Minimal Shrinkage for Noisy Data Recovery Using Schatten-p Norm Objective

Deguang Kong, Miao Zhang and Chris Ding

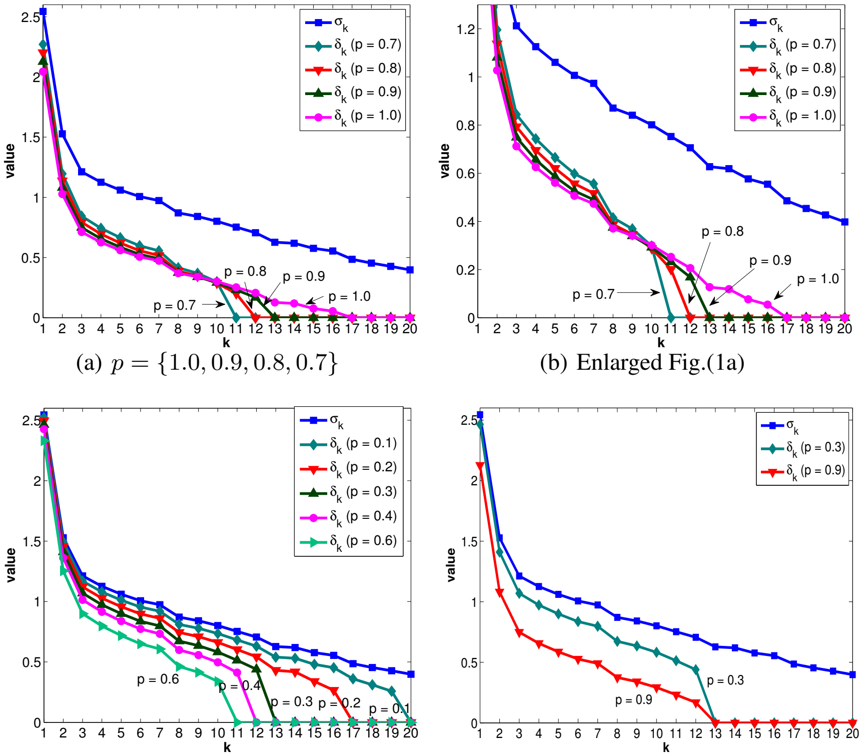
Dept. of Computer Science & Engineering, University of Texas at Arlington, TX, 76013
doogkong@gmail.com, chqding@uta.edu

Abstract. Noisy data recovery is an important problem in machine learning field, which has widely applications for collaborative prediction, recommendation systems, etc. One popular model is to use trace norm model for noisy data recovery. However, it is ignored that the reconstructed data could be shrank (i.e., singular values could be greatly suppressed). In this paper, we present novel noisy data recovery models, which replaces the standard rank constraint (i.e., trace norm) using Schatten-p Norm. The proposed model is attractive due to its suppression on the shrinkage of singular values at smaller parameter p. We analyze the optimal solution of proposed models, and characterize the rank of optimal solution. Efficient algorithms are presented, the convergences of which are rigorously proved. Extensive experiment results on 6 noisy datasets demonstrate the good performance of proposed minimum shrinkage models.

1 Introduction

In big-data era, data is always noisy, development of robust noise tolerant algorithm for data recovery, is always useful and highly demanded. On the other hand, the available of large amount of data makes it more difficult to control the quality the data. The chances of the damaged data or noisy data are increasing. Given input noisy data $\mathbf{X}$, the goal of low rank data recovery problem [1, 2, 3], is to find a low rank approximation $\mathbf{Z}$. Recovered data $\mathbf{Z}$ is expected to be low rank, and retain minimum reconstruction errors (such as least square error) as compared to input data matrix $\mathbf{X}$. In practice, input data can be noisy and also has missing values. This problem has attracted a lot of attentions due to its widely applications in recommendation systems [4], collaborative prediction [5], image/video completion [6], etc.

Data recovery problem has close relations with dimension reduction or low dimension subspace recovery, since for most of high-dimensional data, they may have low-dimensional subspace. Many efforts have been devoted along the direction of principal component analysis (PCA) [7], compressive sensing [8], affine rank minimization [3], etc. For example, Principal component analysis (PCA) seeks for a low-dimensional subspace given data matrix, which can be efficiently computed using singular value decomposition (SVD). However, a major drawback of classical PCA [9] is that, it breaks down under grossly corrupted or noisy observations, such as noises/corruptions in images, and dis-measurement in bio-informatics, etc. In Regularized PCA model (*e.g.*, [10, 11]), it



(a) $p = \{1.0, 0.9, 0.8, 0.7\}$

(b) Enlarged Fig.(1a)

(c) $p = \{0.6, 0.4, 0.3, 0.2, 0.1\}$

(d) $p = \{0.9, 0.3\}$

Fig. 1. Optimal solution δ_k given singular value σ_k of input data $\mathbf{X}$, at different $p = \{1, 0.9, 0.8, \cdots, 0.1\}$ values with fixed $\beta = 0.5$, on dataset Mnist with 20 images, i.e., $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_{20}\}$. To avoid clutter, part of Fig.1a is zoomed in and shown in Fig.1b. In Fig.1d, the solution at $p = 0.3$ is a *faithful* low-rank solution, and the solution at $p = 0.9$ is a *suppressed* low-rank solution.

aims at reducing the rank of the data without explicitly reducing the dimension. However, they do not return the clear representation of subspace and low-dimensional data explicitly.

It is well known that it is a NP-hard problem to directly minimizing the rank of data for recovering input data. Since trace norm can be viewed as a convex envelope of rank function [12], different methods (*e.g.*, [13, 14, 1, 15, 16, 17]), have been proposed by minimizing the trace norm. In this paper, we point out that, standard trace norm model suffers from a serious problem: *shrinkage of reconstructed data and suppression of singular values* (see more details in Figs.(1-2) and §3). We find that the trace norm relaxation may deviate the solution away from the real solution of original rank minimization problem.

The goal of this paper is to develop new methods to solve the approximation of the rank minimization problem. In this paper, we reformulate the noisy data recovery problem using schatten p norm, where efficient algorithms are presented. To summarize, the main contribution of this paper is listed as follows.

- From model construction point of view, we present new models for noisy data recovery, which minimize both data recovery error and rank of recovery data. The proposed models give the minimum shrinkage of recovered data.
- From algorithmic development point of view, we present a complete analysis for proposed model, where the rank of optimal solution is characterized by Theorem 1. Efficient algorithms are developed.
- Extensive experiments on noisy datasets indicate better noisy data recovery performance at smaller p values (p is parameter of our model).

2 Proposed Data Recovery Models

Notation Let $\mathbf{X} = (x_1 \cdots x_n) \in \Re^{d \times n}$ be input n data, each of dimension d. For standard Schatten p norm of matrix $\mathbf{Z}$,

$$||\mathbf{Z}||_{sp} = (\sum_{k=1}^{r} \sigma_k^p)^{\frac{1}{p}} = \left(\mathrm{Tr}[(\mathbf{Z}^T\mathbf{Z})^{\frac{p}{2}}]\right)^{\frac{1}{p}}, \tag{1}$$

where σ_k is the singular value of $\mathbf{Z}$, $r = rank(\mathbf{Z})$.

