Optimal exact least squares rank minimization *

Shuo Xiang[1], Yunzhang Zhu[2], Xiaotong Shen[2] and Jieping Ye[1]

Summary

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 derive a general closed-form for a global minimizer of a nonconvex least squares problem, in lieu of the common practice that seeks a local solution or a surrogate solution based on nuclear-norm regularization. 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 for 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.

Key Words: Exact global minimizer, nonconvex, optimality, rank approximation.

1 Introduction

In multivariate analysis, estimation of lower-dimensional structures has received attention in statistics, signal processing and machine learning. One type of such a structure is low-rank of matrices, 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 [4], multi-class classification [5], matrix completion [8, 16], collaborative filtering [26], computer vision [29, 14], 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 $\boldsymbol{z}_i$

*[1] Department of Computer Science and Engineering, Arizona State University Tempe, AZ 85281; [2]School of Statistics, University of Minnesota, Minneapolis, MN 55455.

follows

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

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) is commonly used in compressed sensing, reduces to the linear model when $k = 1$, and becomes a multivariate autoregressive model with $\boldsymbol{a}_i = \boldsymbol{z}_{i-1}$. In (1), the rank of $\boldsymbol{\Theta}$, denoted by $r(\boldsymbol{\Theta})$, and can be expressed in a matrix form

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

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 pair observation vectors $(\boldsymbol{a}_i, \boldsymbol{z}_i)_{i=1}^n$, with *a priori* knowledge that $r(\boldsymbol{\Theta})$ is low relative to $\min(n, k, p)$, where the number of unknown parameters kp 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, that is, the L_2-norm of all entries of a matrix, and s is an integer-valued tuning parameter with $1 \le s \le \min(n, k, p)$. The problem (3) is nonconvex and NP-hard [**?**], which is like the L_0-function in feature selection. Therefore an exact global solution to (3) has not been available as well as its statistical properties, due primary 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 m, p

held fixed, see, for example, [1, 2, 13, 23, 21]. For (3), the rank constraint is looser, in which two major computational approaches have been proposed. The first involves regularization. A convex envelope of the rank [12] such as the nuclear-norm in (3),is used in a parallel fashion as the L_1 for L_0-norm, which can be solved by efficient algorithms [8, 18, 28, 17]. In some cases, the solution of this convex problem coincides with a global minimizer of (3) under certain isometry assumptions [22]. However, these assumptions can be strong and difficult to check. Recently, [7] obtained a global minimizer of a regularized version of (3). The second attacks (3) by approximating the rank iteratively by calculating the largest singular vectors through greedy search [24], and by singular value projection (SVP) through a local gradient method [16]. Under weaker isometry assumptions [22, 9, 10], these methods guarantee an exact solution of (3) but suffer from the same difficulty as the regularization method [24], although they have achieved promising results on both simulated and real-world data. Theoretically, some loss error bounds of the first approach are obtained in [20] under Frobenius-norm, and rank selection consistency is established in [7]. Unfortunately, to our knowledge, little is known about a solution for (3).

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 an entire regularization solution path at a cost of computing one solution at one regularization parameter value. Second, we establish optimality for a global minimizer of (3). In particular, the proposed method is optimal against any method 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 as well as tuning parameter s. Such a result may not share by its regularized counterpart with a continuum tuning parameter. In addition,

the method enjoys a higher degree of accuracy of parameter estimation than nuclear-norm rank estimation.

After the first draft of this paper was completed, we were aware that [**?**] and [**?**] gave an expression of Theorem 1. However, neither paper considers computational and statistical aspects of the solution. Inevitably, some partial overlaps 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 some simulation results, where several rank minimization methods are compared. Section 7 concludes. The appendix contains the proof.

2 Proposed method: closed-form solution

This section derives a closed-form solution to (3). One strategy is to reduce (3) to a simpler problem through the singular value decomposition (SVD) of matrix $\boldsymbol{A}$ and certain properties of the rank. Before proceeding, we present a motivating lemma, known as the Eckart-Young theorem [11, 27].

