Optimal Exact Least Squares Rank Minimization

Shuo Xiang[1,2], Yunzhang Zhu[3], Xiaotong Shen[3], Jieping Ye[1,2]
[1]Department of Computer Science and Engineering, Arizona State University, AZ 85287
[2]Center for Evolutionary Medicine and Informatics, The Biodesign Institute, Arizona State University, AZ 85287
[3]School of Statistics, University of Minnesota, Minneapolis, MN 55455

ABSTRACT

In multivariate analysis, rank minimization emerges when a low-rank structure of matrices is desired as well as a small estimation error. Rank minimization is nonconvex and generally NP-hard, imposing one major challenge. In this paper, we consider a nonconvex least squares formulation, which seeks to minimize the least squares loss function with the rank constraint. Computationally, we develop efficient algorithms to compute a global solution as well as an entire regularization solution path. Theoretically, we show that our method reconstructs the oracle estimator exactly from noisy data. As a result, it recovers the true rank optimally against any method and leads to sharper parameter estimation over its counterpart. Finally, the utility of the proposed method is demonstrated by simulations and image reconstruction from noisy background.

Categories and Subject Descriptors

H.2.8 [**Database Management**]: Database Applications-Data Mining

General Terms

Algorithms

Keywords

Nonconvex, global optimality, rank minimization

1. INTRODUCTION

In multivariate analysis, estimation of lower-dimensional structures has received attention in statistics, signal processing and machine learning. One type of such structures is the low-rank of matrices [5, 22], where the rank measures the dimension of a multivariate response. Rank minimization approximates multivariate data with the smallest possible rank of matrices. It has many applications in, for instance, multi-task learning [6, 11], multi-class classification [2], matrix completion [8, 17], collaborative filtering [33, 1], clustering [20, 29], computer vision [35, 21, 18], among others. The central topic this article addresses is least squares rank minimization.

Consider multi-response linear regression in which a k-dimensional response vector $\boldsymbol{z}_i$ follows

$$\begin{aligned} &\boldsymbol{z}_i = \boldsymbol{\Theta}^T \boldsymbol{a}_i + \boldsymbol{\varepsilon}_i; \\ &E\boldsymbol{\varepsilon}_i = \boldsymbol{0}, \quad Cov(\boldsymbol{\varepsilon}_i) = \sigma^2 \boldsymbol{I}_{k\times k}; \quad i = 1, \cdots, n, \end{aligned} \tag{1}$$

where $\boldsymbol{a}_i$ is a p-dimensional design vector, $\boldsymbol{\Theta}$ is a $p \times k$ regression parameter matrix, and components of $\boldsymbol{\varepsilon}_i$ are independent. Model (1) reduces to the widely-used linear model in compressed sensing when $k = 1$ and becomes a multivariate autoregressive model with $\boldsymbol{a}_i = \boldsymbol{z}_{i-1}$. Denote the rank of $\boldsymbol{\Theta}$ as $r(\boldsymbol{\Theta})$ and rewrite (1) in the matrix form as follows:

$$\boldsymbol{Z} = \boldsymbol{A}\boldsymbol{\Theta} + \boldsymbol{e}, \tag{2}$$

where $\boldsymbol{Z} = (\boldsymbol{z}_1, \cdots, \boldsymbol{z}_n)^T \in \mathbb{R}^{n\times k}$, $\boldsymbol{A} = (\boldsymbol{a}_1, \cdots, \boldsymbol{a}_n)^T \in \mathbb{R}^{n\times p}$ and $\boldsymbol{e} = (\boldsymbol{\varepsilon}_1, \cdots, \boldsymbol{\varepsilon}_n)^T \in \mathbb{R}^{n\times k}$ are the data, design and error matrices. In (1), we estimate $\boldsymbol{\Theta}$ based on n pairs of observation vectors $(\boldsymbol{a}_i, \boldsymbol{z}_i)_{i=1}^n$, with *a priori* knowledge that $r(\boldsymbol{\Theta})$ is relatively small in comparison to $\min(n, k, p)$, where the number of unknown parameters k, p can greatly exceed the sample size n.

Least squares rank minimization, as described, solves

$$\begin{aligned} \min_{\boldsymbol{\Theta}} \quad & \|\boldsymbol{A}\boldsymbol{\Theta} - \boldsymbol{Z}\|_F^2 \\ \text{s.t.} \quad & r(\boldsymbol{\Theta}) \le s, \end{aligned} \tag{3}$$

where $\|\cdot\|_F$ is the Frobenius-norm and s is an integer-valued tuning parameter taking values from $[1, \min(n, k, p)]$. The general rank minimization is nonconvex and NP-hard [23], which is like the L_0-minimization in univariate analysis. Therefore an exact global solution to (3) is not known as well as its statistical properties, due primarily to discreteness and non-convexity of the rank function.

Estimation under the restriction that $r(\boldsymbol{\Theta}) = r$ has been studied when $n \to \infty$ with k and p held fixed, see [3, 4, 15, 28, 26]. Two major computational approaches have been proposed for approximating the optimal solution of (3). The first one involves regularization using surrogate function, such as the nuclear-norm, which is a convex envelope of the rank function [13] and can be solved by efficient algorithms [8, 19, 34, 25, 16]. In some cases, the solution of this convex problem coincides with a global minimizer of (3) under certain isometry assumptions [27]. However, these as-

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.
KDD'12, August 12–16, 2012, Beijing, China.
Copyright 2012 ACM 978-1-4503-1462-6 /12/08 ...$10.00.

sumptions can be strong and difficult to check. Recently, [7] obtained a global minimizer of a regularized version of (3).

The second approach attacks (3) by approximating the rank function iteratively through calculating the largest singular vector using greedy search [30], or by singular value projection (SVP) through a local gradient method [17]. Under weaker isometry assumptions [27, 10, 9] than that of the nuclear-norm approach, these methods guarantee an exact solution of (3) but suffer from the same difficulties as the regularization method [30], although they have achieved promising results on both simulated and real-world data.

Theoretically, some loss error bounds of the first regularization approach are obtained in [24] under Frobenius-norm, and rank selection consistency is established in [7]. Unfortunately, to our best knowledge, whether similar conclusions hold for our formulation (3) remains largely unknown.

In this paper, we have advanced on two fronts. Computationally, we derive a general closed-form for a global minimizer of (3) in Theorem 1, and give a condition under which (3) and its nonconvex regularized counterpart are equivalent with regard to global minimizers, although these two methods are not generally equivalent. Moreover, we develop an efficient algorithm for computing the entire regularization solution path at the cost of computing only one solution for a single regularization parameter. Theoretically, we establish optimality for a global minimizer of (3). More specifically, the proposed method is optimal against any other ones in that it reconstructs the oracle estimator exactly, thus the true rank, under (1). It is important to note that this exact recovery result is a much stronger property than consistency, which is attributed to the discrete nature of the rank function as well as tuning parameter s. Such a result may not be shared by its regularized counterpart with a continuum tuning parameter. In addition, the method enjoys a higher degree of accuracy for parameter estimation than nuclear-norm rank estimation.

After the first draft of this paper was completed, we were aware that [14] and [32] gave an expression of Theorem 1. However, neither paper considered computational and statistical aspects of the solution. Inevitably, some partial overlaps exist between our Theorem 1 and theirs.

The rest of the paper is organized as follows. Section 2 presents a closed-form solution to (3). Section 3 gives an efficient path algorithm for a regularized version of (3). Section 4 is devoted to theoretical investigation, followed by Section 5 discussing methods for tuning. Section 6 presents proofs for all the theorems we develop and Section 7 presents results of empirical evaluations, where several rank minimization methods are compared. Section 8 concludes the paper.

Notation: Table 1 summarizes the notations used in the rest of the paper.

Table 1: Notation used throughout the paper

Notation	Description
$\boldsymbol{A}$	The design matrix
$\boldsymbol{e}$	Error matrix
Θ^0	The ground truth model
$\hat{\Theta}^s$	Estimator obtained by optimizing (3)
$\hat{\Theta}^*_\lambda$	Estimator obtained by optimizing (7)
$r(\boldsymbol{A})$	The rank of matrix $\boldsymbol{A}$
$\mathcal{P}_s(\boldsymbol{Z})$	The best s-rank approximation of matrix Z in terms of the Frobenius-norm

2. PROPOSED METHOD: CLOSED-FORM SOLUTION

This section derives a closed-form solution to (3). The strategy is to simplify (3) through the singular value decomposition (SVD) of matrix $\boldsymbol{A}$ and utilizing the properties of the rank function. Before proceeding, we present a motivating lemma, known as the Eckart-Young theorem [12].