Given a data matrix $\mathbf{X}$, it is often of interest to compute a matrix $\mathbf{Z}$ that is “close” to $\mathbf{X}$ and satisfies the constraint $rank(\mathbf{Z}) < rank(\mathbf{X})$. Singular value decomposition [18] is the most popular method for such approximations. There are alternative methods that replace this constraint with a more friendly constraint, like, for example, the trace norm. In this paper, we present two models:

Model 1: Schatten p model

We wish to solve the data recovery problem, i.e.,

$$\min_{\mathbf{Z}} \ \frac{1}{2}||\mathbf{Z} - \mathbf{X}||_F^2 + \beta \mathrm{Tr}[(\mathbf{Z}^T\mathbf{Z})^{\frac{p}{2}}], \tag{2}$$

where $\mathrm{Tr}(\mathbf{Z}^T\mathbf{Z})^{\frac{p}{2}} = \sum_{k=1}^{r} \sigma_k^p$, and σ_k is the singular value of $\mathbf{Z}$, β is a parameter to control the scale of schatten p term.

The fact is that the approximation has the same eigen-vectors as the original matrix, and that only eigen-values are shrinked in standard matrix linear algebra. The particular shrinkage of p Schatten norm is better than trace norm ($p = 1$, see Fig. 1), which is corresponding to soft thresholding. At $p = 0$, this is corresponding to hard thresholding (exactly the rank).

Model 2: Robust Schatten p model

We wish to find low-rank data recovery $\mathbf{Z}$ given $\mathbf{X}$, i.e.,

$$\min_{\mathbf{Z}} \ ||\mathbf{Z} - \mathbf{X}||_1 + \beta \mathrm{Tr}[(\mathbf{Z}^T\mathbf{Z})^{\frac{p}{2}}]. \tag{3}$$

This is used for noisy data recovery purpose, which can be viewed as an extension of robust PCA [10].

Motivation

The goal of proposed models is to provide minimum shrinkage of reconstructed data and suppression of singular values. This is the reason, why we replace the trace norm regularization with schatten p regularization. More detailed analysis is provided in §3-4. Our experiment results indicate that proposed models at smaller p values give better recovery performance.

As p becomes small, it is closer to the desired rank constraint:

$$\lim_{p\to 0} \mathrm{Tr}(\mathbf{Z}^T\mathbf{Z})^{\frac{p}{2}} = \lim_{p\to 0} \sum_k \sigma_k^p = rank(\mathbf{Z}).$$

This indicates that the lower p, the better that Schatten norm resembles the rank. Since we wish to do reconstruction with low rank, thus parameter p is usually set to $0 \leq p \leq 1$. In general $p > 1$ case is un-interesting.

Differences of two models The difference of above two models of Eqs.(2, 3) lies in the first term. In Model 1 of Eq.(2), Frobenius norm or the least square error is used to minimize the reconstruction error. In Model 2 of Eq.(3), the L_1-norm is used to minimize the reconstruction error. As is known to us, L_1 error is more robust to noises and outliers, because $||\mathbf{X}-\mathbf{Z}||_1 = \sum_{ij} |\mathbf{X}-\mathbf{Z}|_{ij}$, where residue term is *not* squared. In real world, the observations (like images, text features, etc) can be contaminated by noises or outliers. Model of Eq.(2) is for the data recovery problem polluted by Gaussian noise, while model of Eq.(3) is for data contaminated by Laplacian noises. Both models can be used to solve noisy data recovery, matrix completion problem, etc. For second term, for computational purpose, we add p power to standard term $||\mathbf{Z}||_{sp}$, which plays the same role as standard schatten term for low rank approximation purpose.

Relations with previous methods At $p = 1$, Eq.(3) is equivalent to standard trace-norm model, which optimizes

$$\min_{\mathbf{Z}} ||\mathbf{Z}-\mathbf{X}||_1 + \beta||\mathbf{Z}||_*, \tag{4}$$

where $||\mathbf{Z}||_* = \mathrm{Tr}(\mathbf{Z}^T\mathbf{Z})^{\frac{1}{2}}$ is the trace norm, and σ_k is the singular value of $\mathbf{Z}$. This study is a special case of our model. Note in [10], Schatten p-Norm model at $p = 1$ is called as Robust PCA, because it can correctly recover underlying low-rank structure $\mathbf{Z}$ from the data $\mathbf{X}$ in the presence of gross errors and outlying observations.

3 Illustration of Model 1 and Model 2

Due to the non-smoothness of Schatten norm at $p < 1$, the computational algorithm is challenging. We provide detailed analysis and efficient algorithms of both models in §4, §5 and §6. Here we discuss the general features of the optimal solutions to these two models. The key conclusion is that the solutions at small p are much better than the solution at $p = 1$, which is a previously studied model.

3.1 Illustration of Model 1

To illustrate results of Model 1, we use 20 images from real-world dataset mnist (more details of this dataset is in §7). Let δ_k be the singular values of the optimal solution

$\mathbf{Z}^*$. Let σ_k be the singular values of input data $\mathbf{X}$. We show solution δ_k in Fig.1 along with σ_k. We fix $\beta = 0.5$, but let p vary from $p = 1$ to $p = 0.1$. From Fig.1, we see that at $p = 1$, the optimal solution $\mathbf{Z}^*_{p=1}$, which is represented by $(\delta_1, \delta_2, \cdots, \delta_{20})$, is a simple downshift of $(\sigma_1, \sigma_2, \cdots, \sigma_{20})$. The high rank part ($k = 17 - 20$) is zero. As p decreases, more high rank part of the solution $\{\delta_k\}$ becomes zero, while the lower rank part of $\{\delta_k\}$ moves closer to $\{\sigma_k\}$ of the input data. For example, in Fig.1a, Fig.1b, in optimal solution $\mathbf{Z}^*_{p=0.9}$, the high rank part ($k = 13 - 20$) becomes zero, while the low-rank part ($k = 1 - 7$) is higher than that of $\mathbf{Z}^*_{p=1}$, i.e., this part moves towards corresponding $\{\sigma_k\}$.

In general in low-rank data recovery, we wish the low-rank part of $\mathbf{Z}^*$ is close to those of the input data, while the high-rank part is cut-off (close to zero). Looking in Fig.1d, the solution at $p = 0.3$ is a "faithful" low-rank solution, because the low-rank part is more close or *faithful* to the original data. The solution at $p = 0.9$ is a "suppressed" low-rank solution because the low-rank part is far below the original data, i.e., they are *suppressed*. Clearly, the solution at $p = 0.3$ is more desirable than solution at $p = 0.9$, even though both solutions are low-rank: rank($\mathbf{Z}^*_{p=0.9}$)= rank($\mathbf{Z}^*_{p=0.3}$) = 12.

The Schatten p norm model at small p provides the desirable "faithful" low-rank solution, while the previous work using $p = 1$ also provides a low-rank solution, but the low-rank part is more *suppressed*.

3.2 Illustration of Model 2

Model 2 of Eq.(3) differs from Model 1 by using the L_1 norm in error function. This enables the model to do robust data recovery (e.g., moving outliers back to the correct subspace). However, this model does not change the observed *suppression* in Model 1 at p close to 1 (see Fig.1d). The suppression of singular values leads to the *shrinkage* effect in reconstructed data.