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 unique 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 rank is 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 a simpler problem (4).

When $\boldsymbol{A}$ is nonsingular, (3) has a unique 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, assume, without loss of generality, that $r(\boldsymbol{A}) \geq s$, because $s \leq \min(n, k, p)$ in (3). Given 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 $\|A\boldsymbol{\Theta} - \boldsymbol{Z}\|_F = \|\boldsymbol{U}^T(A\boldsymbol{\Theta} - \boldsymbol{Z})\|_F = \|\boldsymbol{D}\boldsymbol{V}^T\boldsymbol{\Theta} - U^T\boldsymbol{Z}\|_F$. This follows from the fact that the Frobenius-norm is invariant under any orthogonal transformation. Let $\boldsymbol{Y} = \boldsymbol{V}^T\boldsymbol{\Theta}$ and $\boldsymbol{W} = \boldsymbol{U}^T\boldsymbol{Z}$. Then $r(\boldsymbol{Y}) = r(\boldsymbol{\Theta})$. Solving (3) amounts to solving an equivalent form:

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

Consequently, a global minimizer of (3) becomes $V\boldsymbol{Y}^*$, where $\boldsymbol{Y}^*$ is a global minimizer of (5). Next we give a closed form solution $\boldsymbol{Y}^*$ of (5).

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

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

where $\boldsymbol{D}_{r(\boldsymbol{A})}$ is a diagonal matrix consisting of all the nonzero singular value of $\boldsymbol{A}$, $\boldsymbol{a}$ is any vector, and $\boldsymbol{W}_{r(\boldsymbol{A})}$ consists of the first $r(\boldsymbol{A})$ rows of $\boldsymbol{W}$. If $s > r(\boldsymbol{A})$, simply replace the zero part below $\boldsymbol{D}_{r(\boldsymbol{A})}^{-1}\mathcal{P}_s(\boldsymbol{W}_{r(\boldsymbol{A})})$ with a specific matrix to make $r(\boldsymbol{Y}^*)$ to be s. Most importantly, (5) has a unique global minimizer $\hat{\boldsymbol{\Theta}}^s = \boldsymbol{V}\boldsymbol{Y}^*$ when $r(\boldsymbol{A}) = p$.

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 $\boldsymbol{a} = \mathbf{0}$ for $\boldsymbol{Y}^*$ subsequently.

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

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)

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

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

[1] More specifically, for a matrix of dimension $n \times p$, the SVD has a complexity of $O(\min\{s^2p, p^2n\})$, 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.

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 path of regulation solution 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 minimal cost function value $g(\lambda)$ 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 solution path $\boldsymbol{\Theta}^*_\lambda$ as function of λ and the corresponding $\hat{\boldsymbol{\Theta}}^s$ as a function of s. The monotone property is evident together with the equivalence property.

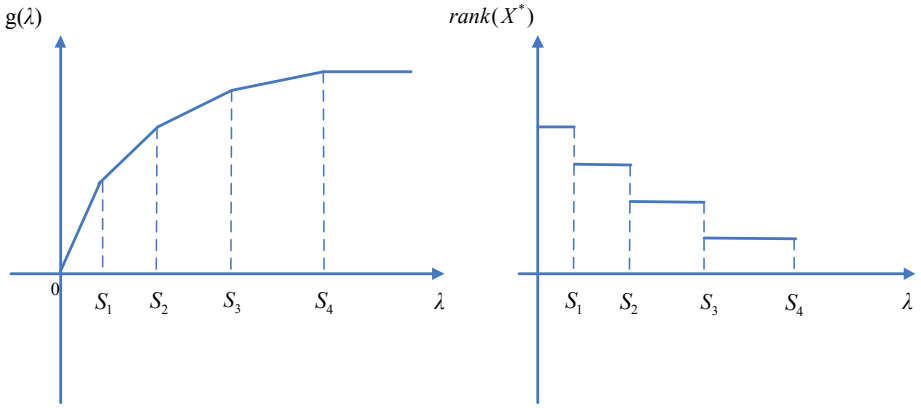


Figure 1: Piecewise linearity of $g(\cdot)$ and the rank of the optimal solution with respect to λ.

