CUR Algorithm for Partially Observed Matrices

Miao Xu XUM@LAMDA.NJU.EDU.CN
National Key Laboratory for Novel Software Technology, Nanjing University
Collaborative Innovation Center of Novel Software Technology and Industrialization, Nanjing, China

Rong Jin RONGJIN@CSE.MSU.EDU
Institute of Data Science and Technologies at Alibaba Group, Seattle, USA

Zhi-Hua Zhou ZHOUZH@LAMDA.NJU.EDU.CN
National Key Laboratory for Novel Software Technology, Nanjing University
Collaborative Innovation Center of Novel Software Technology and Industrialization, Nanjing, China

Abstract

CUR matrix decomposition computes the low rank approximation of a given matrix by using the actual rows and columns of the matrix. It has been a very useful tool for handling large matrices. One limitation with the existing algorithms for CUR matrix decomposition is that they cannot deal with entries in a *partially observed* matrix, while incomplete matrices are found in many real world applications. In this work, we alleviate this limitation by developing a CUR decomposition algorithm for partially observed matrices. In particular, the proposed algorithm computes the low rank approximation of the target matrix based on (i) the randomly sampled rows and columns, and (ii) a subset of observed entries that are randomly sampled from the matrix. Our analysis shows the error bound, measured by spectral norm, for the proposed algorithm when the target matrix is of full rank. We also show that only $O(nr \ln r)$ observed entries are needed by the proposed algorithm to perfectly recover a rank r matrix of size $n \times n$, which improves the sample complexity of the existing algorithms for matrix completion. Empirical studies on both synthetic and real-world datasets verify our theoretical claims and demonstrate the effectiveness of the proposed algorithm.

Proceedings of the 32^{nd} International Conference on Machine Learning, Lille, France, 2015. JMLR: W&CP volume 37.

1. Introduction

In many machine learning applications, it is convenient to represent the data by matrix. Examples include user-item rating matrix in recommender system (Srebro et al., 2004), gene expression matrix in bioinformatics (Mahoney & Drineas, 2008), kernel matrix in kernel learning (Williams & Seeger, 2000), document-term matrix in document retrieval (Mahoney & Drineas, 2008), and instance-label matrix in multi-label learning (Goldberg et al., 2010). An effective approach for handling big matrices is to approximate them by their low rank counterparts which can be computed and stored efficiently. Various methods have been developed for low rank matrix approximation, including truncated singular value decomposition, matrix factorization (Srebro et al., 2004), matrix regression (Koltchinskii, 2011), column subset selection (Boutsidis et al., 2011), the Nyström method (Williams & Seeger, 2000), and random SVD techniques (Halko et al., 2011; Woodruff, 2014).

In this work, we will focus on the CUR algorithm for low rank matrix approximation (Mahoney & Drineas, 2009; Boutsidis & Woodruff, 2014). It is a randomized algorithm that computes the low rank approximation for a given rectangle matrix by randomly sampled columns and rows of the matrix. Compared to other low rank approximation algorithms, CUR is advantageous in that it has (i) an easy interpretation of the approximation result because the subspace is constructed by the actual columns and rows of the target matrix (Mahoney & Drineas, 2009), and (ii) a strong (near-optimal) theoretical guarantee (Bien et al., 2010; Drineas et al., 2006; Mahoney & Drineas, 2008; 2009; Wang & Zhang, 2012; 2013; Boutsidis & Woodruff, 2014). The CUR matrix decomposition algorithm has been successfully applied to many domains, including bioinformatics (Mahoney & Drineas, 2009), collaborative

filtering (Mackey et al., 2011), video background modeling (Mackey et al., 2011), hyperspectral medical image analysis (Mahoney et al., 2006), text data analysis (Mahoney & Drineas, 2008). In the past decade, many variants of the CUR algorithm have been developed (Bien et al., 2010; Drineas et al., 2006; Mackey et al., 2011; Mahoney et al., 2006; Mahoney & Drineas, 2008; 2009; Wang & Zhang, 2012; 2013; Boutsidis & Woodruff, 2014).

Despite the success, one limitation with the existing CUR algorithms is that they either require an access to the *full* matrix (Mahoney & Drineas, 2009), or they just use the sampled rows and columns (Mahoney & Drineas, 2008), ignoring all remaining entries in the matrix. The requirement that the matrix should be fully observed can be difficult to fulfill. For instance, in bioinformatics, it is usually too expensive to acquire the full expression information for hundreds of genes and thousands of individuals; in crowdsourcing, when both the number of workers and instances are large, it becomes impractical to request every worker to label all the instances in study ; in social network analysis, it is often the case that only part of the links between individuals can be accurately detected. In all the above cases, due to the physical or financial constraints, we only have a partial observation of the target matrix, making it difficult to apply the CUR algorithm without ignoring the incomplete part.

One way to deal with the missing entries is to first compute an unbiased estimation of the target matrix based on the observed entries, and then apply the CUR algorithm to the estimated matrix. The main shortcoming of this simple method is that the unbiased estimate can be far from the target matrix when the number of observation is small, as we will show in the empirical study. Another approach is to recover the target matrix from the observed entries using the matrix completion technique (Cai et al., 2010; Candès & Recht, 2012). Since most matrix completion algorithms are developed only for matrices of exactly low rank, they usually work poorly for matrices of full rank (Eriksson et al., 2011). We note that although an adaptive sampling approach is developed in (Krishnamurthy & Singh, 2013) that does apply to matrices of full rank, they use a different sampling strategy, and their bound has a poor dependence on failure probability δ (i.e. $O(1/\delta)$), which significantly limits their applications when both rows and columns are randomly sampled.