LEMMA 1. *The best s-rank approximation, in terms of the Frobenius-norm, for a t-rank matrix $\boldsymbol{Z}$ with $t \geq s$, i.e., a global minimizer $\boldsymbol{\Theta}^*$ of*

$$\begin{aligned} \min_{\boldsymbol{\Theta}} \quad & \|\boldsymbol{\Theta} - \boldsymbol{Z}\|_F^2 \\ s.t. \quad & r(\boldsymbol{\Theta}) \leq s \end{aligned} \tag{4}$$

is given by $\boldsymbol{\Theta}^* = \mathcal{P}_s(\boldsymbol{Z}) = \boldsymbol{U}_z \boldsymbol{D}_s \boldsymbol{V}_z^T$, where $\boldsymbol{D}_s$ consists of the largest s singular values of $\boldsymbol{Z}$ given the SVD of $\boldsymbol{Z} = \boldsymbol{U}_z \boldsymbol{D} \boldsymbol{V}_z^T$.

Intuitively $\mathcal{P}_s(\boldsymbol{Z})$ may be viewed as a projection of $\boldsymbol{Z}$ onto a set of matrices whose ranks are no more than s. Note that (4) is a special case of (3) with matrix $\boldsymbol{A}$ being the identity matrix. This motivates us to solve (3) through the simpler problem (4).

When $\boldsymbol{A}$ is nonsingular, clearly (3) has a global minimizer $\boldsymbol{A}^{-1}\mathcal{P}_s(\boldsymbol{Z})$ by rank preserveness of any nonsingular matrix in matrix multiplication. When $\boldsymbol{A}$ is singular, we first assume that $r(\boldsymbol{A}) \geq s$. However, this assumption is by no means mandatory, which will be discussed later. Given the SVD of $\boldsymbol{A} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^T$, with orthogonal matrices $\boldsymbol{U} \in \mathbb{R}^{n\times n}$ and $\boldsymbol{V} \in \mathbb{R}^{p\times p}$ and diagonal matrix $\boldsymbol{D} \in \mathbb{R}^{n\times p}$, we have

$$\|\boldsymbol{A\Theta} - \boldsymbol{Z}\|_F = \|\boldsymbol{U}^T(\boldsymbol{A\Theta} - \boldsymbol{Z})\|_F = \|\boldsymbol{DV}^T\boldsymbol{\Theta} - \boldsymbol{U}^T\boldsymbol{Z}\|_F.$$

This follows from the fact that the Frobenius-norm is invariant under any orthogonal transformation. Let $\boldsymbol{Y} = \boldsymbol{V}^T\boldsymbol{\Theta}$ and $\boldsymbol{W} = \boldsymbol{U}^T\boldsymbol{Z}$, clearly we have $r(\boldsymbol{Y}) = r(\boldsymbol{\Theta})$. Now solving (3) amounts to solving an equivalent form:

$$\begin{aligned} \min_{\boldsymbol{Y}} \quad & \|\boldsymbol{DY} - \boldsymbol{W}\|_F^2 \\ \text{s.t.} \quad & r(\boldsymbol{Y}) \leq s. \end{aligned} \tag{5}$$

Consequently, a global minimizer of (3) becomes $V\boldsymbol{Y}^*$, where $\boldsymbol{Y}^*$ is a global minimizer of (5) and is given by the following theorem.

THEOREM 1. *Let $\boldsymbol{D}$, $\boldsymbol{Y}$, $\boldsymbol{Z}$ and s be as defined above. If $s \leq r(\boldsymbol{A})$, then a global minimizer of (5) is given by*

$$\boldsymbol{Y}^* = \begin{bmatrix} \boldsymbol{D}^{-1}_{r(\boldsymbol{A})}\mathcal{P}_s(\boldsymbol{W}_{r(\boldsymbol{A})}) \\ \mathbf{a} \end{bmatrix}, \tag{6}$$

where $\boldsymbol{D}_{r(\boldsymbol{A})}$ is a diagonal matrix consisting of all the nonzero singular value of $\boldsymbol{A}$, $\boldsymbol{a}$ can be set as the zero matrix, and $\boldsymbol{W}_{r(\boldsymbol{A})}$ consists of the first $r(\boldsymbol{A})$ rows of $\boldsymbol{W}$.

Here are some remarks regarding the above theorem:

Remark 1. The solution to problem (5) is generally not unique. Specifically, the matrix $\boldsymbol{a}$ in (6) need not be fixed at zero, as long as it does not change the rank of Y^*. However, if $\boldsymbol{A}$ is of full column rank, i.e., when $r(\boldsymbol{A}) = p$, then $\boldsymbol{a}$ vanishes and Y^* can be uniquely determined. In this case, the optimal solution of (3) is also unique.

Remark 2. The optimal Y^* can also be computed for the general matrix A with an arbitrary rank, i.e., when $r(\boldsymbol{A}) < s$. See the proof of Theorem 1 in the Section 6.

It is important to note that the value of $\boldsymbol{a}$ is irrelevant for prediction, but matters for parameter estimation. In other words, when $r(\boldsymbol{A}) < p$, a global minimizer is not unique, hence that parameter estimation is not identifiable: see Section 4 for a discussion. For simplicity, we set $\boldsymbol{a} = \boldsymbol{0}$ for $\boldsymbol{Y}^*$ subsequently.

In what follows, our estimator is defined as $\hat{\boldsymbol{\Theta}}^s$, as well as an estimated rank $\hat{r}$. Algorithm 1 below summarizes the main steps for computing $\hat{\boldsymbol{\Theta}}^s$ with regard to $s \leq \min(n, k, p)$, where the LSRM stands for Least Squares Rank Minimization.

Algorithm 1 Exact solution of (3)

Input: $\boldsymbol{A}$, $\boldsymbol{Z}$, $s \leq r(\boldsymbol{A})$
Output: `A global minimizer` $\boldsymbol{\Theta}$ of (3)
Function: `LSRM`($\boldsymbol{A}$, $\boldsymbol{Z}$, s)

1: **if** $\boldsymbol{A}$ `is nonsingular then`
2: $\quad \boldsymbol{\Theta} = \boldsymbol{A}^{-1}\mathcal{P}_s(\boldsymbol{Z})$
3: **else**
4: $\quad$ `Perform SVD on` $\boldsymbol{A}$: $\boldsymbol{A} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^T$
5: $\quad$ `Extract the first` r `rows of` $\boldsymbol{U}^T\boldsymbol{Z}$ `and denote it as` $\boldsymbol{W}_{r(\boldsymbol{A})}$
6: $\quad \boldsymbol{\Theta} = \boldsymbol{V}\begin{bmatrix} \boldsymbol{D}_{r(\boldsymbol{A})}^{-1}\mathcal{P}_s(\boldsymbol{W}_{r(\boldsymbol{A})}) \\ \boldsymbol{0} \end{bmatrix}$
7: **end if**
8: **return** $\boldsymbol{\Theta}$

The complexity of Algorithm 1 is determined mainly by the most expensive operations–matrix inversion and SVD, specifically, at most one matrix inversion and two SVDs. Such operations roughly require a cubic time complexity [1].

3. REGULARIZATION AND SOLUTION PATH

This section studies a regularized counterpart of (3):

$$\min_{\boldsymbol{\Theta}} \quad \|\boldsymbol{A}\boldsymbol{\Theta} - \boldsymbol{Z}\|_F^2 + \lambda\ r(\boldsymbol{\Theta}), \tag{7}$$

where $\lambda > 0$ is a continuous regularization parameter corresponding to s in (3), and $\boldsymbol{\Theta}_\lambda^*$ is a global minimizer of (7). The next theorem establishes an equivalence between (7) and (3) when $\boldsymbol{\Theta}_\lambda^*$ is unique, occurring when $r(\boldsymbol{A}) = p$. Such a result is not generally anticipated for a nonconvex problem.

THEOREM 2 (EQUIVALENCE). *When $p = r(\boldsymbol{A})$, (7) has a unique global minimizer. Moreover, (7) and (3) are equivalent with respect to their solutions. For any $\boldsymbol{\Theta}_\lambda^*$ with $\lambda \geq 0$, there exists $1 \leq s^* = r(\boldsymbol{\Theta}_\lambda^*)$ such that $\boldsymbol{\Theta}_\lambda^* = \hat{\boldsymbol{\Theta}}^s$, and vice verse.*

Next we develop an algorithm for computing an entire solution path for all values of λ with complexity comparable to that of solving (7) at a single λ-value. For motivation, first consider a special case of the identity $\boldsymbol{A}$ in (7):

$$g(\lambda) = \min_{\boldsymbol{\Theta}} \quad \|\boldsymbol{\Theta} - \boldsymbol{Z}\|_F^2 + \lambda\ r(\boldsymbol{\Theta}). \tag{8}$$

3.1 Monotone property

In (8), $r(\boldsymbol{\Theta})$ decreases as λ increases from 0, where $r(\boldsymbol{\Theta})$ goes through all integer values from $r(\boldsymbol{Z})$ down to 0 when λ becomes sufficiently large. In addition, the value of $g(\lambda)$ is nondecreasing as λ increases. The next theorem summarizes these results.

