Solving the Robust Matrix Completion Problem via a System of Nonlinear Equations

Yunfeng Cai and Ping Li

Cognitive Computing Lab
Baidu Research
No. 10 Xibeiwang East Road, Beijing 100085, China
10900 NE 8th St. Bellevue, WA 98004, USA
{caiyunfeng, liping11}@baidu.com

Abstract

We consider the problem of robust matrix completion, which aims to recover a low rank matrix $L^*$ and a sparse matrix $S^*$ from incomplete observations of their sum $M = L^* + S^* \in \mathbb{R}^{m \times n}$. Algorithmically, the robust matrix completion problem is transformed into a problem of solving a system of nonlinear equations, and the alternative direction method is then used to solve the nonlinear equations. In addition, the algorithm is highly parallelizable and suitable for large scale problems. Theoretically, we characterize the sufficient conditions for when $L^*$ can be approximated by a low rank approximation of the observed $M^*$. And under proper assumptions, it is shown that the algorithm converges to the true solution linearly. Numerical simulations show that the simple method works as expected and is comparable with state-of-the-art methods.

1 Introduction

Robust matrix completion (RMC) (Chen et al., 2011; Tao and Yuan, 2011; Cherapanamjeri et al., 2017; Klopp et al., 2017; Zeng and So, 2018) aims to recover a low rank matrix $L_*$ and a sparse matrix $S_*$ from a sampling of $M = L_* + S_*$. Mathematically, RMC can be formulated as the following optimization problem (Candès et al., 2011; Chandrasekaran et al., 2011):

$$\text{RMC} : \min_{L,S} \text{rank}(L) + \lambda |\text{supp}(S)|,$$

s.t. $L_{ij} + S_{ij} = M_{i,j}, \ (i,j) \in \Omega$,

where $\lambda$ is a tuning parameter, $\Omega$ is a subset of $\{1, \ldots, m\} \times \{1, \ldots, n\}$. When $S = 0$, RMC becomes the matrix completion (MC) problem (Candès and Recht, 2009; Meka et al., 2009; Cai et al., 2010; Candès and Tao, 2010; Jain and Netrapalli, 2015; Liu and Li, 2016); When $\Omega = \{1, \ldots, m\} \times \{1, \ldots, n\}$, RMC becomes the robust principal component analysis (RPCA) (Jolliffe, 2011). Thus, RMC can be taken as a combination/generalization of MC and RPCA (Candès and Plan, 2010; Jain and Netrapalli, 2015; Jain et al., 2013; Keshavan et al., 2010).

In many scientific and engineering problems, people need to recover a low rank matrix from observed data, e.g., the recommender system (Funk, 2006; Candès and Plan, 2010; Hu and Li, 2017, 2018b,a), social network analysis (Huang et al., 2013), machine learning (Candès and Plan, 2010; Davenport and Romberg, 2016), image inpainting (Bertalmío et al., 2000), computer vision (Candès and Plan, 2010), bioinformatics (Kim et al., 2005), etc.

MC/RPCA/RMC has been studied extensively from an optimization point of view, many algorithms are proposed, and exact recovery is discussed under proper assumptions. Most of well-known algorithms are based on convex optimization, in which the rank of a matrix is relaxed to its nuclear norm (the sum of all singular values), and the number of nonzero entries of a matrix is relaxed to its $\ell_1$-norm (the sum of absolute values of all entries), e.g., (Cai et al., 2010; Candès et al., 2011; Candès and Recht, 2009; Candès and Tao, 2010; Recht et al., 2010; Klopp et al., 2017). However, the computation of the nuclear norm of $L$, which requires the computation of its singular value decomposition (SVD), is expensive and unsuitable for parallelization, as a result, algorithms based on the nuclear norm relaxation are often not very realistic for large matrices.

To deal with large problems, the low rank matrix can be represented as the product of a tall-skinny matrix
and a short-fat matrix, so that the low rank property is satisfied automatically. But, the optimization problem becomes nonconvex, which makes it difficult to solve. Bi-convex as the problem is, the alternative minimization can be used to solve it more efficiently, e.g., (Jain et al., 2013). In (Yi et al., 2016), gradient descent method is used to solve RMC, which is shown to be fast. In (Cherapanamjeri et al., 2017), projected gradient method with hard-thresholding is used to solve RMC, with nearly-optimal observation and corruption. We refer the readers to (Zeng and So, 2018) and reference therein for more methods. Besides the optimization based method, quite recently, in (Dutta et al., 2019), the RPCA/RMC is solved via an alternating nonconvex projection method. This method does not require any objective function, convex relaxation or surrogate convex constraint.

**Contribution.** In this paper, we solve the RMC problem via solving a system of nonlinear equations (NLEQ). This method does not require any objective function, convex relaxation or surrogate convex constraint, either. Let \( L_s = X Y^T \), where \( X = [x_1, \ldots, x_m]^T \in \mathbb{R}^{m \times r} \), \( Y = [y_1, \ldots, y_n]^T \in \mathbb{R}^{n \times r} \). The RMC problem can be formulated as the following system of NLEQ:

\[
    x_i^T y_j = M_{ij}, \quad (i, j) \in \Omega \setminus \text{supp}(S_*),
\]

where \( \text{supp}(S_*) \) is the support set of \( S_* \). When the rank \( r \) and the support set of \( S_* \) are both known, solving RMC amounts solving a system of nonlinear equations.

Any numerical method for nonlinear equations can be used to solve it (e.g., the steepest descent method, the Newton method, etc.), among which the simplest one is the alternative direction method (ADM): Fixing \( X \) (or \( Y \)), \( Y \) (or \( X \)) can be updated via solving an (overdetermined) linear system of equations (usually in least square sense). Thus, solving RMC via solving (1) heavily depends on whether we can solving (1) without knowing \( r \) and \( \text{supp}(S_*) \). It is worth mentioning here that in (Meka et al., 2009), such a method was proposed to solve the MC problem. And it was stated there that “... its variants outperform most methods in practice. However, analyzing the performance of alternate minimization is a notoriously hard problem.”