In this work, we address the challenge by developing a novel CUR algorithm, named **CUR+**, for partially observed matrix. More specifically, the proposed algorithm computes a low rank approximation of matrix M based on (i) randomly sampled rows and columns from M, and (ii) randomly sampled entries from M. Unlike most matrix completion algorithms that require solving an optimization problem involving trace norm regularization (Bach, 2008; Cai et al., 2010; Ji & Ye, 2009; Mazumder et al., 2010; Toh & Sangwoon, 2010), the proposed algorithm only needs to solve a standard regression problem and therefore is easy to compute. Although the matrix need to be observed for the worst case, we develop a error bound showing that under minor conditions, the proposed CUR+ works for both low-rank and full-rank matrices. In particular, to perfectly recover a rank-r matrix of size $n \times n$ under the incoherent condition (Candès & Recht, 2012), only $O(nr \ln r)$ observed entries are needed, significantly lower than $O(nr \ln^2 n)$ in standard matrix completion theories (Candès & Recht, 2012; Candès & Tao, 2010; Gross, 2011; Keshavan et al., 2010; Recht, 2011) and lower than $O(nr^{3/2} \ln r)$ for adaptive algorithm for matrix recovery (Krishnamurthy & Singh, 2013). We verify our theoretical claims by empirical studies of low rank matrix approximation.

The rest of the paper is organized as follows: Section 2 briefly reviews the related work on the CUR algorithms and matrix completion. Section 3 presents the proposed algorithm and its theoretical properties. Section 4 gives our empirical study. Section 5 concludes our work with future directions.

2. Related Work

CUR matrix decomposition CUR algorithms compute a low rank approximation of the target matrix using the actual rows and columns of the matrix (Bien et al., 2010; Drineas et al., 2006; Goreinov et al., 1997a;b; Mahoney & Drineas, 2008; 2009; Stewart, 1999; Tyrtyshnikov, 2000; Wang & Zhang, 2012; 2013; Boutsidis & Woodruff, 2014). More specially, let $M \in \mathbb{R}^{n \times m}$ be the given matrix and r be the target rank for approximation. A classical CUR decomposition algorithm (Mahoney & Drineas, 2008; 2009) randomly samples d_1 columns and d_2 rows from M, according to their leverage scores, to form matrices C and R, respectively. The approximated matrix $\widehat{M}$ is then computed using the full matrix M as $\widehat{M} = C(C^\dagger M R^\dagger)R$ (Mahoney & Drineas, 2009), or using the intersection of C and R as $\widehat{M} = C(D_R S_R C)^\dagger R$ (Mahoney & Drineas, 2008), where † is the pseudoinverse, D_R and S_R are rescaling and selection matrix, respectively. (Drineas et al., 2006) gives an additive error bound for the CUR decomposition, and an error bound, a significantly stronger result, is given in (Mahoney & Drineas, 2008). It stated that, with a high probability,

$$\|M - \widehat{M}\|_F \leq (1 + \epsilon)\|M - M_r\|_F \tag{1}$$

where M_r is the best rank-r approximation to M, and $\|\cdot\|_F$ is the Frobenius norm of a matrix.

Various improved versions of CUR have been developed. (Mackey et al., 2011) proposes a divide-and-conquer method to compute the CUR decomposition in parallel. (Wang & Zhang, 2013) proposes an adaptive CUR algorithm with much tighter error bound and much lower time complexity. (Boutsidis & Woodruff, 2014) proposes an input-sparsity CUR algorithm with an optimal lower bound. In (Drineas et al., 2006), the authors suggest a simple uniform sampling of columns and rows for the CUR decomposition when the maximum statistical leverage scores, also referred to as incoherence measure (Candès & Recht, 2012; Candès & Tao, 2010; Recht, 2011), is limited. In (Mahoney et al., 2012), algorithms have been developed to efficiently compute the approximated values of statistical leverage scores without having to calculate the SVD decomposition of a large matrix. As we claimed in the introduction section, all the existing CUR algorithms either require the knowledge of *every* entry in the target matrix and therefore cannot be applied directly to partially observed matrices, or they totally ignore the information contained in those partially observed entries, while our work focus on how to exploit those partially observed entries in CUR algorithm to improve approximation accuracy. More complete list of related work on CUR can be found in (Mahoney & Drineas, 2008; Wang & Zhang, 2013; Boutsidis & Woodruff, 2014).

CUR decomposition is closely related to column subset selection problem (Boutsidis et al., 2011; Deshpande & Rademacher, 2010; Mahoney & Drineas, 2008), which has been studied extensively in theoretical computer science and numerical analysis communities (Mahoney & Drineas, 2008; 2009; Wang & Zhang, 2013). It samples multiple columns from the target matrix M and use them as the basis to approximate M, and is often viewed as a special case of the CUR algorithm. A special case of column subset selection is Nyström methods, which is usually used to approximate Positive Semi-Definitive (PSD) matrix in kernel learning (Williams & Seeger, 2000) while we target general matrix. A more complete list of related Nyström methods can be found in (Jin et al., 2013).

Matrix Completion The objective of matrix completion is to fill out the missing entries of a low-rank matrix based on the observed ones. In the standard matrix completion theory, when entries are missing uniformly at random, it requires $O(nr\ln^2 n)$ observed entries to perfectly recover the target matrix under the incoherence condition (Candès & Recht, 2012; Candès & Tao, 2010; Gross, 2011; Keshavan et al., 2010; Recht, 2011). Multiple improvements have been developed for matrix completion, either to deal with nonuniform missing entries or to develop tighter bounds under more strict coherence conditions. (Krishnamurthy & Singh, 2013) developed an adaptive sensing strategy for matrix completion that removes an $\ln n$ factor from the sample complexity. In (Bhojanapalli & Jain, 2014; Chen et al., 2014), the authors study matrix completion when observed entries are not sampled uniformly at random. (Negahban & Wainwright, 2010; Rhode & Tsybakov, 2011) generalize matrix completion to matrix regression. In (Xu et al., 2013), the authors show that the sample complexity for perfect matrix recovery can be reduced dramatically with appropriate side information.

