A Nearly Unbiased Matrix Completion Approach

Dehua Liu, Tengfei Zhou, Hui Qian*, Congfu Xu, and Zhihua Zhang

College of Computer Science and Technology
Zhejiang University, Hangzhou 310007, China
{dehua,zhoutengfei_zju,qianhui,xucongfu,zhzhang}@zju.edu.cn

Abstract. Low-rank matrix completion is an important theme both theoretically and practically. However, the state-of-the-art methods based on convex optimization usually lead to a certain amount of deviation from the original matrix. To perfectly recover a data matrix from a sampling of its entries, we consider a non-convex alternative to approximate the matrix rank. In particular, we minimize a matrix γ-norm under a set of linear constraints. Accordingly, we derive a shrinkage operator, which is nearly unbiased in comparison with the well-known soft shrinkage operator. Furthermore, we devise two algorithms, non-convex soft imputation (NCSI) and non-convex alternative direction method of multipliers (NCADMM), to fulfil the numerical estimation. Experimental results show that these algorithms outperform existing matrix completion methods in accuracy. Moreover, the NCADMM is as efficient as the current state-of-the-art algorithms.

1 Introduction

Applications of low-rank matrix completion become increasingly popular in machine learning and data mining. For instance, in the system of collaborative filtering, we aim to predict the unknown preference of a user on a set of unrated items, only according to a few submitted rating. In image inpainting problems, large amount of missing pixels should be estimated by exploiting the known content.

Typically, matrix completion is formed as minimizing the rank of matrix when given a few known entries. However, the rank minimization problem is often numerically prohibitive. Thus, many approximation strategies are encouraged. One principled approach is to replace the matrix rank by the nuclear norm, because the nuclear norm is the best convex relaxation of the matrix rank. In the literature [1–3], the authors proved that under certain assumptions on the proportion of the missing entries and locations, most low-rank matrices can be completed exactly by minimizing the nuclear norm under the linear constraints (the completed matrix must be consistent with the observed matrix for the few known entries).

Based on the nuclear norm, Cai *et al.* [4] devised a singular value thresholding (SVT) algorithm for this convex optimization problem. Mazumder *et al.* [5] formed a unconstrained convex optimization problem and developed a soft-impute algorithm. Ma *et al.* [6] devised a fixed point iterative algorithm inspired from the work of Hale *et al.* [7] in the ℓ_1 regularization problem. Lin *et al.* [8, 9] proposed an alternative direction algorithm based on the augmented Lagrangian multipliers. Other efficient

* Corresponding author

algorithms based on convex relaxation include [3, 10, 11]. There are also some other matrix norms that have been considered, e.g., the max-norm which is also a convex approximation of the matrix rank [12].

However, convex relaxation often makes the resulting solution deviate from the original matrix [13]. To address this problem, non-convex approximation to the matrix norm has been also exploited recently. Such treatments include the Schatten ℓ_p-norm ($0 < p < 1$) used by Nie *et al.* [13], the truncated nuclear norm proposed by Hu *et al.* [14] and a so-called matrix γ-norm studied by Wang *et al.* [15]. Note that matrix γ-norm is not really a norm, since it does not satisfy triangle inequality.

The matrix γ-norm is a matrix extension of the MC+ function studied by Mazumder *et al.* [16] and Zhang [17] for variable selection. The γ-norm is characterized by a positive factor γ, and is tighter than the nuclear norm to the matrix rank. Wang *et al.* [15] employed the γ-norm, giving a non-convex approach to robust principle component analysis (RPCA). In this paper we introduce the γ-norm into the matrix completion problem. We develop a shrinkage operator which is nearly unbiased from non-convex rank approximation and put forward two effective algorithms called NCSI and NCADMM.

The remaining parts of the paper are organized as follows. Section 2 reviews the preliminaries for matrix completion. Section 3 presents the NCSI and NCADMM algorithms. Section 4 gives the convergence analysis of our NCSI algorithm. Section 5 conducts the experimental analysis. Finally, we conclude our work in Section 6.

2 Preliminaries

We are given a matrix $\mathbf{M} = [m_{ij}] \in \mathbb{R}^{n\times m}$ with missing entries. Without loss of generality, we assume $m \leq n$. Let $\mathbf{X} = [x_{ij}] \in \mathbb{R}^{n\times m}$ be an unknown low-rank matrix. The matrix completion problem is to address the following rank minimization problem:

$$\begin{aligned} \min_{\mathbf{X}} \quad & \mathrm{rank}(\mathbf{X}) \\ \text{s.t.} \quad & x_{ij} = m_{ij}, \quad \forall (i,j) \in \Omega, \end{aligned}$$

in which $\Omega \subset \{1,\ldots,n\} \times \{1,\ldots,m\}$ is the set of indices of observation entries of $\mathbf{M}$. We denote the indices of the missing entries by $\bar{\Omega} = \{1,\ldots,n\} \times \{1,\ldots,m\} \setminus \Omega$.

This rank minimization problem is generally NP-hard. However, it can be relaxed to a feasible optimization problem via rank approximation. That is, we consider the following alternative:

$$\min_{\mathbf{X}} \quad P(\mathbf{X};\theta) \tag{1}$$

$$\text{s.t.} \quad x_{ij} = m_{ij}, \quad \forall (i,j) \in \Omega, \tag{2}$$

where $P(\mathbf{X};\theta)$ represents the approximation of $\mathrm{rank}(\mathbf{X})$.

It is well known that the nuclear norm, the sum of singular values, is the tightest convex relaxation of the matrix rank. Candès and Tao [2] proved that most low rank matrices can be completed from (1) with $P(\mathbf{X};\theta)$ as the nuclear norm $\|\mathbf{X}\|_*$ if the number

of given entries is greater than $nr\text{polylog}(n)$ up to a positive constant C. Meanwhile, some researchers developed efficient algorithms to solve the above problem such as singular value thresholding SVT [4].

In order to accommodate the small noise in observation, it is better to relax the constraints in (1) by adding a square loss to the objective function, forming an unconstrained problem [5]:

$$\min_{\mathbf{X}} \ \tfrac{1}{2}\|P_\Omega(\mathbf{X}-\mathbf{M})\|_F^2 + \lambda P(\mathbf{X};\theta), \tag{3}$$

where $\|\mathbf{A}\|_F = \sqrt{\sum_{i,j} a_{ij}^2} = \sqrt{\text{tr}(\mathbf{A}\mathbf{A}^\top)} = \sqrt{\sum_i \sigma_i^2(\mathbf{A})}$ is the Frobenius norm of $\mathbf{A} = [a_{ij}]$, and $P_\Omega(\mathbf{A})$ is such an $n\times m$ matrix that its (i,j)th entry is a_{ij} if $(i,j)\in\Omega$ and zero otherwise. $P(\mathbf{X};\theta)$ is usually called a regularization or penalty term.

In order to solve problem (3), a key step is to solve a subproblem of the form:

$$\min_{\mathbf{X}} \ \Big\{\frac{1}{2}\|\mathbf{X}-\boldsymbol{\Phi}\|_F^2 + \lambda P(\mathbf{X};\theta)\Big\}. \tag{4}$$

First of all, we introduce a so-called shrinkage operator.

