no. 441, pp. 120–131, 1998.

[36] L. Breiman *et al.*, “Heuristics of instability and stabilization in model selection,” *Ann. Statist.*, vol. 24, no. 6, pp. 2350–2383, 1996.

[37] L. Breiman, “Bagging predictors,” *Mach. Learn.*, vol. 24, no. 2, pp. 123–140, 1996.

[38] W. Zhao, H. Lian, and S. Ma, “Robust reduced-rank modeling via rank regression,” *J. Stat. Planning Inference*, vol. 180, pp. 1–12, Jan. 2016.

[39] F. Bunea, Y. She, and M. H. Wegkamp, “Optimal selection of reduced rank estimators of high-dimensional matrices,” *Ann. Statist.*, vol. 39, no. 2, pp. 1282–1309, 2011.

[40] R. F. Engle and S. Manganelli, “CAViaR: Conditional autoregressive value at risk by regression quantiles,” *J. Bus. Econ. Statist.*, vol. 22, no. 4, pp. 367–381, 2004.

[41] R. Mazumder, T. Hastie, and R. Tibshirani, “Spectral regularization algorithms for learning large incomplete matrices,” *J. Mach. Learn. Res.*, vol. 11, pp. 2287–2322, Jan. 2010.

[42] H. Liu and B. Yu, “Asymptotic properties of Lasso+mLS and Lasso+Ridge in sparse high-dimensional linear regression,” *Electron. J. Statist.*, vol. 7, pp. 3124–3169, Jun. 2013.

[43] M. Buchinsky, “Estimating the asymptotic covariance matrix for quantile regression models a Monte Carlo study,” *J. Econometrics*, vol. 68, no. 2, pp. 303–338, 1995.

[44] A. Wille *et al.*, “Sparse graphical Gaussian modeling of the isoprenoid gene network in *Arabidopsis thaliana*,” *Genome Biol.*, vol. 5, no. 11, pp. R92-1–R92-13, 2004.

[45] H. L. White, Jr., T.-H. Kim, and S. Manganelli, “Modeling autoregressive conditional skewness and kurtosis with multi-quantile CAViaR,” ECB, Frankfurt, Germany, Working Paper 957, 2008.

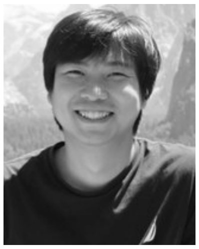
[46] E. F. Fama and K. R. French, “Industry costs of equity,” *J. Financial Econ.*, vol. 43, no. 2, pp. 153–193, 1997.

[47] S. Negahban, P. Ravikumar, M. J. Wainwright, and B. Yu, “A unified framework for high-dimensional analysis of M-estimators with decomposable regularizers,” *Statist. Sci.*, vol. 27, no. 4, pp. 538–557, 2012.

[48] Y. Yang, Y. Feng, and J. A. K. Suykens, “Robust low-rank tensor recovery with regularized redescending M-estimator,” *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 27, no. 9, pp. 1933–1946, Sep. 2016.

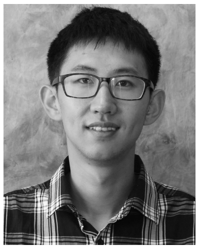
[49] D. L. Donoho *et al.*, “High-dimensional data analysis: The curses and blessings of dimensionality,” *AMS Math Challenges Lect.*, vol. 1, p. 32, Aug. 2000.

[50] J. Fan, F. Han, and H. Liu, “Challenges of big data analysis,” *Nat. Sci. Rev.*, vol. 1, no. 2, pp. 293–314, 2014.



Xiangyu Chang received the Ph.D. degree in applied mathematics from Xi’an Jiaotong University, Xi’an, China, in 2014.

He is currently an Associate Professor with the School of Management, Xi’an Jiaotong University. His current research interests include statistical machine learning, data science, and business statistics.

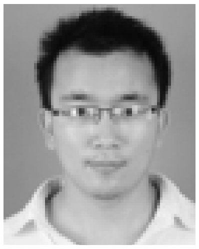


Yan Zhong is currently pursuing the Ph.D. degree with the Department of Statistics, Texas A&M University, College Station, TX, USA.

His current research interests include statistical machine learning, dimensionality reduction, multivariate analysis, and statistics applications in epidemiology.

Yao Wang received the Ph.D. degree in applied mathematics from Xi’an Jiaotong University, Xi’an, China, in 2014.

He is currently an Associate Professor with the Department of Statistics, Xi’an Jiaotong University. His current research interests include statistical signal processing, high-dimensional data analysis, and machine learning.



Shaobo Lin received the Ph.D. degree in applied mathematics from Xi’an Jiaotong University, China, in 2014.

He is currently with the Department of Mathematics, Wenzhou University, Wenzhou, China. His current research interests include massive data analysis, neural networks, and learning theory.