Although it is appealing to directly combine the CUR algorithm with matrix completion to estimate a low rank approximation of a partially observed matrix, it may not work well in practice. One issue is that most matrix completion algorithms are developed for matrix of exactly low rank, significantly limiting its application to low rank matrix application. Although a few studies develop recovery bounds for matrix of full rank, most of them assume the column/row space lies in one or multiple low-rank subspaces even though the observation can be noisy, thus making the matrix full rank (Candès & Plan, 2010; Mackey et al., 2011). There are a few works deal with matrix of exactly full rank (Eriksson et al., 2011; Krishnamurthy & Singh, 2013), but the recovery errors usually deteriorate dramatically when applied to a matrix with a long tail spectrum. In addition, most matrix completion algorithms are computationally expensive, especially for large matrices, since they require, at each iteration of optimization, computing the SVD decomposition of the approximate matrix (Bach, 2008; Cai et al., 2010; Ji & Ye, 2009; Mazumder et al., 2010; Toh & Sangwoon, 2010). In contrast, we focus on the general case where the matrix is of full rank, a significantly more challenging case and the proposed CUR algorithm scales to large matrix and works well for matrix of full rank.

3. CUR+ for Partially Observed Matrices

We describe the proposed CUR+ algorithm, and then present the key theoretical results for it. Due to space limitation, we postpone all the detailed analysis to the supplementary document.

3.1. CUR+ Algorithm

Let $M \in \mathbb{R}^{n\times m}$ be the matrix to be approximated, where $n \geq m$. To approximate M, we first sample uniformly at random d_1 columns and d_2 rows from M, denoted by $A = (\mathbf{a}_1, \ldots, \mathbf{a}_{d_1}) \in \mathbb{R}^{n\times d_1}$, $B = (\mathbf{b}_1, \ldots, \mathbf{b}_{d_2}) \in \mathbb{R}^{m\times d_2}$, respectively, where each $\mathbf{a}_i \in \mathbb{R}^n$ and $\mathbf{b}_j \in \mathbb{R}^m$ is one row and one column of M respectively. We noticed that unifor-

m sampling of rows and columns may not be the best strategy as it does not take into account the difference between individual rows and columns. Other sampling strategies, such as sampling rows/columns based on their statistical leverage scores (Mahoney & Drineas, 2008) and adaptive sampling (Krishnamurthy & Singh, 2013; Wang & Zhang, 2012), can be more effective. We do not choose these sampling methods because they either require an access to the full matrix (Mahoney & Drineas, 2008), introduce serious overhead in computation (Wang & Zhang, 2012), or result in significantly worse bound when matrix is of full rank (Krishnamurthy & Singh, 2013). Finally, for simplicity of discussion, we will assume $d_1 = d_2 = d$ throughout the draft even though our algorithm and analysis can easily be extended to the case when $d_1 \neq d_2$.

Let r be the target rank for approximation, with $r \leq d$. $\widehat{U} = (\widehat{\mathbf{u}}_1, \ldots, \widehat{\mathbf{u}}_r) \in \mathbb{R}^{n\times r}$, $\widehat{V} = (\widehat{\mathbf{v}}_1, \ldots, \widehat{\mathbf{v}}_r) \in \mathbb{R}^{m\times r}$ are the first r eigenvectors of $AA^\top$ and $BB^\top$, respectively. Besides A and B, we furthermore sample, uniformly at random, entries from matrix M. Let Ω include the indices of randomly sampled entries. Our goal is to estimate a low rank approximation of matrix M using A, B, and randomly sampled entries in Ω. To this end, we will solve the following optimization

$$\min_{Z\in\mathbb{R}^{r\times r}} \frac{1}{2}\|\mathcal{R}_\Omega(M) - \mathcal{R}_\Omega(\widehat{U}Z\widehat{V}^\top)\|_F^2 \tag{2}$$

where given Ω, we define a linear operator $\mathcal{R}_\Omega(M) : \mathbb{R}^{n\times m} \mapsto \mathbb{R}^{n\times m}$ as

$$[\mathcal{R}_\Omega(M)]_{i,j} = \begin{cases} M_{i,j} & (i,j) \in \Omega \\ 0 & (i,j) \notin \Omega \end{cases}$$

Let Z_* be an optimal solution to (2). The estimated low rank approximation is given by $\widehat{M} = \widehat{U}Z_*\widehat{V}^\top$. $\widehat{M}$ can also be expressed using standard $C \times U \times R$ formulation by solving a group of linear equations. We note that (2) is a standard regression problem and therefore can be solved efficiently using the standard regression method (e.g. accelerated gradient descent (Nesterov, 2003)). We refer to the proposed algorithm as **CUR+**.

3.2. Guarantee for CUR+

Before presenting the theoretical results, we first describe the notations that will be used throughout the analysis. Let $\sigma_i, i = 1, \ldots, m$ be the singular values of M ranked in descending order, and let $\mathbf{u}_i$ and $\mathbf{v}_i$ be the corresponding left and right singular vectors. Define $U = (\mathbf{u}_1, \ldots, \mathbf{u}_m)$ and $V = (\mathbf{v}_1, \ldots, \mathbf{v}_m)$. Given $r \in [m]$, partitioning the SVD decomposition of M as

$$M = U\Sigma V^\top = \begin{matrix} r & m-r \\ [U_1 & U_2] \end{matrix} \begin{bmatrix} \Sigma_1 & \\ & \Sigma_2 \end{bmatrix} \begin{bmatrix} V_1^\top \\ V_2^\top \end{bmatrix} \tag{3}$$

Table 1. Current results of sample complexity for matrix completion (including matrix regression). Comparing methods including Sequential Matrix Completion (SMC) in (Krishnamurthy & Singh, 2013), Universal Matrix Completion (UMC) in (Bhojanapalli & Jain, 2014), AltMinSense in (Jain et al., 2013) and all the other trace norm minimization methods (Candès & Recht, 2012; Candès & Tao, 2010; Chen et al., 2014; Keshavan et al., 2010; Recht, 2011).