The contributions of this paper are three folds. First, for both full and partial observation cases, we characterize some sufficient conditions for when the low rank approximation of the observed \( M \) is approximate \( L_s \). Second, we propose to solve RMC via solving a system of NLEQ rather than optimization, and an ADM method, which carefully handles the unknown \( r \) and \( \text{supp}(S_*) \) issue, is developed. Third, we carefully analyze the convergence of the ADM, and it is shown that under proper assumptions, the ADM converges to the true solution linearly, i.e., exact recovery can be achieved. So, we give an answer to a problem which is even more difficult than the aforementioned “notoriously hard problem”. In addition, the algorithm is highly parallelizable and naturally suitable for large scale problems. It is also worth mentioning here that the results of this paper are applicable to the MC problem as well as the RPCA problem.

The rest of this paper is organized as follows. In Section 2, we first develop the algorithm, followed by its convergence analysis in Section 3. Numerical experiments are presented in Section 4. Concluding remarks are given in Section 5.

**Notation.** We shall adopt the MATLAB style convention to access the entries of vectors and matrices. The set of integers from \( i \) to \( j \) inclusive is \([i:j]\). For a matrix \( A \), its submatrices \( A_{(k:t, i:j)}, A_{(k:t, :)}, A_{(:, i:j)} \) consist of intersections of row \( k \) to \( row \ell \) and column \( i \) to \( row \ell \) to \( and column j \), \( row k \) to row \( \ell \) and all columns, \( all rows and column i \) to \( column j \), respectively. \( A_{(j,:)} \) and \( A_{(:, k)} \) denote the \( j \)th row and \( k \)th column of \( A \), respectively. \( ||A|| \) stands for the spectral norm of \( A \). \( ||A||_F \) denotes the Frobenius norm, \( ||A||_1 = \sum_{ij} |a_{ij}| \), \( ||A||_{\text{max}} = \max_{ij} |a_{ij}|, ||A||_{2,\infty} = \max_i \|A(:,i)\| \). \( A^\dagger \) stands for the Moore–Penrose inverse, and \( \kappa(A) = \|A\|\|A^\dagger\| \) denote the condition number of \( A \). Denote by \( \sigma_j(A) \) for \( 1 \leq j \leq \min\{m,n\} \) the singular values of \( A \) and they are always arranged in a non-increasing order: \( \sigma_1(A) \geq \sigma_2(A) \geq \cdots \geq \sigma_{\min\{m,n\}}(A) \). \( R(A) \) stands for the range space of \( A \), i.e., \( R(A) = \text{span}\{y \in \mathbb{R}^m \mid y = Ax, x \in \mathbb{R}^n\} \). The vector \( e_j \) stands for the \( j \)th column of the identity matrix \( I \). Furthermore, for an index set \( \Omega \subset \{1, \ldots, n\} \times \{1, \ldots, m\} \), \( |\Omega| \) denotes the cardinality of \( \Omega \), \( |\Omega| A = [A_{(i,j) \in \Omega}] \in \mathbb{R}^{m \times n} \), where \( I \) is the indicator function.

## 2 Algorithm

In this section, we first reformulate the RMC problem as a problem of solving a system of nonlinear equations (NLEQ), show how to solve the NLEQ via an alternative direction method (ADM), then the overall algorithm is summarized.

### 2.1 Problem Reformulation, Difficulty and Solution

Let \( M = X Y^T \), where \( X = [x_1, \ldots, x_m]^T \in \mathbb{R}^{m \times r} \), \( Y = [y_1, \ldots, y_n]^T \in \mathbb{R}^{n \times r} \). The MC problem can be formulated as

\[
    x_i^T y_j = M_{ij}, \quad (i, j) \in \Omega.
\]

Now recall (1), the task of RMC becomes solving (2) with unknown \( r \) and minimum violators. Here by a “violator”, denoted by \((i', j')\), we mean that \( x_{i'}^T y_{j'} \neq M_{i'j'} \), it will be also referred to as an “outlier” hereafter.
The ADM is the simplest method to solve the NLEQ of form (2) – given an initial guess for $X$, we fix $X$, (2) becomes a linear system in $Y$ (which is assumed to be overdetermined), we can solve the linear system for $Y$; similarly, we fix $Y$ to solve $X$: the iteration continues until convergence. As $r$ and $\text{supp}(S)$ are unknown, our task is to determine them during the iteration of ADM. Next, we first show how to determine $r$, then $\text{supp}(S)$.

**Determine the rank adaptively** In the $t$th iteration of ADM, let the current rank estimation be $r_t$, $Y_t$ be current estimation for $Y$ and $Y_t$ have orthonormal columns, i.e., $Y_t^T Y_t = I_{r_t}$. By ADM, the estimation of $X$ can be obtained, denote it by $\hat{X}_{t+1}$. In order to update $r_t$, we need to compute the singular values of $\hat{X}_{t+1} Y_t^T$. Noticing that $Y_t$ is orthonormal, then the top $r_t$ singular values of $\hat{X}_{t+1} Y_t^T$ will be the singular values of $\hat{X}_{t+1}$. Then the singular values can be obtained as follows: first, compute the QR decomposition of $\hat{X}_{t+1}$:

$$\tilde{X}_{t+1} = \hat{X}_{t+1} R_{x,t+1},$$

where $\tilde{X}_{t+1} \in \mathbb{R}^{m \times r_t}$ is orthonormal, $R_{x,t+1} \in \mathbb{R}^{r_t \times r_t}$; second, compute the SVD of $R_{x,t+1}$:

$$R_{x,t+1} = Q_z \Sigma Q_y^T,$$

where $Q_z \in \mathbb{R}^{r_t \times r_t}$, $Q_y \in \mathbb{R}^{n \times r_t}$ are orthogonal, $\Sigma = \text{diag}(\hat{\sigma}_1, \ldots, \hat{\sigma}_{r_t})$ with $\hat{\sigma}_1 \geq \cdots \geq \hat{\sigma}_{r_t} \geq 0$. Then the singular values of $\tilde{X}_{t+1} Y_t^T$ are $\hat{\sigma}_1, \ldots, \hat{\sigma}_{r_t}$. Similarly, when $X_t$ is the current estimation for $X$, and $X_t^T X_t = I_{r_t}$, we can compute an estimation for $Y$, its singular values can be obtained.