THEOREM 3 (MONOTONE PROPERTY). *Let $r(\boldsymbol{Z})$ be r. Then the following properties hold:*

(1) There exists a solution path vector $\mathcal{S}$ of length $r+2$ satisfying the following:

$$\mathcal{S}_0 = 0, \quad \mathcal{S}_{r+1} = +\infty, \quad \mathcal{S}_{k+1} > \mathcal{S}_k, \quad k = 0, 1, \cdots, r$$
$$\boldsymbol{\Theta}_\lambda^* = \mathcal{P}_{r-k}(\boldsymbol{Z}), \quad \text{if } \mathcal{S}_k \leq \lambda < \mathcal{S}_{k+1},$$

(2) Function $g(\lambda)$ is nondecreasing and piecewise linear.

The monotone property leads to an efficient algorithm for calculating the pathwise solution of (8). Figure 1 displays the solution path by illustrating the function value $g(\lambda)$ and $r(\boldsymbol{\Theta}_\lambda^*)$ as a function of λ.

3.2 Pathwise algorithm

Through the monotone property, we compute the optimal solution of (8) at a particular λ by locating λ in the correct interval in the solution path vector $\mathcal{S}$, which can be achieved efficiently via a simple binary search. Algorithm 2 describes the main steps.

Algorithm 2 Pathwise solution of (8)

Input: $\boldsymbol{\Theta}$, $\boldsymbol{Z}$
Output: `Solution path vector` $\mathcal{S}$, `pathwise solution` $\boldsymbol{\Theta}$
Function: `pathwise`($\boldsymbol{\Theta}$,$\boldsymbol{Z}$)

1: `Initialize`: $\mathcal{S}_0 = 0$, $\boldsymbol{\Theta}_0 = \boldsymbol{Z}$, $r = r(\boldsymbol{Z})$
2: `Perform SVD on` $\boldsymbol{Z}$: $\boldsymbol{Z} = U\boldsymbol{D}V^T$
3: **for** $i = r$ **down to** 1 **do**
4: $\quad \mathcal{S}_{r-i+1} = \sigma_i^2$
5: $\quad \boldsymbol{\Theta}_{r-i+1} = \boldsymbol{\Theta}_{r-i} - \sigma_i u_i v_i^T$
6: **end for**
7: **return** $\mathcal{S}$, $\boldsymbol{\Theta}$

Algorithm 2 requires only one SVD operation, therefore its complexity is of the same order as that of Algorithm 1 at a single s-value. When $\boldsymbol{Z}$ is a low-rank matrix, existing software for SVD computation such as PROPACK is applicable to further improve computational efficiency.

3.3 Extension to general $\boldsymbol{A}$

For a general design matrix $\boldsymbol{A}$, note that

$$\|A\boldsymbol{\Theta} - \boldsymbol{Z}\|_F^2 = \|\boldsymbol{D}\boldsymbol{Y} - \boldsymbol{W}\|_F^2 = \|\boldsymbol{W}'\|_F^2 + \|\boldsymbol{D}_r\boldsymbol{Y}_r - \boldsymbol{W}_r\|_F^2,$$

where $\boldsymbol{W} = \begin{bmatrix} \boldsymbol{W}_r \\ \boldsymbol{W}' \end{bmatrix}$. After ignoring the constant term $\|\boldsymbol{W}'\|_F^2$, we solve

$$\min_{\boldsymbol{Y}_r} \quad \|\boldsymbol{D}_r\boldsymbol{Y}_r - \boldsymbol{W}_r\|_F^2 + \lambda\ r(\boldsymbol{Y}_r).$$

[1] More specifically, for a matrix of dimension $n \times p$, the SVD has a complexity of $O(\min\{n^2p, p^2n\})$, whereas the matrix inversion has a complexity of $O(r(\boldsymbol{A})^3)$, which can be improved to $O(r(\boldsymbol{A})^{2.807})$ when the Strassen Algorithm is utilized.

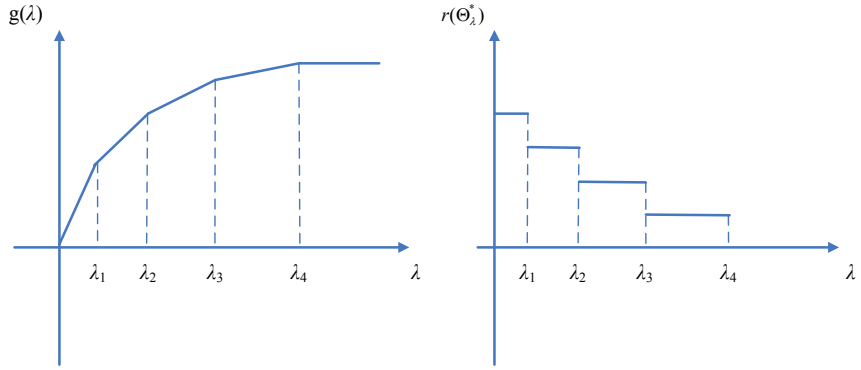


Figure 1: Piecewise linearity of $g(\cdot)$ and the rank of the optimal solution with respect to λ.

Note that $\boldsymbol{D}_r$ is nonsingular, then the problem reduces to the simple case

$$\min_{\hat{\boldsymbol{Y}}} \quad \|\hat{\boldsymbol{Y}} - \boldsymbol{W}_r\|_F^2 + \lambda\, r(\hat{\boldsymbol{Y}}),$$

where $\hat{\boldsymbol{Y}} = \boldsymbol{D}_r \boldsymbol{Y}_r$. The solution path can be obtained directly from Algorithm 2.

4. STATISTICAL PROPERTIES

This section is devoted to the theoretical investigation of least squares rank minimization, which remains largely unexplored, although nuclear-norm regularization has been studied. In particular, we will reveal what is the best performance for prediction as well as the optimal risk for parameter estimation. Moreover, we will establish the optimality of the proposed method. In fact, the proposed method reconstructs the oracle estimator, the optimal one as if the true rank were known in advance. Here the oracle estimator $\hat{\boldsymbol{\Theta}}^0$ is defined as a global minimizer of $\|\boldsymbol{A}\boldsymbol{\Theta} - \boldsymbol{Z}\|_F^2$ subject to $r(\boldsymbol{\Theta}) = r_0$, where $\boldsymbol{\Theta}^0$ and $r_0 = r(\boldsymbol{\Theta}^0) \geq 1$ denote the true parameter matrix and the true rank respectively. This leads to the exact rank recovery, in addition to the reconstruction of the optimal performance of the oracle estimator. In other words, the proposed method is optimal against any other methods such as the nuclear-norm rank regularization.

Given the design matrix $\boldsymbol{A}$, we study the accuracy of rank recovery as well as prediction and parameter estimation. Let $\mathbb{P}$ and $\mathbb{E}$ be the true probability and expectation under $\boldsymbol{\Theta}^0$ given $\boldsymbol{A}$. For rank recovery, we use the metric $\mathbb{P}(\hat{r} \neq r_0)$. For prediction and parameter estimation, we employ the risk $\mathbb{E}K(\hat{\boldsymbol{\Theta}}^s, \boldsymbol{\Theta}^0)$ and $\mathbb{E}\|\hat{\boldsymbol{\Theta}}^s - \boldsymbol{\Theta}^0\|_F^2$ respectively, where

$$\begin{aligned} K(\hat{\boldsymbol{\Theta}}^s, \boldsymbol{\Theta}^0) &= (2\sigma^2 n)^{-1} \sum_{i=1}^{n} \|\boldsymbol{a}_i^T(\hat{\boldsymbol{\Theta}}^s - \boldsymbol{\Theta}^0)\|_2^2 \\ &= (2\sigma^2 n)^{-1} \|\boldsymbol{A}(\hat{\boldsymbol{\Theta}}^s - \boldsymbol{\Theta}^0)\|_F^2 \end{aligned}$$

is the Kullback-Leibler loss and $\|\cdot\|_2$ is the L_2-norm for a vector. Note that the predictive risk equals to $2\sigma^2 \mathbb{E}K(\hat{\boldsymbol{\Theta}}^s, \boldsymbol{\Theta}^0)$ and parameter estimation is considered when it is identifiable, i.e., $r(\boldsymbol{A}) = p$.

Now we present the risk bounds under (1) without a Gaussian error assumption.