Definition 1 (Shrinkage operator). $S_{\lambda,\theta}(\boldsymbol{\Phi}) = \text{argmin}_{\mathbf{X}}\{\frac{1}{2}\|\mathbf{X}-\boldsymbol{\Phi}\|_F^2 + \lambda P(\mathbf{X};\theta)\}$ *is a shrinkage operator if it shrinks the small singular value of* $\boldsymbol{\Phi}$ *to* 0.

In this paper, we would like to consider the special penalty $P(\mathbf{X};\theta)$ which is constructed from a single variable function. Suppose $p(x;\theta)$ is a function of single variable function with domain $\mathbb{R}_+$, then $P(\mathbf{X};\theta) = \sum_i p(\sigma_i(\mathbf{X}))$. We can construct many penalty on matrix by this way. In this case we define the overloading of shrinkage operator on $\mathbb{R}_+$ as $S_{\lambda,\theta}(z) = \text{argmin}_{x\geq 0}\{\frac{1}{2}(x-z)^2 + \lambda p(\mathbf{x};\theta)\}$.

For example, the popular used nuclear norm $P(\mathbf{X};\theta) = \|\mathbf{X}\|_*$ is derived from function $p(x) = x$. And $S_\lambda(z) = \text{argmin}_{x\geq 0}\{\frac{1}{2}(x-z)^2 + \lambda x\}$ for $z \geq 0$.

Let $\boldsymbol{\Phi} = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^\top$ be the thin SVD of $\boldsymbol{\Phi}$. It has been proved that in the case of $P(\mathbf{X};\theta) = \|\mathbf{X}\|_*$ the shrinkage operator has a simple form which is given by $S_\lambda(\boldsymbol{\Phi}) = \mathbf{U}\boldsymbol{\Sigma}_\lambda\mathbf{V}^\top$ with $\boldsymbol{\Sigma}_\lambda = \text{diag}(S_\lambda(\sigma_1),\ldots,S_\lambda(\sigma_m))$ [4–6], where $S_\lambda(\cdot)$ defined for a single variable is the overloading operator which is given by

$$S_\lambda(z) = [z-\lambda]_+ = \begin{cases} z-\lambda & \text{if} \quad z > \lambda, \\ 0 & \text{if} \quad z \leq \lambda. \end{cases}$$

We call $S_\lambda(\cdot)$ the soft shrinkage operator.

Observe that problem (4) can be viewed as the extreme case of (3), when Ω is the set of subscript indices of all entries of matrix. We find that soft shrinkage operator derived from the nuclear norm may lead to deviation for large λ, since a same positive number is subtracted from all the singular values of a matrix. This encourages us to use a non-convex penalty which results in a shrinkage operator keeping the large singular values unchanged while shrinking the small ones to zero. We establish this thought the following definition.

Definition 2 (Nearly unbiasedness). *We say that the operator* $S_{\lambda,\theta}(\boldsymbol{\Phi})$ *is nearly unbiased, if it keeps the sufficiently large singular value of* $\boldsymbol{\Phi}$ *unchanged.*

The non-convex penalty has been mentioned in [5] and further explored by some researchers in matrix recovery as well as matrix completion problems [14, 15]. Here we employ the treatment of Wang *et al.* [15] who devised a so-called γ-norm in their work of matrix recovery problem. We will show that this non-convex penalty can make a nearly unbiased estimator for the matrix completion problem.

3 Methodology

For a matrix $\mathbf{X} \in \mathbb{R}^{n\times m}$ with $m \leq n$, $\mathbf{X} = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^\top$ is the SVD factorization with $\boldsymbol{\Sigma} = \mathrm{diag}\{\sigma_1, ..., \sigma_m\}$. According to Wang *et al.* [15], γ-norm is defined as

$$\|\mathbf{X}\|_\gamma = \sum_{i=1}^{m} p(\sigma_i; \gamma),$$

where $p(x;\gamma) = \int_0^x (1 - \frac{u}{\gamma})_+ du = (x - \frac{x^2}{2\gamma})I(x < \gamma) + \frac{\gamma}{2} I(x \geq \gamma)$. A key step to construct an algorithm is an optimization problem whose solution is summarized as below.

Theorem 1. *Suppose $\boldsymbol{\Phi} = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^\top$ is the SVD factorization. The minimizer of $\phi(\mathbf{X}) = \frac{1}{2}\|\boldsymbol{\Phi} - \mathbf{X}\|_F^2 + \lambda\|\mathbf{X}\|_\gamma$ with $\gamma > \lambda$ is $S_{\lambda,\gamma}(\boldsymbol{\Phi}) = \mathbf{U}\boldsymbol{\Sigma}_{\lambda,\gamma}\mathbf{V}^\top$, where $\boldsymbol{\Sigma}_{\lambda,\gamma} = \mathrm{diag}(S_{\lambda,\gamma}(\sigma_1), \cdots, S_{\lambda,\gamma}(\sigma_m))$ is a diagonal matrix with the diagonal elements*

$$S_{\lambda,\gamma}(\sigma_i) = \underset{x\geq 0}{\mathrm{argmin}}\{\frac{1}{2}(x-\sigma_i)^2 + \lambda p(x;\gamma)\} = \begin{cases} \sigma_i & \text{if} \quad \sigma_i \geq \gamma, \\ \frac{\sigma_i - \lambda}{1 - \frac{\lambda}{\gamma}} & \text{if} \quad \lambda \leq \sigma_i < \gamma, \\ 0 & \text{if} \quad \sigma_i < \lambda. \end{cases}$$

Proof. Let the thin SVD of$\mathbf{X}$ be of $\mathbf{X} = \mathbf{U}\boldsymbol{\Lambda}\mathbf{V}^T$, where $\mathbf{U} = [\mathbf{u}_1, \ldots, \mathbf{u}_m]$ has orthonormal columns, $\mathbf{V} = [\mathbf{v}_1, \ldots, \mathbf{v}_m]$ is orthogonal, and $\boldsymbol{\Lambda} = \mathrm{diag}(\lambda_1, \ldots, \lambda_m)$ is arranged as $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_m \geq 0$.

$$\begin{aligned} \phi(\mathbf{X}) &= \frac{1}{2}\|\boldsymbol{\Phi} - \mathbf{X}\|_F^2 + \lambda \sum_{i=1}^{m} \int_0^{\lambda_i} (1 - \frac{u}{\gamma})_+ \, du \\ &= \frac{1}{2}\|\boldsymbol{\Phi} - \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^T\|_F^2 + \lambda \sum_{i=1}^{m} \int_0^{\lambda_i} (1 - \frac{u}{\gamma})_+ \, du \\ &= \frac{1}{2}\|\boldsymbol{\Phi}\|_F^2 + \frac{1}{2}(\sum_{i=1}^{m} \lambda_i^2 - 2\sum_{i=1}^{m} \lambda_i \mathbf{u}_i^T \boldsymbol{\Phi} \mathbf{v}_i) + \lambda \sum_{i=1}^{m} \int_0^{\lambda_i} (1 - \frac{u}{\gamma})_+ \, du. \end{aligned}$$

Fixing $\mathbf{u}_i$ and $\mathbf{v}_i$, and then differentiating $\phi(\mathbf{X})$ with respect to λ_i yields

$$\lambda_i - \mathbf{u}_i^T \boldsymbol{\Phi} \mathbf{v}_i + \lambda(1 - \frac{\lambda_i}{\gamma})_+ = 0.$$