Let $L_t$ be the current estimation for $L_s$. When the rank is underestimated, i.e., $r_t < r$, the residual $\tau_t = \| \Pi_t (L_t + S_t - M) \|_F$ will stagnate, in such case, we increase the estimated rank $r_t$. When the rank is overestimated, i.e., $r_t > r$, we expect to observe rank deficiency from the singular values $\hat{\sigma}_1, \ldots, \hat{\sigma}_{r_t}$. In such a case, we decrease the estimated rank $r_t$. For the RMC problem, we prefer an overestimated rank over an underestimated rank due to the following reason. The residual $\tau_t$ stagnates for two reasons: one is that the estimated rank is smaller than the true rank; the other is that $| \text{supp}(S_s) \setminus \text{supp}(S_t) |$ is large. Then when the residual stagnates, it is difficult for us to make a good choice – to increase the estimated rank or to drop some equalities (of course, those equalities need to be carefully selected) in (2). Increasing the estimated rank when $| \text{supp}(S_s) \setminus \text{supp}(S_t) |$ is large or dropping equalities in (2) when the rank is underestimated will both lead to catastrophic consequences, such as the estimated rank exceeds a prescribed limit, too many “correct” equalities are dropped which will probably result in underdetermined linear systems when updating $X$ (or $Y$). With an overestimated rank, when the residual stagnates, we decrease the estimated rank via the singular values of $L_t$; if there is no rank deficiency in $L_t$, we drop some equalities in (2).

When an overestimated rank decreases to the actual rank, it is expected that the estimated rank will remain unchanged in the follow-up iterations. Therefore, we do not need to check the singular values of $L_t$ in each iteration for the sake of efficiency.

**Determine $\text{supp}(S)$ via outlier detection** When a good approximation $\tilde{L}$ of $L_s$ is obtained, $S_s = M - \tilde{L}_s \approx M - \tilde{L}$. Thus, it is reasonable to detect $(i, j) \in \Omega \cap \text{supp}(S_s)$ from the residual $(R_{ij} = M_{ij} - \tilde{L}_{ij})_{(i,j)\in\Omega}$. Outlier detection has been used for centuries to remove abnormal data. Various outlier detection techniques have been used (Ester et al., 1996; Hodge and Austin, 2004; Xu et al., 2010; Rahmani and Li, 2019; Slawski et al., 2019). In our implementation, we simply determine the outliers as follows: find the top-$k$ values in each row and column of $| \Omega |$ (unavailable entries of $| \Omega |$ are set to zero), and the entries in the intersection are taken as outliers. Alternatively, simply find the top $k'$ values among all entries of $| \Omega |$. Here $k, k'$ are two parameters which can be tuned. In what follows, we denote

$$T_s(A) = [b_{ij}],$$

where $s$ is the number of the removed outliers, $b_{ij} = A_{(i,j)}$ if $A_{(i,j)}$ is an outlier, $b_{ij} = 0$, otherwise. Of course, one can also try other outlier detection techniques.

### 2.2 Algorithm details

Now we present Algorithm 1, which summarizes the ADM for RMC described in the previous subsection. Some implementation details follows.

**Initializing $Y_0$** According to Theorem 2 below, good initial guesses for $X$ and $Y$ can be obtained by computing the SVD of $\Pi_{\Omega_0}(M)$. An iterative procedure (e.g., Krylov subspace method) is usually adopted to accomplish the task, in which matrix vector products $\Pi_{\Omega_0}(M)v$ and $\Pi_{\Omega_0}(M)^Tv$ are called several times. A simpler way, which is more efficient and numerically proven to be reliable, is the following: compute $W = \Pi_{\Omega_0}(M)^T \Pi_{\Omega_0}(M) N$, compute an orthonormal basis for $W$, and set the columns of $Y_1$ as the basis. Here $N \in \mathbb{R}^{n \times n}$ is a random matrix with entries drawn from the standard normal distribution. Such a procedure is essentially one iteration of the subspace method (a generalization of power method to compute several dominant eigenvectors). Since an initial guess for $X$ or $Y$ is sufficient for ADM to run in Algorithm 1, it is indeed unnecessary to compute the estimations for both $X$ and $Y$. 

Yunfeng Cai and Ping Li
Algorithm 1 ADM for RMC via NLEQ

Input: The observed matrix $\Pi_\Omega(M)$, a sparsity level parameter $s$, an estimated rank $r_0$, an upper bound $κ$ for the condition number of $L_*$, and a tolerance tol.

Output: $X ∈ \mathbb{R}^{m \times r}, \ Y ∈ \mathbb{R}^{n \times r}$ and $S ∈ \mathbb{R}^{m \times n}$ such that $\|\Pi_\Omega(XY^T + S - M)\|_F ≤ tol, \|S\|_0 ≤ s$.