We demonstrate the robust data recovery and the shrinkage effects for Model 2 at different p values on a simple toy data in Fig.(2a). The original data $\mathbf{X}$ are shown as black circles. Reconstructed data $\mathbf{z}_i$ are shown as red-squares. We show the reconstructed results at $p = 0.2$ Fig.(2b, 2e, 2f), $p = 0.5$ (Fig.2c, 2g), $p = 1$ (Fig.2d, 2h). We have two observations.

First, at $0 \le p \le 1$, outliers $(\mathbf{x}_{13}, \mathbf{x}_{14}, \mathbf{x}_{15})$ all move towards the correct subspace, indicating the desired denoising data recovery effects.

Second, for non-outlier data, the reconstructed data shrink strongly at $p = 1$, but they shrink much less at $p = \{0.2, 0.5\}$. This shrinkage is result of the singular value suppression in computed $\mathbf{Z}$. At $p = \{0.2, 0.5, 1\}$, the largest singular value are $\{5.35, 4.49, 2.93\}$, while the second singular values are very small, i.e., $\{$1.7e-8, 1.7e-16, 9.8e-9$\}$, respectively.

In summary, the Schatten model at small p enables us to do robust data recovery but without significant shrinkage in previous models which use $p = 1$.

To our knowledge the singular value suppression and shrinkage (both at $p = 1$ and smaller p values) have not been studied previously.

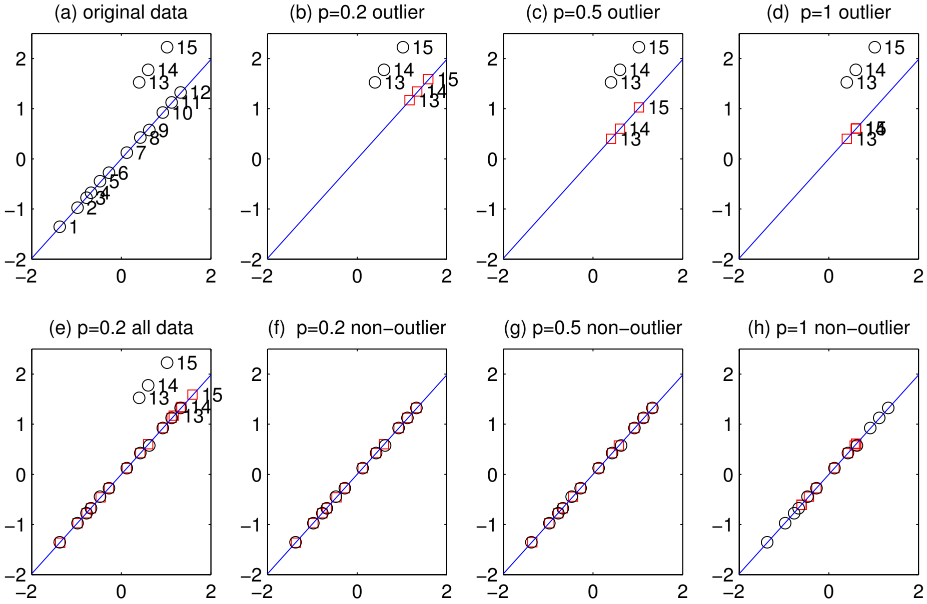


Fig. 2. Demonstration of robust Schatten-p model of Eq.(3) on a toy data shown in panel (a): original data shown as black circles. $(\mathbf{x}_1 \cdots \mathbf{x}_{12})$ are non-outliers and $(\mathbf{x}_{13} \cdots \mathbf{x}_{15})$ are outliers. Reconstructed data $\mathbf{z}_i$ are shown as red-diamonds. Blue line indicates the subspace computed from standard PCA on non-outlier data. Results of Schatten model at $p = 0.2$ are shown in (e). This $p = 0.2$ results are split to outliers and non-outliers as shown in (b) and (f). Similarly, results for $p = 0.5$ shown in (c) and (g); results for $p = 1$ shown in (d) and (h). At $p = 1$, non-outliers shrink towards coordinate (0,0). At smaller p, non-outliers shrink far less.

4 Analysis and Algorithm of Model 1

We show how to solve Model 1 of Eq.(2) at different p values. This also serves as the basic step in solving Model 2 of Eq.(3) using the ALM of §5. To our knowledge, this problem has not been studied before.

Property 1 The global optimal solution for Eq.(2) at all $0 \le p \le 1$, can be efficiently computed, even though it is non-convex at $p < 1$.

Property 2 Rank of the optimal solution $\mathbf{Z}^*$ has a closed form solution:

Theorem 1. *Let the singular value decomposition (SVD) of* $\mathbf{X}$ *be* $\mathbf{X} = \sum_k \sigma_k \mathbf{u}_k \mathbf{v}_k^T$. *Then rank of optimal solution* $\mathbf{Z}^*$*:* $rank(\mathbf{Z}^*) =$ *largest* k*, such that*

$$\sigma_k \le \left(\frac{\beta p (2-p)^{(2-p)}}{(1-p)^{(1-p)}} \right)^{\frac{1}{2-p}}, \quad 0 < p \le 1. \tag{5}$$

In particular, $p = 1, \sigma_k \le \beta; p = \frac{1}{2}, \sigma_k \le (\sqrt{\frac{27}{16}}\beta)^{\frac{2}{3}}$.

Property 3 Optimal solution $\mathbf{Z}^*$ has a closed form solution at $p = \frac{1}{2}$.

Property 4 Optimal solution $\mathbf{Z}^*$ at $0 < p < 1$ can be obtained using Newton's method.

To prove above 4 properties for Model 1 of Eq.(2), we need the following useful lemma.

Lemma 1. *Let the singular value decomposition (SVD) of* $\mathbf{X}$ *be* $\mathbf{X} = \sum_k \sigma_k \mathbf{u}_k \mathbf{v}_k^T$. *The global optimal* $\mathbf{Z}$ *for Eq.(2) is given by* $\mathbf{Z} = \sum_k \delta_k \mathbf{u}_k \mathbf{v}_k^T$, *where* δ_k *is given by solving,*

$$\min_{\delta_1,\cdots,\delta_r} \sum_{k=1}^{r} \Big[\frac{1}{2}(\delta_k - \sigma_k)^2 + \beta\delta_k^p\Big], \quad s.t.\ \delta_k \geq 0, k = 1\cdots r. \tag{6}$$

4.1 Proof of Lemma 1

Proof. Let the optimal solution of $\mathbf{Z}$ have the SVD $\mathbf{Z} = \mathbf{F}\Delta\mathbf{G}^T$ where $\mathbf{F} = (\mathbf{f}_1 \cdots \mathbf{f}_r)$ and $\mathbf{G} = (\mathbf{g}_1 \cdots \mathbf{g}_r)$ are the singular vectors of $\mathbf{Z}$, and $\Delta = \mathrm{diag}(\delta_1 \cdots \delta_r)$ be their singular values. The key is to prove that the singular vectors of $\mathbf{Z}^*$ are the same as those of the input data $\mathbf{X}$. Using von Neumann's trace inequality

$$|\mathrm{Tr}(\mathbf{Z}^T\mathbf{X})| \leq \mathrm{Tr}\Delta\Sigma = \sum_{k=1}^{r} \delta_k\sigma_k. \tag{7}$$