THEOREM 4. *Under* (1)*, the oracle estimator is exactly reconstructed by our method in that* $\hat{\boldsymbol{\Theta}}^{r_0} = \boldsymbol{\Theta}^0$ *under* $\mathbb{P}$*, when* $r_0 \leq min(r(\boldsymbol{A}), p)$*. As a result, exact reconstruction of the optimal performance is achieved by our estimator* $\hat{\boldsymbol{\Theta}}^{r_0}$*. In particular,*

$$\mathbb{E}K(\hat{\boldsymbol{\Theta}}^{r_0}, \boldsymbol{\Theta}^0) \begin{cases} = & \frac{r_0 k}{2n} \quad \text{if } r_0 = r(\boldsymbol{A}) \\ \leq & \frac{2\mathbb{E}(\sum_{j=1}^{r_0} \sigma_j^2)}{n} \quad \text{if } r_0 < r(\boldsymbol{A}) \end{cases}$$

and

$$\mathbb{E}\|\hat{\boldsymbol{\Theta}}^{r_0} - \boldsymbol{\Theta}^0\|_F^2 \begin{cases} = & \frac{r_0 k}{\sigma_{\min}^2(\boldsymbol{A}) n} \quad \text{if } r_0 = r(\boldsymbol{A}) = p \\ \leq & \frac{\mathbb{E}(\sum_{j=1}^{r_0} \sigma_j^2)}{\sigma_{\min}^2 n} \quad \text{if } r_0 < r(\boldsymbol{A}) = p \end{cases},$$

where σ_j *and* $\sigma_{\min} > 0$ *are the* j*th largest and the smallest nonzero singular value of* $\boldsymbol{e}^\star = \boldsymbol{U}_{r(A)}^T \boldsymbol{e}$ *and* $n^{-1/2}\boldsymbol{A}$*, respectively, and* $\boldsymbol{U}_{r(A)}$ *denotes the first* $r(A)$ *rows of* $\boldsymbol{U}$*.*

Remark: In general, $\mathbb{E}\sum_{j=1}^{r_0} \sigma_j^2 \leq r_0 \mathbb{E}\sigma_1^2$.

Theorem 4 says that the optimal oracle estimator is exactly reconstructed by our method. Interestingly, the true rank is exactly recovered from noisy data, which is attributed to discreteness of the rank and is analogous to the maximum likelihood estimation over a discrete parameter space. Concerning prediction and parameter estimation, the optimal Kullback-Leibler risk is $\frac{r_0 k}{n}$ but the risk under the Frobenius-norm is $\frac{r_0 k}{\sigma_{\min}^2 n}$. For prediction, only the effective degrees of freedom $\sum_{j=1}^{r_0} \sigma_j^2$ matters as opposed to p, which is in contrast to a rate $\frac{r_0 kp}{n}$ without a rank restriction. This permits p to be much larger than n, or $k, p \gg n$. For estimation, however, p enters the risk through $\sigma_{\min}^2$, where p can not be larger than n, or $\max(k, p) \leq n$.

5. TUNING

As shown in Section 4, theoretically, exact rank reconstruction can be accomplished through tuning. Practically, we employ a predictive measure for rank selection.

The predictive performance of $\hat{\boldsymbol{\Theta}}^s$ is measured by

$$\text{MSE}(\hat{\boldsymbol{\Theta}}^s) = \frac{1}{n}\mathbb{E}\|\boldsymbol{A}\hat{\boldsymbol{\Theta}}^s - \boldsymbol{Z}\|_F^2,$$

which is proportional to the risk $R(\hat{\boldsymbol{\Theta}}^s)$, where the expectation is taken with respect to $(\boldsymbol{Z}, \boldsymbol{A})$.

To estimate s over integer values in $[0, \cdots, \min(n, p, k)]$, one may cross-validate through a tuning data set, which estimates the MSE. Alternatively, one may use the generalized degrees of freedom [31] through data perturbation without a tuning set.

$$\widehat{\mathrm{GDF}}(\hat{\boldsymbol{\Theta}}^s) = n^{-1}\|\boldsymbol{Z}-\boldsymbol{A}\hat{\boldsymbol{\Theta}}^s\|^2+2n^{-1}\sum_{i=1}^{n}\sum_{l=1}^{k}\widehat{\mathrm{Cov}}(\boldsymbol{z}_{il},(\boldsymbol{a}_i^T\hat{\boldsymbol{\Theta}}^s)_l), \tag{9}$$

where $\widehat{\mathrm{Cov}}(\boldsymbol{z}_{il},(\boldsymbol{a}_i^T\hat{\boldsymbol{\Theta}}^s)_l)$ is the estimated covariance between the lth component of $\boldsymbol{z}_i$ and the lth component of $\boldsymbol{a}_i^T\hat{\boldsymbol{\Theta}}^s$. In the case that $\boldsymbol{e}_i$ in (1) follows $N(\boldsymbol{0}, \sigma^2 I_{k\times k})$, the method of data perturbation of [31] is applicable. Specifically, sample $\boldsymbol{e}_i^*$ from $N(\boldsymbol{0}, \sigma^2 I_{k\times k})$ and let

$$\boldsymbol{Z}^* = (1-\tau)\boldsymbol{Z} + \tau\tilde{\boldsymbol{Z}}$$
$$\widehat{\mathrm{Cov}}(\boldsymbol{z}_{il},(\boldsymbol{a}_i^T\hat{\boldsymbol{\Theta}}^s)_l) = \tau^{-1}\mathrm{Cov}^*(\boldsymbol{z}_{il}^*,(\boldsymbol{a}_i^T\hat{\boldsymbol{\Theta}}^{*s})_l)$$

where $\mathrm{Cov}^*(\boldsymbol{z}_{il}^*,(\boldsymbol{a}_i\hat{\boldsymbol{\Theta}}^{*s})_l)$ is the sample covariance with the Monte Carlo size T. For the types of problems we consider, we fixed T to be n.

6. PROOF OF THEOREMS

In this section we present detailed proofs to the theorems developed in the previous sections. We first present a technical lemma to be used in the proof of Theorem 4.

LEMMA 2. *Suppose $\boldsymbol{A}$ and $\boldsymbol{B}$ are two $n_1 \times n_2$ matrices. Then,*

$$\langle \boldsymbol{A}, \boldsymbol{B}\rangle \le \|\boldsymbol{A}\|_F\|\mathcal{P}_{r(\boldsymbol{A})}(\boldsymbol{B})\|_F, \tag{10}$$

where $\langle A, B\rangle = Tr(A^TB) = Tr(B^TA)$, Tr denotes the trace, and $r(\boldsymbol{A})$ is the rank of $\boldsymbol{A}$.

PROOF. Let the singular value decomposition of $\boldsymbol{A}$ and $\boldsymbol{B}$ be $\boldsymbol{A} = \boldsymbol{U}_1\boldsymbol{\Sigma}_1\boldsymbol{V}_1^T$ and $\boldsymbol{B} = \boldsymbol{U}_2\boldsymbol{\Sigma}_2\boldsymbol{V}_2^T$ where $\boldsymbol{U}_i$ and $\boldsymbol{V}_i, i = 1, 2$, are orthogonal matrices, and Σ_1 and Σ_2 are diagonal matrices with their diagonal elements being the singular values of $\boldsymbol{A}$ and $\boldsymbol{B}$, respectively. Then

$$\begin{aligned}\langle \boldsymbol{A}, \boldsymbol{B}\rangle &= Tr(\boldsymbol{V}_1\boldsymbol{\Sigma}_1^T\boldsymbol{U}_1^T\boldsymbol{U}_2\boldsymbol{\Sigma}_2\boldsymbol{V}_2^T)\\ &= Tr(\boldsymbol{\Sigma}_1^T\boldsymbol{U}_1^T\boldsymbol{U}_2\boldsymbol{\Sigma}_2\boldsymbol{V}_2^T\boldsymbol{V}_1)\\ &\equiv Tr(\boldsymbol{\Sigma}_1^T\boldsymbol{U}\boldsymbol{\Sigma}_2\boldsymbol{V}^T),\end{aligned}$$

where $\boldsymbol{U} = \boldsymbol{U}_1^T\boldsymbol{U}_2$ and $\boldsymbol{V} = \boldsymbol{V}_1^T\boldsymbol{V}_2$ continue to be orthogonal. Let the ordered singular values of $\boldsymbol{A}$ be $\sigma_1 \ge \cdots \ge \sigma_{r(\boldsymbol{A})}$ and $\widetilde{\boldsymbol{B}} = (\tilde{b}_{ij}) = \boldsymbol{U}\boldsymbol{\Sigma}_2\boldsymbol{V}^T$. By Cauchy-Schwarz's inequality,

$$\begin{aligned}Tr(\boldsymbol{\Sigma}_1^T\boldsymbol{U}\boldsymbol{\Sigma}_2\boldsymbol{V}^T) &= Tr(\boldsymbol{\Sigma}_1^T\widetilde{\boldsymbol{B}})\\ &= \sum_{i=1}^{r(\boldsymbol{A})}\sigma_i\tilde{b}_{ii}\\ &\le \sqrt{\sum_{i=1}^{r(\boldsymbol{A})}\sigma_i^2}\sqrt{\sum_{i=1}^{r(\boldsymbol{A})}\tilde{b}_{ii}^2}\\ &= \|\boldsymbol{A}\|_F\sqrt{\sum_{i=1}^{r(\boldsymbol{A})}\tilde{b}_{ii}^2}.\end{aligned} \tag{11}$$