1: Set $S_0 = T_0(M), \ X_0 = 0, \ Y_0 = 0, \ Σ_0 = 0, \ t = 1$;
2: Compute $[X_1, Σ_1, Y_1] = SVD_{\tau_0}((M - S_0)/p'),$ where $p' = (\|Ω\| - s)/mn$;
3: Compute $R_t = Π_\Omega(M - X_tΣ_tY_t^T)$;
4: Set $S_t = T_t(R_t)$. $Ω_t = Ω \setminus supp(S_t)$;
5: Compute $τ_t = \|Π_\Omega(M - X_tΣ_tY_t^T - S_t)\|_F$;
6: while $τ_t > tol$ do
7: Set $t = t + 1$;
8: Solve $Π_{Ω_{t-1}}(X_tY_t^T - 1) = Π_{Ω_{t-1}}(M)$ for $\tilde{X}_t$;
9: Compute the QR decomposition $\tilde{X}_t = \tilde{X}_tR_{x,t}$, where $\tilde{X}_t$ has orthonormal columns, $R_{x,t}$ is upper triangular;
10: Compute the SVD $R_{x,t} = Q_{x,t}\tilde{Σ}Q_{x,t}^T$, where $\tilde{Σ} = \text{diag}(\hat{σ}_1, \ldots, \hat{σ}_{r_t}), \ Q_{x,t}, Q_{y,t}$ are orthogonal;
11: Set $r_t = r_{t-1} - \{j \mid κ\hat{σ}_j < σ_1\}$;
12: Set $\tilde{X}_t = [\tilde{X}_tQ_{x,t}]_{(1:1:r_t)}$;
13: Solve $Π_{Ω_{t-1}}(\tilde{X}_t\tilde{Y}_t^T) = Π_{Ω_{t-1}}(M)$ for $\tilde{Y}_t$;
14: Compute the QR decomposition $\tilde{Y}_t = \tilde{Y}_tR_{y,t}$, where $\tilde{Y}_t$ has orthonormal columns, $R_{y,t}$ is upper triangular;
15: Compute the SVD $R_{y,t}^T = Q_{y,t}\tilde{Σ}Q_{y,t}^T$, where $\tilde{Σ} = \text{diag}(\hat{σ}_1, \ldots, \hat{σ}_{r_t}), \ Q_{x,t}, Q_{y,t}$ are orthogonal;
16: Set $r_t = r_t - \{j \mid κ\hat{σ}_j < σ_1\}$;
17: Set $X_t = [\tilde{X}_tQ_{x,t}]_{(1:1:r_t)}, \ Y_t = [\tilde{Y}_tQ_{y,t}]_{(1:1:r_t)}, \ Σ_t = \tilde{Σ}_{(1:1:r_t)}$;
18: Compute $R_t = Π_\Omega(M - X_tΣ_tY_t^T)$;
19: Set $S_t = T_t(R_t)$. $Ω_t = Ω \setminus supp(S_t)$;
20: Compute $τ_t = \|Π_\Omega(M - X_tΣ_tY_t^T - S_t)\|_F$;
21: end while

Solving $X_t$ and $Y_t$ On Lines 8 and 13, $\tilde{X}_t$ and $\tilde{Y}_t$ can both be solved row by row or simultaneously. And to obtain one row of $\tilde{X}_t$ or $\tilde{Y}_t$, a small linear system needs to be solved. When the linear system is underdetermined, Algorithm 1 may break down. Therefore, in each row and column, Algorithm 1 requires the number of observed entries (after the removal of the corrupted entries) must be larger than the rank. To be more precise, we need the small linear system to be good conditioned. In general, it is difficult to determine how many rows/columns are needed to ensure the linear system to be good conditioned. Numerically, for a random matrix $A ∈ \mathbb{R}^{m \times r}$ (generated from a standard normal distribution) with $s = O(r) > 2r$ is usually good conditioned. So, we may declare that $O(r) > 2r$ observations in each row and column are sufficient.

In our implementation, the linear systems are solved in the least square sense. One may also choose to minimize $\ell_p$-norm ($p ≥ 0$) of the residual as in (Zeng and So, 2018).

Computational complexity When the number of observations in each row and column is $O(r)$, each linear system can be solved in $O(r^3)$ FLOPS. So, in each iteration, the computational complexity of the linear system solving on Lines 8 and 13 is $O(mn + n^2 r^3)$. The computational complexity of the QR decomposition on lines 9 and 14 is $O((m + n)r^3)$. The computational complexity of the SVD is $O(r^3)$. So, the overall computational complexity of Algorithm 1 is dominated by the linear system solving. When the number of observations in certain row/column is much larger than $r$, we may randomly choose $O(r)$ observations from the row/column, then solve a much smaller linear system of equations. Again, the overall computational complexity in each iteration is $O((m + n)r^3)$.

Also, note that the linear systems on Line 8 and 13 can be solved in parallel. Therefore, Algorithm 1 are suitable for large scale problems.

Remark 1. When $Ω_t$ is fixed, Algorithm 1 essentially minimizes $\|Π_\Omega(XY^T - M)\|_F$ via ADM. If gradient method is used to minimize $\|Π_\Omega(XY^T - M)\|_F$, Algorithm 1 is similar to the GD method in (Yi et al., 2016), except the regularization term $\|U_t^TV_t - V_t^TV_t\|_F$ in the loss function.

3 Convergence

This section analyzes the convergence of Algorithm 1. We first study the full observation case, which serves as a motivation for the partial observation case next.