From this, we have

$$\mathrm{Tr}(\mathbf{U}\Delta\mathbf{V}^T)^T\mathbf{X} = \mathrm{Tr}\Delta\Sigma \geq \mathrm{Tr}(\mathbf{Z}^T\mathbf{X}) = \mathrm{Tr}(\mathbf{F}\Delta\mathbf{G}^T)^T\mathbf{X}, \tag{8}$$

where the inequality comes from Eq.(7). The inequality

$$\mathrm{Tr}(\mathbf{U}\Delta\mathbf{V}^T)^T\mathbf{X} \geq \mathrm{Tr}(\mathbf{F}\Delta\mathbf{G}^T)^T\mathbf{X}$$

implies

$$\frac{1}{2}\|\mathbf{U}\Delta\mathbf{V}^T - \mathbf{X}\|^2 + \beta\mathrm{Tr}\Delta^p \leq \frac{1}{2}\|\mathbf{F}\Delta\mathbf{G}^T - \mathbf{X}\|^2 + \beta\mathrm{Tr}\Delta^p.$$

This indicates $(\mathbf{U}, \mathbf{V})$ are better singular vectors for $\mathbf{Z}$ than $(\mathbf{F}, \mathbf{G})$. This proves that the optimal singular vectors for $\mathbf{Z}$ must be the same singular vectors of $\mathbf{X}$. Setting $\mathbf{Z} = \mathbf{U}\Delta\mathbf{V}^T$ in Eq.(2), we obtain Eq.(6).

4.2 Analysis of Property 1

Due to Lemma 1, we now solve the simpler problem of Eq.(6) instead of the original harder problem of Eq.(2). Clearly the optimization of Eq.(6) decouples into r independent subproblems, each for one δ_k:

$$\min_{\delta_k} \frac{1}{2}(\delta_k - \sigma_k)^2 + \beta\delta_k^p, \quad s.t.\ \delta_k \geq 0. \tag{9}$$

KKT complementarity slackness condition for $\delta_k \geq 0$ leads to $\Big[(\delta_k - \sigma_k) + p\beta\delta_k^{p-1}\Big]\delta_k = 0$. The optimization of Eq.(9) decouples into r independent subproblems, and each of them is of the type:

$$\min_{x \geq 0} \; J(x) = \frac{1}{2}(x-a)^2 + \beta x^p, \tag{10}$$

where $x, a \in \Re$. Here the correspondence between Eq.(10) and Eq.(9) is $a = \sigma_k$, $x = \delta_k$.$J(x)$ is a weight sum over two functions: $J(x) = f_1(x) + \beta f_2(x)$, where $f_1(x) = \frac{1}{2}(x-a)^2, f_2(x) = x^p$. $f_1(x)$ has a local minima at $x_1 = a$. $f_2(x)$ is a singular function, $p \leq 1$ with singularity at $x_2 = 0$, which is also a local minima.

Therefore, $J(x)$ in general has two local minima (x_1^*, x_2^*). Because $f_2(x)$ is singular at x_2, for Eq.(10), the singular point (local minima) does not change with different weight β. Thus $x_2^* = 0$ is always a local minima.

When β is small, $x_1^* = a$. As β increases, x_1^* moves towards 0. At certain (β, p), this local minima disappears, $J(x)$ has only one local minima $x_2^* = 0$. This condition is determined by the same condition as in Theorem 1 or Eq.(12) with $\sigma_k = a$. x_1^* is easily computed using Property 4.

In summary, the optimal solution of Eq.(10) is either the trivial one $x_2^* = 0$ or $\min(x_1^*, x_2^*)$, when x_1^* exits. This means Eq.(9) can be easily solved. Thus Eq.(6) can be easily solved for each rank one at a time.

4.3 Proof of Theorem 1

Proof. First, optimization of Eq.(2) is equivalent to optimizing Eq.(10), which can be further written as,

$$\min_{z \geq 0} \; g(z) = \frac{1}{2}(z-1)^2 + \hat{\beta} z^p, \tag{11}$$

where $z = x/a, \hat{\beta} = \beta a^{(p-2)}$. First, we note a key quantity, the zero crossing point z_0 exists, where the second derivative $g''(z)$ changes its sign, i.e., $g''(z_0) = 0$. We need two lemmas.

Lemma 2. *This cross point z_0 always exists at any β.*

Lemma 3. *If the slope of cost function of Eq.(11) at the crossing point z_0 is negative, i.e., $g'(z_0) < 0$, there exists two distinct local minima: $z_2 = 0$ and $z_1 > 0$. If $g'(z_0) \geq 0$, $z_2 = 0$ is the global optimal solution.*

Lemmas 2 and 3 give the key properties of optimization of Eq.(11). Set $g''(z_0) = 0$, we obtain $z_0 = [\hat{\beta} p(1-p)]^{\frac{1}{2-p}}$. Lemma 2 states that $z_2 = 0$ is the global solution, $g'(z_0) = z_0 - 1 + \hat{\beta} p z_0^{p-1} \geq 0$, i.e., $[\hat{\beta} p(1-p)]^{\frac{1}{2-p}} - 1 + \hat{\beta} p [\hat{\beta} p(1-p)]^{\frac{p-1}{2-p}} \geq 0$. Solving for β, we have,

$$\beta \geq \frac{1(1-p)^{(1-p)}}{p(2-p)^{(2-p)}} \cdot {\sigma_k}^{(2-p)}, \quad 0 < p \leq 1. \tag{12}$$

This indicates that the optimal solution δ_k of Eq.(11) is zero (i.e., $\delta_k = 0$), if Eq.(12) holds. This completes the proof.

4.4 Analysis of Property 3

Clearly, at $p = \frac{1}{2}$, from Eq.(9), we need to solve $\delta_k - \sigma_k + (\beta/2)\delta_k^{-1/2} = 0, \; s.t. \; \delta_k \geq 0$. Let $\rho_k = (\frac{\delta_k}{\sigma_k})^{1/2}, \mu = \frac{\beta}{2{\sigma_k}^{\frac{3}{2}}}$, this becomes $\rho_k^3 - \rho_k + \mu = 0$, where $\rho_k \geq 0$. The analytic solution of this cubic equation can be solved in closed form.

Algorithm 1 ALM algorithm to solve Eq.(3)

```
Input: data matrix $\mathbf{X}$, parameter $\rho > 1$.
Output: low rank approximation $\mathbf{Z}$.
Procedure:
Initialize $\mathbf{E}$, $\mathbf{Z}$, $\Omega$, $\mu > 0, t = 0, ;\ \ \rho = 1.1$
while Not converge do
    Updating $\mathbf{E}$ according to Eq.(16)
    Updating $\mathbf{Z}$ according to Eq.(17)
    Updating $\Omega$: $\Omega := \Omega + \mu(\mathbf{Z} - \mathbf{X} - \mathbf{E})$
    Updating $\mu$: $\mu := \rho\mu$
end while
```

4.5 Analysis of Property 4

From analysis of property 1, the optimization of Eq.(10) has two local optima: $x_1^* > 0, x_2^* = 0$. Our algorithm is: (b1) to use Newton's method to compute x_1^*; (b2) compare $J(x_1^*), J(x_2^*)$, and pick the smaller one. It is easy to see $J'(x) = x - a + \beta p x^{p-1}$, $J''(x) = 1+\beta p(p-1)x^{p-2}$. Using standard Newton's method, we can update x through $x \leftarrow x - \frac{J'(x)}{J''(x)}$. This algorithm has quadratic convergence. In practical applications, we found this Newton's algorithm typically converges to local minima within a few iterations.

