Carnegie Mellon University
Mellon College Of Science

THESIS

SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF DOCTOR OF PHILOSOPHY

TITLE : Adaptive Matrix Completion

PRESENTED BY : Ilqar Ramazanli

ACCEPTED BY THE DEPARTMENT OF: Mathematical Sciences

Barnabás Póczos	May 2020
MAJOR PROFESSOR	DATE
Thomas Bohman	May 2020
DEPARTMENT HEAD	DATE

APPROVED BY THE COLLEGE COUNCIL

Rebecca W. Doerge	May 2020
DEAN	DATE

Adaptive Matrix Completion

Ilqar Ramazanli

A DISSERTATION

in

Mathematical Sciences

Presented to the faculties of Carnegie Mellon University in Partial Fulfillment of the Requirements for the Degree of Doctor of Philosophy

May 2020

Barnabás Póczos, advisor

Dissertation Committee:
Barnabás Póczos, Alan Frieze, Tomasz Tkocz, Jafar Jafarov

Acknowledgments

I am very grateful to my advisor Barnabás Póczos for all of the time he spends to discuss research and guide me whenever I face obstacles. Besides research, I am very thankful to him for his enlightening advice regarding my career choice and how to design my Ph.D. education. His patience and calmness always gave me motivation in my work and made my research experience entertaining and enjoyable. My thanks to prof. Eunsu Kang for her welcoming attitude always and her contributions to discussions.

I want to thanks to prof Jian Ma, Gary Miller, Dmitry Kramkov and Sashank Reddi for their priceless research discussions during these years. It's my pleasure to acknowledge prof Alan Frieze, Tomasz Tzock, and Dr. Jafar Jafarov for being on my committee.

My endless thanks to my fellow graduate students, Abutalib Aghayev, Mehmet Berat Aydemir, Vijay Bhattiprolu, Kayla Bollinger, Joseph Briggs, Debsoumya Chakraborti, Xiao Chang, Da Qi Chen, Mochong Duan, Yuanyuan Feng, Likhit Ganedi, Giovanni Gravina, Mihir Hasabnis, Tetsuya Ishikawa, David Itkin, Sittinon Jirattikansakul, Paul Liang, Shilin Ma, Han Nguyen, Yue Niu, Hai Pham, Yue Pu, Aditya Raut, Antoine Remond-Tiedrez, Amirbehshad Shahrasbi, Xiaofei Shi, Kerrek Stinson, Ananya Uppal, Son Van, Zoe Wellner, Weicheng Ye, Seven Yue, Jing Zhang, Tingting Zhang, Shirley Zhu for insightful discussion, fun conversations, and laughter moments.

It's my pleasure to thank my senior team members and managers for their guidance and patience during my internships in Google - Graphite and Face-

book - Ads Measurement Science teams. Without insightful discussions from Jedrzej Jablonski, Ira Kones, Xinfeng Li, Harry Mavroforakis, and Eric Tsai, it would be very challenging to build a bridge between theoretical knowledge to the application to real-life problems.

I am very thankful to my family for their continuous support during these years. Finally, I want to thank my friends all over the world, especially my friends in New York City. Sanjar Ahmadov, Khayyam Guliyev, Jahangir Najafov, and Yurika Ueno have always welcomed and accommodated me and organized exciting events while I am in the city.

ABSTRACT

In this thesis we study the problem of exact completion for $m \times n$ sized matrix of rank r and the problem of low-rank estimation with the adaptive sampling method. We introduce a relation of the exact completion problem with the sparsest vector of column and row spaces. Using this relation, we propose matrix completion algorithms that exactly recovers the target matrix. These algorithms are superior to previous works in two important ways. First, our algorithms exactly recovers μ_0-coherent column space matrices by probability at least $1 - \epsilon$ using much smaller observations complexity than - $\mathcal{O}(\mu_0 rn \log \frac{r}{\epsilon})$—the state of art. Specifically, many of the previous adaptive sampling methods require to observe the entire matrix when the column space is highly coherent. However, we show that our method is still able to recover this type of matrices by observing a small fraction of entries under many scenarios. Second, we propose an exact completion algorithm, which requires minimal pre-information as either row or column space is not being highly coherent. We provide an extension of these algorithms that is robust to sparse random noise. Besides, we propose an additional low-rank estimation algorithm that is robust to any small noise by adaptively studying the shape of column space. At the end of the thesis, we provide experimental results that illustrate the strength of the algorithms proposed here.