Denoting $\xi_i = \mathbf{u}_i^T \boldsymbol{\Phi} \mathbf{v}_i$, we obtain

$$\hat{\lambda_i} = S_{\lambda,\gamma}(\xi_i) = \begin{cases} \xi_i & \text{if} \quad \gamma < \xi_i, \\ \frac{\xi_i - \lambda}{1 - \frac{\lambda}{\gamma}} & \text{if} \quad \lambda < \xi_i \leq \gamma, \\ 0 & \text{if} \quad \xi_i \leq \lambda. \end{cases}$$

Substituting the $\hat{\lambda_i}$ back into $g(\mathbf{X})$ yields

$$\phi(\mathbf{X}) = \frac{1}{2}\|\boldsymbol{\Phi}\|_F^2 + \frac{1}{2}\sum_{i=1}^{m}\hat{\lambda_i}^2 - \sum_{i=1}^{m}\hat{\lambda_i}\xi_i + \lambda\sum_{i=1}^{m}\Big\{(\hat{\lambda_i} - \frac{\hat{\lambda_i}^2}{2\gamma})I(\hat{\lambda_i} \le \gamma) + \frac{\gamma}{2}I(\hat{\lambda_i} > \gamma)\Big\}.$$

We now see that minimizing $\phi(\mathbf{X})$ w.r.t. $\mathbf{X}$ is equivalent to the minimization of ψ w.r.t. the ξ_i. Here

$$\begin{aligned}\psi &= \frac{1}{2}\sum_{i=1}^{m}\hat{\lambda_i}^2 - \sum_{i=1}^{m}\hat{\lambda_i}\xi_i + \lambda\sum_{i=1}^{m}\Big\{\Big(\hat{\lambda_i} - \frac{\hat{\lambda_i}^2}{2\gamma}\Big)I(\hat{\lambda_i} \le \gamma) + \frac{\lambda\gamma}{2}I(\hat{\lambda_i} > \gamma)\Big\} \\ &= \frac{1}{2}\sum_{\lambda<\xi_i\le\gamma}\mu^2(\xi_i-\lambda)^2 + \frac{1}{2}\sum_{\xi_i>\gamma}\xi_i^2 - \sum_{\lambda<\xi_i\le\gamma}\mu(\xi_i-\lambda)\xi_i - \sum_{\xi_i>\gamma}\xi_i^2 \\ &\quad +\lambda\sum_{\lambda<\xi_i\le\gamma}\Big\{\mu(\xi_i-\lambda) - \frac{\mu^2(\xi_i-\lambda)^2}{2\gamma}\Big\} + \sum_{\xi_i>\gamma}\frac{\lambda\gamma}{2},\end{aligned}$$

where $\mu = \frac{1}{1-\frac{\lambda}{\gamma}} = \frac{\gamma}{\gamma-\lambda}$. Since the ξ_i are partitioned into the three parts, we consider the corresponding terms of ψ separately. The term of ψ corresponding to $\lambda < \xi_i \le \gamma$ is

$$\begin{aligned}\psi_1 &= \sum_{\lambda<\xi_i\le\gamma}\Big\{\frac{1}{2}\mu^2(\xi_i-\lambda)^2 - \mu(\xi_i-\lambda)\xi_i + \lambda\mu(\xi_i-\lambda) - \frac{\lambda\mu^2(\xi_i-\lambda)^2}{2\gamma}\Big\} \\ &= -\sum_i\frac{\mu}{2}(\xi_i-\lambda)^2.\end{aligned}$$

The term of ψ corresponding to $\xi_i > \gamma$ is

$$\psi_2 = \sum_{\xi_i>\gamma}\Big\{-\frac{1}{2}\xi_i^2 + \frac{\lambda\gamma}{2}\Big\}.$$

Recall that $\xi_i = \mathbf{u}_i^T\boldsymbol{\Phi}\mathbf{v}_i$. In order to minimize ψ w.r.t. ξ_i, $\mathbf{u}_i$ and $\mathbf{v}_i$ should be the singular vectors corresponding the singular values of $\boldsymbol{\Phi}$ and $\xi_i = \sigma_i$. Thus, it is necessary that the optimal solution to $\min\phi(\mathbf{X})$ is $\mathbf{X} = S_{\lambda,\gamma}(\boldsymbol{\Phi})$. □

Now we have introduced a so-called non-convex soft shrinkage operator $S_{\lambda,\gamma}(\cdot)$. Compared to the popular soft shrinkage operator, it has an advantage of nearly unbiasedness, since it keeps large singular values of a matrix unchanged, see Figure 1. It is expected that our algorithms have higher accuracy compared to the state-of-the-art algorithms.

Based on this non-convex soft shrinkage operator, we develop two algorithms. One is to directly solve an unconstrained problem and the other is to solve an equivalent form but with explicit constraints.

3.1 The Non-convex Soft Imputation Algorithm

The γ-norm regularization problem is

$$\min_{\mathbf{X}}\{J(\mathbf{X}) = \frac{1}{2}\|P_\Omega(\mathbf{X}-\mathbf{M})\|_F^2 + \lambda\|\mathbf{X}\|_\gamma\}. \tag{5}$$

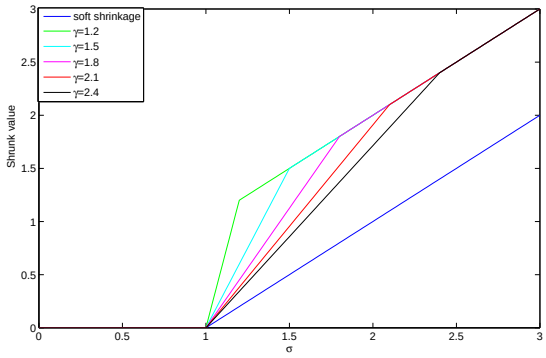

Fig. 1. Non-convex soft shrinkage operator vs. soft shrinkage operator ($\lambda = 1$)

We now derive an iterative process to minimize $J(\mathbf{X})$. Suppose we have obtained $\mathbf{X}^k$ at iteration k for a fixed λ, we bound $J(\mathbf{X})$ from above by

$$Q(\mathbf{X}|\mathbf{X}^k) = \frac{1}{2}\|P_\Omega(\mathbf{X}^k - \mathbf{M})\|_F^2 + \langle P_\Omega(\mathbf{X}^k - \mathbf{M}), \mathbf{X} - \mathbf{X}^k\rangle + \lambda\|\mathbf{X}\|_\gamma + \frac{1}{2\alpha}\|\mathbf{X} - \mathbf{X}^k\|_F^2,$$

where $0 < \alpha < 1$. It is obvious that $J(\mathbf{X}) \leq Q(\mathbf{X}|\mathbf{X}^k)$, and the equality hold only when $\mathbf{X} = \mathbf{X}^k$. Then we set

$$\begin{aligned}
\mathbf{X}^{k+1} &= \operatorname*{argmin}_{\mathbf{X}} Q(\mathbf{X}|\mathbf{X}^k) \\
&= \operatorname*{argmin}_{\mathbf{X}} \{\frac{1}{2}\|\mathbf{X} - (\mathbf{X}^k + \alpha P_\Omega(\mathbf{M} - \mathbf{X}^k))\|_F^2 + \alpha\lambda\|\mathbf{X}\|_\gamma\} \\
&= S_{\alpha\lambda,\gamma}(\mathbf{X}^k + \alpha P_\Omega(\mathbf{M} - \mathbf{X}^k)).
\end{aligned}$$

The above described iterations constitute inner loop for a fixed λ. In the outer loop, we decrease λ every time and use the previous solution as warm start for next iteration. We call this algorithm as non-convex soft imputation (NCSI), see Algorithm 1.