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 $\mathcal{S}$, 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`($\boldsymbol{\Theta}$,$\boldsymbol{Z}$)

`Initialize`: $\mathcal{S}_0 = 0$, $\boldsymbol{\Theta}_0 = \boldsymbol{Z}$, $r = r(\boldsymbol{Z})$

`Perform SVD on` $\boldsymbol{Z}$: $\boldsymbol{Z} = U\boldsymbol{D}V^T$

for $i = r$ **down to** 1 **do**

  $\mathcal{S}_{r-i+1} = \sigma_i^2$

  $\boldsymbol{\Theta}_{r-i+1} = \boldsymbol{\Theta}_{r-i} - \sigma_i u_i v_i^T$

end for

return $\mathcal{S}$, $\boldsymbol{\Theta}$

Algorithm 2 requires only one SVD operation, hence that its complexity is of 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 $\|A\boldsymbol{\Theta} - \boldsymbol{Z}\|_F^2 = \|\boldsymbol{D}\boldsymbol{Y} - \boldsymbol{W}\|_F^2 = \|\boldsymbol{W}'\|_F^2 + \|\boldsymbol{D}_r\boldsymbol{Y}_r - \boldsymbol{W}_r\|_F^2$. After ignoring constant term $\|\boldsymbol{W}'\|_F^2$, we solve $\min_{\boldsymbol{Y}_r}(\|\boldsymbol{D}_r\boldsymbol{Y}_r - \boldsymbol{W}_r\|_F^2 + \lambda\ r(\boldsymbol{Y}_r))$. 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 theoretical investigation of least squares rank minimization, which remain largely unexplored, although nuclear-norm regularization has been studied. In particular, we will reveal what is the best performance for estimating rank as well as the optimal risk for parameter estimation. Moreover, we will establish 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 $\min_{\boldsymbol{\Theta}} \|\boldsymbol{A}\boldsymbol{\Theta} - \boldsymbol{Z}\|_F^2$ 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 exact rank recovery, in addition to reconstruction of the optimal performance of the oracle estimator. In other words, the proposed method is optimal against any method such as nuclear norm

rank minimization.

Given the design matrix $\boldsymbol{A}$, we study 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 $K(\hat{\boldsymbol{\Theta}}^s, \boldsymbol{\Theta}_0) = (2\sigma^2 n)^{-1} \sum_{i=1}^n \|\boldsymbol{a}_i^T(\hat{\boldsymbol{\Theta}}^s - \boldsymbol{\Theta}_0)\|_2^2 = (2\sigma^2 n)^{-1}\|\boldsymbol{A}(\hat{\boldsymbol{\Theta}}^s - \boldsymbol{\Theta}_0)\|_F$ is the Kullback-Leibler loss, under (3), 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, or $r(\boldsymbol{A}) = p$.

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* $\hat{\boldsymbol{\Theta}}^{r_0} = \hat{\boldsymbol{\Theta}}^0$ *under* $\mathbb{P}$*, when* $r_0 \leq r(\boldsymbol{A})$*. 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 for noisy data, which is attributed to discreteness of the rank and is analogous to 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 $\frac{r_0 k}{\sigma^2_{\min} n}$. 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 $kp >> n$. For estimation, however, p enters the risk through $\sigma^2_{\min}$, where p can not be larger than n, or $\max(k, p) \leq n$.

5 Tuning

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

The predictive performance of $\hat{\boldsymbol{\Theta}}^s$ is measured by $MSE(\hat{\boldsymbol{\Theta}}^s) = n^{-1} E \|\boldsymbol{Z} - \boldsymbol{A}\hat{\boldsymbol{\Theta}}^s\|_F^2$, which is proportional to the risk $R(\hat{\boldsymbol{\Theta}}^s)$ where the expectation E 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 [25] through data perturbation without a tuning set.