To present the results, we need to define the $k$ canonical angles. Let $X, Y$ be two $k$-dimensional subspaces of $\mathbb{R}^n$. Let $X, Y ∈ \mathbb{R}^{n \times k}$ be the orthonormal basis matrices of $X$ and $Y$, respectively, i.e.,

$R(X) = X, \quad X^TX = I_k, \quad$ and $\quad R(Y) = Y, \quad Y^TY = I_k.$

Denote $ω_j$ for $1 ≤ j ≤ k$ the singular values of $Y^TX$ in ascending order, i.e., $ω_1 ≤ · · · ≤ ω_k$. The $k$ canonical angles $θ_j(X, Y)$ between $X$ and $Y$ are defined by

$0 ≤ θ_j(X, Y) := \arccos ω_j ≤ \frac{π}{2}$ for $1 ≤ j ≤ k$. (4)

They are in descending order, i.e., $θ_1(X, Y) ≥ · · · ≥ θ_k(X, Y)$. Set

$Θ(X, Y) = \text{diag}(θ_1(X, Y), \ldots, θ_k(X, Y)).$ (5)

In what follows, we sometimes place a vector or matrix in one or both arguments of $θ_j(·, ·)$ and $Θ(·, ·)$ with the meaning that it is about the subspace spanned by the vector or the columns of the matrix argument.
3.1 Full observation case

\textbf{Theorem 1.} Let $M = L_s + S_s \in \mathbb{R}^{m \times n}$ $(m \geq n)$, where $L_s$ is low rank, i.e., $r = \text{rank}(L_s) \ll n$. Let the SVD of $M$ be $M = U \Sigma V^T$, where $U = [U_1 | U_2] = [u_1, \ldots, u_r | u_{r+1}, \ldots, u_m]$, $V = [V_1 | V_2] = [v_1, \ldots, v_r | v_{r+1}, \ldots, v_n]$ are orthogonal matrices, $\Sigma = \begin{bmatrix} \text{diag}(\Sigma_1, \Sigma_2) \\ 0 \end{bmatrix}$, $\Sigma_1 = \text{diag}(\sigma_1, \ldots, \sigma_r)$, $\Sigma_2 = \text{diag}(\sigma_{r+1}, \ldots, \sigma_n)$, and $\sigma_1 \geq \cdots \geq \sigma_n$. Let $M_r = U_1 \Sigma_1 V_1^T$ be the best rank $r$ approximation of $M$. Let the economy sized SVD of $L$ be $L_s = U_s \Sigma_s V_s^T$, where $U_s \in \mathbb{R}^{m \times r}$ and $V_s \in \mathbb{R}^{n \times r}$ both have orthonormal columns, $\Sigma_s = \text{diag}(\sigma_1, \ldots, \sigma_s)$ with $\sigma_1 \geq \cdots \geq \sigma_s > 0$. If

\begin{align}
\| (I - U_s U_s^T) S_s (I - V_s V_s^T) \| &< \sigma_{r+1}, \\
\max \{ \| S_s V_s \|, \| S_s^T U_s \| \} &< \sigma_s - \sigma_{r+1},
\end{align}

(6a) (6b)

then

\begin{align}
\frac{\| L_s - M_r \|_{\text{max}}}{\| L_s \|} \leq (\| U_s \|_{2,\infty} \theta_\nu + \| V_s \|_{2,\infty} \theta_\nu) + (1 + 3 \| U_s \|_{2,\infty} \| V_s \|_{2,\infty}) \theta_\nu \theta_v,
\end{align}

where $\theta_\nu = \| \sin \Theta(U_s, U_s) \|$, $\theta_v = \| \sin \Theta(V_s, V_s) \|$ and $\eta = \frac{\max \{ \| S_s V_s \|, \| S_s^T U_s \| \}}{\sigma_s - \sigma_{r+1} \max \{ \| S_s V_s \|, \| S_s^T U_s \| \}}$.

Theorem 1 tells that when (6) holds and $\eta$ is small, the principal angles between $U_1$ and $U_s$, $V_1$ and $V_s$ will be small, and the best rank-$r$ approximation of $M$ is a good approximation of $L_s$. Notice that, (6) does not necessarily implies $\| S_s \|$ is small (compared with $\| L_s \|$). In fact, we have the following example, in which $\| S_s \|$ is comparable with $\| L_s \|$ and $\eta = 0$.

\textbf{Example 1.} Let $L = L_s + S_s = \frac{1}{n} 1_n 1_n^T$,

\begin{equation}
S = \frac{1}{4} \begin{bmatrix}
2 & -1 & 0 & \cdots & 0 & -1 \\
-1 & 2 & -1 & \cdots & 0 & 0 \\
0 & -1 & 2 & -1 & \cdots & \vdots \\
\vdots & \ddots & \ddots & \ddots & \ddots & \ddots \\
0 & \cdots & 0 & -1 & 2 & -1 \\
-1 & \cdots & 0 & 0 & -1 & 2
\end{bmatrix}, \quad \text{where } 1_n \text{ is an } n\text{-by-1 vector of ones, } n \text{ is even, } \rho \in (-1, 1) \text{ is a real parameter.}
\end{equation}

One can verify that $\| L \| = 1$, $\| S \| = \rho$, $\| S_1 \| = 0$, the first two singular values of $M$ are $\sigma_1 = 1$ and $\sigma_2 = |\rho|$, and the economy sized SVD of $L$ can be given by $L = U_s \Sigma_s V_s^T$, where $U_s = \frac{1}{\sqrt{2}} 1_n$, $\Sigma_s = \sigma_s = 1$. Then $\| (I - U_s U_s^T) S (I - V_s V_s^T) \| = \| S \| = |\rho| < 1 = \sigma_1$, and $\max \{ \| S V_s \|, \| S^T U_s \| \} = 0 < 1 - |\rho| = \sigma_2 - \sigma_1$. In other words, the assumption (6) holds. Noticing that $\eta = \frac{\max \{ \| S V_s \|, \| S^T U_s \| \}}{\sigma_s - \sigma_{r+1} \max \{ \| S V_s \|, \| S^T U_s \| \}} = 0$, by Theorem 1, we can conclude that $\| \sin \Theta(U_1, U_s) \| = \| \sin \Theta(V_1, V_s) \| = 0$, $M_1 = L$, where $U_1, V_1$ are the top left and right singular vectors of $M$, respectively, and $M_1$ is the best rank-1 approximation of $M$.