Similarly, let the ordered singular values of $\boldsymbol{B}$ be $\eta_1 \ge \cdots \ge \eta_{r(\boldsymbol{B})}$. Then it suffices to show that $\sum_{i=1}^{r(\boldsymbol{A})}\tilde{b}_{ii}^2 \le \sum_{i=1}^{r(\boldsymbol{A})}\eta_i^2$. Assume, without loss of generality, that $\eta_i = 0$ if $i > r(\boldsymbol{B})$. Let $n = \min(n_1, n_2)$. By Cauchy-Schwarz's inequality,

$$\begin{aligned}\sum_{i=1}^{r(\boldsymbol{A})}\tilde{b}_{ii}^2 &= \sum_{i=1}^{r(\boldsymbol{A})}\big(\sum_{k=1}^{n}u_{ik}\eta_k v_{ik}\big)^2\\ &\le \sum_{i=1}^{r(\boldsymbol{A})}\big(\sum_{k=1}^{n}u_{ik}^2\eta_k^2\big)\big(\sum_{k=1}^{n}v_{ik}^2\big)\\ &\le \sum_{i=1}^{r(\boldsymbol{A})}\big(\sum_{k=1}^{n}u_{ik}^2\eta_k^2\big)\\ &= \sum_{k=1}^{n}\eta_k^2\big(\sum_{i=1}^{r(\boldsymbol{A})}u_{ik}^2\big)\\ &\le \sum_{k=1}^{r(\boldsymbol{A})}\eta_k^2,\end{aligned}$$

where the last step uses the fact that $\sum_{i=1}^{r(\boldsymbol{A})}u_{ik}^2 \le 1$ and $\sum_{k=1}^{n}\sum_{i=1}^{r(\boldsymbol{A})}u_{ik}^2 = \sum_{i=1}^{r(\boldsymbol{A})}\sum_{k=1}^{n}u_{ik}^2 \le r(\boldsymbol{A})$. A combination of the above bounds lead to the desired results. This completes the proof. □

6.1 Proof of Theorem 1

First partition $\boldsymbol{D}$ and $\boldsymbol{W}$ as follows:

$$\boldsymbol{D} = \begin{bmatrix}\boldsymbol{D}_{r(\boldsymbol{A})} & \boldsymbol{0}\\ \boldsymbol{0} & \boldsymbol{0}\end{bmatrix}, \quad \boldsymbol{W} = \begin{bmatrix}\boldsymbol{W}_{r(\boldsymbol{A})}\\ \boldsymbol{W}'\end{bmatrix},$$

then

$$\begin{aligned}\boldsymbol{D}\boldsymbol{Y} - \boldsymbol{W} &= \begin{bmatrix}\boldsymbol{D}_{r(\boldsymbol{A})}\boldsymbol{Y}_{r(\boldsymbol{A})}\\ \boldsymbol{0}\end{bmatrix} - \begin{bmatrix}\boldsymbol{W}_{r(\boldsymbol{A})}\\ \boldsymbol{W}'\end{bmatrix}\\ &= \begin{bmatrix}\boldsymbol{D}_{r(\boldsymbol{A})}\boldsymbol{Y}_{r(\boldsymbol{A})} - \boldsymbol{W}_{r(\boldsymbol{A})}\\ -\boldsymbol{W}'\end{bmatrix}.\end{aligned}$$

Evidently, only the first $r(\boldsymbol{A})$ rows of $\boldsymbol{Y}$ are involved in minimizing $\|\boldsymbol{D}\boldsymbol{Y} - \boldsymbol{W}\|_2^2$, which amounts to computing the global minimizer $\boldsymbol{Y}_{r(\boldsymbol{A})}^*$ of

$$\arg\min_{\boldsymbol{Y}_{r(\boldsymbol{A})}}\|\boldsymbol{D}_{r(\boldsymbol{A})}\boldsymbol{Y}_{r(\boldsymbol{A})} - \boldsymbol{W}_{r(\boldsymbol{A})}\|_F^2.$$

Then $\boldsymbol{Y}_{r(\boldsymbol{A})}^* = \boldsymbol{D}_{r(\boldsymbol{A})}^{-1}\mathcal{P}_s(\boldsymbol{W}_{r(\boldsymbol{A})})$ by non-singularity of $\boldsymbol{D}_{r(\boldsymbol{A})}$ and Lemma 1 with $s \le r(\boldsymbol{A})$. For $s > r(\boldsymbol{A})$, recall that, only the upper part of $\boldsymbol{Y}^*$ is relevant in minimizing (5). The result then follows. This completes the proof.

6.2 Proof of Theorem 2

For any $\boldsymbol{\Theta}_\lambda^*$ with $\lambda > 0$, let $s^* = r(\boldsymbol{\Theta}_\lambda^*)$. Next we prove by contradiction that $\boldsymbol{\Theta}_\lambda^* = \hat{\boldsymbol{\Theta}}^{s^*}$. Suppose $\boldsymbol{\Theta}_\lambda^* \ne \hat{\boldsymbol{\Theta}}^{s^*}$. By uniqueness of $\hat{\boldsymbol{\Theta}}^{s^*}$ given in Theorem 1 and the definition of minimization, $\|\boldsymbol{A}\hat{\boldsymbol{\Theta}}^{s^*}-\boldsymbol{Z}\|_F^2 < \|\boldsymbol{A}\boldsymbol{\Theta}_\lambda^*-\boldsymbol{Z}\|_F^2$. This, together with $r(\hat{\boldsymbol{\Theta}}^{s^*}) = r(\boldsymbol{\Theta}_\lambda^*)$, implies that

$$\|\boldsymbol{A}\hat{\boldsymbol{\Theta}}^{s^*} - \boldsymbol{Z}\|_F^2 + \lambda r(\hat{\boldsymbol{\Theta}}^{s^*}) < \|\boldsymbol{A}\boldsymbol{\Theta}_\lambda^* - \boldsymbol{Z}\|_F^2 + \lambda r(\boldsymbol{\Theta}_\lambda^*).$$

This contradicts to the fact that $\boldsymbol{\Theta}_\lambda^*$ is a minimizer of (7). The converse is revealed in the proof of Theorem 3.

6.3 Proof of Theorem 3

We prove the first conclusion by constructing such a solution path vector $\mathcal{S}$. Let $\mathcal{S}_0 = 0$, $\mathcal{S}_{r+1} = +\infty$. Define $\mathcal{S}_k$ for $1 \le k \le r$ as the solution of equation:

$$\|\mathcal{P}_{r-k+1}(\boldsymbol{Z})-\boldsymbol{Z}\|_F^2+\mathcal{S}_k(r-k+1) = \|\mathcal{P}_{r-k}(\boldsymbol{Z})-\boldsymbol{Z}\|_F^2+\mathcal{S}_k(r-k).$$

It follows that

$$\begin{aligned}\mathcal{S}_k &= \|\mathcal{P}_{r-k}(\boldsymbol{Z}) - \boldsymbol{Z}\|_F^2 - \|\mathcal{P}_{r-k+1}(\boldsymbol{Z}) - \boldsymbol{Z}\|_F^2 \\ &= \sum_{j=r-k+1}^{r} \sigma_j^2 - \sum_{j=r-k+2}^{r} \sigma_j^2 \\ &= \sigma_{r-k+1}^2,\end{aligned} \tag{12}$$

where σ_j is the jth largest non-zero singular value of $\boldsymbol{Z}$. By (12), $\mathcal{S}_k$ is increasing. In addition, by definition of S_k and S_{k+1}, whenever λ falls into the interval $[\mathcal{S}_k, \mathcal{S}_{k+1})$, the rank of a global minimizer $\boldsymbol{\Theta}^*$ of (8) would be no more than $r-k$ and larger than $r-k-1$. In other words, $\boldsymbol{\Theta}_\lambda^*$ is always of rank $r-k$ and is given by $\mathcal{P}_{r-k}(\boldsymbol{Z})$. Therefore, the constructed solution path vector $\mathcal{S}$ satisfies all the requirements in the theorem.