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

where $\widehat{\mathrm{Cov}}(\boldsymbol{z}_{il}, (\boldsymbol{a}_i \hat{\boldsymbol{\Theta}}^s{}_l)$ is the estimated covariance between the lth component of $\boldsymbol{z}_i$ and the lth component of $\boldsymbol{a}_i \hat{\boldsymbol{\Theta}}^s$. 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 [25] is applicable. Specifically, sample $\boldsymbol{e}_i^*$ from $N(\boldsymbol{0}, \sigma^2 I_{k\times k})$ and let $\boldsymbol{Z}^* = (1-\tau)\boldsymbol{Z} + \tau\tilde{\boldsymbol{Z}}$, $\widehat{\mathrm{Cov}}(\boldsymbol{z}_{il}, (\boldsymbol{a}_i \hat{\boldsymbol{\Theta}}^s)_l) = \tau^{-1}\mathrm{Cov}^*(\boldsymbol{z}_{il}^*, (\boldsymbol{a}_i \hat{\boldsymbol{\Theta}}^{*s})_l)$ 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 Numerical examples

This section examines effectiveness of the proposed exact methods, and compares against the nuclear-norm regularization as well as the SVP method [20]. One benefit of our method is that it can evaluate the approximation quality of non-exact methods Note that nuclear-norm regularization is not in the form of (3). For a fair comparison, we run the nuclear-norm regularization with multiple regularization values and select an suitable regularization parameter value giving the solution satisfying the rank constraint in (3).

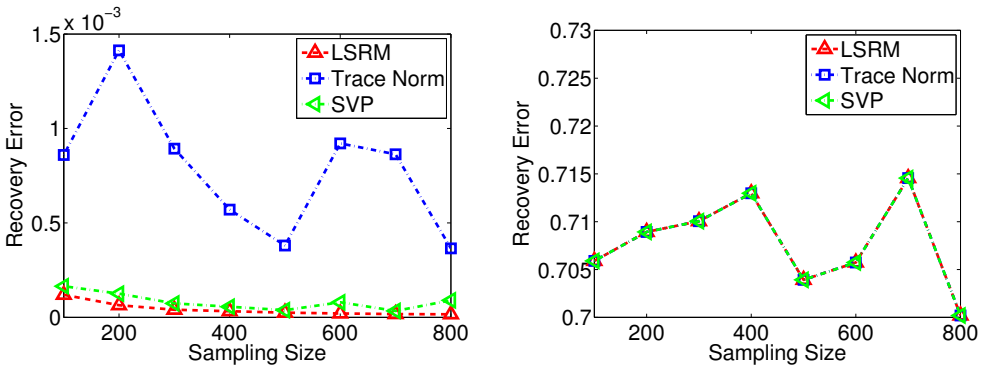
Synthetic Data. Simulations are performed under (2). First, the $n \times p$ design matrix $\boldsymbol{A}$ is sampled, with each entry being iid $N(0,1)$. Second, the $p \times k$ truth $\boldsymbol{\Theta}^0$ is generated by multiplying a $p \times r$ matrix and a $r \times k$ matrix, each entry of which has a uniform distribution over $(0,1)$ and $r = r(\boldsymbol{A})$. 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 rank recovery and predictive performance, we compute the absolute difference $|\hat{r} - r_0|$ and the MSE $\|\boldsymbol{A}(\boldsymbol{\Theta}^0 - \hat{\boldsymbol{\Theta}}^{\hat{r}})\|_F^2$, 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 n. For each competing method, several cases are examined with ??$n = 0.5p$ and ??$n = 2p$, corresponding to full row rank and full column rank $\boldsymbol{A}$, respectively. The results are summarized in Table 1 and displayed in Figure 2.

Table 1: Average recovery error for all three algorithms.