3.2 The Non-convex Alternating Direction Method of Multipliers

The above NCSI algorithm is designed to iteratively minimize an unconstrained problem (5). We can equivalently reform it as a optimization problem with linear constraints:

$$\begin{aligned}
&\min_{\mathbf{X},\mathbf{E}} \|\mathbf{X}\|_\gamma + \frac{\tau}{2}\|P_\Omega(\mathbf{E})\|_F^2 \\
&\text{s.t.} \quad P_\Omega(\mathbf{M}) = \mathbf{X} + \mathbf{E}.
\end{aligned} \tag{6}$$

where $\tau = \frac{1}{\lambda}$. We employ alternating direction method of multipliers (ADMM) to solve the optimization problem. ADMM was originally proposed in [18], and has been applied to a number of convex optimization problems [19]. Recently ADMM algorithm

Algorithm 1 The NCSI algorithm

input: ν, $P_\Omega(\mathbf{M})$ and tolerance ϵ, $0 < \alpha < 1$, $0 < \rho < 1$, γ
Initialize: $\mathbf{Z}_0 = \mathbf{0}$, λ_0
while $\lambda_i > \nu$ **do**
  $\mathbf{X}^0 = \mathbf{Z}_i$
  repeat
    $\mathbf{X}^{k+1} = S_{\alpha\lambda_i,\gamma}(\mathbf{X}^k + \alpha P_\Omega(\mathbf{M} - \mathbf{X}^k))$
  until $\frac{\|P_\Omega(\mathbf{X}^k - \mathbf{M})\|_F}{\|P_\Omega(\mathbf{M})\|_F} < \epsilon$
  $\mathbf{Z}_{i+1} = \mathbf{X}^k$
  $\lambda_{k+1} = \rho\lambda_k$
end while
output $\mathbf{X}_{sol} = \mathbf{Z}_i$

have been used in the minimization of a non-convex function [13, 20]. Here we use ADMM to solve the non-convex problem (6). The derived algorithm is called non-convex alternating direction method of multipliers (NCADMM). This algorithm is similar to [8], while it has an advantage of taking noise into consideration. Thus it is expected to have a higher accuracy.

The augmented Lagrangian function of problem (6) is

$$L(\mathbf{X}, \mathbf{E}, \mathbf{Y}, \mu) = \|\mathbf{X}\|_\gamma + \frac{\tau}{2}\|P_\Omega \mathbf{E}\|_F^2 + \langle \mathbf{Y}, P_\Omega(\mathbf{M}) - \mathbf{X} - \mathbf{E}\rangle + \frac{\mu}{2}\|P_\Omega(\mathbf{M}) - \mathbf{X} - \mathbf{E}\|_F^2.$$

The NCADMM optimize w.r.t. one variable while keeping the others fixed. Specifically the optimization problem can be solved efficiently by the following iterations.

$$\begin{aligned}
\mathbf{X}^{k+1} &= \operatorname*{argmin}_{\mathbf{X}} L(\mathbf{X}, \mathbf{E}^k, \mathbf{Y}^k, \mu) \\
&= \operatorname*{argmin}_{\mathbf{X}} \|\mathbf{X}\|_\gamma + \frac{\mu}{2}\|\mathbf{X} - (P_\Omega(\mathbf{M}) - \mathbf{E}^k + \frac{1}{\mu}\mathbf{Y}^k)\|_F^2 \\
&= S_{\frac{1}{\mu},\gamma}(P_\Omega(\mathbf{M}) - \mathbf{E}^k + \frac{1}{\mu}\mathbf{Y}^k),
\end{aligned}$$

$$\begin{aligned}
\mathbf{E}^{k+1} &= \operatorname*{argmin}_{\mathbf{E}} L(\mathbf{X}^{k+1}, \mathbf{E}, \mathbf{Y}^k, \mu) \\
&= \operatorname*{argmin}_{P_\Omega(\mathbf{E})} \frac{\tau + \mu}{2}\|P_\Omega(\mathbf{E})\|_F^2 - \langle P_\Omega(\mathbf{Y}^k + \mu(P_\Omega(\mathbf{M}) - \mathbf{X}^{k+1})), P_\Omega(\mathbf{E})\rangle \\
&+ \operatorname*{argmin}_{P_{\bar{\Omega}}(\mathbf{E})} \frac{\mu}{2}\|P_{\bar{\Omega}}(\mathbf{E})\|_F^2 - \langle P_{\bar{\Omega}}(\mathbf{Y}^k + \mu(P_\Omega(\mathbf{M}) - \mathbf{X}^{k+1})), P_{\bar{\Omega}}(\mathbf{E})\rangle \\
&= P_\Omega(\mathbf{E}^{k+1}) + P_{\bar{\Omega}}(\mathbf{E}^{k+1}),
\end{aligned}$$

and

$$\mathbf{Y}^{k+1} = \mathbf{Y}^k + \mu(P_\Omega(\mathbf{M}) - \mathbf{X}^{k+1} - \mathbf{E}^{k+1}),$$

where

$$\begin{aligned} P_{\Omega}(\mathbf{E}^{k+1}) &= \frac{\mu}{\tau+\mu}P_{\Omega}(\mathbf{M}-\mathbf{X}^{k+1}) + \frac{1}{\mu+\tau}P_{\Omega}(\mathbf{Y}^{k}), \\ P_{\bar{\Omega}}(\mathbf{E}^{k+1}) &= P_{\bar{\Omega}}(-\mathbf{X}^{k+1}) + \frac{1}{\mu}P_{\bar{\Omega}}(\mathbf{Y}^{k}). \end{aligned}$$

Note that if we set $\mathbf{Y}^0 = \mathbf{0}$, then during the iterations $P_{\bar{\Omega}}(\mathbf{Y}^k) = \mathbf{0}$ for all k. Using this property and eliminating the variable E leads to the iteration as follows

$$\mathbf{X}^{k+1} = S_{\frac{1}{\mu},\gamma}\left(\mathbf{X}^{k} + \frac{\tau}{\tau+\mu}P_{\Omega}(\mathbf{M}-\mathbf{X}^{k}) - \frac{1}{\mu+\tau}P_{\Omega}(\mathbf{Y}^{k-1}) + \frac{1}{\mu}\mathbf{Y}^{k}\right), \quad (7)$$

$$\mathbf{Y}^{k+1} = \mathbf{Y}^{k} + \mu\left(\frac{\tau}{\tau+\mu}P_{\Omega}(\mathbf{M}-\mathbf{X}^{k+1}) - \frac{1}{\mu+\tau}P_{\Omega}(\mathbf{Y}^{k})\right). \quad (8)$$

In previously described process, the penalty parameter μ is fixed. It is found that a small constant μ may lead to slow convergence, while large μ may make the algorithm ill-conditioned. Thus a dynamic μ is preferred in practice. Inspired by [9] we use the following update rule for μ.

$$\mu_{k+1} = \min(\mu_{max}, \rho\mu_k), \quad (9)$$

where μ_{max} is the upper bound on the penalty parameter μ. The value of ρ is determined by

$$\rho = \begin{cases} \rho_0, & \mu_k\frac{\|\mathbf{X}^{k+1}-\mathbf{X}^{k}\|}{\|P_{\Omega}(\mathbf{M})\|_F^2} < \zeta, \\ 1, & \text{otherwise}, \end{cases} \quad (10)$$

where $\rho_0 > 1$ and $\zeta > 0$ is a threshold fixed in advance. We summarize the entire procedure in Algorithm 2.