Moreover, when $\mathcal{S}_k \le \lambda < \mathcal{S}_{k+1}$,

$$\begin{aligned}g(\lambda) &= \|\mathcal{P}_{r-k}(\boldsymbol{Z}) - \boldsymbol{Z}\|_F^2 + \lambda\ r(\mathcal{P}_{r-k}(\boldsymbol{Z})) \\ &= \|\mathcal{P}_{r-k}(\boldsymbol{Z}) - \boldsymbol{Z}\|_F^2 + (r-k)\lambda.\end{aligned} \tag{13}$$

Since $\mathcal{P}_{r-k}(\boldsymbol{Z})$ is independent of λ, $g(\lambda)$ is a nondecreasing linear function of λ in each interval $[\mathcal{S}_k, \mathcal{S}_{k+1})$. Combined with the definition of the solution path vector $\mathcal{S}$, we conclude that $g(\lambda)$ is nondecreasing and piecewise linear with each element of $\mathcal{S}$ as a kink point, as shown in Figure 1. The proof is completed.

6.4 Proof of Theorem 4

The proof uses direct calculations. First we bound the Kullback-Leibler loss. By Theorem 1,

$$\boldsymbol{A}\hat{\boldsymbol{\Theta}}^{r_0} = \boldsymbol{U}_{r(\boldsymbol{A})}\mathcal{P}_{r_0}(\boldsymbol{W}_{r(\boldsymbol{A})}),$$

where

$$\boldsymbol{W}_{r(\boldsymbol{A})} = \boldsymbol{D}_{r(\boldsymbol{A})}\big(\boldsymbol{V}^T\boldsymbol{\Theta}^0\big)_{r(\boldsymbol{A})} + \big(\boldsymbol{U}^T\boldsymbol{e}\big)_{r(\boldsymbol{A})}.$$

Denote $\boldsymbol{B} = \boldsymbol{D}_{r(\boldsymbol{A})}\big(\boldsymbol{V}^T\boldsymbol{\Theta}^0\big)_{r(\boldsymbol{A})}$ with rank $r(\boldsymbol{B}) \le r_0$. Since the Frobenius-norm is invariant under orthogonal transformation, it follows that

$$\begin{aligned}&\|\boldsymbol{A}\hat{\boldsymbol{\Theta}}^{r_0} - \boldsymbol{A}\boldsymbol{\Theta}^0\|_F^2 \\ =&\|\boldsymbol{U}_{r(\boldsymbol{A})}\mathcal{P}_{r_0}(\boldsymbol{W}_{r(\boldsymbol{A})}) - \boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^T\boldsymbol{\Theta}^0\|_F^2 \\ =&\|\boldsymbol{U}_{r(\boldsymbol{A})}\mathcal{P}_{r_0}(\boldsymbol{W}_{r(\boldsymbol{A})}) - \boldsymbol{U}_{r(\boldsymbol{A})}\boldsymbol{D}_{r(\boldsymbol{A})}\big(\boldsymbol{V}^T\boldsymbol{\Theta}^0\big)_{r(\boldsymbol{A})}\|_F^2 \\ =&\|\mathcal{P}_{r_0}(\boldsymbol{W}_{r(\boldsymbol{A})}) - \boldsymbol{D}_{r(\boldsymbol{A})}\big(\boldsymbol{V}^T\boldsymbol{\Theta}^0\big)_{r(\boldsymbol{A})}\|_F^2 \\ =&\|\mathcal{P}_{r_0}\big(\boldsymbol{B} + \big(\boldsymbol{U}^T\boldsymbol{e}\big)_{r(\boldsymbol{A})}\big) - \boldsymbol{B}\|_F^2, \\ =&\|\mathcal{P}_{r_0}(\boldsymbol{B} + \boldsymbol{e}^\star) - \boldsymbol{B}\|_F^2,\end{aligned}$$

where $\boldsymbol{e}^\star = \big(\boldsymbol{U}^T\boldsymbol{e}\big)_{r(\boldsymbol{A})}$. From the definition of $\mathcal{P}_{r_0}(\boldsymbol{B}+\boldsymbol{e}^\star)$ we can conclude that:

$$\|\mathcal{P}_{r_0}(\boldsymbol{B}+\boldsymbol{e}^\star) - \boldsymbol{B} - \boldsymbol{e}^\star\|_F^2 \le \|\boldsymbol{B} - \boldsymbol{B} - \boldsymbol{e}^\star\|_F^2 = \|\boldsymbol{e}^\star\|_F^2,$$

which implies that,

$$\begin{aligned}\|\mathcal{P}_{r_0}(\boldsymbol{B}+\boldsymbol{e}^\star) - \boldsymbol{B}\|_F^2 &\le 2\langle \mathcal{P}_{r_0}(\boldsymbol{B}+\boldsymbol{e}^\star) - \boldsymbol{B}, \boldsymbol{e}^\star\rangle \\ &\le 2\|\mathcal{P}_{r_0}(\boldsymbol{B}+\boldsymbol{e}^\star) - \boldsymbol{B}\|_F\|\mathcal{P}_{r_0}(\boldsymbol{e}^\star)\|_F,\end{aligned}$$

where the last inequality follows from Lemma 2. Thus,

$$\|\mathcal{P}_{r_0}(\boldsymbol{B}+\boldsymbol{e}^\star) - \boldsymbol{B}\|_F^2 \le 4\|\mathcal{P}_{r_0}(\boldsymbol{e}^\star)\|_F^2 = 4\sum_{j=1}^{r_0}\sigma_j^2. \tag{14}$$

The risk bounds then follow.

Second we bound $\|\hat{\boldsymbol{\Theta}}^{r_0} - \boldsymbol{\Theta}^0\|_F^2$, which is equal to

$$\begin{aligned}&\|\boldsymbol{V}\begin{bmatrix}\boldsymbol{D}_{r(\boldsymbol{A})}^{-1}\mathcal{P}_{r_0}(\boldsymbol{W}_{r(\boldsymbol{A})}) \\ \boldsymbol{0}\end{bmatrix} - \boldsymbol{\Theta}^0\|_F^2 \\ =&\|\begin{bmatrix}\boldsymbol{D}_{r(\boldsymbol{A})}^{-1}\mathcal{P}_{r_0}(\boldsymbol{W}_{r(\boldsymbol{A})}) \\ \boldsymbol{0}\end{bmatrix} - \boldsymbol{V}^T\boldsymbol{\Theta}^0\|_F^2 \\ =&\|\boldsymbol{D}_{r(\boldsymbol{A})}^{-1}\mathcal{P}_{r_0}(\boldsymbol{W}_{r(\boldsymbol{A})}) - (\boldsymbol{V}^T\boldsymbol{\Theta}^0)_{r(\boldsymbol{A})}\|_F^2 + \|(\boldsymbol{V}^T\boldsymbol{\Theta}^0)_{r(\boldsymbol{A})^c}\|_F^2 \\ \le&\frac{1}{\sigma_{\min}^2 n}\|\mathcal{P}_{r_0}(\boldsymbol{W}_{r(\boldsymbol{A})}) - \boldsymbol{D}_{r(\boldsymbol{A})}(\boldsymbol{V}^T\boldsymbol{\Theta}^0)_{r(A)}\|_F^2 \\ &+ \|(\boldsymbol{V}^T\boldsymbol{\Theta}^0)_{r(A)^c}\|_F^2,\end{aligned}$$

where $\sigma_{r(\boldsymbol{A})}(n^{-1/2}\boldsymbol{A}) = n^{1/2}\sigma_{\min}$. If $p = r(\boldsymbol{A})$, then the last term vanishes. Thus

$$\begin{aligned}\|\hat{\boldsymbol{\Theta}}^{r_0} - \boldsymbol{\Theta}^0\|_F^2 &\le \frac{1}{\sigma_{r(\boldsymbol{A})}^2 n}\|\mathcal{P}_{r_0}(\boldsymbol{W}_{r(\boldsymbol{A})}) - \boldsymbol{D}_{r(\boldsymbol{A})}(\boldsymbol{V}^T\boldsymbol{\Theta}^0)_{r(\boldsymbol{A})}\|_F^2 \\ &\le \frac{4\sum_{j=1}^{r_0}\sigma_j^2}{\sigma_{\min}^2 n}.\end{aligned}$$

Finally, if $r(\boldsymbol{A}) \ge r_0$, $\mathbb{E}\|\hat{\boldsymbol{\Theta}}^{r_0} - \boldsymbol{\Theta}^0\|_F^2 \le \frac{4}{\sigma_{r(\boldsymbol{A})}^2}\mathbb{E}\big(\sum_{j=1}^{r_0}\sigma_j^2\big)$ and $\mathbb{E}\|\boldsymbol{A}\hat{\boldsymbol{\Theta}}^{r_0} - \boldsymbol{A}\boldsymbol{\Theta}^0\|_F^2 \le 4\mathbb{E}\big(\sum_{j=1}^{r_0}\sigma_j^2\big)$. In particular, if $r(\boldsymbol{A}) = r_0$, then

$$\begin{aligned}\mathbb{E}\|\hat{\boldsymbol{\Theta}}^{r_0} - \boldsymbol{\Theta}^0\|_F^2 &= \frac{1}{\sigma_{\min}^2}\frac{r_0 k}{n} \\ \mathbb{E}\|\boldsymbol{A}\hat{\boldsymbol{\Theta}}^{r_0} - \boldsymbol{A}\boldsymbol{\Theta}^0\|_F^2 &= \frac{r_0 k}{n}.\end{aligned}$$

7. EMPIRICAL EVALUATIONS

This section examines effectiveness of the proposed method and compares it with the nuclear-norm regularization as well as the SVP method [17]. One benefit of our method is that it can evaluate the approximation quality of the inexact ones. For the SVP, we choose a default initial value $\boldsymbol{0}$ for this local method since no other choices are guaranteed to deliver better performance. Our numerical experiment, which is not reported in here, suggests that the SVP method is indeed sensitive to the choice of the initial value. For nuclear-norm regularization, we select a suitable regularization parameter value giving the solution satisfying the rank constraint in (3).