	SVP	Trace Norm	LSRM
$n = 2p$	7.0780e-5	7.6482e-4	3.6072e-5
$n = 0.5p$	0.7067	0.7078	0.7067

In the case of $n = 2p$, the recovery result is significantly better than the other case and SVP is comparable to the exact LSRM. While in the case of $n = 0.5p$, all three algorithms perform nearly identically. Our results verify the effectiveness of the SVP algorithm. Recall



(a) $n = 2p$ (b) $n = 0.5p$

Figure 2: Recovery error on synthetical data with growing sample size n.

Figure 3: Original MIT logo image.

that SVP is based on the direct rank constraint, which is not guaranteed to find the global optimum. However our results show that it achieves near-optimal results, demonstrating its potential in other applications. Our results also imply that nuclear-norm regularization is more effective for the high-dimensional case where p is larger than n.

MIT logo Recovery. Next, we evaluate the performance of the three algorithms for recovering the MIT logo image, which naturally yields a low-rank structure and has been utilized in [22, 16]. The original logo is shown in Figure 3, which is of size 44×85 and has rank 7, and we only use the gray image in this experiment.

We generate the design matrix $\boldsymbol{A}$ at random with the sample size of n, ranging from 20 to 80. Gaussian noises with zero mean and 0.5 standard deviation are added to each element of the sampled data. The results are listed in Figure 4, from which we can observe that: 1) As expected, for all three algorithms, better reconstruction results can be achieved when the design matrix becomes larger, i.e., when $n = 60$ and $n = 80$. 2) Both SVP and nuclear-norm

regularization produce comparable results as our method in the case of $n = 20$ and $n = 40$. This is consistent with the observation from the last experiment. We also observe that in the last two cases ($n = 60$ and $n = 80$), the recovery by the exact LSRM is nearly perfect. This is also consistent with the analysis in Remark 2 of Section 2.

(a) m = 20 (b) m = 40

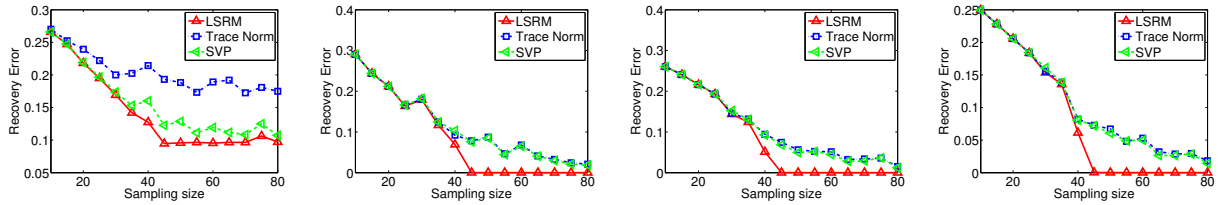
(c) m = 60 (d) m = 80

Figure 4: Recovery result of the MIT logo with varying sampling size m. From left to right (for each case of n): SVP; nuclear-norm regularization; our method-LSRM.

Note that in practice the exact rank of the matrix to be estimated is unknown. Next, we vary the value of the rank r in the constraint and show the trend of the relative error with a growing sampling size in Figure 5. Again we can see a clear transition of the exact LSRM around the sampling size of 44, after which a perfect recovery is achieved. In the case of an under-determined $\boldsymbol{A}$, i.e., $\boldsymbol{A}$ is not of full column rank, all three algorithms produce similar recovery result. These results further demonstrate that in the high-dimensional case, both SVP and the trace norm produce reasonable approximations. Our experimental study partially validates the current practice of using these approximation algorithms.

7 Conclusion

This paper derives an exact closed-form global minimization for nonconvex least square rank minimization. In addition, an efficient pathwise algorithm is also developed for its regularized counterpart. Moreover, we establish optimality of the global solution, against



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

Figure 5: Relative recovery error of the MIT logo image with different rank constraints.

any other solutions.

An exact global solution seems rare for nonconvex minimization. However, it is possible to expand the present work to a general loss function, by solving a sequence of weighted least squares rank minimization problems iteratively. We will pursue alone this direction for a general rank minimization problem.