Method	#Observation	Method	#Observation
CUR+	$nr \ln r$	AltMinSense	$nr^{4.5} \ln n$
SMC	$nr \ln^2 r$	Others	$nr \ln^2 n$
UMC	nr^2		

Let $\widetilde{\mathbf{u}}_i, i \in [n]$ be the ith row of U_1 and $\widetilde{\mathbf{v}}_i, i \in [m]$ be the ith row of V_1. The *incoherence* measure for U_1 and V_1 is defined as

$$\mu(r) = \max\left(\max_{i\in[n]} \frac{n}{r}|\widetilde{\mathbf{u}}_i|^2, \max_{i\in[m]} \frac{m}{r}|\widetilde{\mathbf{v}}_i|^2\right) \tag{4}$$

Similarly, we can have the *incoherence* measure for matrices $\widehat{U}$ and $\widehat{V}$ that include the first r eigenvectors of $AA^\top$ and $BB^\top$, respectively. Let $\widehat{\mathbf{u}}'_i, i \in [n]$ be the ith row of $\widehat{U}$ and $\widehat{\mathbf{v}}'_i, i \in [m]$ be the ith row of $\widehat{V}$. Define the incoherence measure for $\widehat{U}$ and $\widehat{V}$ as

$$\widehat{\mu}(r) = \max\left(\max_{i\in[n]} \frac{n}{r}|\widehat{\mathbf{u}}'_i|^2, \max_{i\in[m]} \frac{m}{r}|\widehat{\mathbf{v}}'_i|^2\right) \tag{5}$$

Define projection operators $P_U = UU^\top$, $P_V = VV^\top$, $P_{\widehat{U}} = \widehat{U}\widehat{U}^\top$, and $P_{\widehat{V}} = \widehat{V}\widehat{V}^\top$. We will use $\|\cdot\|_2$ and $\|\cdot\|_F$ respectively for the spectral norm and Frobenius norm of a matrix.

We first present the theoretical guarantee for the CUR+ algorithm when the rank of the target matrix M is no greater than r.

Theorem 1. ***(Low-Rank Matrix Approximation)*** *Assume rank$(M) \leq r$, $d \geq 7\mu(r)r(t + \ln r)$, and $|\Omega| \geq 7\mu^2(r)r^2(t + 2\ln r)$. Then, with a probability at least $1 - 5e^{-t}$, we have $M = \widehat{M}$, where $\widehat{M}$ is a low rank approximation estimated by the CUR+ algorithm.*

Remark Theorem 1 shows that a rank-r matrix can be perfectly recovered from $2dn + |\Omega| = O(nr \ln r)$ observed entries with $t = \Omega(\ln r)$, under the incoherent condition, which is a common assumption to perfectly recover an incomplete matrix (Candès & Recht, 2012; Candès & Tao, 2010; Gross, 2011; Keshavan et al., 2010; Recht, 2011). In Table 1, we compare the sample complexity of the CUR+ algorithm with the sample complexity of the other matrix completion algorithms. We observe that our result significantly improves the sample complexity from previous

work. We should note that unlike (Krishnamurthy & Singh, 2013) where the incoherence measure is only assumed for column vectors, we assume a small incoherence measure for both row and column vectors here. It is this stronger assumption that allows us to sample both rows and columns, leading to the improvement from previous work (Krishnamurthy & Singh, 2013) in the sample complexity $O(nr^{3/2}\ln r)$ to $O(nr\ln r)$.

We now consider a more general case where matrix M is of full rank. Theorem 2 bounds the difference between M and $\widehat{M}$, measured in spectral norm,

Theorem 2. *Let $r \leq m$ be an integer that is no larger than m. Assume (i) $d \geq 7\mu(r)r(t+\ln r)$, and (ii) $|\Omega| \geq 7\widehat{\mu}^2(r)r^2(t+2\ln r)$. Then with a probability at least $1-3e^{-t}$*

$$\|M-\widehat{M}\|_2^2 \leq 8\sigma_{r+1}^2(1+2mn)\left(1+\frac{m+n}{d}\right).$$

As indicated by Theorem 2, when both $\mu(r)$ and $\widehat{\mu}(r)$, the incoherence measure for the first r singular/eigen vectors of M and the sampled columns/rows, are small, we have

$$\|M-\widehat{M}\|_2 \leq O\left(\sqrt{\frac{mn^2}{d}}\|M-M_r\|_2\right)$$

provided that $d \geq O(r\ln r)$ and $|\Omega| \geq O(r^2\ln r)$.

One limitation with Theorem 2 is that $\widehat{\mu}(r)$ is a random variable depending on the sampled columns and rows. Since $\widehat{\mu}(r)$ can be as high as n/r, $|\Omega|$, the number of observed entries required by Theorem 2, can be as large as $O(n^2)$, making it practically meaningless. Below, we develop a result that explicitly bounds $\widehat{\mu}$ with a high probability. Using the high probability bound for $\widehat{\mu}$, we are able to show that under appropriate conditions, we need at most $O(n^2/d^2)$ observed entries in order to establish a error bound for $\|M-\widehat{M}\|$.

To make our analysis simple, we focus on the case when M is of full rank but with skewed singular value distribution. In particular, we assume $\sigma_r \geq \sqrt{2}\sigma_{r+1}$. In order to effectively capture the skewed singular value distribution, we introduce the concept of *numerical rank* $r(M,\eta)$ (Golub & Loan, 1996) with respect to non-negative constant $\eta > 0$

$$r(M,\eta) = \sum_{i=1}^{m}\frac{\sigma_i^2}{\sigma_i^2+mn\eta}$$

Note that when $\eta = 0$, the numerical rank is equivalent to the true rank of the matrix. The larger η is , the smaller it compared to the true rank. In the following analysis, we will replace rank r with numerical rank $r(M,\eta)$.