Algorithm 2 The NCADMM algorithm

input: $\tau = 1/\lambda$, $P_{\Omega}(\mathbf{M})$, tolerance ϵ , threshold ζ, μ_{max}, $\rho_0 > 1$, γ
Initialize: $\mathbf{X}_1^0 = \mathbf{0}$, $\mathbf{Y}^0 = \mathbf{0}$,
repeat

$$\mathbf{X}^{k+1} = S_{\frac{1}{\mu},\gamma}\left(\mathbf{X}^{k} + \frac{\tau}{\tau+\mu}P_{\Omega}(\mathbf{M}-\mathbf{X}^{k}) - \frac{1}{\mu+\tau}P_{\Omega}(\mathbf{Y}^{k-1}) + \frac{1}{\mu}\mathbf{Y}^{k}\right)$$

$$\mathbf{Y}^{k+1} = \mathbf{Y}^{k} + \mu\left(\frac{\tau}{\tau+\mu}P_{\Omega}(\mathbf{M}-\mathbf{X}^{k+1}) - \frac{1}{\mu+\tau}P_{\Omega}(\mathbf{Y}^{k})\right)$$

Update μ_{k+1} according to (9) and (10)
until $\frac{\|P_{\Omega}(\mathbf{X}^{k+1}-\mathbf{M})\|_F}{\|P_{\Omega}(\mathbf{M})\|_F} < \epsilon$
Output $\mathbf{X}_{sol} = \mathbf{X}^k$

4 Convergence Analysis of NCSI Algorithm

We need to further explore the γ-norm before proving the convergence property of NCSI algorihtm. First we need a definition called absolutely symmetric function [21].

Definition 3 (Absolutely symmetric). *Suppose f is a mapping from $\mathbf{R}^m$ to $\mathbf{R}$. We say that f is absolutely symmetric if $f(x_1, x_2, \ldots, x_m) = f(|x_{\pi(1)}|, |x_{\pi(2)}|, \ldots, |x_{\pi(m)}|)$ for any permutation π.*

Lemma 1. *The gamma-norm $\|\mathbf{X}\|_\gamma$ of a $n \times m$ matrix $\mathbf{X}$ can be decomposed as the difference of two convex functions $f(\boldsymbol{\sigma}(\mathbf{X}))$ and $g(\boldsymbol{\sigma}(\mathbf{X}))$ of matrix $\mathbf{X}$, where*

$$f(\boldsymbol{\sigma}(\mathbf{X})) = \sum_{i=1}^{m} \sigma_i(\mathbf{X}), \tag{11}$$

$$g(\boldsymbol{\sigma}(\mathbf{X})) = \sum_{i=1}^{m} \frac{\sigma_i^2(\mathbf{X})}{2\gamma} \mathbb{I}\{\sigma_i(\mathbf{X}) < \gamma\} + (\sigma_i(\mathbf{X}) - \frac{\gamma}{2}) \mathbb{I}\{\sigma_i(\mathbf{X}) \geq \gamma\}. \tag{12}$$

The above Lemma can be inferred from [22]: If a mapping f is absolutely symmetric and convex on $\mathbb{R}^m$, then $f(\boldsymbol{\sigma}(\mathbf{X}))$ is convex w.r.t. matrix $\mathbf{X}$. In Lemma 1, both $f(\boldsymbol{\sigma})$ and $g(\boldsymbol{\sigma})$ are absolutely symmetric and convex, so $f(\boldsymbol{\sigma}(\mathbf{X}))$ and $g(\boldsymbol{\sigma}(\mathbf{X}))$ are convex functions on matrix $\mathbf{X}$.

Definition 4. *We say that two matrices $\mathbf{X}$ and $\mathbf{Y}$ in $\mathbf{R}^{n\times m}$ have a simultaneous ordered singular value decomposition if there exist two orthonormal matrices $\mathbf{U} \in \mathbf{R}^{n\times m}$ and $\mathbf{V} \in \mathbf{R}^{m\times m}$ such that $\mathbf{X} = \mathbf{U}\mathrm{diag}(\boldsymbol{\sigma}(\mathbf{X}))\mathbf{V}^\top, \mathbf{Y} = \mathbf{U}\mathrm{diag}(\boldsymbol{\sigma}(\mathbf{Y}))\mathbf{V}^\top$.*

Theorem 2. *let a function f be absolutely symmetric and convex. Consider the corresponding convex function $f(\boldsymbol{\sigma}(\mathbf{X}))$. The matrix Y is a subgradient of $f(\boldsymbol{\sigma}(\mathbf{X}))$ at $\mathbf{X}$ if and only if $\boldsymbol{\sigma}(\mathbf{Y})$ is a subgradient of f at $\boldsymbol{\sigma}(X)$ and the two matrices $\mathbf{X}$ and $\mathbf{Y}$ admit simultaneous ordered singular value decomposition.*

Detailed proof of this theorem can be found in [21]. We can compute the subgradient of function $f(\boldsymbol{\sigma}(\mathbf{X}))$ and $g(\boldsymbol{\sigma}(\mathbf{X}))$ w.r.t. $\mathbf{X}$ by applying this theorem directly.

Corollary 1. *(1) Let $f(\boldsymbol{\sigma}(\mathbf{X}))$ and $g(\boldsymbol{\sigma}(\mathbf{X}))$ be defined as Eqn. (11) and (12). Suppose $\mathbf{X} \in \mathbb{R}^{n\times m}$.The matrix $\mathbf{Y}_f$ is a subgradient of $f(\boldsymbol{\sigma}(\mathbf{X}))$ if and only if $\sigma_i(\mathbf{Y}_f) = \begin{cases} 1 & \text{if } \sigma_i(\mathbf{X}) > 0 \\ \alpha & \text{if } \sigma_i(\mathbf{X}) = 0 \end{cases}$ where $0 \leq \alpha \leq 1$, and the two matrices admit simultaneous ordered singular value decomposition.*

(2) Suppose $h_i(\boldsymbol{\sigma}(\mathbf{X})) = \frac{\sigma_i(\mathbf{X})}{\gamma} \mathbb{I}\{\sigma_i(\mathbf{X}) < \gamma\} + \mathbb{I}\{\sigma_i(\mathbf{X}) \geq \gamma\}$. $\mathbf{Y}_g$ is a subgradient of $g(\boldsymbol{\sigma}(\mathbf{X}))$ if and only if $\sigma_i(\mathbf{Y}_g) = h_i(\boldsymbol{\sigma}(\mathbf{X}))$ and the two matrices $\mathbf{X}$ and $\mathbf{Y}_g$ admit simultaneous ordered singular value decomposition.

The following theorem shows that our algorithm NSCI decreases the objective function at every iteration.

Theorem 3. *For every fixed* $0 < \alpha < 1$ *and* $\lambda > 0$*, define a sequence* $\mathbf{X}^k$

$$\mathbf{X}^{k+1} = S_{\alpha\lambda,\gamma}(\mathbf{X}^k + \alpha P_\Omega(\mathbf{M} - \mathbf{X}^k)) \tag{13}$$

with a starting point $\mathbf{X}^0$*. The sequence* $\mathbf{X}^k$ *satisfies*

$$J(\mathbf{X}^{k+1}) \leq J(\mathbf{X}^k) - \frac{1-\alpha}{2\alpha}\|\mathbf{X}^{k+1} - \mathbf{X}^k\|_F^2. \tag{14}$$

.