This thesis have been written mainly based on the paper [12].

Contents

Chapter 1

Introduction

In this thesis, we illustrate how adaptivity helps us to reach theoretical lower bounds concerning observation count in the matrix completion problem. In modern data analysis, it has been presented that in many scenarios, adaptive sensing and sampling can work more efficiently than passive methods [10, 27]. Our main objective is to further optimize adaptive sampling by minimizing the number of observations needed to recover the target matrix entirely. We show how to recover the entire low-rank matrix by observing information-theoretically least number of entries in various settings.

Low-rank matrix completion plays a significant role in many real-world applications, including camera motion inferring, multi-class learning, positioning of sensors, and gene expression analysis [2, 15]. In gene expression analysis, the target matrix represents expression levels for various genes across several conditions. Measuring gene expression, however, is expensive, and we would like to estimate the target matrix with a few observations as possible. Here, we provide an algorithm that can be used for matrix

completion from limited data. Roughly speaking, to find each unknown expression level, we are supposed to do multiple measurements. Each of the additional measurements has its extra cost. Naturally, we aim to solve the entire problem using the least possible measurement cost.

Krishnamurthy and Singh ([15], [16]) illustrated how adaptive sampling reduces observation complexity compared to passive sampling. These results were two of the earliest algorithms that were robust against coherent row space. Like many other results in the literature, these algorithms also heavily rely on the incoherence of the column space. At first, authors showed for an $n \times n$ sized, rank r matrix, with column space coherence of μ_0, can be exactly recovered using just $\mathcal{O}(n\mu_0 r^{1.5} \log r)$ observations ([15]), then this result optimized to $\mathcal{O}(n\mu_0 r \log^2 r)$ in the later work ([16]). Recently, [2] further improved previous results by proposing algorithm that performs $\mathcal{O}(n\mu_0 r \log r)$ observations to accomplish the task.

The main goal of this work is to give a new approach to the exact recovery problem using the sparsest vector of column and row spaces instead of coherence. Finding sparsest vector has been in the focus of the research attention for a long time ([20], [22], [7]). However, to the best of our knowledge, it is the first time applied to active matrix completion problem.

In this particular work, we approach the exact completion problem in various given pre-information settings. Our first algorithm requires the precise value of the rank and no other information. Our second algorithm does not request any information except knowing that either column or row space

is not highly coherent. For comparison, as we discuss in the next sections, previous adaptive sampling recovery algorithms require the value of $nr\mu_0 \text{polylog}\frac{r}{\epsilon}$, which implicitly requests estimation or exact value of r and μ_0.

1.1 Main Results

In light of the above discussion, we state the main contributions of this work.

- Relation of the sparsest vector of the column and row space and the problem of exact recovery has been studied in detail. An exact completion algorithm is proposed with respect to these vectors. Moreover, using the relation of the sparsest vector to coherence number, we show that the proposed method exactly recovers the underlying low-rank matrix using less observation than the state of the art.

- We provide efficient algorithms that require minimal information as $\mathbf{ERR}$ (rank), $\mathbf{ERRE}$ (either row or column space is not coherent). Moreover, we show the observation complexity of these algorithms is upper bounded by $\mathcal{O}(nr\mu_0 \log \frac{r}{\epsilon})$. (the expression for observation complexity is provided in the next sections after all the necessary definitions are given).

- To the best of our knowledge, all previous adaptive sampling methods need to observe entire matrix if the underlying matrix has a highly coherent column space. In the algorithm $\mathbf{EREI}$ we show that having incoherent row space can be a backup and we can still recover these

matrices using a small fraction of entries even it has highly coherent column space.

- We extend our exact recovery method to make it robust to sparse random noise in columns similar to [2]. Besides, we provide a low-rank estimation method that is robust to any small noise, relying on the adaptive estimation of the angle between the underlying subspace and estimated subspace.

1.2 Related Work

The power of adaptive sampling had been illustrated even earlier than [15]. [10, 19, 1, 25] showed that under certain hypothesis, adaptive sampling outperforms all passive schemes.