5 ALM Algorithm to Solve Model 2

Augmented lagrange multipliers(ALM) have been widely used to solve different kinds of optimization problems ([10], [19]). Here we adapt standard ALM method [20, 19] to solve Schatten-p model of Eq.(3). It is worth noting that it is not trivial to solve Eq.(3) using ALM method. One challenging step is to solve the associated Schatten-p term shown in §4.

According to ALM algorithm, by imposing constraint variable $\mathbf{E} = \mathbf{Z} - \mathbf{X}$, the problem of Eq.(3) is equivalent to solve,

$$\min_{\mathbf{E},\, \mathbf{Z}} \|\mathbf{E}\|_1 + \beta \mathrm{Tr}(\mathbf{Z}^T\mathbf{Z})^{\frac{p}{2}}, \qquad s.t. \qquad \mathbf{Z} - \mathbf{X} - \mathbf{E} = 0. \tag{13}$$

According to ALM algorithm, we need to solve,

$$\min_{\mathbf{E},\, \mathbf{Z}} \|\mathbf{E}\|_1 + \langle \Omega, \mathbf{Z} - \mathbf{X} - \mathbf{E} \rangle + \frac{\mu}{2}||\mathbf{Z} - \mathbf{X} - \mathbf{E}||_F^2 + \beta \mathrm{Tr}(\mathbf{Z}^T\mathbf{Z})^{\frac{p}{2}}, \tag{14}$$

where Lagrange multiplier is Ω and μ is penalty constant. For this problem, Ω and μ updated in a specified pattern:

$$\Omega \leftarrow \Omega + \mu(\mathbf{Z} - \mathbf{X} - \mathbf{E}), \quad \mu \leftarrow \rho\mu.$$

We need to search for optimal $\mathbf{E}$, $\mathbf{Z}$ iteratively until the algorithm converges. Now we discuss how we solve $\mathbf{E}$, $\mathbf{Z}$ in each step.

Update E To update the error matrix $\mathbf{E}$, we derive Eq.(15) with fixed $\mathbf{Z}$ and obtain the following form:

$$\min_{\mathbf{E}} \frac{\mu}{2}||\mathbf{E} - \mathbf{A}||_F^2 + ||\mathbf{E}||_1 \tag{15}$$

where $\mathbf{A} = \mathbf{X} - \mathbf{Z} + \frac{\Omega}{\mu}$. It is well-known that the solution to the above LASSO type problem [21] is given by,

$$\mathbf{E}_{ij} = sign(\mathbf{A}_{ij}) \max(|\mathbf{A}_{ij}| - \frac{1}{\mu}, 0). \tag{16}$$

Update Z To update $\mathbf{Z}$ while fixing $\mathbf{E}$, we minimize the relevant part of Eq.(14), which is

$$\min_{\mathbf{Z}} \beta \mathrm{Tr}(\mathbf{Z}\mathbf{Z}^T)^{\frac{p}{2}} + \frac{\mu}{2}||\mathbf{Z} - \mathbf{X} - \mathbf{E} + \frac{\Omega}{\mu}||_F^2. \tag{17}$$

Setting $\mathbf{B} = \mathbf{X} + \mathbf{E} - \frac{\Omega}{\mu}, \hat{\beta} = \frac{\beta}{\mu}$, this optimization becomes Eq.(2), which has been solved in §4.

6 Iterative Algorithm to Solve Model 2

We present another efficient iterative algorithm to solve Eq.(3), where the variable matrix $\mathbf{Z}$ is updated iteratively. Suppose $\mathbf{Z}_t$ is the value of $\mathbf{Z}$ at t-th step. At step t, the key step of our algorithm is to iteratively update j-th column ($\mathbf{z}_j$) of $\mathbf{Z}$ one at a time, according to

$$\mathbf{z}_j = \mathbf{A}^{-1}(\mathbf{A}^{-1} + p\lambda \mathbf{D}_j^{-1})^{-1}\mathbf{x}_j, \tag{18}$$

where $\mathbf{A} = (\mathbf{Z}_t\mathbf{Z}_t^T)^{p/2-1}, \mathbf{W}_{ij} = 1/|(\mathbf{Z}_t - \mathbf{X})_{ij}|, \mathbf{D}_j = \mathrm{diag}(\mathbf{w}_j)$, $\mathbf{w}_j$ is the j-th column of $\mathbf{W}$. This process is iteratively done for $1 \leq j \leq n$. Then $\mathbf{Z}$ is updated until the algorithm converges. More detailed algorithm is summarized in Algorithm 2. In computing $\mathbf{z}_j$ of Eq.(18), we first use conjugate gradient method to compute $\tilde{\mathbf{z}}_j$, where $(\mathbf{A}^{-1} + p\lambda\mathbf{D}_j^{-1})\tilde{\mathbf{z}}^j = \mathbf{x}_j$, and then $\mathbf{z}_j = \mathbf{A}^{-1}\tilde{\mathbf{z}}_j$.

Algorithm 2 An iterative algorithm to solve Eq.(3)

```
Input: $\mathbf{X}$, $\lambda$
Output: $\mathbf{Z}$
while not converge do
    compute $\mathbf{A}^{-1}$
    for $j = 1 : n$ do
        compute $\mathbf{D}_j^{-1}$, solve $\mathbf{z}_j$ according to Eq.(18)
    end for
end while
```

6.1 Convergence of Algorithm

Let $J(\mathbf{Z}) = \|\mathbf{Z} - \mathbf{X}\|_1 + \beta \mathrm{Tr}(\mathbf{Z}^T\mathbf{Z})^{\frac{p}{2}}$, we have

Theorem 2. *Updating* $\mathbf{Z}$ *using Eq.(18),* $J(\mathbf{Z})$ *decreases monotonically.*

The proof requires the following two Lemmas.

Lemma 4. *Define the objective function*

$$J_2(\mathbf{Z}) = \|\mathbf{Z} - \mathbf{X}\|^2_{\mathbf{W}} + p\beta Tr(\mathbf{Z}^T\mathbf{A}\mathbf{Z}). \tag{19}$$

where $\|\mathbf{A}\|^2_{\mathbf{W}} = \sum_{ij} A^2_{ij} W_{ij}$. *The updated* $\mathbf{Z}_{t+1}$ *using Eq.(18) satisfies*

$$J_2(\mathbf{Z}_{t+1}) \leq J_2(\mathbf{Z}_t) \tag{20}$$

Lemma 5. *The updated* $\mathbf{Z}_{t+1}$ *using Eq.(18) satisfies*

$$J(\mathbf{Z}_{t+1}) - J(\mathbf{Z}_t) \leq \frac{1}{2}\big[J_2(\mathbf{Z}_{t+1}) - J_2(\mathbf{Z}_t)\big] \tag{21}$$

6.2 Proof of Theorem 2

Proof. From Eq.(20), clearly, LHS of Eq.(21) is LHS ≤ 0.