Proof. Suppose

$$L(\mathbf{X}, \mathbf{Z}) = \frac{1}{2}\|P_\Omega(\mathbf{Z} - \mathbf{M})\|_F^2 + \langle P_\Omega(\mathbf{Z} - \mathbf{M}), \mathbf{X} - \mathbf{Z}\rangle + \lambda\|\mathbf{X}\|_\gamma.$$

Since

$$\mathbf{X}^{k+1} = \underset{\mathbf{X}}{\operatorname{argmin}}\, L(\mathbf{X}, \mathbf{X}^k) + \frac{1}{2\alpha}\|\mathbf{X} - \mathbf{X}^k\|_F^2, \tag{15}$$

then we have

$$\begin{aligned} J(\mathbf{X}^{k+1}) &\leq L(\mathbf{X}^{k+1}, \mathbf{X}^k) + \frac{1}{2}\|\mathbf{X}^{k+1} - \mathbf{X}^k\|_F^2 \\ &= L(\mathbf{X}^{k+1}, \mathbf{X}^k) + \frac{1}{2\alpha}\|\mathbf{X}^{k+1} - \mathbf{X}^k\|_F^2 - \frac{1-\alpha}{2\alpha}\|\mathbf{X}^{k+1} - \mathbf{X}^k\|_F^2 \\ &\leq L(\mathbf{X}^k, \mathbf{X}^k) + \frac{1}{2\alpha}\|\mathbf{X}^k - \mathbf{X}^k\|_F^2 - \frac{1-\alpha}{2\alpha}\|\mathbf{X}^{k+1} - \mathbf{X}^k\|_F^2 \\ &= J(\mathbf{X}^k) - \frac{1-\alpha}{2\alpha}\|\mathbf{X}^{k+1} - \mathbf{X}^k\|_F^2. \end{aligned}$$

□

The inequality (14) tells us that $J(\mathbf{X}^k)$ monotonously decrease to its limit point since $J(\mathbf{X}) \geq 0$. Meanwhile the sequence $\{\|\mathbf{X}^{k+1} - \mathbf{X}^k\|_F^2\}$ converges to 0.

The next theorem states that any limit point generated by Algorithm 1 is a critical point of objective function (5).

Theorem 4. *For every fixed* $0 < \alpha < 1$*,* $\lambda > 0$ *and* $\gamma > \lambda$*. Each limit point of* $\mathbf{X}^k$ *generated by Eqn. (13) is a critical point of* $J(\mathbf{X})$*.*

Proof. Suppose there is a subsequence $\{\mathbf{X}^k\}_{k\in\mathbb{K}}$ converging to $\mathbf{X}^\infty$. According to the minimization problem (15) and Lemma 1 we have

$$0 \in P_\Omega(\mathbf{X}^k - \mathbf{M}) + \frac{1}{\alpha}(\mathbf{X}^{k+1} - \mathbf{X}^k) + \lambda(\partial f(\boldsymbol{\sigma}(\mathbf{X}^{k+1})) - \partial g(\boldsymbol{\sigma}(\mathbf{X}^{k+1}))).$$

Suppose $\mathbf{S}^{k+1} \in \partial f(\boldsymbol{\sigma}(\mathbf{X}^{k+1}))$, $\mathbf{T}^{k+1} \in \partial g(\boldsymbol{\sigma}(\mathbf{X}^{k+1}))$ satisfying

$$P_\Omega(\mathbf{X}^k - \mathbf{M}) + \frac{1}{\alpha}(\mathbf{X}^{k+1} - \mathbf{X}^k) + \lambda(\mathbf{S}^{k+1} - \mathbf{T}^{k+1}) = 0. \tag{16}$$

Since $\mathbf{S}^{k+1}$ and $\mathbf{T}^{k+1}$ are subgradient of $f(\boldsymbol{\sigma}(\mathbf{X}))$ and $g(\boldsymbol{\sigma}(\mathbf{X}))$ at $\mathbf{X}^{k+1}$, according to Corollary 1, there exist a $n \times m$ orthonormal matrix $\mathbf{U}^{k+1}$ and a $m \times m$ orthogonal matrix $\mathbf{V}^{k+1}$ such that

$$\mathbf{S}^{k+1} = \mathbf{U}^{k+1}\mathrm{diag}\{\boldsymbol{\sigma}(\mathbf{S}^{k+1})\}\mathbf{V}^{(k+1)\top}, \tag{17}$$
$$\mathbf{T}^{k+1} = \mathbf{U}^{k+1}\mathrm{diag}\{\mathbf{h}(\boldsymbol{\sigma}(X^{k+1}))\}\mathbf{V}^{(k+1)\top}, \tag{18}$$
$$\mathbf{X}^{k+1} = \mathbf{U}^{k+1}\mathrm{diag}\{\boldsymbol{\sigma}(X^{k+1})\}\mathbf{V}^{(k+1)\top}. \tag{19}$$

Since orthogonal matrices $\mathbf{U}^{k+1}$, $\mathbf{V}^{k+1}$ and the singular values $\boldsymbol{\sigma}(S^{k+1})$ are bounded, without loss of generality we suppose they converging to $\mathbf{U}^{\infty}$, $\mathbf{V}^{\infty}$ and $\boldsymbol{\sigma}(S^{\infty})$. According to (17), (18) and (19) we infer that $\mathbf{S}^{\infty}$, the limit point of $\mathbf{S}^{k+1}$ is subgradient of $f(\boldsymbol{\sigma}(\mathbf{X}^{\infty}))$ and $\mathbf{T}^{\infty}$, the limit point of $\mathbf{T}^{k+1}$ is subgradient of $g(\boldsymbol{\sigma}(\mathbf{X}^{\infty}))$.

Make $k \to \infty, k \in \mathbb{K}$ and use $\lim_{k\to\infty} \mathbf{X}^{k+1} - \mathbf{X}^{k} = 0$, the Eqn. (16) transfers to

$$P_{\Omega}(\mathbf{X}^{\infty} - \mathbf{M}) + \lambda(\mathbf{S}^{\infty} - \mathbf{T}^{\infty}) = 0.$$

So $\mathbf{X}^{\infty}$ is a critical point of $J(\mathbf{X})$. □

5 Experiments

In this section, we conduct experiments on synthetic data, image data and three standard collaborative filtering datasets. To show the effectiveness of NCSI and NCADMM, we compare them with the following matrix completion solvers: ALM [8], SVT [4], SoftImpute [5], and OptSpace [3]. Particularly, ALM, SVT, and SoftImpute are based on the nuclear norm, while OptSpace represent matrix as its factors and optimize a nonconvex objective function. Besides, in collaborative filtering experiment, we also add PMF [23] and GECO [24] into our comparison list.

5.1 Synthetic Data

We generate synthetic data $\mathbf{X}$ by $\mathbf{X} = \mathbf{M}+\sigma\mathbf{Z}$, where $\mathbf{X}, \mathbf{M}, \mathbf{Z} = [z_{ij}] \in \mathbb{R}^{m\times n}$. $z_{i,j}$ is Gaussian white noise with zero mean and standard deviation of one. And $\mathbf{M}$ is a matrix with rank of r produced by $\mathbf{M} = \mathbf{L}\mathbf{R}^{\top}$, in which both $\mathbf{L} \in \mathbb{R}^{m\times r}$ and $\mathbf{R} \in \mathbb{R}^{n\times r}$ have i.i.d. Guassian entries. The set of observed entries Ω is uniformly sampled among the $m \times n$ indices. Suppose that the degree of freedom of matrix with rank r is d_r. We fixed the number of observed entries to $5d_r$ and σ to 10^{-6}.