Exact recovery and matrix completion has been studied extensively under passive schemes as well. Nuclear norm minimization is one of the most popular methods [9]. [4] and [23] showed that $\Omega((m+n)r\max(\mu_0^2, \mu_1^2)\log^2 n_2)$ observations are enough to recover an $m \times n$ matrix of rank r using nuclear norm minimization, where μ_0 and μ_1 correspond to column and row space coherence parameters. Using the same technique, [5] showed that under the uniform sampling setting we need at least $\Omega(mr\mu_0 \log n)$ observations to recover the matrix exactly. This result implies the near optimality of nuclear norm minimization. Another work using nuclear norm minimization is due to [6], in which they show how to recover coherent $n \times n$ sized matrix of rank r using just $\mathcal{O}(nr\log^2 r)$ observations.

[14] showed how to use a nuclear norm minimization approach to approximate noisy low-rank matrices under some global information conditions. This work assumes that μ_0 (coherence of column space) is below a given threshold. This result has a similar flavor to ours in that it works even when there is less initial knowledge about the target matrix. Later, this result was extended to a point where without any assumption on μ_0 the target matrix could be approximated. However, the reconstruction error of approximation becomes worse in this case [18]. There are other approaches for noisy matrix completion which they mainly focus on parameters that describe how much information an observation reveals [13, 21].

1.3 Preliminaries

Let $\mathbf{M}$ denote the underlying $m \times n$ sized rank-r matrix that we target to recover. For any positive integer n, let $[n]$ represent the set $\{1, 2, ..., n\}$. For any vector $x = (x_1, x_2, ..., x_n)$ of size n, $\|x\|_p$ will denote the L_p norm of it. We call x_i the i'th coordinate of x. For any, $\Omega \subset [n]$ let x_Ω denote the induced subvector of x from coordinates Ω. For instance, for the vector $x = (1, 2, 4, 8, 9)$ and $\Omega = \{1, 3\}$, x_Ω represents the vector $(1, 4)$. For any $\mathbf{R} \subset [m]$, $\mathbf{M}_{\mathbf{R}:}$ stands for an $|\mathbf{R}| \times n$ sized submatrix of $\mathbf{M}$ that rows are restricted by $\mathbf{R}$. We define $\mathbf{M}_{:\mathbf{C}}$ in a similar way for restriction with respect to columns. Intuitively, $\mathbf{M}_{\mathbf{R}:\mathbf{C}}$ defined for $|\mathbf{R}| \times |\mathbf{C}|$ sized submatrix of $\mathbf{M}$ with rows restricted to $\mathbf{R}$ and columns restriced to $\mathbf{C}$. Moreover, for the special case $\mathbf{M}_{i:}$ stands for i-th row and $\mathbf{M}_{:j}$ stands for the j'th column. Similarly, $\mathbf{M}_{i:\mathbf{C}}$ will represent the restriction of the row i by C and

$\mathbf{M}_{\mathbf{R}:j}$ represents restriction of the column j by $\mathbf{R}$. $\theta(u, v)$ stands the angle between vectors u and v. Moreover, $\theta(u, \mathbb{V}) = \min\{\theta(u, v)|v \in \mathbb{V}\}$ and $\theta(\mathbb{U}, \mathbb{V}) = \max\{\theta(u, \mathbb{V})|u \in \mathbb{U}\}$ for subspaces $\mathbb{U}$ and $\mathbb{V}$. The projection operator to subspace $\mathbb{U}$ will be represented by $\mathcal{P}_{\mathbb{U}}$.

1.4 Problem Setup

One of the critical factors in the matrix completion problem is due to the coherence parameter of the target matrix [11, 3]. We define the coherence of an r-dimensional subspace $\mathbb{U}$ of $\mathbb{R}^n$ in the following way:

$$\mu(\mathbb{U}) = \frac{n}{r} \max_{1 \leq j \leq n} ||\mathcal{P}_{\mathbb{U}} e_j||^2,$$

where e_j denotes the j-th standard basis element and $\mathcal{P}_{\mathbb{U}}$ represents the orthogonal projection operator onto the subspace $\mathbb{U}$. It is easy to see that if $e_j \in \mathbb{U}$ for some $j \in [n]$, then the coherence will attain its maximum value: $\mu(\mathbb{U}) = \frac{n}{r}$. We can see that if $\mathbb{U}$ is equally distant from each standard basis vectors, then $\mu(\mathbb{U})$ will be close to 1, and additionally, it is lower bounded by 1.