6.3 Proof of Lemma 4

Proof. Setting $\partial J_2(\mathbf{Z})/\partial Z_{ij} = 0$, we have $(\mathbf{Z} - \mathbf{X})_{ij} W_{ij} + p\lambda(\mathbf{A}\mathbf{Z})_{ij} = 0$. This can be written as $Z_{ij}W_{ij} + p\lambda(\mathbf{A}\mathbf{Z})_{ij} = X_{ij}W_{ij}$. In matrix form, $\mathbf{D}_j\mathbf{z}_j + p\lambda\mathbf{A}\mathbf{z}_j = \mathbf{D}_j\mathbf{x}_j$. Thus we have

$$\mathbf{z}_j = (\mathbf{D}_j + p\lambda\mathbf{A})^{-1}\mathbf{D}_j\mathbf{x}_j = [\mathbf{D}_j(\mathbf{A}^{-1} + p\lambda\mathbf{D}_j^{-1})\mathbf{A}]^{-1}\mathbf{D}_j\mathbf{x}_j, \tag{22}$$

which gives Eq.(18).

6.4 Proof of Lemma 5

Proof. Let $\varDelta$ = LHS - RHS of Eq.(21). We have $\varDelta$ = $\alpha + \beta$ where

$$\begin{aligned}\alpha &= \sum_{ij}\Big[|(\mathbf{Z}_{t+1} - \mathbf{X})_{ij}| - |(\mathbf{Z}_t - \mathbf{X})_{ij}| - \frac{(\mathbf{Z}_{t+1} - \mathbf{X})^2_{ij}}{2|(\mathbf{Z}_t - \mathbf{X})_{ij}|}\frac{(\mathbf{Z}_t - \mathbf{X})^2_{ij}}{2|(\mathbf{Z}_t - \mathbf{X})_{ij}|}\Big] \\ &= \sum_{ij}\frac{-1}{2|(\mathbf{Z}_t - \mathbf{X})_{ij}|}\Big[|(\mathbf{Z}_{t+1} - \mathbf{X})_{ij}| - |(\mathbf{Z}_t - \mathbf{X})_{ij}|\Big]^2 \leq 0.\end{aligned}$$

and

$$\begin{aligned}\beta &= \lambda\Big[\mathrm{Tr}(\mathbf{Z}_{t+1}\mathbf{Z}^T_{t+1})^{\frac{p}{2}} - \mathrm{Tr}(\mathbf{Z}_t\mathbf{Z}^T_t)^{\frac{p}{2}}\Big] - \frac{p}{2}\lambda\Big[\mathrm{Tr}\mathbf{Z}^T_{t+1}(\mathbf{Z}_t\mathbf{Z}^T_t)^{\frac{p}{2}}\mathbf{Z}_{t+1} - \mathrm{Tr}\mathbf{Z}^T_t(\mathbf{Z}_t\mathbf{Z}^T_t)^{\frac{p}{2}}\mathbf{Z}_t\Big] \\ &= \lambda\Big[\mathrm{Tr}(\mathbf{Z}_{t+1}\mathbf{Z}^T_{t+1})^{\frac{p}{2}} - \mathrm{Tr}(\mathbf{Z}_t\mathbf{Z}^T_t)^{\frac{p}{2}}\Big] - \frac{p}{2}\lambda\mathrm{Tr}\Big[\big(\mathbf{Z}_{t+1}\mathbf{Z}^T_{t+1} - \mathbf{Z}_t\mathbf{Z}^T_t\big)(\mathbf{Z}_t\mathbf{Z}^T_t)^{\frac{p}{2}}\Big] \\ &\leq 0,\end{aligned} \tag{23}$$

where in the last inequality, we set $\mathbf{A} = \mathbf{Z}_{t+1}\mathbf{Z}^T_{t+1}$, $\mathbf{B} = \mathbf{Z}_t\mathbf{Z}^T_t$ and used Lemma 6 below. Clearly $\varDelta = \alpha + \beta \leq 0$.

Lemma 6. *[24] For any two symmetric positive definite matrices* $\mathbf{F}, \mathbf{G}$ *and* $0 < p \leq 2$,

$$Tr\,[\mathbf{F}^{p/2} - \mathbf{G}^{p/2}] \leq \frac{p}{2}Tr\,[(\mathbf{F} - \mathbf{G})\mathbf{G}^{p/2-1}] \tag{24}$$

Due to space limit, we omit the proofs of Lemma 6 here.

Table 1. Description of Data sets

Dataset	#data	#dimension	#class
AT&T_oc	400	2576	40
Binalpha_oc	1404	320	36
Umist_oc	360	644	20
YaleB	256	2016	4
CMUPIE_oc	680	1024	68
Mnist_oc	150	784	10

7 Connection to Related Works

We note [22] proposes an algorithm to solve squared schatten p model, i.e., $\min_{\mathbf{Z}} f(\mathbf{Z}) + \beta\Big(\text{Tr}(\mathbf{Z}^T\mathbf{Z})^{\frac{p}{2}}\Big)^{\frac{2}{p}}$, which cannot be directly applied here. [23] proposes an iterative reweighted algorithm for trace norm minimization problem, in the similar vein as what has been proposed for adaptive lasso. However, it cannot be directly applied to solve Eq.(3). As compared to [24, 25, 26, 27], our goal is for noisy data recovery problem raised in computer vision, instead of for matrix completion problems with missing values.

8 Experiments

We use six widely used image data sets, including four face datasets: AT&T Umist, YaleB [28] and CMUPIE; and two digit datasets: Mnist [29] and Binalpha [1]. We generate **occluded** image datasets corresponding to 5 original data sets (except YaleB). For YaleB dataset, the images are taken under different poses with different illumination conditions. The shading parts of the images play the similar role of occlusion (noises). Thus we use the original YaleB data with first 4 persons in our experiments. For the other 5 datasets, half of the images are selected from each category for occlusion with block size of wxw pixels (e.g., $w = 10$). The locations of occlusions are random generated without overlaps among the images from the same category. *Occluded* images (with occlusion size 7×7) generated from Umist data sets are shown in Fig. 4. Table 1 summarizes the characteristics of these occluded data sets.

We did all experiments using Eq.(3). At $p < 1$, objective function in Eq.(3) is not convex any more, and we cannot get global minima. We initialize $\mathbf{Z}$ using trace norm minimization solution, i.e., set $p = 1$ in Eq.(3). In the following experiments, we did both algorithms proposed in §5-6, and reported the results using the one achieving smaller objectives.