We only compare our methods with ALM since none of the algorithms mentioned before claimed to outperform ALM in terms of accuracy or efficiency on large synthetic matrices. Additionally, We set the parameter γ to 4 in both NCSI and NCADMM. And the same stop criterion is adopted for all algorithms:

$$\frac{\|P_{\Omega}(\mathbf{X} - \mathbf{M})\|_F}{\|P_{\Omega}(\mathbf{M})\|_F} < \epsilon,$$

in which ϵ is set to 0.3σ. We evaluate the accuracy of the solution $\mathbf{X}_{sol}$ of our algorithm by the relative error (RE), which is a widely used metric in matrix completion, defined

by

$$RE = \frac{\|\mathbf{X}_{sol} - \mathbf{M}\|_{\mathbf{F}}}{\|\mathbf{M}\|_{\mathbf{F}}}.$$

We report the RE and #SVD (number of doing SVD) in Table 1. Experimental results demonstrate that NCSI and NCADMM consistently outperform ALM in accuracy; NCADMM achieve higher accuracy with nearly the same time cost as ALM; and NCSI and NCADMM have almost the same accuracy (since they solve matrix completion problem using same γ-norm based scheme).

Table 1. Comparisons among NCSI, NCADMM, ALM on the synthetic data.

(rank ratio)	RE(10^{-7})			#SVD			CPU-time(minutes)		
(r, $\frac{\|\Omega\|}{(m\times n)}$)	NCSI	NCADMM	ALM	NCSI	NCADMM	ALM	NCSI	NCADMM	ALM
	$m = 10000, n = 10000$								
(10,0.012)	5.412	5.412	6.230	1818	322	325	10.45	4.36	4.7
(20,0.024)	3.853	3.790	4.040	850	179	180	13.78	5.73	6.18
(30,0.035)	3.026	3.020	4.106	522	149	150	17.73	8.58	8.68
(50,0.057)	2.794	2.794	3.864	313	122	115	26.58	11.32	14.78
	$m = 20000, n = 20000$								
(10,0.006)	5.976	5.963	6.410	3755	800	683	48.41	37.02	32.28
(20,0.012)	4.428	4.416	5.246	1606	283	246	76.85	30.21	28.95
(30,0.018)	3.797	3.801	3.952	1034	199	201	76.85	28.24	28.61
(50,0.030)	2.839	2.839	3.957	613	173	164	114.3	60.13	57.41

5.2 Experiment on Image Data

In the image inpainting experiment, we aim to estimate missing (or masked) pixels by exploiting the known content. As colored image is commonly represented as three matrices(containing red, green and blue components respectively) we simply deal with each of three matrix and combine them together to obtain the final results.

Performance of different algorithms are evaluated by the $PSNR$ (Peak Signal-to-Noise Ratio) metric. Suppose that the total number of missing pixel is T and the total squared error TSE is defined by $TSE = error_r^2 + error_g^2 + error_b^2$, then the total mean squared error MSE is defined by $MSE = TSE/3T$. And the $PSNR$ can be evaluated as $PSNR = 10\log_{10} 255^2/MSE$.

In our experiments, the parameters of ALM, SVT, Soft-Impute, and OptSpace are carefully chosen to achieve the best performance. For NCSI and NCADMM we fix $\gamma = 100$ and empirically set $\lambda = 0.001$. Since large μ in NCADMM will make the minimization problem ill-conditioned, we set $\mu_{max} = 10^{10}$.

Two experiments using different image masks are reported. The first is a relatively easy matrix completion problem with random mask. We report the results in Fig. 2 and

Fig. 3. We see that the γ-Norm minimization scheme always achieve larger $PSNR$ compared with other five methods from Fig. 2. Second experiment uses text mask. It is generally agreed that image inpainting with text mask is more difficult since the observed pixels are not randomly sampled and text mask may result in loss of important image information. We report our results in Fig. 4. The results of NCSI and NCADMM are also encouraging.

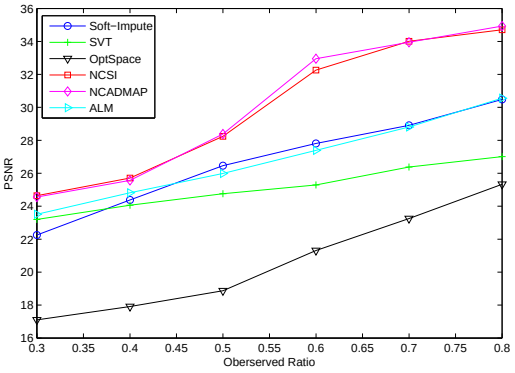

Fig. 2. Comparison of matrix completion algorithms for recovery of a image under different observed ratios

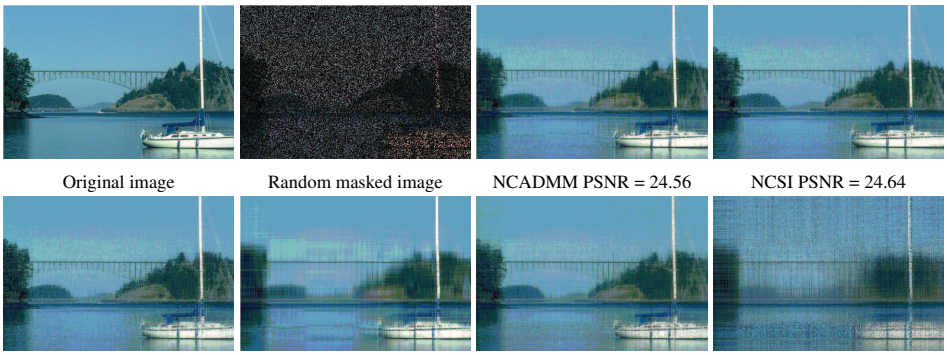

Fig. 3. Comparison of matrix completion algorithms for recovery of a image only 30% of its pixels are observed.

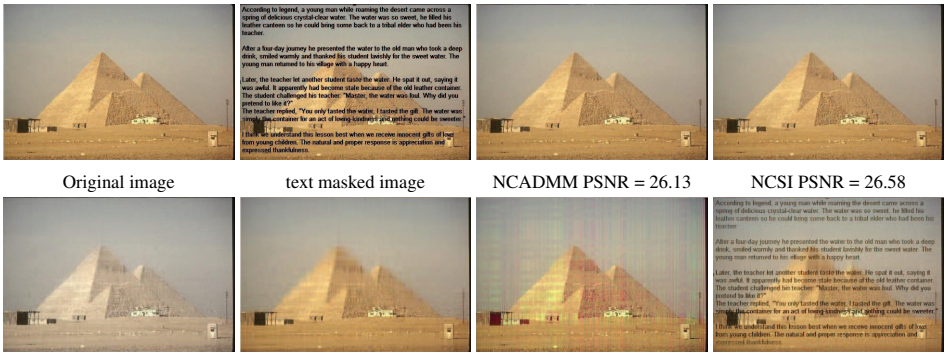

Fig. 4. Comparison of matrix completion algorithms for recovery of a image masked by text.

