Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence (AAAI-16)

Scalable Completion of Nonnegative Matrices with the Separable Structure

Xiyu Yu, Wei Bian, Dacheng Tao
Center for Quantum Computation and Intelligent Systems, University of Technology Sydney
xiyu.yu@student.uts.edu.au
{wei.bian, dacheng.tao}@uts.edu.au

Abstract

Matrix completion is to recover missing/unobserved values of a data matrix from very limited observations. Due to widely potential applications, it has received growing interests in fields from machine learning, data mining, to collaborative filtering and computer vision. To ensure the successful recovery of missing values, most existing matrix completion algorithms utilise the low-rank assumption, i.e., the fully observed data matrix has a low rank, or equivalently the columns of the matrix can be linearly represented by a few numbers of basis vectors. Although such low-rank assumption applies generally in practice, real-world data can process much richer structural information. In this paper, we present a new model for matrix completion, motivated by the separability assumption of nonnegative matrices from the recent literature of matrix factorisations: there exists a set of columns of the matrix such that the resting columns can be represented by their convex combinations. Given the separability property, which holds reasonably for many applications, our model provides a more accurate matrix completion than the low-rank based algorithms. Further, we derives a scalable algorithm to solve our matrix completion model, which utilises a randomised method to select the basis columns under the separability assumption and a coordinate gradient based method to automatically deal with the structural constraints in optimisation. Compared to the state-of-the-art algorithms, the proposed matrix completion model achieves competitive results on both synthetic and real datasets.

Introduction

In many practical problems, people would like to recover missing/observed values of a data matrix from a small subset of observations. For example, in the famous Netflix challenge, one has to predict users' preferences in a huge and sparse matrix according to a very small fraction of ratings. This matrix completion process is critical for many tasks, such as, reconstruction, prediction and classification. Therefore, it has received great interests recently, in a wide range of fields from machine learning, data mining, to collaborative filtering and computer vision (Shamir and Shalev-Shwartz 2011; Xiao and Guo 2014; Huang, Nie, and Huang 2013; Liu et al. 2015; Yao and Kwok 2015).

Matrix completion seems to be an impossible task due to the large amount of unknown variables but limited information. In order to make it tractable, more assumptions should be made on the data matrix. A reasonable and commonly used assumption is that the matrix has a low-rank, i.e., its columns reside in a subspace spanned by several basis vectors. To incorporate the low-rank assumption, directly minimising the rank function is straightforward, but it is NP-hard (Fazel, Hindi, and Boyd 2004). Alternatively, inspired by ℓ_1 norm minimisation for sparse signal recovery, minimising the nuclear norm of the data matrix imposes sparsity on the singular values of the data matrix and thus fulfils the low-rank assumption. Indeed, the nuclear norm is known as the best convex lower bound of the rank function (Fazel 2002). Theoretical results have established for successful matrix completion by using nuclear norm minimisation (Candès and Recht 2009; Recht 2011; Candès and Tao 2010; Bhojanapalli and Jain 2014; Chen et al. 2013; Cai, Candès, and Shen 2010), and even in the cases where the data matrix is contaminated with noise (Candes and Plan 2010; Keshavan, Montanari, and Oh 2009). Another way to incorporate low-rank assumption is to use the matrix production idea. Any matrix $\mathbf{X} \in \mathbb{R}^{m\times n}$ of a low rank r could be factorized into two small matrices, i.e., $\mathbf{X} = \mathbf{UV}$, where $\mathbf{U} \in \mathbb{R}^{m\times r}$ and $\mathbf{V} \in \mathbb{R}^{r\times n}$. Based on this idea, we can directly apply this factorization form into matrix completion model so as to obtain a low-rank matrix (Wen, Yin, and Zhang 2012; Aravkin et al. 2014; Recht and Ré 2013). There are many other works inherently based on low-rank assumption but using some different views, such as, probabilistic model (Todeschini, Caron, and Chavent 2013; Lafond et al. 2014), summation of rank-one matrices (Wang et al. 2014), etc.

However, data from real-world applications can process much richer structural information than the low-rank assumption. In particular, the recent literature of matrix factorizations shown that it is reasonable to assume that there exists a set of columns of the data matrix such that the resting columns can be represented by their convex combinations (Recht et al. 2012; Kumar, Sindhwani, and Kambadur 2013; Gillis, Vavasis, and others 2014; Arora et al. 2012) . Such assumption is called separability and has been utilised in developing efficient algorithms for nonnegative matrix factorization (Benson et al. 2014; Zhou, Bian, and Tao 2013). Mo-

tivated by this, we propose a Nonnegative Matrix Completion model under Separability Assumption (NMCSA).We derives a scalable algorithm to solve our matrix completion model. First, a randomised method is applied to select the basis columns of the data matrix, which overcomes the difficulty of the original combinatorial optimisation over the basis index. We show that under mild regularity conditions, the randomised method is able to successfully select the basis columns with high probabilities. Next, we developed a coordinate gradient based method to optimise data matrix. One advantages of our method is that it automatically deal with these structural constraints of separability. In addition, it has a closed form solution in each iteration, with only linear complexity. These features show the potentials of our algorithm for solving large problems.

We present our model NMCSA and the optimisation algorithm in the next two sections. Then, empirical evaluations on both synthetic and real-world data are report.