We furthermore generalize the definition of *incoherence* measure to matrix with *numerical rank*, that is, we further define incoherence measure $\mu(\eta)$ as

$$\mu(\eta) = \max\left(\max_{1\leq i\leq m}\frac{m}{r(M,\eta)}|V_{i,*}\Sigma|^2, \max_{1\leq i\leq n}\frac{n}{r(M,\eta)}|U_{i,*}\Sigma|^2\right) \quad (6)$$

It is easy to verify that $\mu(\eta) \geq 1$. Compared to the standard incoherence measure defined in (4), the key difference is that (6) introduces singular values Σ into the definition of incoherence measure, making it appropriate for matrix of full rank.

The following two lemmas relate $r\mu(r)$ and $r\widehat{\mu}(r)$, respectively, with $r(M,\eta)\mu(\eta)$,

Lemma 1. *If we choose $\eta = \sigma_r^2/mn$, we have*

$$r\mu(r) \leq 2r(M,\eta)\mu(\eta)$$

Lemma 2. *Assume that $d \geq 16(\mu(\eta)r(M,\eta)+1)(t+\ln n)$, and $\sigma_r \geq \sqrt{2}\sigma_{r+1}$. Set $\eta = \sigma_r^2/mn$. With a probability $1-4e^{-t}$, we have*

$$r\widehat{\mu}(r) \leq 2r(M,\eta)\mu(\eta) + 18n\delta^2/r$$
$$\text{where } \delta^2 = \frac{4}{d}(\mu(\eta)r(M,\eta)+1)(t+\ln n)$$

Using Theorem 2, Lemma 1 and 2, we have the result for full-rank matrix with skewed singular value distribution,

Theorem 3. ***(Full Rank Matrix Approximation)*** *Assume $d \geq 16(\mu(\eta)r(M,\eta)+1)(t+\ln n)$ and $\sigma_r \geq \sqrt{2}\sigma_{r+1}$. Set $\eta = \sigma_r^2/mn$. We have, with a probability $1-7e^{-t}$,*

$$\|M-\widehat{M}\|_2^2 \leq 8\sigma_{r+1}^2(1+2mn)\left(1+\frac{m+n}{d}\right) \text{ if}$$

$$|\Omega| \geq 7F^2(t+2\ln r) = O\left(\frac{n^2}{d^2}\right) \text{ where}$$

$$F = \left(2\mu(\eta)r(M,\eta) + 72\frac{n}{d}(\mu(\eta)r(M,\eta)+1)(t+\ln n)\right)^2$$

As indicated by Theorem 3, we will have a bound similar to that of Theorem 2 if $|\Omega| \geq O(n^2/d^2)$. The key difference between Theorem 2 and 3 is that in Theorem 2, the requirement for $|\Omega|$ depends on $\widehat{\mu}(r)$, a random variable depending on the sampled rows and columns. In contrast, in Theorem 3, we remove $\widehat{\mu}$ and bound $|\Omega|$ directly. We finally note that the result $|\Omega| \geq O(n^2/d^2)$ requires nearly the entire matrix for accurately estimating the low rank approximation of the target matrix. This is due to the challenge to recover a full-rank matrix even when the spectrum decays in a realistic way. It remains an open question whether it is possible to reduce the number of observed entries for CUR-type low rank approximation.

4. Experiments

We first verify the theoretical result in Theorem 1, i.e. the dependence of sample complexity on r and n, using synthetic data. We then evaluate the performance of the proposed CUR+ algorithm by comparing it to the state-of-the-art algorithms for low rank matrix approximation. We implement the proposed algorithm using Matlab, and all the experiments were run on a Linux server with CPU 2.53GHz and 48GB memory.

4.1. Experiment (I): Verifying the Dependence on r

We will verify the sample complexity result in Theorem 1, i.e. $d \geq O(r \ln r)$ and $|\Omega| \geq O(r^2 \ln r)$. We note both the requirements on d and $|\Omega|$ are independent from matrix size.

Settings Here we study square matrices of different sizes and ranks, with n varied in $\{1,000; 2,000; l4,000; 8,000; 10,000\}$, and r varied in $\{10, 20, 30, 50\}$. For each special n and r, we search for the smallest d and $|\Omega|$ that can lead to almost perfect recovery of the target matrix (i.e. $\|M-\widehat{M}\|_F/\|M\|_F \leq 2\times 10^{-4}$) in all 10 independent trials. To create the rank-r matrix $M \in \mathbb{R}^{n\times n}$, we first randomly generate matrix $M_L \in \mathbb{R}^{n\times r}$ and $M_R \in \mathbb{R}^{r\times n}$ with each entry of M_L and M_R drawn independently at random from $\mathcal{N}(0,1)$, and M is given by $M = M_L \times M_R$. To create A and B, we sample uniformly at random d rows and columns. We further sample $|\Omega|$ entries from M to be partially observed. Under this construction scheme, the difference between the incoherence $\mu(r)$ for different sized matrices are relatively small (from minimum 1.4127 to maximum 2.4885). Although we will plot d and $|\Omega|$'s dependence on $\mu(r)$, we will ignore their impact in discussion of the results.

Results The dependence of minimal d on r and n is given in Figure 1(a) and (b), where (a) plots d against $r \ln r$ and (b) shows d versus $r^2 \ln r$. We can see clearly that d has a linear dependence on $r \ln r$. We also observed from Figure 1(a) that d is almost independent from n, the matrix size. Figure 1(c) and (d) plot the $|\Omega|$, the minimum number of observed entries, against $r \ln r$ and $r^2 \ln r$. The result in Figure 1 (d) confirms our theoretical finding, i.e. $|\Omega| \propto r^2 \ln r$.