5.3 Collaborative Filtering

Collaborative filtering (CF) is a technique used by some recommender systems. One of the CF's main purposes is to predict the unknown preference of a user on a set of unrated items, according to other similar users or similar items. In order to validate the performance of our methods, we compare our NCSI and NCADMM with three algorithms using nuclear norm: ALM, SVT and Soft-Impute, and three other non-nuclear-norm algorithms: OptSpace, GECO and PMF.

We use three standard MovieLens Data Sets[1]:

MovieLens-100K contains 100,000 ratings for 1682 movies by 943 users;
MovieLens-1M contains 1million ratings for 3,900 movies by 6,040 users;
MovieLens-10M contains 10 million ratings for 10,681 movies by 71,567 users.

For each data set, we randomly select 70% ratings as known samples, and use the rest ratings to test the performance of the methods. Then, we run 5 repeats for each data set and each method, and report the average results in table 2.

In our experiment, We fix $\gamma = \|P_\Omega(\mathbf{M})\|_F$ and use the commonly accepted CF metric $RMSE$ (Root Mean Square Error) to evaluate the eight methods. $RMSE$ is defined by

$$RMSE = \frac{1}{|T|}\sqrt{\sum_{(i,j)\in T}^{|T|}(X_{ij} - M_{ij})^2},$$

where T is the test set.

Our results in Table 2 show that γ-norm based algorithms outperform other matrix completion algorithms and are competitive to the state-of-the-art collaborative filtering method PMF.

[1] http://www.grouplens.org

Table 2. Performance of deference matrix completion methods on real collaborative filtering data sets.

Data set	NCSI	NCADMM	ALM	SVT	Soft-Impute	OptSpace	GECO	PMF
MovieLens-100k	0.9710	0.9710	1.083	1.536	1.071	1.583	0.9810	0.9790
MovieLens-1M	0.8670	0.8670	0.9037	0.9498	0.9185	1.007	0.8808	0.8683
MovieLens-10M	0.8250	0.8250	0.8843	0.9731	0.8854	too long	0.8402	0.8247

6 Conclusion

In this paper we have employed the matrix γ-norm as a non-convex relaxation to the matrix rank and devised two algorithms: non-convex soft imputation (NCSI) and non-convex alternative direction method of multipliers algorithm (NCADMM), to solve the matrix completion problem. The algorithms are effective, because they can achieve high accuracy in the simulated datasets and real world datasets. Moreover, the NCADMM is quite efficient as its running CPU-time is comparable with the current state-of-the-art algorithms.

Acknowledgments

This work is partially supported by the National Natural Science Foundation of china (No. 61070239 and No. 61272303) and the National Program on Key Basic Research Project of China (973 Program, No. 2010CB327903).

References

1. Candès, E. J., Recht. B.: Exact matrix completion via convex optimization. Foundations of Computational mathematics. vol.9, no.6, pp. 717–772. (2009)
2. Candès, E. J., Tao, T.: The power of convex relaxation: Near-optimal matrix completion. IEEE Transactions on Information Theory. vol.56, no.5, pp. 2053–2080. (2010)
3. Keshavan, R., Montanari, A., Sewoong, O.: Matrix Completion from a Few Entries. IEEE Transactions on Information Theory. vol.56, no.6, pp. 2980–2998. (2010)
4. Cai, J., Candès, E. J., Shen, Z.: A singular value thresholding algorithm for matrix completion. SIAM Journal on Optimization vol.20, no.4, pp. 1956–1982. (2010)
5. R. Mazumder and T. Hastie and R. Tibshirani: Spectral Regularization Algorithms for Learning Large Incomplete Matrices. The Journal of Machine Learning Research. vol.99, pp. 2287–2322. (2010)
6. Ma, S., Goldfarb, D., Chen, L.: Fixed point and Bregman iterative methods for matrix rank minimization. Mathematical Programming. vol. 128, no.1-2, pp. 321–353. (2011)
7. Hale, E.T., Yin, W., Zhang, Y.: A fixed-point continuation method for ℓ_1-regularized minimization with applications to compressed sensing. Technical Report, CAAM TR07-07 (2007)
8. Lin, Z., Chen, M., Wu, L., Ma, Y.: The augmented Lagrange multiplier method for Exact Recovery of Corrupted Low-Rank Matrices. arXiv preprint arXiv:1009.5055. (2010)
9. Lin, Z., Liu, R., Su, Z.: Linearized alternating direction method with adaptive penalty for low-rank representation. Advances in neural information processing systems. (2011)

10. Hu, Y. Zhang, D., Liu, J., Ye, J., He, X.: Accelerated singular value thresholding for matrix completion. Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining. (2012)
11. Recht, B., Fazel, M., Parrilo, Pablo A.: Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. SIAM review. vol.52, no.3, pp. 471-501. (2010)
12. Cai, T. T., and Zhou, We.: Matrix completion via max-norm constrained optimization. arXiv preprint arXiv:1303.0341. (2013)
13. Nie, Feiping and Wang, Hua and Cai, Xiao and Huang, Heng and Ding, Chris.:Robust Matrix Completion via Joint Schatten p-Norm and ℓ_p-Norm Minimization. 2012 IEEE 12th International Conference on Data Mining (ICDM). pp. 566–574. (2012)
14. Hu, Y., Zhang, D., Ye, J., Li, X., He, X.: Fast and Accurate Matrix Completion via Truncated Nuclear Norm Regularization. IEEE Transactions on Pattern Analysis and Machine Intelligence. (2012)
15. Wang, S., Liu, D., Zhang, Z.: Nonconvex Relaxation Approaches to Robust Matrix Recovery. International Joint Conference on Artificial Intelligence. (2013)
16. Mazumder, R., Friedman, J. H., Hastie, T.: SparseNet: Coordinate descent with nonconvex penalties. Journal of the American Statistical Association. vol.106, no.495 (2011)
17. Zhang, C. H.: Nearly Unbiased Variable Selection Under Minimax Concave Penalty. The Annals of Statistics. vol. 38, no.2, pp. 894–942. (2010)
18. Gabay, D., and Bertrand M.: A dual algorithm for the solution of nonlinear variational problems via finite element approximation. Computers & Mathematics with Applications vol.2, no.1, pp.1–40. (1976)
19. Boyd, S., Parikh, N., Chu, E., Peleato, B., Eckstein, J.: Distributed optimization and statistical learning via the alternating direction method of multipliers. Foundations and Trends in Machine Learning. vol.3, no.1, pp.1–122. (2011)
20. Xiang, S., Shen, X., Ye, J.: Efficient Sparse Group Feature Selection via Nonconvex Optimization. arXiv preprint arXiv:1205.5075. (2012)
21. Lewis, A. S., Hristo S. S.: Nonsmooth analysis of singular values. Part I: Theory. Set-Valued Analysis. vol.13, no.3, pp. 213–241. (2005)
22. Lewis, A. S.: Group invariance and convex matrix analysis. SIAM Journal on Matrix Analysis and Applications. vol.17, no.4, pp. 927-949. (1996)
23. Salakhutdinov, R., Mnih, A.: Probabilistic matrix factorization. Advances in neural information processing systems. pp. 1257–1264. (2008)
24. Shalev-Shwartz, S., Gonen, A., Shamir, O.: Large-scale convex minimization with a low-rank constraint. International Conference on Machine Learning. (2011)