Notations: Throughout this paper, boldface uppercase (resp. lowercase) letters, such as $\mathbf{X}$ (resp. $\mathbf{x}$), denote a matrix (resp. vector); letters which are not bold denote scalars. $\mathbf{\Pi}$ is a permutation matrix. For a matrix $\mathbf{X}$, $X_{ij}, \mathbf{X}_{i\cdot}$and $\mathbf{X}_{\cdot j}$ respectively denote the $(i,j)^{\text{th}}$ element, i^{th} row vector and j^{th} column vector; we may also use $\mathbf{X}_i$ to denote a column vector $\mathbf{X}_{\cdot i}$ of a matrix $\mathbf{X}$ whenever it is appropriate. $\|\cdot\|_F$ denotes the Frobenius norm of a matrix; $\|\cdot\|_2$ denotes the ℓ_2 norm of a vector. The set of nonnegative real number is denoted by $\mathbb{R}^+$. For a matrix, $\mathbf{X} \in \mathbb{R}_+^{m\times n}$ (or $\mathbf{X} \geq 0$) indicates that all elements in $\mathbf{X}$ are nonnegative. $[n]$ denotes the set $\{1, 2, \cdots, n\}$, and $|\cdot|$ is the cardinality of a set. If S is a subset of an arbitrary set, then $\bar{S}$ is its complement set.

Matrix Completion Model

Given a nonnegative data matrix $\mathbf{X} \in \mathbb{R}_+^{m\times n}$, each column of which corresponds to a data point in $\mathbb{R}_+^m$, we say $\mathbf{X}$ has the separable property if the columns of $\mathbf{X}$ can be represented by convex combinations of its few columns, which are called basis columns, while the basis columns cannot be represented by the resting columns. Mathematically, we have the following definition for separability.

Definition 1 (Separability)**.** *For nonnegative matrix $\mathbf{X} \in \mathbb{R}_+^{m\times n}$ and a submatrix $\mathbf{X}_S$ composed by its columns with index $S \in [n]$, $\mathbf{X}$ is separable if and only if it resides in a convex hull generated by $\mathbf{X}_S$, i.e.,*

$$\forall i \in [n], \mathbf{X}_i \in \mathbf{conv}(\mathbf{X}_S), \mathbf{X}_S = \{\mathbf{X}_i\}_{i\in S},$$

or equivalently,

$$\mathbf{X\Pi} = \mathbf{X}_S [\mathbf{I} \quad \mathbf{F}] \quad \text{and} \quad \mathbf{F}^T\mathbf{1}_r = \mathbf{1}_{(n-r)},$$

where $r = |S|$ and $\mathbf{F} \in \mathbb{R}_+^{r\times(n-r)}$.

The separability can be interpreted that a rich dataset could be represented by a small subset of its few exemplars therein (Mahoney and Drineas 2009; Yang et al. 2015). Compared to the low-rank assumption, it offers additional advantages: 1) representing dataset using several exemplar data points results in more natural and interpretable models; 2) instead of searching an undetermined model with the basis and weights in a high-dimensional space, directly anchoring representative exemplars much reduces the complexity of learning space. Indeed, a series of recent works have been carried out by utilising the separability to learn succinct representations of large high-dimensional data, especially in the direction of nonnegative matrix factorisations (Benson et al. 2014; Zhou, Bian, and Tao 2013).

However, in many application scenarios, we have only incomplete data with a very small portion of observations, like collaborative filtering, and need to recover the unobserved values from such limited information. In such cases, separability is still helpful as the structural constraints to restrict the freedom of unobserved values and therefore offer chances for successful completion. More importantly, the separable property holds rationally for practical datasets. Taking the MovieLens dataset as an example, the ratings of an ordinary person could be represented by combination of the ratings by typical users. Thus, albeit a large amount of movies unrated by most users, we can still predict these ratings by limited observed information.

Specifically, in matrix completion tasks, we are given a data matrix $\mathbf{X}$ that is only partially observed within a support set Ω. Denote by $\widetilde{\mathbf{X}}$ is incomplete version of $\mathbf{X}$, i.e.,

$$\mathbf{X}_{ij} = \widetilde{\mathbf{X}}_{ij}, \ \forall\, (i,j) \in \Omega,$$

we intend to obtain a full recovery of $\mathbf{X}$ from $\widetilde{\mathbf{X}}$. The following uniform sampling condition on the observed entries of $\mathbf{X}$ is commonly needed for matrix completion (Candès and Recht 2009). It excludes the cases where a few columns of $\mathbf{X}$ are mostly observed while the rest are almost empty. Suppose $\rho = 0.1$, it implies roughly 10% of the entries of $\mathbf{X}$ are observed.

Condition 1 (Uniform sampling)**.** *The incomplete version $\widetilde{\mathbf{X}}$ is generated by sampling the entries of $\mathbf{X}$ uniformly, with Bernoulli distribution $B(1,\rho)$.*

Accordingly, the completion of a separable nonnegative matrix $\mathbf{X}$ from incomplete observation $\widetilde{\mathbf{X}}$ can be formulated as the optimisation below,

$$\begin{aligned} \min_{S,\mathbf{\Pi},\mathbf{X},\mathbf{F}} \ & \frac{1}{2}\|\mathbf{X\Pi} - \mathbf{X}_S [\mathbf{I} \quad \mathbf{F}]\|_F^2 \\ \text{s.t.} \ & \mathbf{X} \in \mathbb{R}_+^{m\times n}, \mathbf{X}_{ij} = \widetilde{\mathbf{X}}_{ij}, \forall (i,j) \in \Omega, \\ & \mathbf{F}^T\mathbf{1}_r = \mathbf{1}_{n-r}, \mathbf{F} \in \mathbb{R}_+^{r\times(n-r)}, \\ & |S| = r. \end{aligned} \tag{1}$$

Remarks. The matrix completion model (1) takes most advantages of separability by using several representative data points to recover missing values. Besides, given completed $\mathbf{X}$, it also gives rise to a unique matrix factorisation $\mathbf{X}_S [\mathbf{I} \quad \mathbf{F}]$, which can be used for further analysis, such data clustering and visualisations.