8 Appendix

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,*

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

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

Proof of Lemma 2: 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

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

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 $\sigma_1 \geq \cdots \geq \sigma_{r(\boldsymbol{A})}$ and $\widetilde{\boldsymbol{B}} = (\tilde{b}_{ij}) = \boldsymbol{U}\boldsymbol{\Sigma}_2\boldsymbol{V}^T$. By Cauchy-Schwarz's inequality,

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

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

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

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

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

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

Evidently, only the first $r(\boldsymbol{A})$ rows of $\boldsymbol{Y}$ are involved in minimizing $\|\boldsymbol{DY}-\boldsymbol{W}\|_2^2$, 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 $\boldsymbol{Y}^*_{r(\boldsymbol{A})}=\boldsymbol{D}^{-1}_{r(\boldsymbol{A})}\mathcal{P}_s(\boldsymbol{W}_{r(\boldsymbol{A})})$ 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.

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 $\boldsymbol{\Theta}^*_\lambda\neq\hat{\boldsymbol{\Theta}}^{s^*}$. By uniqueness of $\hat{\boldsymbol{\Theta}}^{s^*}$ given in Theorem 2 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 minimized. This establishes the result. The converse can be proved similarly using the proof of **Theorem 3**. This completes the proof.

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

$$\mathcal{S}_k=\|\mathcal{P}_{r-k}(\boldsymbol{Z})-\boldsymbol{Z}\|_F^2-\|\mathcal{P}_{r-k+1}(\boldsymbol{Z})-\boldsymbol{Z}\|_F^2=\sum_{j=r-k+1}^{r}\sigma_j^2-\sum_{j=r-k+2}^{r}\sigma_j^2=\sigma^2_{r-k+1}. \tag{12}$$

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 $[\mathcal{S}_k,\mathcal{S}_{k+1})$, $r(\boldsymbol{\Theta}^*)$ for 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 $\mathcal{S}_k \leq \lambda < \mathcal{S}_{k+1}$,

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

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. This completes the proof.

Proof of Theorem 4: The proof uses direct calculations.

First we bound the Kullback-Leibler loss. By Theorem 1, $\boldsymbol{A}\hat{\boldsymbol{\Theta}}^{r_0} = \boldsymbol{U}_{r(\boldsymbol{A})}\mathcal{P}_{r_0}(\boldsymbol{W}_{r(\boldsymbol{A})})$, with $\boldsymbol{W}_{r(\boldsymbol{A})} = \boldsymbol{D}_{r(\boldsymbol{A})}\big(\boldsymbol{V}^T\boldsymbol{\Theta}^0\big)_{r(\boldsymbol{A})} + \big(\boldsymbol{U}^T\boldsymbol{e}\big)_{r(\boldsymbol{A})}$. It follows from that orthogonal matrices are invariant under $\|\cdot\|_F$ that

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

where $\boldsymbol{B} = \boldsymbol{D}_{r(\boldsymbol{A})}\big(\boldsymbol{V}^T\boldsymbol{\Theta}^0\big)_{r(\boldsymbol{A})}$ with rank $r(\boldsymbol{B}) \leq r_0$, $\boldsymbol{e}^\star = \big(\boldsymbol{U}^T\boldsymbol{e}\big)_{r(\boldsymbol{A})}$. By definition of $\mathcal{P}_{r_0}(\boldsymbol{B} + \boldsymbol{e}^\star)$,

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

which implies that,

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

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

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 $r(\boldsymbol{A}) = r_0$, $\mathbb{E}\|\hat{\boldsymbol{\Theta}}^{r_0}-\boldsymbol{\Theta}^0\|_F^2 = \frac{1}{\sigma_{\min}^2}\frac{r_0 k}{n}$ and $\mathbb{E}\|\boldsymbol{A}\hat{\boldsymbol{\Theta}}^{r_0}-\boldsymbol{A}\boldsymbol{\Theta}^0\|_F^2 = \frac{r_0 k}{n}$.