4.2. Experiment(II): Comparison with Baseline Methods for Low Rank Approximation

We evaluate the performance of the proposed CUR+ algorithm on several benchmark data sets that have been used in the recent studies of the CUR matrix decomposition algorithm, including Enron emails ($39,861 \times 28,102$), Dexter ($20,000 \times 2,600$), Farm Ads ($54,877 \times 4,143$) and Gisette ($13,500 \times 5,000$), where each row of the matrix corresponds to a document and each column corresponds to a term/word. Detailed information of these data sets can be found in (Wang & Zhang, 2013). All four matrices are of full rank and have skewed singular value distribution, as shown in Figure 2

Baselines Since both the rows/columns and entries observed in the proposed algorithm are sampled uniformly at random, we only compare our approach to the standard CUR algorithm using uniformly sampled rows and columns. Although the adaptive sampling based approaches (Krishnamurthy & Singh, 2013) usually yield lower errors than the standard CUR algorithm, they do not choose observed entries randomly and therefore are not included in the comparison. Let C be a set of d_1 sampled columns and R be the set of d_2 sampled rows. The low rank approximation by the CUR algorithm is given by $\widehat{M} = CZR$, where $Z \in \mathbb{R}^{d_1\times d_2}$. Two methods are adopted to estimate Z. We first estimated Z by $Z = C^\dagger M R^\dagger$. Since this estimation requires an access to the full matrix, we refer to it in this section as **CUR-F** (Mahoney & Drineas, 2009). In the second method, we estimate Z by the intersection of C and R, and then calculate $\hat{M} = CZR$. Since this method exploits the intersection of C and R, we refer to it as **CUR-I** (Mahoney & Drineas, 2008). Evidently, CUR-F is expected to work better than our proposal and will provide a lower bound for the CUR algorithm for partially observed matrices. Note that we also construct an unbiased estimator M_e by using the randomly observed entries in Ω, and then estimate matrix Z by $Z = C^\dagger M_e R^\dagger$. We call this algorithm CUR-E. Its performance deteriorates a lot compared to other algorithms. Due to space limitation, we will present the results in the supplementary document.

Settings To make our result comparable to the previous studies, we adapted the same experiment strategy as in (Wang & Zhang, 2012; 2013). More specially, for each data set, we set $d_1 = \alpha r$ and $d_2 = \alpha d_1$, with rank r varied in the range of $\{10, 20, 50\}$ and α is set to be 5. To create partial observations, we randomly sample $|\Omega| = \Omega_0 = nmr^2/nnz(M)$ entries from the target matrix M, where $nnz(M)$ is the number of non-zero entries of M. We measure the performance of low rank matrix approximation by the related spectral-norm difference $\ell_s = \|M - \widehat{M}\|/\|M - M_r\|$ which has solid theoretical guarantee according to Theorem 3. To make a fair comparison with previous work measured by Frobenius norm, we also report the results measured by relative Frobenius norm, that is $\ell_F = \|M - \widehat{M}\|_F/\|M - M_r\|_F$. Finally, we follow the experimental protocol specified in (Wang & Zhang, 2012) by repeating every experiment 10 times and reporting the mean value.

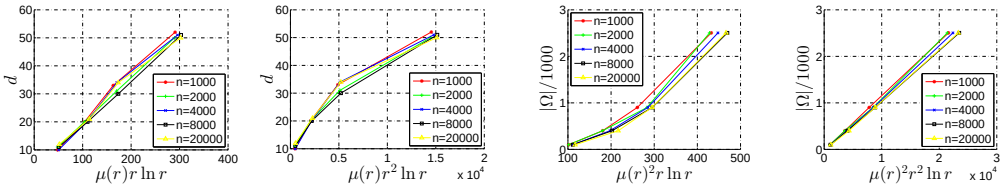

Figure 1. Experiment results on the synthetic data. (a)(b) plot the minimum d for perfect matrix recovery against $r\ln r$ and $r^2\ln r$ respectively, and (c)(d) plot the minimum $|\Omega|$ for perfect matrix recovery against $r\ln r$ and $r^2\ln r$. The results confirm the theoretical finding in Theorem 1, i.e. $d = O(r\ln r)$ and $|\Omega| = O(r^2\ln r)$.

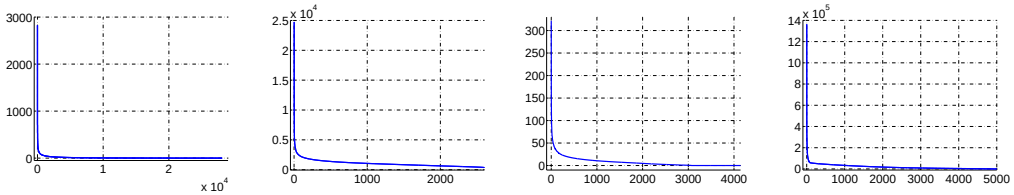

Figure 2. Singular values of real data ranked in descending order. All these four data sets are full-rank and have skewed singular value distribution.

Results Figure 3 shows the results of low rank matrix approximation. We observe that in most cases, with increasing number of observed entries, CUR+ shows much more similar performance as CUR-F that has an access to the full target matrix M. Note that because CUR-I do not use those partially observed entries, their performance do not change with increasing $|\Omega|$. On the other hand, on most datasets, although CUR+'s performance is similar to that of CUR-I at the beginning when the number of observed entries is small, their performance diverges a lot with increasing number of observed entries. We observe that there are several exceptions, for example, Enron data when $r = 10$, We plan to examine this unusual phenomenon in the future.

Figure 4 shows the results measured by Frobenius norm for $r = 10, 20$ and 50. We found the results are similar to that measured by spectral norm, that is, CUR+ yields similar performance as CUR-F with increasing number of observed entries, and performs significantly better than CUR-I when more entries are observed.

5. Conclusion

In this paper, we propose a CUR-style low rank approximation algorithm for partially observed matrix. Our analysis shows that the proposed algorithm only needs $O(nr\ln r)$ number of observed entries to perfectly recover a low-rank matrix, improving the results of the existing algorithms for matrix completion (of course under a slightly stronger condition). We also show the the spectral error bound for the proposed algorithm when the target matrix is of full rank. Empirical studies on both synthetic data and real datasets verify our theoretical claims and furthermore, demonstrate that the proposed algorithm is more effective in handling partially observed matrix than the existing CUR algorithms. Since adaptive sampling has shown promising results for low rank matrix approximation (Krishnamurthy & Singh, 2013), in the future, we plan to combine the proposed algorithm with adaptive sampling strategy to further reduce the error bound. We also plan to exploit the recent studies on matrix approximation/completion with non-uniform sampling and extend the CUR algorithm to the case when observed entries are non-uniform sampled.