7.1 Synthetic Data

Simulations are performed under (2). First, the $n \times p$ design matrix $\boldsymbol{A}$ is sampled, with each entry being iid $N(10,1)$. Second, the $p \times k$ truth $\boldsymbol{\Theta}^0$ is generated by multiplying a $p \times r$ matrix and an $r \times k$ matrix, each entry of which has a normal distribution over $N(10,1)$. The data matrix $\boldsymbol{Z}$ is then sampled according to (2) with $\boldsymbol{e}$ following iid $N(0,\sigma^2)$ with $\sigma = 0.5$.

For predictive performance and rank recovery, we compute the MSE $\|\boldsymbol{A}(\boldsymbol{\Theta}^0 - \hat{\boldsymbol{\Theta}}^{\hat{r}})\|_F^2$, the absolute difference $|\hat{r} - r_0|$ and record the training time for each method, averaged over 100 simulation replications on a test set of size $10n$, where $\hat{r}$ is tuned over integers in $[0, \min(n,p)]$ by an independent tuning set of size $2n$. We consider three possible situations, i.e., when $k = r_0 < p$, $k > p > r_0$ and $p > k > r_0$. To illustrate our theoretical conclusion in Theorem 4 that the prediction error bound does not depend on the value of p, we add one more case of $p > k > r_0$ with a larger value of p. The simulation results are summarized in Tables 2-4.

Table 2: Prediction results for the synthetic data: averaged MSEs as well as their standard deviations, for three competing methods based on the selected tuning parameters over 100 simulation replications. Our, SVP, and Nuclear-Norm refer to our method, the SVP method and nuclear-norm regularization method.

Set-up	Our	SVP	Nuclear-Norm
$k = r_0 < p, 5 = 5 < 99$	0.0110 (0.0065)	0.4591 (0.0386)	0.4130 (0.0360)
$k > p > r_0, 50 > 40 > 5$	12.0010 (0.6125)	10166.5052 (1236.3237)	7709.0779 (960.6982)
$p > k > r_0, 50 > 40 > 5$	13.1945 (0.5682)	10185.7699 (1132.3515)	8116.2653 (943.6058)
$p > k > r_0, 99 > 40 > 5$	121.4019 (70.3054)	14597.0728 (1185.7200)	13153.3217 (1056.5155)

Table 3: Rank recovery for the synthetic data: averaged values of $|\hat{r} - r_0|$ as well as their standard deviations, for three competing methods based on the selected tuning parameters over 100 simulation replications. Our, SVP, and Nuclear-Norm refer to our method, the SVP method and nuclear-norm regularization method.

Set-up	Our	SVP	Nuclear-Norm
$k = r_0 < p, 5 = 5 < 99$	0.8400 (0.8005)	0.0000 (0.0000)	0.0000 (0.0000)
$k > p > r_0, 50 > 40 > 5$	0.0000 (0.0000)	3.5300 (1.6358)	1.6400 (0.6439)
$p > k > r_0, 50 > 40 > 5$	0.0000 (0.0000)	4.0800 (1.3830)	1.6500 (0.5925)
$p > k > r_0, 99 > 40 > 5$	0.0000 (0.0000)	2.9800 (1.9121)	0.2200 (0.4399)

As suggested in Tables 2 and 3, our exact method is much more precise in prediction in all cases and rank recovery except one case of $k = r_0 = 5 < p = 99$, than the other two methods. This is in agreement with the theoretical results in Theorem 4 that exact reconstruction of the oracle estimator is achieved through tuning. Note that the best MSE value does not necessarily yield best rank recovery, as in the case of $k = r_0 = 5 < p = 99$, which is due to the bias/variance trade-off phenomenon. As indicated in Table 4, our method is, on average, 10-20 times faster than the other two.

7.2 MIT logo Recovery

Next we examine the three competing methods for reconstructing the MIT logo image, which was studied in [27, 17]. The original logo is displayed in Figure 2, where we use the gray image of size 44×85 and the image matrix has rank 7.

Figure 2: Original MIT logo image.

Our objective is to reconstruct this image from its noisy version and examine the quality of reconstruction. Towards this end, we sample the design matrix $\boldsymbol{A}$ with each entry being iid $N(0, 1)$, where the sample size n ranges from 20 to 80. To generate a noisy version, we add random error sampled from $N(0, 0.5^2)$ to each element of the sample data. The reconstruction results are displayed in Figure 3 for the three methods with the default initial value $\mathbf{0}$ for the SVP.

Visually, our method delivers highly competitive performance as compared to the other two methods, as displayed in Figure 3, and yields nearly perfect reconstruction when the size of the design matrix $\boldsymbol{A}$ becomes larger, say $n = 60$ and $n = 80$. For all the methods, better reconstruction can be reached as n increases, and comparable results are achieved when n is small, say $n = 20$ and $n = 40$. This conclusion is consistent with our theoretical analysis.

In practice, the exact rank of the matrix to be estimated is unknown but may reasonably be assumed to be small. In this sense, s needs to be tuned or estimated. Next we investigate the effect of estimation of s with regard to the reconstruction quality, measured by the MSE and the recovery rate. The latter is defined as the ratio $\|\hat{\boldsymbol{\Theta}}^s - \boldsymbol{\Theta}^0\|_F / \|\boldsymbol{\Theta}^0\|_F$, commonly used to measure the quality of parameter estimation. In particular, we display both of them as a function of s in Figures 4 and 5, where s is defined in (3).

For the relative recovery error, a clear transition of our solution occurs around $s = 44$, after which a perfect recovery is achieved, whereas no improvement occurs for the other two methods when s increases. In the case of an underdetermined $\boldsymbol{A}$, i.e., $\boldsymbol{A}$ is not of full column rank, all three methods produce similar recovery results. For the MSEs, our method yields the perfect result zero when $s \geq 7$ and a reasonably small value when $s = 3, 5$, whereas the other two methods lead to elevated MSEs as s increases. This is in accordance with our theory which suggests that our method constructs the oracle estimator giving perfect image reconstruction when $s \leq 7$.

8. CONCLUSION

This paper considers a nonconvex least squares formulation based on the rank constraint/regularization. We establish the optimality of the global solution, against any other ones. Experimental results on synthetic and real data demonstrate the efficiency and effectiveness of the proposed algorithm. In future work, we plan to expand the present work to a general loss function.

9. ACKNOWLEDGEMENT

This work was supported by NSF grants IIS-0953662 and DMS-0906616, NIH grants R01LM010730, 2R01GM081535-01, and R01HL105397.

10. REFERENCES

[1] J. Abernethy, F. Bach, T. Evgeniou, and J. Vert.

Table 4: Run time for the synthetic data: average training time in seconds for three competing methods based on the selected tuning parameters over 100 simulation replications. Our, SVP, and Nuclear-Norm refer to our method, the SVP method and nuclear-norm regularization.

Set-up	Our	SVP	Nuclear-Norm
$k = r_0 < p, 5 = 5 < 99$	0.0224 (0.0045)	14.1523 (0.3933)	0.9766 (0.0074)
$k > p > r_0, 50 > 40 > 5$	0.0349 (0.0031)	41.2122 (1.8511)	5.8937 (0.2945)
$p > k > r_0, 50 > 40 > 5$	0.0368 (0.0025)	29.5856 (1.1786)	5.7447 (0.1173)
$p > k > r_0, 99 > 40 > 5$	0.0542 (0.0039)	79.9850 (3.6724)	7.6046 (0.1011)

(a) n = 20 (b) n = 40

(c) n = 60 (d) n = 80

Figure 3: Reconstruction of a noisy version of the MIT logo with varying sampling size n. From left to right (for each case of n): SVP; nuclear-norm regularization; our method.