Optimisation

The separable structure makes the optimisation (1) nontrivial, and inapplicable the existing algorithms for low-rank assumption based matrix completions. The index set S of the

convex basis contains discrete variables that are inherently hard to deal with; the nonnegative constraint on $\mathbf{X}$ and the additional constraint to restrict $\mathbf{F}$ onto an $r-1$ dimensional simplex (column-wise), $\mathcal{S}_F = \{\mathbf{F} : \mathbf{F}^T\mathbf{1}_r = \mathbf{1}_{n-r},\ \mathbf{F} \in \mathbb{R}_+^{r\times(n-r)}\}$ convey further difficulties for the optimisation.

We propose to break down the whole problem into two steps. First, we apply a randomised method to determine the index set S of the convex basis. Although the randomised feature gives the optimal solution in terms of probability, it offers an extreme efficient way to address the discrete optimisation over S. Besides, we provide a theoretical justification to show that the probability of successful selection of S is overwhelming, given mild regularity conditions. Second, we derive scalable methods to optimise $\{\mathbf{X}, \mathbf{F}\}$. The methods are built upon coordinate gradient based optimisation, and thus efficient and scalable to large problems. In addition, they are able to handle the structural constraints over $\{\mathbf{X}, \mathbf{F}\}$ automatically, without any projections onto the feasible set required as in general gradient based methods for constrained optimisations.

Randomised Method for Convex Basis Selection

Our randomised method for selecting the index set S of the convex basis is motivated by the following proposition on the separable nonnegative matrix $\mathbf{X}$. In the following analysis, we can ignore the permutation matrix for simplicity.

Proposition 1 (Projection Property)**.** *Given a separable nonnegative matrix $\mathbf{X} = \mathbf{X}_S\,[\mathbf{I} \quad \mathbf{F}] \in \mathbb{R}^{m\times n}$, where $\mathbf{X}_S \subset \mathbf{X}$ is the convex basis with index set S, and its projection $\mathbf{y} = \mathbf{X}^T\beta$ onto any random direction (vector) $\beta \in \mathbb{R}^m$, if $i = \arg\max_{1\le j\le n} \mathbf{y}_j$, then it must hold that $j \in S$.*

Actually, Proposition 1 is the corollary of the much general facts that the projection of a high-dimensional convex set into a low-dimensional subspace is still a convex set and any vertex of the latter must correspond to a vertex of the original. Most recently, quite a few works have been done (Zhou, Bian, and Tao 2013), by utilizing Proposition 1 to design efficient algorithms for separable nonnegative matrix factorisations, as well as other related problems such as learning topic models and latent variable models.

Our study further extends this direction of research to matrix completions. Note that in the completion task, matrix $\mathbf{X}$ is only partially observed (with generally very few entries) on the support set Ω. *Such sparse character could lead to a considerably poor projection property, compared to that stated by Proposition 1 for a fully observed matrix.* The key reason is that for $\mathbf{X}$ with missing values, specially when such missing-value patterns for rows of $\mathbf{X}$ are distinct and random, a dense random projection (i.e., with a dense random vector $\beta \in \mathbf{R}^m$) will partially malfunction and thus unable to capture the separable structure of $\mathbf{X}$. Considering this, we prefer a sparse random projection and in the extreme case, the projection onto a randomly chosen standard base vector $\mathbf{e}_j$ of $\mathbb{R}^m$. And the following regularity conditions are needed for the success of our randomised method.

Condition 2 (Minimal Probability for Basis Selection)**.** *For separable nonnegative matrix $\mathbf{X} = \mathbf{X}_S\,[\mathbf{I} \quad \mathbf{F}] \in \mathbb{R}^{m\times n}$, denoting by $p_i^* = \Pr(i = \arg\max_{j\in[n]} \mathbf{X}^T\mathbf{e} \text{ and } i \in S)$, where $\mathbf{e}$ is a randomly chosen standard basis vector of $\mathbb{R}^m$, it holds $\sum_{i\in S} p_i^* = 1$. Assume $\min_{i\in S} p_i^* \ge \frac{1-\rho}{\rho(n-|S|)} + \gamma$, for some $\gamma > 0$.*

Condition 3 (Data Generation)**.** *For separable nonnegative matrix $\mathbf{X} = \mathbf{X}_S\,[\mathbf{I} \quad \mathbf{F}] \in \mathbb{R}^{m\times n}$, the columns $\{\mathbf{F}_j, j \in \bar{S}\}$ are i.i.d. instances sampled from some distribution $p(\mathbf{f})$ over $\{\mathbf{f} : \mathbf{f}^T\mathbf{1}_r = 1, \mathbf{f} \in \mathbb{R}_+^r\}$.*

Note that both conditions are mild and to be hold in general. In particular, the probability requirement in Condition 2 is easy to be satisfied given large enough n, and Condition 3 nearly imposes no harsh restrictions on $\mathbf{X}$. However, the necessity of the two conditions can be interpreted as below: Condition 2 guarantees that even with incomplete matrix $\widetilde{\mathbf{X}}$, any column $\{\widetilde{\mathbf{X}}_j, j \in S\}$ has a better chance to be selected, while Condition 3 ensures that none of the columns $\{\widetilde{\mathbf{X}}_j, j \in \bar{S}\}$ has an absolutely large probability to be selected, so as to malfunction our randomised method. With such conditions, we have Proposition 2 for the (random) projection property for a separable nonnegative matrix with missing entries.