Now let us assume all entries of $M$ are observed, $S_s$ is sufficiently small, and can be taken as a perturbation to $L_s$. Let $Y_t$ be guess for $Y$ and have orthonormal columns, let us perform one iteration of ADM. For the case of illustration, also assume that $r_{t+1} = r_t$. Then the iteration reads: (S1) $\hat{X}_{t+1} = MY_t$; (S2) $\hat{X}_{t+1} = MY_t G_x$, where $G_x$ is such that $\hat{X}_{t+1}$ is orthonormal; (S3) $\hat{Y}_{t+1} = M^T \hat{X}_{t+1}$; (S4) $\hat{Y}_{t+1} = M^T \hat{X}_{t+1} G_y$, where $G_y$ is such that $\hat{Y}_{t+1}$ is orthonormal. So, we get

\begin{equation}
\hat{Y}_{t+1} = (M^T M) Y_t G_x G_y,
\end{equation}

(7)

which is just one iteration of subspace iteration (a generation of power iteration) for computing the dominant eigenspace of $M^T M$ (e.g., (Demmel, 1997; Stewart, 2001; Van Loan and Golub, 2012)). In fact, (S2) and (S4) are iterations for subspaces spanned by the dominant right and left singular vectors of $M$, respectively. Classic results tell that the subspaces $R(X_t)$ and $R(Y_t)$ converge to the subspaces spanned by the right and left singular vectors of $M$ corresponding with the dominant singular values. When the perturbation is small, $R(X_t)$ and $R(Y_t)$ are good approximations for $R(L_s)$ and $R(L_x^2)$, respectively. In particular, when $S_s$ is small and $V_s^T Y_t$ is nonsingular, we have $\| \sin \Theta(X_{t+1}, U_s) \| = 0$, $\| \sin \Theta(Y_{t+1}, V_s) \| = 0$, i.e., one iteration of ADM gives the true solution.

3.2 Partial observation case

For the partial observation case, we study the convergence of Algorithm 1 under the following assumptions:

(A1) For $L_s$, the column and row incoherence conditions with parameter $\mu$ hold, i.e.,

\begin{equation}
\max_{1 \leq i \leq m} \| U_s^T e_i \| \leq \frac{\mu r}{m}, \quad \max_{1 \leq j \leq n} \| V_s^T e_j \| \leq \frac{\mu r}{n}.
\end{equation}

(A2) $S_s$ has at most $\varrho$-fraction nonzero entries per row and column, i.e.,

\begin{equation}
\| S_{s(i,:)} \| \leq \varrho m, \quad \| S_{s(:,j)} \| \leq \varrho n, \quad \text{for all } i, j.
\end{equation}

(A3) Each entry of $M$ is observed independently with probability $p$.

Besides the notations in Algorithm 1, we also adopt the following notations:

$\theta_{x,t} = \| \sin \Theta(X_t, U_s) \|, \quad \theta_{y,t} = \| \sin \Theta(Y_t, V_s) \|.$

The following theorem tells that the SVD of the partial observed matrix (after the removal of outlier) indeed
gives good approximation for $U_*, V_*$. Furthermore, $X_1$, $Y_1$ satisfy an incoherence condition with parameter $\mu_1$, $\|L_* - X_1\Sigma Y_1^T\|_{\max}$ is bounded.

**Theorem 2.** Assume (A1), (A2), (A3) and $m \geq n$. Let $M = L_* + S_* \in \mathbb{R}^{m \times n}$ with $\text{rank}(L_*) = r$, $S_*$ be obtained as in Algorithm 1. Denote $r'_s = \frac{\|S_* - S_i\|_2}{\|S_* - S_i\|_2}^2$, $\gamma = \frac{2}{1-\delta} \sqrt{\frac{2\rho}{r'_s}}$. If

$$\xi = \frac{6\sqrt{\frac{\alpha}{p'n}}}{\rho} \kappa = \frac{6\sqrt{\alpha}}{\rho \sigma_x}. \text{ Further assume } \mu < n \text{ and that there exists a positive constant } \mu'_1 < n \text{ such that}$$

$$\xi = \frac{6\sqrt{\frac{\alpha}{p'n}}}{\rho} \kappa = \frac{6\sqrt{\alpha}}{\rho \sigma_x}.$$  

then with probability $\geq 1 - 1/m^{10 + \log_2 \alpha}$, it holds that

$$\max\{\theta_{x,1}, \theta_{y,1}\} \leq 3(\xi + \gamma)\mu_1 \kappa,$$

$$\|X_1\|_{2,\infty} \leq \sqrt{\frac{\mu_1 r}{m}}, \quad \|Y_1\|_{2,\infty} \leq \sqrt{\frac{\mu_1 r}{n}},$$

$$\|L_* - X_1\Sigma Y_1^T\|_{\max} \leq \|L_*\| \left(\frac{\mu_1}{m}\theta_{y,1} + \sqrt{\frac{\mu_1}{n}}\theta_{x,1} + \Theta_{x,1}\theta_{y,1}\right) + \mathcal{O}(n^{-3/2}),$$

where $\mu_1 = 2(\mu + \mu'_1)$.

**Remark 2.** When there is no corruption, i.e., $q \equiv 0$, then $\gamma = 0$. Furthermore, when $m = \mathcal{O}(n) \gg 1$, since $p'r \equiv \mathcal{O}(1)\tau$, we know that $\xi = \sqrt{\frac{\alpha}{p'n}}$ is small, the larger $p'$ is, the smaller $\xi$ is. By Theorem 1, $\theta_{x,1}$ and $\theta_{y,1}$ will be small. In other words, the SVD $\frac{1}{p'}\Pi_{\Omega_0}(M - S_0) = X_1\Sigma Y_1^T$, gives good approximation for both $U_*$ and $V_*$, by $X_1$ and $Y_1$, respectively.

The following theorem, which is motivated by (Drineas and Mahoney, 2018, Lemma 55), establish the bridge between the full observation case and the partial observation case. This gives an upper bound for the distance between the least square solutions between the full observation case and the partial observation case.