Low-rank matrix factorization with attributes. *Arxiv preprint cs/0611124*, 2006.

[2] Y. Amit, M. Fink, N. Srebro, and S. Ullman. Uncovering shared structures in multiclass classification. In *Proceedings of the 24th Annual International Conference on Machine Learning*, pages 17–24. ACM, 2007.

[3] T. Anderson. Estimating linear restrictions on regression coefficients for multivariate normal distributions. *The Annals of Mathematical Statistics*, 22(3):327–351, 1951.

[4] T. Anderson. Asymptotic distribution of the reduced rank regression estimator under general conditions. *The Annals of Statistics*, 27(4):1141–1154, 1999.

[5] T. André, R. Nowak, and B. Van Veen. Low-rank estimation of higher order statistics. *Signal Processing, IEEE Transactions on*, 45(3):673–685, 1997.

[6] A. Argyriou, T. Evgeniou, and M. Pontil. Multi-task feature learning. *Advances in Neural Information Processing Systems*, 19:41, 2007.

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

[8] J. Cai, E. Candès, and Z. Shen. A singular value thresholding algorithm for matrix completion. *SIAM Journal of Optimization*, 20(4):1956–1982, 2010.

[9] E. Candes and Y. Plan. Matrix completion with noise. *Arxiv preprint arXiv:0903.3131*, 2009.

[10] E. Candès and B. Recht. Exact matrix completion via convex optimization. *Foundations of Computational Mathematics*, 9(6):717–772, 2009.

[11] J. Chen, J. Liu, and J. Ye. Learning incoherent sparse and low-rank patterns from multiple tasks. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1179–1188. ACM, 2010.

[12] C. Eckart and G. Young. The approximation of one matrix by another of lower rank. *Psychometrika*, 1(3):211–218, 1936.

[13] M. Fazel, H. Hindi, and S. Boyd. A rank minimization heuristic with application to minimum order system approximation. In *American Control Conference, 2001. Proceedings of the 2001*, volume 6, pages 4734–4739. IEEE, 2001.

[14] S. Friedland and A. Torokhti. Generalized rank-constrained matrix approximations. *SIAM Journal on Matrix Analysis and Applications*, 29(2):656–659, 2007.

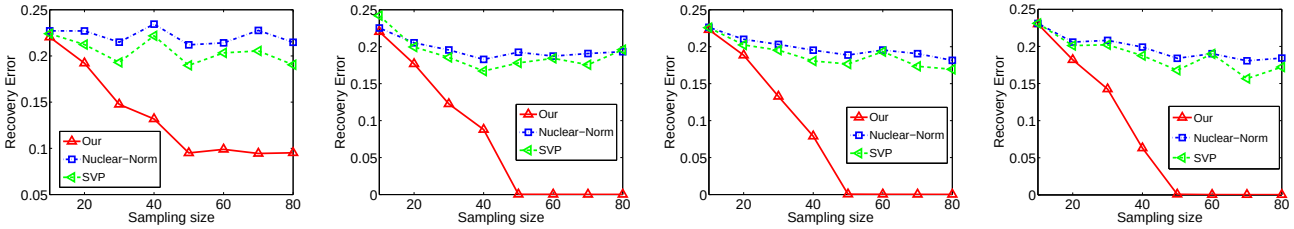
[15] A. Izenman. Reduced-rank regression for the multivariate linear model. *Journal of multivariate analysis*, 5(2):248–264, 1975.

[16] M. Jaggi and M. Sulovskỳ. A Simple Algorithm for Nuclear Norm Regularized Problems. In *Proceedings of the 27th Annual International Conference on Machine Learning*, 2010.

[17] P. Jain, R. Meka, and I. Dhillon. Guaranteed rank minimization via singular value projection. *Advances in Neural Information Processing Systems*, 23:937–945, 2010.

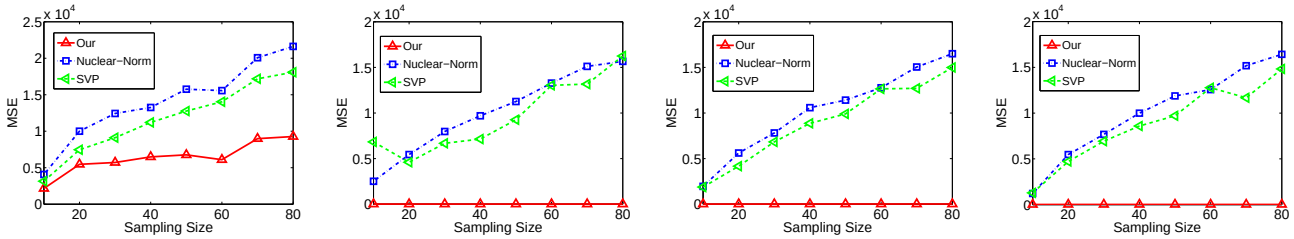
[18] H. Ji, C. Liu, Z. Shen, and Y. Xu. Robust video denoising using low rank matrix completion. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 1791–1798. IEEE, 2010.

[19] S. Ji and J. Ye. An accelerated gradient method for trace norm minimization. In *Proceedings of the 26th*



(a) s = 3 (b) s = 5 (c) s = 7 (d) s = 10

Figure 4: Relative recovery error as a function of the sampling size for the MIT logo image under different rank constraints.



(a) s = 3 (b) s = 5 (c) s = 7 (d) s = 10

Figure 5: MSE as a function of the sampling size for the MIT logo image under different rank constraints. Note that the MSE of our method in (b) is not zero but indistinguishable from zero, which is unlike (c) and (d) in which the MSEs are identical to zero.

Annual International Conference on Machine Learning, pages 457–464, 2009.

[20] B. Kulis, A. Surendran, and J. Platt. Fast low-rank semidefinite programming for embedding and clustering. In *Eleventh International Conference on Artifical Intelligence and Statistics, AISTATS 2007*, 2007.

[21] G. Liu, Z. Lin, S. Yan, J. Sun, Y. Yu, and Y. Ma. Robust recovery of subspace structures by low-rank representation. *Arxiv preprint arXiv:1010.2955*, 2010.

[22] X. Luo. High dimensional low rank and sparse covariance matrix estimation via convex minimization. *Arxiv preprint arXiv:1111.1133*, 2011.

[23] B. K. Natarajan. Sparse approximation solutions to linear systems. *SIAM J. Comput.*, 24(2):227–234, 1995.

[24] S. Negahban and M. J. Wainwright. Estimation of (near) low-rank matrices with noise and high-dimensional scaling. *Annals of Statistics*, 39(2):1069–1097, 2011.

[25] T. Pong, P. Tseng, S. Ji, and J. Ye. Trace norm regularization: Reformulations, algorithms, and multi-task learning. *SIAM Journal on Optimization*, 20:3465–3489, 2010.

[26] C. Rao. Matrix approximations and reduction of dimensionality in multivariate statistical analysis. *Multivariate analysis*, 5:3–22, 1980.

[27] B. Recht, M. Fazel, and P. Parrilo. Guaranteed Minimum-Rank Solutions of Linear Matrix Equations via Nuclear Norm Minimization. *SIAM Review*, 52(471), 2010.

[28] G. Reinsel and R. Velu. *Multivariate reduced-rank regression: theory and applications*. Springer New York, 1998.

[29] B. Savas and I. Dhillon. Clustered low rank approximation of graphs in information science applications. In *Proceedings of the SIAM Data Mining Conference.*, 2011.

[30] S. Shalev-Shwartz, A. Gonen, and O. Shamir. Large-Scale Convex Minimization with a Low-Rank Constraint. In *Proceedings of the 28th Annual International Conference on Machine Learning*, 2011.

[31] X. Shen and H. Huang. Optimal model assessment, selection, and combination. *Journal of the American Statistical Association*, 101(474):554–568, 2006.

[32] D. Sondermann. Best approximate solutions to matrix equations under rank restrictions. *Statistical Papers*, 27(1):57–66, 1986.

[33] N. Srebro, J. Rennie, and T. Jaakkola. Maximum-margin matrix factorization. *Advances in Neural Information Processing Systems*, 17:1329–1336, 2005.

[34] K. Toh and S. Yun. An accelerated proximal gradient algorithm for nuclear norm regularized linear least squares problems. *Pacific J. Optim*, 6:615–640, 2010.

[35] J. Wright, A. Ganesh, S. Rao, and Y. Ma. Robust principal component analysis: Exact recovery of corrupted low-rank matrices via convex optimization. *Advances in Neural Information Processing Systems*, 2009.