Proposition 2 (Projection Property with Missing Entries)**.** *For separable nonnegative matrix $\mathbf{X} = \mathbf{X}_S\,[\mathbf{I} \quad \mathbf{F}] \in \mathbb{R}^{m\times n}$ satisfying Conditions 2 and 3, let $\widetilde{\mathbf{X}}$ be its incomplete version with observation probability ρ. Then, given the projection $\mathbf{y} = \widetilde{\mathbf{X}}^T\mathbf{e}$ onto a (uniformly selected) random standard direction of $\mathbb{R}^m$, if $i' = \arg\max_{1\le j'\le n} \mathbf{y}_{j'}$, it holds that $\Pr(i' = i; i \in S) \ge \rho p_i^*$, and $\Pr(i' = j; \forall j \in \bar{S}) \le (1-\rho)/(n-|S|)$.*

Now, we are ready to present the basis selection algorithm and main theorem for the identifiability of the basis of an incomplete separable nonnegative matrix $\widetilde{\mathbf{X}}$ by random projections.

Theorem 1 (Basis Selection by Random Projections)**.** *Given incomplete version $\widetilde{\mathbf{X}}$ of a separable nonnegative matrix $\mathbf{X} = \mathbf{X}_S\,[\mathbf{I} \quad \mathbf{F}] \in \mathbb{R}^{m\times n}$, with observation probability ρ, let $\boldsymbol{\pi}$ be the statistics defined via T times projections on randomly selected standard basis vector $\mathbf{e}$ of $\mathbb{R}^m$, i.e.,*

$$\boldsymbol{\pi}_i = \frac{1}{T}\sum_{t=1}^{T} 1\big(i = \arg\max_{j\in[n]} \widetilde{\mathbf{X}}_j^T\mathbf{e}\big) \tag{2}$$

It holds that

$$\min_{j\in S} \boldsymbol{\pi}_j > \max_{j\in\bar{S}} \boldsymbol{\pi}_j, \tag{3}$$

with probability at least $1 - |S|(n-|S|)\exp(-T\gamma^2\rho^2/16)$.

By applying Theorem 1, we have the following algorithm for our randomised method to select the basis set S given an incomplete separable nonnegative matrix $\widetilde{\mathbf{X}}$.

Scalable Optimisation for F and X

Given the index set S of the convex basis, the problem of matrix completion (1) reduces to

$$\begin{aligned} \min_{\mathbf{Y},\mathbf{Z},\mathbf{F}} \quad & \tfrac{1}{2}\|\mathbf{Y} - \mathbf{Z}\mathbf{F}\|_F^2 \\ \text{subject to} \quad & \mathbf{X} \in \mathbb{R}_+^{m\times n}, X_{ij} = \widetilde{X}_{ij}, \forall (i,j) \in \Omega \\ & \mathbf{F}^T\mathbf{1}_r = \mathbf{1}_{n-r}, \mathbf{F} \in \mathbb{R}_+^{r\times(n-r)}. \end{aligned} \tag{4}$$

Algorithm 1 Basis Selection by Random Projections

Input: $\widetilde{\mathbf{X}}, r, T$
Output: S

```
Initialization: I = ∅
for k = 1 : T do
    randomly select standard basis vector e
    i = arg max_{j∈[n]} X̃_j^T e
    I = I ∪ {i}
end for
S ← r unique elements of I with largest occurrences
```

where $\mathbf{Y} = \mathbf{X}_{\bar{S}}, \mathbf{Z} = \mathbf{X}_S, \mathbf{X} = [\mathbf{Y} \quad \mathbf{Z}]\,\mathbf{\Pi}$.

Clearly, this is a constrained minimisation over triplet $\{\mathbf{F}, \mathbf{Y}, \mathbf{Z}\}$ and can be solved by alternating optimisation methods. However, as the problem size (m, n) scales up quickly, standard off-the-shelf algorithms can be considerably inefficient on solving (4). This can be understood by the fact that least squares with the nonnegative and/or the simplex constraints are nontrivial even with a moderate problem size. Therefore, we intend to derive a scalable algorithm that solves (4) in a most probably efficient way.

Updating F: When $(\mathbf{Z}, \mathbf{Y})$ are fixed, the optimisation over $\mathbf{F}$ reads

$$\begin{aligned} \min_{\mathbf{F}} \quad & \tfrac{1}{2}\|\mathbf{Y} - \mathbf{Z}\mathbf{F}\|_F^2 \\ \text{subject to} \quad & \mathbf{F}^T\mathbf{1}_r = \mathbf{1}_{n-r}, \mathbf{F} \in \mathbb{R}_+^{r\times(n-r)} \end{aligned} \tag{5}$$

This is a least square problem with a feasible set defined by the $r-1$ dimensional simplex (column-wise), $\mathcal{S}_F = \{\mathbf{F} : \mathbf{F}^T\mathbf{1}_r = \mathbf{1}_{n-r},\ \mathbf{F} \in \mathbb{R}_+^{r\times(n-r)}\}$. Again, the gradient descent method combined with projection onto $\mathcal{S}_F$ will not work preferably. Following a similar strategy of solving (9), we intend to optimise the rows of $\mathbf{F}$ via the coordinate gradient descent method. Unfortunately, the constraint $\mathbf{F}^T\mathbf{1}_r = \mathbf{1}_{n-r}$ makes no freedom for an individual row of $\mathbf{F}$ given the rest. We overcome this problem by using a randomised coordinate optimisation strategy, which randomly selects two rows of $\mathbf{F}$ to optimise jointly. For any two rows $\mathbf{F}_{i\cdot}$ and $\mathbf{F}_{j\cdot}$, letting $\mathbf{E} = \mathbf{Y} - \mathbf{Z}_{\cdot\overline{ij}}\mathbf{F}_{\overline{ij}\cdot}$, where $\mathbf{Z}_{\cdot\overline{ij}}$ is the submatrix of $\mathbf{Z}$ excluding the i,j-th columns, and analogically $\mathbf{F}_{\overline{ij}\cdot}$. We have the following minimisation