**Theorem 3.** Let $m \geq n$, and denote

$$X_{\text{opt}} = \arg \min_X \|XY_i^T - (M - S_i)\|,$$

$$\tilde{X}_{\text{opt}} = \arg \min_X \|\Pi_{\Omega_0}(X Y_i^T - (M - S_i))\|.$$  

Assume that $\Omega_i$ can be obtained by sampling each entry of $M$ with probability $q = p(1 - q)$, $\|Y_i\|_{2,\infty} \leq \sqrt{\frac{\mu_1 r}{n}}$ for some $\mu' > 0$, and

$$\inf_{X \in \mathbb{R}^{m \times r}} \frac{\|\Pi_{\Omega_i}(XY_i^T)\|}{\|X\|} \geq \sigma$$

for some constant $\sigma > 0$. Then w.p. $\geq 0.99$, it holds

$$\|\tilde{X}_{\text{opt}} - X_{\text{opt}}\| \leq \left(\frac{2}{3} \log(m + n) + 5\right) \frac{\sqrt{\mu_1 r p'}}{\sqrt{\alpha}} \|R\|_{\max},$$

where $R = (M - S)(I - Y_i Y_i^T)$.

**Remark 3.** The requirement (11) is critical. The parameter $\sigma$ reflects the condition number of the least square problem on line 8 of Algorithm 1. What’s more, the larger $p'$ is, the larger $\sigma$ is (in particular, if $p' = 1$, $\sigma = 1$), the smaller the distance between $\tilde{X}_{\text{opt}}$ and $X_{\text{opt}}$ is, which agrees with our intuition. $R$ is the residual for the full observation case, i.e., $R = X_{\text{opt}}Y_i^T - (M - S_i)$. If the residual is small, the distance between $\tilde{X}_{\text{opt}}$ and $X_{\text{opt}}$ will be small, too.

**Definition 1.** Define $\mu' \equiv \max\{\mu_u, \mu_v\}$, where

$$\mu_u \equiv \sup_{U \in \mathbb{R}^{m \times r}} \frac{\|U\|_{2,\infty}}{\|\sin \Theta(U, U)\|} \|\theta_{x,1}\|,$$

$$\mu_v \equiv \sup_{V \in \mathbb{R}^{n \times r}} \frac{\|V\|_{2,\infty}}{\|\sin \Theta(U, V)\|} \|\theta_{y,1}\|.$$  

By definition of $\mu'$, we know that if $\theta_{x,t} \leq \theta_{x,1}$, $\theta_{y,t} \leq \theta_{y,1}$ for all $t$, then $X_t$, $Y_t$ satisfy the incoherence condition with parameter $\mu'$. Recall Theorem 1, under the assumption of (10), $\theta_{x,1}$ and $\theta_{y,1}$ are quite small (at the order of $\frac{1}{\sqrt{n}}$), then we can show that $\mu' \leq \mu_1$, which implies that $\mu'$ is not large.

The next theorem establishes the convergence rate for the ADM, which is the key in our proof of Theorem 5.

**Theorem 4.** Assume that $\Omega_i$ can be obtained by sampling each entry of $M$ with probability $q = p(1 - q)$, $\|Y_i\|_{2,\infty} \leq \sqrt{\frac{\mu_1 r}{n}}$ for some $\mu' > 0$, (11) and

$$\|L_* - X_1\Sigma Y_1^T\|_{\max} \leq c\|L_*\| \theta_{y,t} \sqrt{\frac{\mu_1}{m}},$$

for some constants $c > 0$. Denote

$$r_s = \inf_i \frac{\|S_i - S_i^2\|_F}{\|S_i - S_i^2\|_F^2}, \quad \zeta = \sqrt{\frac{2\mu_1 r}{mr_s}},$$

$$C_{LS} = \left(\frac{2}{3} \log(m + n) + 5\right) \frac{\sqrt{\mu_1 r p'}}{\sqrt{\alpha}},$$

$$C = C_{LS}(1 + 2c\sqrt{2p'n}) \sqrt{\mu_1},$$

$$\epsilon = \epsilon \kappa, \quad \phi = \frac{8c(\tau + \sqrt{2}\kappa) + \sqrt{2C\kappa^2}}{1 - 2c - C/\sqrt{m}}.$$  

Further assume $\theta_{y,t} \leq \frac{1}{\sqrt{n}}$, then w.p. $\geq 0.99$, $\theta_{x,t+1} \leq \phi \theta_{y,t}$. 

Solving the Robust Matrix Completion Problem via a System of Nonlinear Equations

Solving the Robust Matrix Completion Problem via a System of Nonlinear Equations
Remark 4. The assumption (12) is not a strong requirement as it looks. By Theorem 1, (12) is natural for \( t = 1 \). For general \( t > 1 \), it can be shown that there exists a constant \( c > 0 \) such that (12) holds (see supplementary for details), as long as \( X_t, Y_t \) satisfy the incoherence condition. In general, the constant \( c \) is at the order of \( \mathcal{O}(1) \).

Remark 5. The constant \( r_* \) is the infimum of the stable rank of \( S_t - S_* \). If we take a random matrix, whose entries are i.i.d. drawn from a normal distribution \( N(0, \sigma^2) \), to approximate \( S_t - S_* \). Numerically, we found that \( r_* = \mathcal{O}(n) \). Therefore, \( \epsilon \approx c_0 \sqrt{\sigma_p \mu r} \) is a small number as long as \( \phi \) is small, \( \mu \) and \( \kappa \) are not large. When \( \phi' \) is sufficiently large, \( C = \mathcal{O}(1) \). Consequently, when \( m = \mathcal{O}(n) \) is large, \( \phi' \ll 1 \), which agrees with our conclusion in the full observation case.