We want to present an algorithm due to [15] here before providing our main results in the next section. Authors proposed an adaptive algorithm that can recover $n \times n$ sized rank-r matrices using $\mathcal{O}(n\mu_0 r^{1.5} \log r)$ observations, which was indeed better than known state of the art $\mathcal{O}(n\mu_0 r^2 \log^2 n)$ for passive algorithms ([23]). The algorithm studies column space by deciding whether the partially observed column is linearly independent with

previously fully observed columns. Authors show observing $\mathcal{O}(\mu_0 r^{1.5} \log \frac{r}{\epsilon})$ observations for each column is enough to make the decision for linear independence with probability $1 - \epsilon$. The algorithm below describes the details of the proposed algorithm.

KS2013: Exact recovery [15].

Input: $d = \mathcal{O}(\mu_0 r^{1.5} \log \frac{r}{\epsilon})$

Initialize: $k = 0, \widehat{\mathbf{U}}^0 = \emptyset$

1: Draw uniformly random entries $\Omega \subset [m]$ of size d
2: **for** i from 1 to n **do**
3: **if** $\|\mathbf{M}_{\Omega:i} - \mathcal{P}_{\widehat{\mathbf{U}}^{\mathbf{k}}_{\mathbf{\Omega}}} \mathbf{M}_{\Omega:i}\| > 0$
4: Fully observe $\mathbf{M}_{:i}$
5: $\widehat{\mathbf{U}}^{k+1} \leftarrow \widehat{\mathbf{U}}^k \cup \mathbf{M}_{:\mathbf{i}}$,
6: Orthogonalize $\widehat{\mathbf{U}}^{k+1}$
7: $k = k + 1$
8: **otherwise:**
9: $k = k + 1$ $\widehat{\mathbf{M}}_{:i} = \widehat{\mathbf{U}}^k \widehat{\mathbf{U}}_{\Omega}^{k+} \widehat{\mathbf{M}}_{\Omega:i}$

Output: $\widehat{\mathbf{M}}$

Later, authors improved the observation complexity to $\mathcal{O}(n\mu_0 r \log^2 \frac{r}{\epsilon})$ in a proceeding work ([16]). Then, another improvement due to [2] further reduced this complexity to $\mathcal{O}(n\mu_0 r \log \frac{r}{\epsilon})$-current state of the art using similar setting and algorithm.

Chapter 2

Exact Completion Problem

In this chapter, we provide theoretical results for the exact completion problem for $m \times n$ sized rank-r matrices. We first start with defining the *sparsity-number* and study its properties. Then, in the following sections, we provide exact recovery algorithms under different pre-information using the idea of *sparsity-number*. The first algorithm uses the precise value of the rank as the only pre-information. The second algorithm assumes that either row or column space sparsity number is not low (which this algorithm can be treated as a rank estimation algorithm as well). Finally, the third algorithm will assume that we have an estimation of rank and sparsity numbers of coherence numbers.

2.1 Exact Recovery with sparsity-number

In this section, we define the *sparsity-number* and discuss its properties. *Sparsity-number* of vectors is directly related to ℓ_0 norm and for matrices and subspaces it is directly related to ℓ_0 semi-norm of basis columns. Formal

definitions for each case after defining *nonsparsity-number* below:

Definition 1. *We represent nonsparsity-number of a vector $x \in \mathbb{R}^m$ by $\psi(x)$ and define as $\psi(x) = \|x\|_0$. Moreover, we extend the definition to matrices and subspaces in the following way: for $\mathbf{M} \in \mathbb{R}^{m\times n}$ of rank r and subspace $\mathbb{U} \subseteq \mathbb{R}^m$ of dimension r we have*

$$\psi(\mathbf{M}) = \min\{\psi(x) | x = \mathbf{M}z \text{ and } z \notin \text{null}(\mathbf{M})\}$$
$$\psi(\mathbb{U}) = \min\{\psi(x) | x \in \mathbb{U} \text{ and } x \neq 0\}$$

Then *sparsity-number* is just completion of the *nonsparsity-number*:

Definition 2. *Sparsity-number is denoted by $\overline{\psi}$ and for vector $x \in \mathbb{R}^m$, for matrix $\mathbf{M} \in \mathbb{R}^{m\times n}$ and for subspace $\mathbb{U} \subseteq \mathbb{R}^m$ it is defined as:*

$$\overline{\psi}(x) = m - \psi(x)$$
$$\overline{\psi}(\mathbf{M}) = m - \psi(\mathbf{M})$$
$$\overline{\psi}(\mathbb{U}) = m - \psi(\mathbb{U})$$

The space *sparsity-number* for matrices provides a novel and easy way to analyze adaptive matrix completion algorithms. In many adaptive matrix completion methods, a crucial step is to decide whether a column is (or is not) contained in a given subspace. Ideally, we would like to make this decision as soon as possible before observing the entire column vector. Here, we show that *sparsity-number* helps us to decide whether a partially observed vector can be contained in a given subspace or not.