$$\begin{aligned} \min_{\mathbf{F}_{i\cdot},\mathbf{F}_{j\cdot}} \quad & \tfrac{1}{2}\|\mathbf{E} - \mathbf{Z}_{\cdot i}\mathbf{F}_{i\cdot} - \mathbf{Z}_{\cdot j}\mathbf{F}_{j\cdot}\|_F^2 \\ \text{subject to} \quad & \mathbf{F}_{i\cdot} + \mathbf{F}_{j\cdot} = \mathbf{f}, \\ & \mathbf{F}_{i\cdot} \geq 0, \mathbf{F}_{j\cdot} \geq 0. \end{aligned} \tag{6}$$

where $\mathbf{f} = \mathbf{1} - \Sigma_{k\neq i,j}\mathbf{F}_{k\cdot}$.The optimal solution of (6) is given by

$$\begin{aligned} \mathbf{F}_{j\cdot} &= \Big[\min\Big\{\frac{(\mathbf{Z}_{\cdot j} - \mathbf{Z}_{\cdot i})^T(\mathbf{E} - \mathbf{Z}_{\cdot i}\mathbf{f})}{\|\mathbf{Z}_{\cdot j} - \mathbf{Z}_{\cdot i}\|^2}, \mathbf{f}\Big\}\Big]_+ \\ \mathbf{F}_{i\cdot} &= \mathbf{f} - \mathbf{F}_{j\cdot} \end{aligned} \tag{7}$$

Updating Y: Firstly, the updating of $\mathbf{Y}$ is trivial given $(\mathbf{Z}, \mathbf{F})$. We can fill the missing/unobserved entries of $\mathbf{Y}$ using corresponding entries in $\mathbf{ZF}$, i.e.,

$$\mathbf{Y}(\bar{\Omega}) = \{\mathbf{ZF}\}(\bar{\Omega}). \tag{8}$$

where $\bar{\Omega}$ is the complement of support set of $\mathbf{Y}$.

Updating Z: When $(\mathbf{Y}, \mathbf{F})$ are fixed, the updating of $\mathbf{Z}$ can be achieved by solving the following minimisation

$$\begin{aligned} \min_{\mathbf{Z}} \quad & \tfrac{1}{2}\|\mathbf{Y} - \mathbf{Z}\mathbf{F}\|_F^2 \\ \text{subject to} \quad & \mathbf{X} \in \mathbb{R}_+^{m\times n}, X_{ij} = \widetilde{X}_{ij}, \forall (i,j) \in \Omega. \end{aligned} \tag{9}$$

Note that this is basically a nonnegative least squares problem, which can be solved by the standard gradient descent plus projection onto the feasible set method. Albeit the theoretical guarantees for convergence, such method is considerably inefficient, as the projection step does not favour at all the decreasing of the objective function. Therefore, we propose to solve (9) by a coordinate gradient descent method that simultaneously deals with the nonnegative constraint. Specifically, we optimise each of the columns of $\mathbf{Z}$ in a sequential way. For the t-th column of $\mathbf{Z}$, we have the following minimisation,

$$\begin{aligned} \min_{\mathbf{Z}_{\cdot t}} \quad & \tfrac{1}{2}\|\mathbf{Y} - \mathbf{Z}_{\cdot\bar{t}}\mathbf{F}_{\bar{t}\cdot} - \mathbf{Z}_{\cdot t}\mathbf{F}_{t\cdot}\|_F^2 \\ \text{subject to} \quad & \mathbf{Z}_{\cdot t} \in \mathbb{R}_+^m, X_{ij} = \widetilde{X}_{ij}, \forall (i,j) \in \Omega. \end{aligned} \tag{10}$$

where $\mathbf{Z}_{\cdot\bar{t}}$ is the submatrix of $\mathbf{Z}$ excluding the t-th column, and analogically $\mathbf{F}_{\bar{t}\cdot}$. Further, due to the equality constraint in support set Ω, we only need to optimise the corresponding entries of $\mathbf{Z}_{\cdot t}$ in $\bar{\Omega}_t$. Let $\mathcal{I}$ be the index set of unconstrained entries of $\mathbf{Z}_{\cdot t}$, i.e., $\bar{\Omega}_t$, and $\mathbf{z} = \mathbf{Z}_{\cdot t}(\mathcal{I})$ and $\mathbf{A} = \mathbf{Y}(\mathcal{I}, :) - \mathbf{Z}_{\cdot\bar{t}}(\mathcal{I}, :)\mathbf{F}_{\bar{t}\cdot}$. Then, the optimisation of $\mathbf{z}$ is given by

$$\min_{\mathbf{z}\geq 0} \ \frac{1}{2}\|\mathbf{A} - \mathbf{z}\mathbf{F}_{t\cdot}\|_F^2 \tag{11}$$

which enjoys a close-form optimal solution

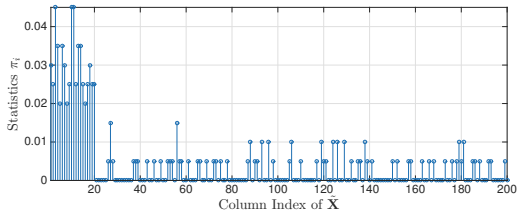
$$\mathbf{z}^* = \max\{\mathbf{A}\mathbf{F}_{t\cdot}^T/\|\mathbf{F}_{t\cdot}\|_2^2, 0\} \tag{12}$$

Given the updating rules of each element of $\{\mathbf{F}, \mathbf{Y}, \mathbf{Z}\}$, problem (4) could be alternatively optimised by Algorithm 2. In Algorithm 2, convergence of each subproblem in each iteration is time-consuming and not necessary. We need only an improvement on the optimisation result and a decrease of the objective function by our coordinate gradient based method. This can also ensure the final convergence of the whole optimisation process.