Illustrative examples To visualize the denoising effect of proposed method, we apply our model on YaleB dataset. YaleB contains images with different shading which plays similar role of occlusion (noises). Thus we did not add occlusion and use the original data. In this demonstration and following experiment, each data (image) is linearized into a vector each $\mathbf{x}_i$, and the input matrix $\mathbf{X}$ is constructed as $\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n)$. We typically set the rank k equal to the number of classes in the dataset. Due to space limit, computed $\mathbf{Z}$ at different p values for the two persons are shown. In Fig.(3),

[1] http://www.kyb.tuebingen.mpg.de/ssl-book/ benchmarks.html

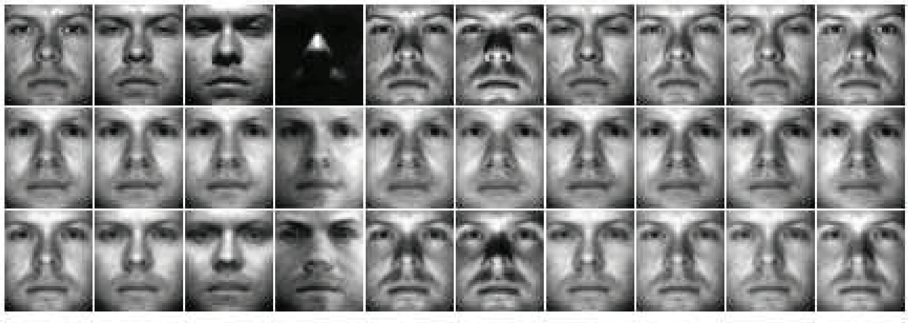


Fig. 3. Reconstructed images ($\mathbf{Z}$) of YaleB dataset using Model 2 of Eq.(3) shown in 1 panel. First line: original images of one person, Second line: reconstructed images $\mathbf{Z}$ at $p = 1$, Third line: reconstructed images at $p = 0.2$. One can see $p = 1$ images are very similar to each other (most fine details are lost), while $p = 0.2$ images retain some fine details and are closer to original images.

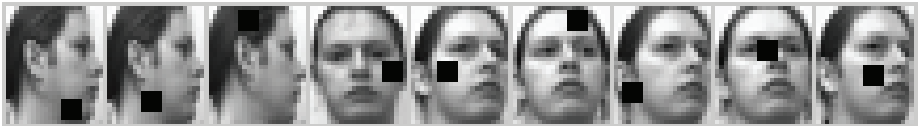


Fig. 4. Occluded image dataset Umist.

20 images are shown as 2 panels, each panel for one person. On each panel, the first line images are original images $\mathbf{X}$, the 2nd line are computed $\mathbf{Z}$ at $p = 1$, and 3rd line are computed $\mathbf{Z}$ at 0.2.

Clearly, at different p values (such as $p = \{1, 0.2\}$), Schatten p-Norm model can effectively recover the original data by removing the shadings. See 2nd line on each panel in Fig.(3), almost every person is recovered to same template, not any difference any more. In contrast, we have much better visualization results (with more details) when $p = 0.2$ (see 3rd line on each panel). Moreover, these fine details are expected to be helpful for classification on images from different persons.

Table 2. True data recovery: True signal reconstruction error at different p on six datasets

dataset	$\frac{\|\|\mathbf{X}_E\|\|_F}{\|\|\mathbf{X}_0\|\|_F}$	Noise-free reconstruction error at different p				
		$p = 1$	$p = 0.75$	$p = 0.5$	$p = 0.2$	$p = 0$
AT&T	0.3657	0.2672	0.2240	0.2199	0.2159	**0.2132**
Binalpha	0.2359	0.2023	0.1974	0.1845	**0.1594**	0.1729
Umist	0.3123	0.2816	0.2290	0.2199	0.2153	**0.2151**
YaleB	N/A	0.2304	0.2264	0.2174	**0.1912**	0.2126
CMUPIE	0.2542	0.2012	0.1925	0.1845	**0.1594**	0.1729
Mnist	0.5574	0.5123	0.4993	0.4814	**0.4542**	0.4553

True data recovery: true signal reconstruction error Given noisy data $\mathbf{X}$, $\mathbf{X} = \mathbf{X}_0 + \mathbf{X}_E$, where $\mathbf{X}_0$ is the true signal and $\mathbf{X}_E$ is the noise. Our goal is to recover $\mathbf{X}_0$ using Eq.(3). We did experiments on above 6 datasets. To evaluate the performance, we define the true signal data recovery error, $E_{\text{true-signal}} = \frac{||\mathbf{Z}-\mathbf{X}_0||_F}{||\mathbf{X}_0||_F}$. Clearly, smaller $E_{\text{true-signal}}$ values indicate better recovery. Computed true signal reconstruction error are shown in Table.2. The experiment results indicate that true signal reconstruction errors are *smaller* at smaller p values. We also list $\frac{||\mathbf{X}_E||_F}{||\mathbf{X}_0||_F}$ values in Table.2 to indicate the level of occlusions. Interestingly, $E_{\text{true-signal}} < \frac{||\mathbf{X}_E||_F}{||\mathbf{X}_0||_F}$ on all datasets at different p values. This further confirms "de-noisy" effects of proposed data recovery model.

Table 3. Loss of fine-details: variance of reconstructed $\mathbf{Z}$ on six datasets, original images: $\mathbf{X}_0$, occluded images: $\mathbf{X}$

dataset	$\mathbf{X}_0$	$\mathbf{X}$	Variance of $\mathbf{Z}$ at different p				
			$p = 1$	$p = 0.75$	$p = 0.5$	$p = 0.2$	$p = 0$
AT&T	8.89	9.03	5.83	7.13	7.45	8.11	7.80
Binalpha	27.90	31.13	13.40	22.89	25.38	26.89	26.73
Umist	7.01	7.42	3.87	5.31	5.71	6.38	6.01
YaleB	9.75	9.75	7.28	8.22	8.59	9.19	8.76
CMUPIE	12.09	13.16	8.12	10.07	10.54	11.30	10.87
Mnist	9.24	10.26	0.49	4.41	5.45	7.04	5.85

Loss of fine details in recovered data and its measure Due to suppression of higher order/frequency terms associated with smaller singular values, fine details of original data $\mathbf{X}$ are lost in the recovered $\mathbf{Z}$. As a consequence, recovered individual images are very similar to each other. One numeric measure is the variance of reconstructed images. We therefore define $var(\mathbf{Z}) = \sum_{i=1}^{n} ||\mathbf{z}_i - \bar{\mathbf{z}}||^2, \ \bar{\mathbf{z}} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{z}_i$, where $\mathbf{z}_i \in \Re^{d\times 1}$ is the reconstructed image corresponding to each original image $\mathbf{x}_i$. Larger variance values indicate more fine details are preserved in the solution $\mathbf{Z}$. Computed variance of $\mathbf{Z}$ are shown in Table.3. Clearly, reconstructed images preserve more detailed information at small p (say $p = 0.2$). One demonstrating example is shown in Fig.3, where fine details of individual images are mostly suppressed at $p = 1$, but are generally preserved/reetained at $p = 0.2$.