Acknowledgments

The authors want to thank all the people especially all the reviewers providing helpful comments and suggestions to improve the paper. This research was partially supported by the 973 Program (2014CB340501) and National Science Foundation of China NSFC (61333014, 61305067).

References

Bach, F. Consistency of trace norm minimization. *JMLR*, 9:1019–1048, 2008.

Bhojanapalli, S. and Jain, P. Universal matrix completion. In *ICML*, 2014.

Bien, J., Xu, Y., and Mahoney, M. Cur from a sparse optimization viewpoint. In *NIPS*, 2010.

Boutsidis, C. and Woodruff, D. P. Optimal CUR matrix decompositions. In *STOC*, 2014.

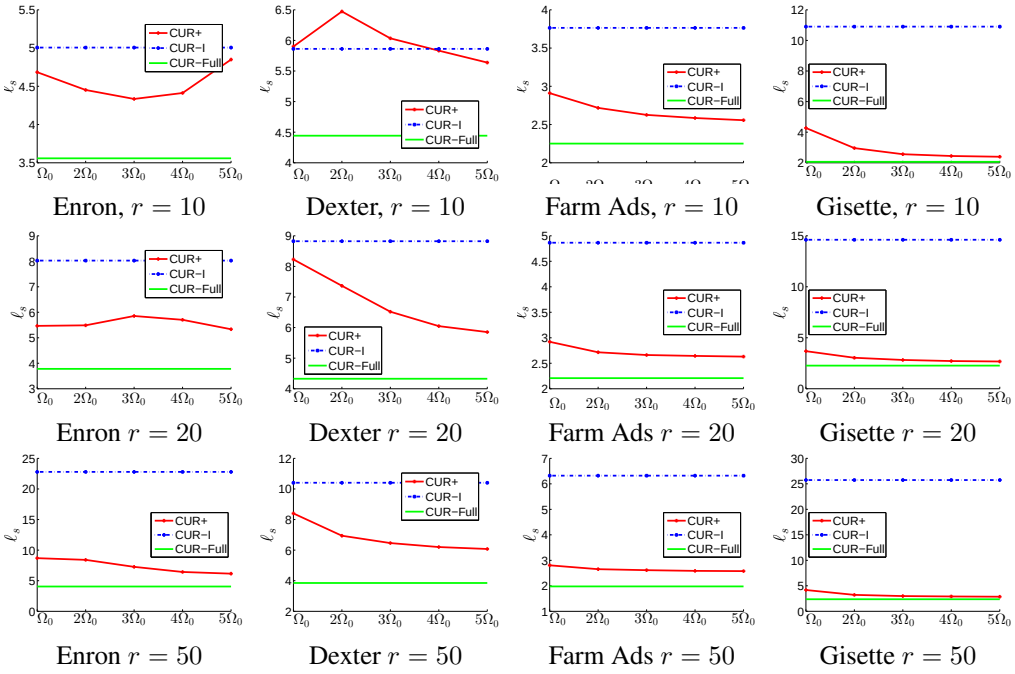

Figure 3. Comparison of CUR algorithms with the number of sampled columns (rows) fixed as $d_1 = 5r$ ($d_2 = 5d_1$), where $r = 10, 20, 50$. The number of observed entries $|\Omega|$ is varied from Ω_0 to $5\Omega_0$. The results are measured by relative spectral norm.

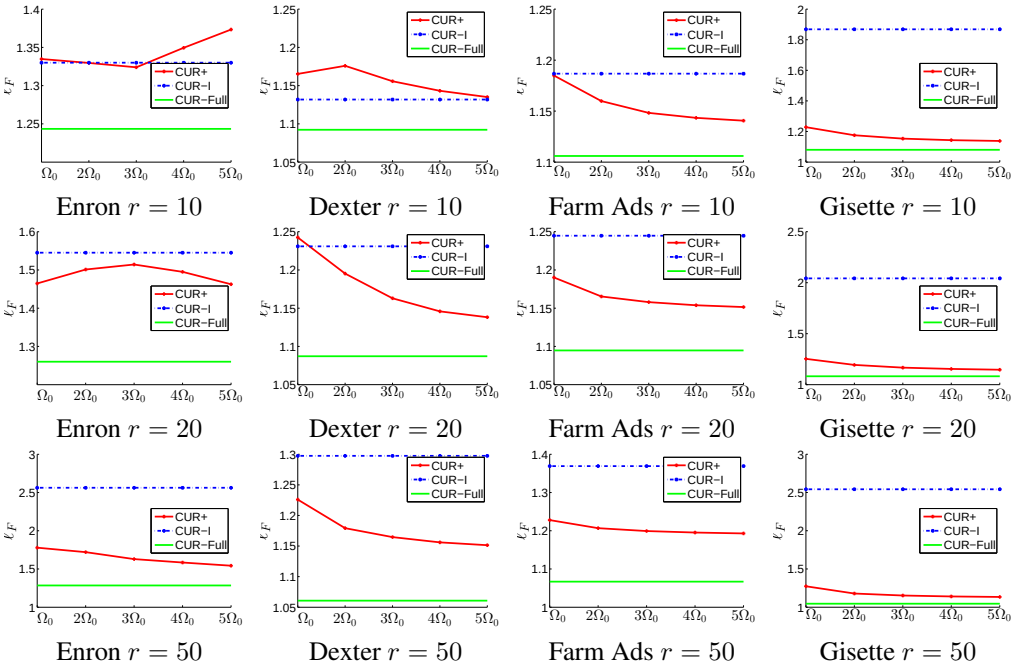

Figure 4. Comparison of CUR algorithms with the number of sampled columns (rows) fixed as $d_1 = 5r$ ($d_2 = 5d_1$), where $r = 10, 20, 50$. The number of observed entries $|\Omega|$ is varied from Ω_0 to $5\Omega_0$. The results are measured by relative Frobenius norm.