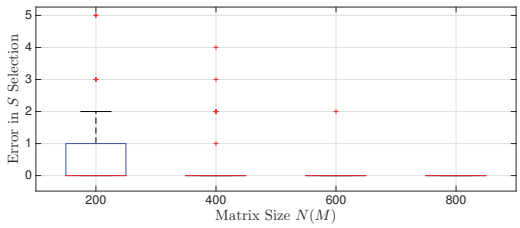
Algorithm 2 Scalable Optimisation for $\mathbf{F}$ and $\mathbf{X}$

Input: $\widetilde{\mathbf{X}}, S$
Output: $\mathbf{X}, \mathbf{F}$

```
Initialization: Y = X̃_S̄, Z = X̃_S
while unconvergence do
    update two randomly selected rows of F by (7)
    update Y by (8)
    for k = 1 : r do
        update k-th column of Z by (12)
    end for
end while
X_S̄ = Y, X_S = Z, X = [Y  Z] Π
```



(a) An example of basis selection



(b) Basis selection errors

Figure 1: Properties of Randomised Method for Selection of Basis Columns.

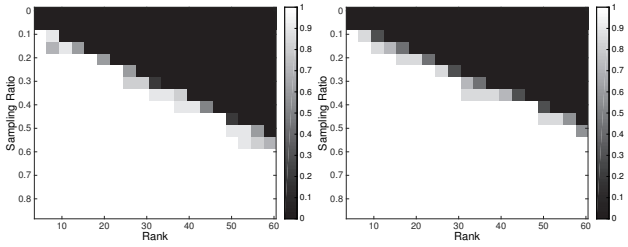
Empirical Evaluations

To demonstrate the effectiveness of NMCSA for completing matrices with separable structures, we conduct empirical valuations both on synthetic and real datasets. We also compare NMCSA with state-of-the-art methods for low-rank based matrix completions, including APGL (Toh and Yun 2010) and LMaFit (Wen, Yin, and Zhang 2012).

Synthetic Data Experiments

In this section, we use synthetic data to verify the effectiveness of randomised method to find convex basis and evaluate the recovery performance of the proposed algorithm. The synthetic data matrices in these experiments are generated in the form $\mathbf{X} = \mathbf{X}_S \begin{bmatrix} \mathbf{I} & \mathbf{F} \end{bmatrix} \mathbf{\Pi}$. The entries of $\mathbf{X}_S$ and $\mathbf{F}$ are generated by i.i.d. uniform distribution in [0,1] at first, and then their column are normalized to have unit l_1 norm. In this case, the data matrix automatically has normalized rows.

Basis Selection In the first experiment, we try to justify the validity of the randomised method for basis selection, i.e., with high probabilities, the method is able to find the convex basis of a separable nonnegative matrix. We randomly generate an incomplete matrix of size 200×200, with parameters rank $r = 20$ and sampling ratio $\rho = 0.15$. We count the frequency of each column being identified as basis columns in all projections and check whether the true basis vectors occupy with highest frequencies. Figure 1(a) shows an example from one experiment. We can see that the basis columns, which correspond to the first 20 columns of the matrix, have the highest frequencies, and thus the randomised method success in selecting the basis columns. Further, we increase the size N and M of the matrix from 200 to 800, and rerun the experiment 50 times for different size settings. Figure 1(b) shows the statistics of errors for basis selection. We can see that for size $N(M) = 200$, on most of the experiments, the randomised method correctly selects the basis, and only on two experiments, it gives an error larger than 1. However, as the matrix size increases, the errors reduce significantly. For example, for $N(M) = 600$, only on 1 out of the 50 experiments, the randomised method gives error 1, while for $N(M) = 800$, it successes on all experiments.



(a) LMaFit (b) NMCSA

Figure 2: Phase Transistion diagrams for matrix completion recoverability.

Phase Transition Analysis A benchmark method to evaluate the performance of a matrix completion algorithm is the phase transition analysis (Candès and Recht 2009; Wen, Yin, and Zhang 2012). In the second experiment, by using a single phase diagram, we can test the influence of sampling ratio and rank individually or simultaneously on the recovery results. Here, we fix the size of matrices to be 500×500, and vary the rank and sampling ratio in the following ranges, i.e., $r \in \{5, 8, 11, \cdots, 59\}$ and $\rho \in \{0.01, 0.06, \cdots, 0.86\}$, according to (Wen, Yin, and Zhang 2012). Then, 50 independent matrix completion experiments are performed for each pair (r, ρ). We declare that a matrix is successfully recovered if the relative error $\frac{\|\mathbf{X} - \mathbf{X}_{opt}\|_F}{\|\mathbf{X}\|_F}$ is less than 1e-3, where $\mathbf{X}$ is the ground truth and $\mathbf{X}_{opt}$ the result of completion. The experimental results are shown in Figure 2. Each color cell of the phase diagrams corresponds to the recovery rate for each pair (r, ρ). White means perfect recovery of all matrices of 50 experiments while black means all failed. As shown in this figure, compared to LMaFit (Wen, Yin, and Zhang 2012), NMCSA has a better capability of recovering separable matrices with a wider range of ranks and sampling ratios.