Classification results using recovered $\mathbf{Z}$ So far we have discussed low rank recovery capability of computed $\mathbf{Z}$. Reconstructed low rank $\mathbf{Z}$ is expected to have much clear structure after removing noises and outliers. As a by-product of solving low-rank data recovery problem, computed $\mathbf{Z}$ can be used for classification tasks. We compare the classification results by using the occluded images $\mathbf{X}$ and recovered data $\mathbf{Z}$ at different p. The experiments are done on two widely used classifiers: k nearest neighbor (kNN) and support vector machine[1] using 5 fold cross validation. Since the regularization coefficient is also a hyper-parameter, the performance of each Schatten-p norm model is evaluated at an optimal value of β (which is determined by cross validation). The experiment results are shown in Table.4. We have two important observations from

[1] http://www.csie.ntu.edu.tw/ cjlin/libsvm/

Table 4. classification accuracy(shown as percentage) on six occluded datasets using input corrupted data $\mathbf{X}$ and reconstructed $\mathbf{Z}$ at different p values

dataset	method	$\mathbf{X}$	Reconstructed $\mathbf{Z}$ at different p				
			$p = 1$	$p = 0.75$	$p = 0.5$	$p = 0.2$	$p = 0$
AT&T	SVM	29.75	30.52	33.75	34.25	**36.78**	35.53
	KNN	25.75	28.63	**30.25**	28.75	29.31	28.33
Binalpha	SVM	38.35	44.78	42.74	43.84	**48.53**	47.43
	KNN	52.09	56.78	55.10	54.65	**58.23**	57.87
Umist	SVM	59.83	65.89	63.17	64.35	**68.33**	67.67
	KNN	89.12	93.89	92.67	93.75	**94.23**	93.01
YaleB	SVM	46.11	52.12	51.89	53.78	54.67	**54.96**
	KNN	85.43	90.89	90.36	91.15	**91.76**	91.40
CMUPIE	SVM	29.24	33.57	**36.74**	34.21	35.39	34.98
	KNN	58.12	64.03	65.38	64.27	**66.39**	65.64
Mnist	SVM	49.38	51.93	53.24	**57.18**	56.79	54.67
	KNN	76.63	81.35	80.75	81.56	**82.47**	82.34

experiment results. **(1)** Performances for image categorization tasks are improved by using computed $\mathbf{Z}$ at different p values; **(2)** Classification accuracy is consistently better at smaller p values on both SVM and kNN classifiers, as compared to that at large p values. All above results suggest us to use Schatten p-Norm at small p values.

9 Conclusion

We present novel models for low-rank data recovery, where efficient algorithms are proposed. Extensive experiment results indicate schatten p model gives relatively better reconstructed results at small p values. In the next step, we will further explore how to scale our model for large-size problems.

Acknowledgement. This work is partially supported by NSF-CCF-0939187, NSF-DMS-0915228, NSF-CCF-0917274.

References

1. Cai, J., Candès, E., Shen, Z.: A singular value thresholding algorithm for matrix completion. SIAM Journal on Optimization **20** (2010) 1956–1982
2. Candès, E., Recht, B.: Exact matrix completion via convex optimization. Communications of the ACM **55** (2012) 111–119
3. Recht, B., Fazel, M., Parrilo, P.: Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. SIAM Review **52** (2010) 471–501
4. Rennie, J.M., Srebro, N.: Fast maximum margin matrix factorization for collaborative prediction. In: ICML. (2005) 713–719
5. Abernethy, J., Bach, F., Evgeniou, T., M.Paristech, Jaakkola, T.: A new approach to collaborative filtering: Operator estimation with spectral regularization. Journal of Machine Learning Research **10** (2009) 803–826

6. Liu, J., Musialski, P., Wonka, P., Ye, J.: Tensor completion for estimating missing values in visual data. In: ICCV. (2009) 2114–2121
7. Jolliffe, I.: Principal Component Analysis. Springer (1986)
8. Donoho, D.: Compressed sensing. IEEE Transactions on Information Theory **52** (2006) 1289–1306
9. Ding, C.H.Q., Zhou, D., He, X., Zha, H.: R_1-pca: rotational invariant l_1-norm principal component analysis for robust subspace factorization. In: ICML. (2006) 281–288
10. Wright, J., Ganesh, A., Rao, S., Peng, Y., Ma, Y.: Robust principal component analysis: Exact recovery of corrupted low-rank matrices via convex optimization. In: NIPS. (2009)
11. Peng, Y., Ganesh, A., Wright, J., Xu, W., Ma, Y.: Rasl: Robust alignment by sparse and low-rank decomposition for linearly correlated images. In: CVPR. (2010) 763–770
12. Fazel, M.: Matrix rank minimization with applications. PhD thesis, Stanford University (2002) 1–130
13. Candès, E., Tao, T.: The power of convex relaxation: near-optimal matrix completion. IEEE Transactions on Information Theory **56** (2010) 2053–2080
14. Chandrasekaran, V., Sanghavi, S., Parrilo, P., Willsky, A.: Sparse and low-rank matrix decompositions. In: Allerton'09 Proceedings of the 47th annual Allerton conference on Communication, control, and computing. (2009) 962–967
15. Ma, S., Goldfarb, D., Chen, L.: Fixed point and bregman iterative methods for matrix rank minimization. Mathematical Programming: Series A and B **128** (2011) 321–353
16. Toh, K., Yun, S.: An accelerated proximal gradient algorithm for nuclear norm regularized linear least squares problems. Pacific Journal of Optimization (2010)
17. Pong, T., Tseng, P., Ji, S., Ye, J.: Trace norm regularization: Reformulations, algorithms, and multi-task learning. SIAM Journal on Optimization **20** (2010) 3465–3489
18. Golub, G.H., Reinsch, C.: Singular value decomposition and least squares solutions. Numerische Mathematik **14** (1970) 403 – 420
19. Lin, Z., Chen, M., Ma, Y.: The augmented lagrange multiplier method for exact recovery of corrupted low-rank matrices. In: UIUC Tech. Rep., UILU-ENG-09-2214. (2010)
20. Bertsekas, D.: Constrained Optimization and Lagrange Multiplier Methods. Athena Scientific (1996)
21. Tibshirani, R.: Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society, Series B **58** (1994) 267–288
22. Argyriou, A., Micchelli, C., Pontil, M., Ying, Y.: A spectral regularization framework for multi-task structure learning. In: NIPS. (2007)
23. Jain, P., Meka, R., Dhillon, I.: Guaranteed rank minimization via singular value projection. In: NIPS. (2010)
24. Nie, F., Huang, H., Ding, C.: Low-rank matrix recovery via efficient schatten p-norm minimization. In: Twenty-Sixth AAAI Conference on Artificial Intelligence. (2012)
25. Mohan, K., Fazel, M.: Iterative reweighted least squares for matrix rank minimization. In: Communication, Control, and Computing (Allerton), 2010 48th Annual Allerton Conference on, IEEE (2010) 653–661
26. Keshavan, R., Montanari, A., Oh, S.: Matrix completion from noisy entries. The Journal of Machine Learning Research **11** (2010) 2057–2078
27. Wipf, D.: Non-convex rank minimization via an empirical bayesian approach. In: UAI. (2012) 914–923
28. Georghiades, A., Belhumeur, P., Kriegman, D.: From few to many: Illumination cone models for face recognition under variable lighting and pose. PAMI (2001) 643–660
29. Lecun, Y., Bottou, L., Bengio, Y., P.Haffner: Gradient-based learning applied to document recognition. In: Proceedings of the IEEE. (1998) 2278–2324