Remark 6. Under similar assumptions as in Theorem 4, it can also be shown that \( \theta_{t,t+1} \leq \phi \theta_{x,t+1} \). Then \( \theta_{t,t+1} \leq \phi^2 \theta_{y,t} \). When \( \phi < 1 \), \( \{\theta_{y,t}\}_t \) is a monotonically decreasing sequence. Combining it with the definition of \( \mu' \), we know that \( Y_t \) satisfies the incoherence condition with parameter \( \mu' \). Similarly, \( X_t \) also satisfies the incoherence condition.

Theorem 5. Assume (A1), (A2), (A3) and \( m \geq n \). Assume that \( \Omega_t \) can be obtained by sampling each entry of \( M \) with probability \( p = \rho(1 - \rho) \), \( r_i \equiv r \) and

\[
\inf_{X \in \mathbb{R}^{m \times r}} \| \Pi_{\Omega_t}(XY^T) \|_F \geq \sigma, \quad \inf_{Y \in \mathbb{R}^{n \times r}} \| \Pi_{\Omega_t}(X^TY^T) \|_F \geq \sigma
\]

for some \( \sigma > 0 \). Let \( r_* \), \( \zeta \), \( C \), \( \epsilon \) be the same as in Theorem 4. Then with high probability, it holds that

\[
\| M - S_{t+1} + X_{t+1}X_{t+1}^T \|_F \leq \psi \| M - S_t - X_tX_t^T \|_F
\]

where

\[
\psi = \frac{2\sqrt{2}(\kappa + 2\epsilon \sqrt{\frac{\mu}{m}} + \frac{C_0}{\sqrt{m}})(8\kappa + \sqrt{2}\kappa + \sqrt{2}C_0)}{(1 - 4\epsilon \sqrt{\frac{\mu}{m}})(1 - 2\epsilon - \frac{C_0}{\sqrt{m}})}.
\]

Remark 7. If \( \phi < 1 \), then by Theorem 5, \( \{\| M - S_t - X_tX_t^T \|_F \}_t \) is a monotonically decreasing sequence. And in limit, with high probability, it holds

\[
\lim_{t \to \infty} \| M - S_t - X_tX_t^T \|_F = 0.
\]

Recall the way we determine \( S_t \), we get \( (L_s - X_tX_t^T)_{ij} = 0 \) for any \( (i, j) \notin \text{supp}(S_t) \). Then we can show that

\[
X_tX_t^T \to L_s, \quad S_t \to S_* \quad \text{as} \quad t \to \infty,
\]

i.e., exact recovery is achieved, with high probability.

4 Experiments

We compare Algorithm 1 (NLEQ) with the gradient descent (GD) method (Yi et al., 2016) and the PG-RMC method (Cherapanamjeri et al., 2017). The codes of GD and PG-RMC are obtained from the IISlibrary (Sobral et al., 2016) in Github. \footnote{https://github.com/andrewssobral/lrslibrary/tree/master/algorithms/mc}

4.1 Synthetic Data

We generate the data matrix \( M \in \mathbb{R}^{d \times d} \) as follows. The low rank matrix \( L_s \) is given by \( L_s = U_sV_s^T \), where the entries of \( U_s, V_s \in \mathbb{R}^{d \times \rho} \) are drawn independently from the Gaussian distribution with mean zero, variance \( 1/d \). Each entry of the sparse matrix \( S \) are nonzero with probability \( \rho \), and the nonzero entries of \( S \) are uniformly drawn from \([-\frac{r}{d}, \frac{r}{d}] \). Each entry of \( M = L_s + S \) is observed independently with probability \( p \).

The results are presented in Figure 1. In Figure 1-(a), we compare Algorithm 1 (NLEQ) with the gradient descent (GD) method (Yi et al., 2016) and the PG-RMC method (Cherapanamjeri et al., 2017). The codes of GD and PG-RMC are obtained from the IISlibrary (Sobral et al., 2016) in Github. \footnote{https://github.com/andrewssobral/lrslibrary/tree/master/algorithms/mc}

4.2 Foreground-background separation

The next task is foreground-background separation. By stacking up the vectorized video frames, we get a full data matrix. The static background will form a low rank matrix while the foreground can be taken as
the sparse component. We apply our method NLEQ, and also GD and PG-RMC to two public benchmarks, the Bootstrap and ShoppingMall. Each entry of the data matrix is observed independently w.p. \( p = 0.05 \). As presented in Figure 2, all three methods are able to separate the foreground from the background, and the backgrounds obtained by three methods are similar.

Figure 1: Results on synthetic data. Up to down: (a)-(d). (a) \( d = 1 \times 5 \), \( r = 10 \), \( p = 0.0015 \), \( \rho = 0.1 \). (b) \( d = 1 \times 4 \), \( r = 10 \), \( p = 0.02, 0.1, 0.5 \). (c) \( d = 1 \times 3, 5 \times 3, 1 \times 4, 5 \times 4, 1 \times 5 \), \( r = 10 \), \( p = 0.15 r^2 \log(m)/m \), \( \rho = 0.1 \). (d) \( d = 1 \times 5 \), \( r = 10 \), \( p = 0.002 \), \( \rho = 0.1 \).

Figure 2: Foreground-background separation.

5 Conclusion

In this paper, we study the RMC problem from an algebraic point of view – transform the RMC problem into a problem of solving an overdetermined nonlinear system of equations (with outliers). This method does not require any objective function, convex relaxation or surrogate convex constraint. Algorithmically, we propose to solve the NLEQ via ADM, in which the true rank and support set of the corruption are determined during the iteration. The algorithm is highly parallelizable and suitable for large scale problems. Theoretically, we characterize the sufficient conditions for when \( L^* \) can be approximated by the low rank approximation of \( M \) or \( \frac{1}{p} \Pi_L(M) \). We establish sufficient conditions for \( M_r \approx L_\ast \), where \( M_r \) is the best rank \( r \) approximation of the observed \( M \). The convergence of the algorithm is guaranteed, and exact recovery is achieved under proper assumptions. Numerical simulations show that the algorithm is comparable with state-of-the-art methods in terms of efficiency and accuracy.
References