On Large Problems Next, we evaluate the performance of the proposed NMCSA on larger matrix completion problems, and compare it with the state-of-the-art algorithms, LMaFit (Wen, Yin, and Zhang 2012) and APGL (Toh and Yun 2010). Following the same experimental settings (Cai, Candès, and Shen 2010), we fix the sampling ratio $\rho = 0.2$, and vary the matrix size $N(M)$ from 1000, to 5000 and 10000, and the rank from 10, to 20 and 50. The parameters for different algorithms are set as below: for APGL, $tol = 10^{-4}, \mu = 10^{-4}, truncation = 1$, and $truncation\_gap = 100$; for LMaFit, $est\_rank = 1, tol = 10^{-4}$, and K be $\lfloor 1.25r \rfloor$ or $\lfloor 1.5r \rfloor$ (Wen, Yin, and Zhang 2012). All the experiments are performed in Matlab on a desktop computer. Table 1 shows the results for performance comparison. One

can see that NMCSA outperforms its competitors LMaFit and APGL consistently on all experiments.

Table 1: Recovery Error for Large Problems.

Incomplete Matrix		Computational Results		
Size $N(M)$	**Rank** r	**APGL**	**LMaFit**	**NMCSA**
1000	10	3.645e-04	3.149e-04	2.649e-04
	20	6.496e-04	8.208e-04	6.718e-04
	50	1.245e-01	1.567e-01	3.512e-02
5000	10	3.922e-04	2.154e-04	9.989e-05
	20	3.477e-04	3.139e-04	1.804e-04
	50	4.023e-03	3.330e-03	3.184e-04
10000	10	1.599e-04	1.792e-04	9.998e-05
	20	1.852e-04	2.247e-04	1.226e-04
	50	4.321e-04	4.615e-04	2.518e-04

Table 2: Recommendation Dataset Statistics.

Dataset	Dim	Sampling ratio
Jester-1	24983×101	0.7247
Jester-2	23500×101	0.7272
Jester-3	24938×101	0.2474
Jester-all	73421×101	0.5634
MovieLens-100K	943×1682	0.0630
MovieLens-1M	6040×3706	0.0447
MovieLens-10M	71567×10677	0.0131

Real Data Experiments

We further evaluate NMCSA on two real datasets, Jester[1] and MovieLens, for collaborative filtering. Both datasets are benchmarks and have been commonly used in the literature for matrix completions. It has been noticed that completing the missing/unobserved entries of these datasets are considerably challenging, because of the very large problem scales and the relatively low sampling ratio. Table 2 summaries the basic information of the two datasets.

As no test data are available in these datasets, a common choice is to sample the available ratings by 50% for training and use the resting 50% for test (Wang et al. 2014; Aravkin et al. 2014). To evaluate the performance of completion, we use two measures, the Normalized Mean Absolute Error (NMAE) and the Root Mean Square Error (RMSE), calculated on the support set Ω,

$$\text{NMAE} = \frac{\sum_{(i,j)\in\Omega} |X_{ij} - M_{ij}|}{(r_{max} - r_{min})|\Omega|}$$

and

$$\text{RMSE} = \sqrt{\frac{\Sigma_{(i,j)\in\Omega}(X_{ij} - M_{ij})^2}{|\Omega|}}$$

where r_{max} and r_{min} are the lower and upper bounds for the ratings. For Jester dataset, $r_{max} = 20$, and $r_{min} = 0$; for MovieLens, $r_{max} = 5$, and$r_{min} = 1$. $|\Omega|$ is the cardinality of the support set.

The parameters for the used matrix completion methods are set as below (Wen, Yin, and Zhang 2012): for APGL, $tol = 10^{-4}, \mu = 10^{-4}, truncation = 1$, and $truncation\_gap = 20$, while for LMaFit, $est\_rank = 2, tol = 10^{-3}, K = 1$, and $rk\_inc = 2$. Table 3 and 4 show the experimental results on the Jester and the MovieLens datasets. Note that our method NMCSA outperforms consistently its competitor, APGL and LMaFit, on both datasets. We believe this is because the separable structure offers more information than the low-rank assumption for the recovery of missing/unobserved data in these two datasets.

Table 3: Performance of Matrix Completion on the Jester Dataset (NMSE/RMSE).

	APGL	**LMaFit**	**NMCSA**
Jester-1	0.0919/3.3708	0.1149/3.9858	0.0874/3.1797
Jester-2	0.0921/3.3948	0.1133/3.7868	0.0899/3.2931
Jester-3	0.0969/3.4945	0.1157/3.8446	0.0928/3.7397
Jester-all	0.1568/4.388	0.1151/3.9750	0.0900/3.3115

Table 4: Performance of Matrix Completion on the MovieLens Dataset (NMSE/RMSE).

	APGL	**LMaFit**	**NMCSA**
100K	0.1204/0.8707	0.1504/1.0949	0.1011/0.7493
1M	0.1415/0.9615	0.1479/0.9850	0.0973/0.9541
10M	0.1245/0.8581	0.1355/0.8338	0.0986/0.9632

Conclusions

In this paper, we have proposed a novel matrix completion model for recovering matrices with the separable structure. By using the separability rather than the low-rank assumption, our model exploits richer structural information of real-world data, and achieves better matrix completion results when the separability applies. A scalable algorithm is derived to optimise the proposed model. We use a randomised method to select basis columns of the data matrix and a coordinate gradient based method to automatically deal with the structural constraints from the separability. On both synthetic and real-world datasets, our model achieves competitive performances compared to the state-of-the-art matrix completion methods.

[1]The ratings of Jester dataset are from -10 to 10, which is not nonnegative. In the experiments, we make a shift by adding 10 to each entry of the data matrix. Note that such manipulation does not affect the geometric structure of the dataset, and thus the separability should still hold.

Acknowledgments

This research is supported by Australian Research Council Projects (No: FT-130101457 & No: DP-140102164); and the Chancellors Postdoctoral Research Fellowship of the University of Technology Sydney.