References

[1] Anderson, T.W. (1951). Estimating linear restrictions on regression coefficients for multivariate normal distributions. *Annals of Mathematical Statistics*, **22**, 327-351.

[2] Anderson, T.W. (1999). Asymptotic distribution of the reduced rank regression estimator under general conditions. *Ann. Statist.*, **27**, 1141-1154.

[3] Aharon, M., Elad, M. and Bruckstein, A. (2009). An algorithm for designing overcomplete dictionaries for sparse representation, *IEEE Transactions on Signal Processing*,

11(54), 4311-4322.

[4] Argyriou, A., Evgeniou, T. and Pontil, M. (2007). Multi-task feature learning, *Advances in neural information processing systems*, **19**, 41-41.

[5] Amit, Y., Fink, M., Srebro, N. and Ullman, S. (2007). Uncovering shared structures in multiclass classification, *Proceedings of the 24th Annual International Conference on Machine learning*, 17–24.

[6] Bai, Z.D. (1999) Methodologies in spectral analysis of large dimensional random matrices, A review. *Statistica Sinica*, **9**, 611-677.

[7] Bunea, F., She, Y., and Wegkamp, M. (2011). Optimal selection of reduced rank estimators of high-dimensional matrices. *Ann. Statist.*, **39**, 1282-1309.

[8] Cai, J.F., Candès, E.J. and Shen, Z., (2008). A singular value thresholding algorithm for matrix completion. Arxiv preprint arXiv:0810.3286.

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

[10] Candes, E.J. and Plan, Y. (2009). Matrix completion with noise. Arxiv preprint arXiv:0903.3131.

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

[12] Fazel, M., Hindi, H. and Boyd, S.P. (2001). A rank minimization heuristic with application to minimum order system approximation. *American Control Conference, 2001. Proceedings of the 2001*, **6**, 4734–4739.

[13] Izenman, A.J. (1975). Reduced-rank regression for the multivariate linear model. *J. Multi. Analy.*, **5**, 248262.

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

[15] Liu, J. and Ye, J. (2009). Efficient euclidean projections in linear time. *Proceedings of the 26th Annual International Conference on Machine Learning*, 657–664.

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

[17] Jaggi, M. and Sulovskỳ, M. (2010). A simple algorithm for nuclear norm regularized problems. *Proceedings of the 27th Annual International Conference on Machine Learning.*

[18] Ji, S. and Ye, J. (2009). An accelerated gradient method for trace norm minimization. *Proceedings of the 26th Annual International Conference on Machine Learning*, 457–464.

[19] Konstantinides, K., Yao, K. (1988). Statistical analysis of effective singular values in matrix rank determination. *IEEE Transactions on Acoustics, Speech and Signal Processing*, **36(5)**, 757-763.

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

[21] Rao, C.R. (1978). Matrix approximations and reduction of dimensionality in multivariate statistical analysis. In *Proceedings of the fifth international symposium of multivariate analysis*; P. R. Krishnaiah Editor, North-Holland Publishing.

[22] Recht, B., Fazel, M. and Parrilo, P.A. (2010). Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization, *SIAM Review*, **52**, 471-471.

[23] Reinsel, G.C., Velu, R.P. (1998). Multivariate reduced-rank regression: Theory and Applications. *Lecture Notes in Statistics*, Springer, New York.

[24] Shalev-Shwartz, S., Gonen, A. and Shamir, O. (2011). Large-scale convex minimization with a low-rank constraint. *Proceedings of the 28th Annual International Conference on Machine Learning*.

[25] Shen, X. and Huang, H-C. (2006). Optimal model assessment, selection and combination. *J. Amer. Statist. Assoc.*, **101**, 554-68.

[26] Srebro, N., Rennie, J.D.M. and Jaakkola, T.S. (2005). Maximum-margin matrix factorization. *Advances in neural information processing systems*, **17**, 1329–1336.

[27] Stewart, G.W. (1993). On the early history of the singular value decomposition. *SIAM review*, **35(4)**, 551–566.

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

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