Boutsidis, C., Drineas, P., and Magdon-Ismail, M. Near optimal column-based matrix reconstruction. In *FOCS*, 2011.

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

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

Candès, E. and Recht, B. Exact matrix completion via convex optimization. *Commun. ACM*, 2012.

Candès, E. and Tao, T. The power of convex relaxation: near-optimal matrix completion. *TIT*, 2010.

Chen, Y., Bhojanapalli, S., Sanghavi, S., and Ward, R. Coherent matrix completion. In *ICML*, 2014.

Deshpande, A. and Rademacher, L. Efficient volume sampling for row/column subset selection. In *FOCS*, 2010.

Drineas, P., Kannan, R., and Mahoney, M.W. Fast Monte Carlo algorithms for matrices III: Computing a compressed approximate matrix decomposition. *SIAM J. Comput.*, 36:184–206, 2006.

Eriksson, B., Balzano, L., and Nowak, R. High-rank matrix completion and subspace clustering with missing data. *CoRR*, 2011.

Goldberg, A., Zhu, X., Recht, B., Xu, J.-M., and Nowak, R. Transduction with matrix completion: Three birds with one stone. In *NIPS*, 2010.

Golub, G. and Loan, C. *Matrix computations (3rd ed.)*. Johns Hopkins University Press, 1996.

Goreinov, S., Zamarashkin, N., and Tyrtyshnikov, E. seudo-skeleton approximations by matrices of maximal volume. *Mathematical Notes*, 62(4):515–519, 1997a.

Goreinov, S. A., Tyrtyshnikov, E. E., and Zamarashkin, N. L. A theory of pseudoskeleton approximations. *Linear Algebra and Its Applications*, 261(1-3):1–21, 1997b.

Gross, D. Recovering low-rank matrices from few coefficients in any basis. *TIT*, 57(3):1548–1566, 2011.

Halko, N., Martinsson, P.-G., and Tropp, J. A. Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions. *SIAM Review*, 53(2):217–288, 2011.

Jain, P., Netrapalli, P., and Sanghavi, S. Low-rank matrix completion using alternating minimization. In *STOC*, 2013.

Ji, S. and Ye, J. An accelerated gradient method for trace norm minimization. In *ICML*, 2009.

Jin, R., Yang, T., Mahdavi, M., Li, Y.-F., and Zhou, Z.-H. Improved bounds for the nyström method with application to kernel classification. *TIT*, 59(10):6939–6949, 2013.

Keshavan, R., Montanari, A., and Oh, S. Matrix completion from a few entries. *TIT*, 2010.

Koltchinskii, V. Low rank matrix recovery: nuclear norm penalization. In *Oracle Inequalities in Empirical Risk Minimization and Sparse Recovery Problems*. Springer, 2011.

Krishnamurthy, A. and Singh, A. Low-rank matrix and tensor completion via adaptive sampling. In *NIPS*, 2013.

Mackey, L., Talwalkar, A., and Jordan, M. Divide-and-conquer matrix factorization. In *NIPS*, 2011.

Mahoney, M., Maggioni, M., and Drineas, P. Tensor-cur decompositions for tensor-based data. In *KDD*, 2006.

Mahoney, M., Drineas, P., Magdon-Ismail, M., and Woodruff, D. Fast approximation of matrix coherence and statistical leverage. In *ICML*, 2012.

Mahoney, M. W. and Drineas, P. Relative-error CUR matrix decompositions. *SIAM J. Matrix Anal. Appl.*, 30: 844–881, 2008.

Mahoney, M. W. and Drineas, P. CUR matrix decompositions for improved data analysis. *Proc. Natl. Acad. Sci. USA*, 106:697–702, 2009.

Mazumder, R., Hastie, T., and Tibshirani, R. Spectral regularization algorithms for learning large incomplete matrices. *JMLR*, 11:2287–2322, 2010.

Negahban, S. and Wainwright, M. Estimation of (near) low-rank matrices with noise and high-dimensional scaling. In *ICML*, 2010.

Nesterov, Y. *Introductory Lectures on Convex Optimization: A Basic Course*. Springer, 2003.

Recht, B. A simpler approach to matrix completion. *JMLR*, 12:3413–3430, 2011.

Rhode, A. and Tsybakov, A. Estimation of high dimensional low rank matrices. *Annual of Statistics*, 39(2): 887–930, 2011.

Srebro, N., Rennie, J., and Jaakkola, T. Maximum-margin matrix factorization. In *NIPS*, 2004.

Stewart, G. Four algorithms for the the efficient computation of truncated pivoted qr approximations to a sparse matrix. *Numerische Mathematik*, 1999.

Toh, K.-C. and Sangwoon, Y. An accelerated proximal gradient algorithm for nuclear norm regularized linear least squares problems. *Pacific Journal of Optimization*, 2010.

Tyrtyshnikov, E. Incomplete cross approximation in the mosaic-skeleton method. *Computing*, 2000.

Wang, S. and Zhang, Z. A scalable cur matrix decomposition algorithm: Lower time complexity and tighter bound. In *NIPS*, 2012.

Wang, S. and Zhang, Z. Improving cur matrix decomposition and the nyström approximation via adaptive sampling. *JMLR*, 14(1):2729–2769, 2013.

Williams, C. and Seeger, M. Using the nyström method to speed up kernel machines. In *NIPS*, 2000.

Woodruff, D. P. Sketching as a tool for numerical linear algebra. *Foundations and Trends in Theoretical Computer Science*, 10(1-2):1–157, 2014.

Xu, M., Jin, R., and Zhou, Z.-H. Speedup matrix completion with side information: Application to multi-label learning. In *NIPS*, 2013.