References

Aravkin, A.; Kumar, R.; Mansour, H.; Recht, B.; and Herrmann, F. J. 2014. Fast methods for denoising matrix completion formulations, with applications to robust seismic data interpolation. *SIAM Journal on Scientific Computing* 36(5):S237–S266.

Arora, S.; Ge, R.; Kannan, R.; and Moitra, A. 2012. Computing a nonnegative matrix factorization–provably. In *Proceedings of the forty-fourth annual ACM symposium on Theory of computing*, 145–162. ACM.

Benson, A. R.; Lee, J. D.; Rajwa, B.; and Gleich, D. F. 2014. Scalable methods for nonnegative matrix factorizations of near-separable tall-and-skinny matrices. In *Advances in Neural Information Processing Systems*, 945–953.

Bhojanapalli, S., and Jain, P. 2014. Universal matrix completion. In *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*, 1881–1889.

Cai, J.-F.; Candès, E. J.; and Shen, Z. 2010. A singular value thresholding algorithm for matrix completion. *SIAM Journal on Optimization* 20(4):1956–1982.

Candes, E. J., and Plan, Y. 2010. Matrix completion with noise. *Proceedings of the IEEE* 98(6):925–936.

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

Candès, E. J., and Tao, T. 2010. The power of convex relaxation: Near-optimal matrix completion. *Information Theory, IEEE Transactions on* 56(5):2053–2080.

Chen, Y.; Bhojanapalli, S.; Sanghavi, S.; and Ward, R. 2013. Coherent matrix completion. *arXiv preprint arXiv:1306.2979*.

Fazel, M.; Hindi, H.; and Boyd, S. 2004. Rank minimization and applications in system theory. In *American Control Conference, 2004*, volume 4, 3273–3278. IEEE.

Fazel, M. 2002. *Matrix rank minimization with applications*. Ph.D. Dissertation, Stanford University.

Gillis, N.; Vavasis, S.; et al. 2014. Fast and robust recursive algorithms for separable nonnegative matrix factorization. *Pattern Analysis and Machine Intelligence, IEEE Transactions on* 36(4):698–714.

Huang, J.; Nie, F.; and Huang, H. 2013. Robust discrete matrix completion. In *Twenty-Seventh AAAI Conference on Artificial Intelligence*.

Keshavan, R.; Montanari, A.; and Oh, S. 2009. Matrix completion from noisy entries. In *Advances in Neural Information Processing Systems*, 952–960.

Kumar, A.; Sindhwani, V.; and Kambadur, P. 2013. Fast conical hull algorithms for near-separable non-negative matrix factorization. In *Proceedings of The 30th International Conference on Machine Learning*, 231–239.

Lafond, J.; Klopp, O.; Moulines, E.; and Salmon, J. 2014. Probabilistic low-rank matrix completion on finite alphabets. In *Advances in Neural Information Processing Systems*, 1727–1735.

Liu, M.; Luo, Y.; Tao, D.; Xu, C.; and Wen, Y. 2015. Low-rank multi-view learning in matrix completion for multi-label image classification. In *Twenty-Ninth AAAI Conference on Artificial Intelligence*.

Mahoney, M. W., and Drineas, P. 2009. Cur matrix decompositions for improved data analysis. *Proceedings of the National Academy of Sciences* 106(3):697–702.

Recht, B., and Ré, C. 2013. Parallel stochastic gradient algorithms for large-scale matrix completion. *Mathematical Programming Computation* 5(2):201–226.

Recht, B.; Re, C.; Tropp, J.; and Bittorf, V. 2012. Factoring nonnegative matrices with linear programs. In *Advances in Neural Information Processing Systems*, 1214–1222.

Recht, B. 2011. A simpler approach to matrix completion. *The Journal of Machine Learning Research* 12:3413–3430.

Shamir, O., and Shalev-Shwartz, S. 2011. Collaborative filtering with the trace norm: Learning, bounding, and transducing. In *COLT*, 661–678.

Todeschini, A.; Caron, F.; and Chavent, M. 2013. Probabilistic low-rank matrix completion with adaptive spectral regularization algorithms. In *Advances in Neural Information Processing Systems*, 845–853.

Toh, K.-C., and Yun, S. 2010. An accelerated proximal gradient algorithm for nuclear norm regularized linear least squares problems. *Pacific Journal of Optimization* 6(615-640):15.

Wang, Z.; Lai, M.-J.; Lu, Z.; Fan, W.; Davulcu, H.; and Ye, J. 2014. Rank-one matrix pursuit for matrix completion. In *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*, 91–99.

Wen, Z.; Yin, W.; and Zhang, Y. 2012. Solving a low-rank factorization model for matrix completion by a nonlinear successive over-relaxation algorithm. *Mathematical Programming Computation* 4(4):333–361.

Xiao, M., and Guo, Y. 2014. Semi-supervised matrix completion for cross-lingual text classification. In *Twenty-Eighth AAAI Conference on Artificial Intelligence*.

Yang, T.; Zhang, L.; Jin, R.; and Zhu, S. 2015. An explicit sampling dependent spectral error bound for column subset selection. In *Proceedings of The 32th International Conference on Machine Learning (ICML-15)*, 135–143.

Yao, Q., and Kwok, J. T. 2015. Colorization by patch-based local low-rank matrix completion. In *Twenty-Ninth AAAI Conference on Artificial Intelligence*.

Zhou, T.; Bian, W.; and Tao, D. 2013. Divide-and-conquer anchoring for near-separable nonnegative matrix factorization and completion in high dimensions. In *Data Mining (ICDM), 2013 IEEE 13th International Conference on*, 917–926. IEEE.