Optimal Observation for Each Column Before proceeding to more advanced algorithms, we target to answer one fundamental question. What is the specific number of entries in $\mathbf{KS2013}$ to observe in a column that allows us to deterministically decide whether it is independent or dependent on previous columns? The following lemma helps us to answer this fundamental question.

Lemma 1. *Let $\mathbb{U}$ be a subspace of $\mathbb{R}^m$ and $x^1, x^2, ..., x^n$ be any set of vectors selected from $\mathbb{U}$. Then the linear dependence of $x^1_\Omega, x^2_\Omega, ..., x^n_\Omega$ implies linear dependence of $x^1, x^2, ..., x^n$, for any $\Omega \subset [m]$ such that $|\Omega| > \overline{\psi}(\mathbb{U})$.*

Proof. By the hypothesis of linear dependence, there are coefficients $\alpha_1, ..., \alpha_n$, not all zero, such that

$$\alpha_1 x^1_\Omega + ... + \alpha_n x^n_\Omega = 0.$$

To show linear dependence of $x^1, \ldots, x^n$ we prove the following equation also satisfies

$$\alpha_1 x^1 + ... + \alpha_n x^n = 0$$

Assume by contradiction $y = \sum_{i=1}^{k} \alpha_i x_i$ is a nonzero vector. But, we have

$$y_\Omega = \sum_{i=1}^{k} \alpha_i x_{i\Omega} = 0$$

which implies $\overline{\psi}(\mathbb{U}) \geq |\Omega|$ from the definition of *space sparsity number*. However, $|\Omega| > \overline{\psi}(\mathbb{U})$ from the hypothesis of the lemma which concludes a

contradiction. Therefore, the assumption the vector y being nonzero vector cannot be true, then the following satisfies :

$$y = \sum_{i=1}^{k} \alpha_i x_i = 0$$

□

In the following example we show that the statement of the lemma 1 is tight. Specifically, we show that there is a matrix which linear dependence in $\overline{\psi}(\mathbb{U})$ many coordinates does not imply linear dependence of vectors.

$$\mathbf{M} = \begin{bmatrix} 1 & 2 & 5 \\ 1 & 2 & 4 \\ 1 & 0 & 4 \\ 1 & 0 & 4 \end{bmatrix}$$

First observation here is columns of $\mathbf{M}$ is linearly independent. Then, the next observation is that $e_1 = (1, 0, 0, 0)$ is contained in the column space. Therefore, the *space sparsity number* of the column space of $\mathbf{M}$ is at least equal to 3. Using the fact that *space sparsity number* is less then 4 we conclude that column *space sparsity number* is exactly equal to 3. Lets check the submatirx $\mathbf{M}_{\Omega:}$ where $\Omega = \{2, 3, 4\}$:

$$\mathbf{M}_{\mathbf{\Omega}:} = \begin{bmatrix} 1 & 2 & 4 \\ 1 & 0 & 4 \\ 1 & 0 & 4 \end{bmatrix}$$

Columns of $\mathbf{M}_{\mathbf{\Omega}:}$ is linearly dependent (first and third column), however columns of $\mathbf{M}$ is not. Then, it follows that there is an example that when $\|\Omega\| = \overline{\psi}(\mathbb{U})$ but the hypothesis of the lemma 1 not satisfied. Therefore, the

statement of the lemma 1 is tight.

Following lemma shows the trivial reverse statement of the lemma 1.

Lemma 2. *Let $\mathbb{U}$ be a subspace of $\mathbb{R}^m$ and $x^1, x^2, ..., x^n$ be any set of vectors from $\mathbb{U}$. Then the linear dependence of $x^1, x^2, ..., x^n$ implies linear dependence of $x^1_\Omega, x^2_\Omega, ..., x^n_\Omega$, for any $\Omega \subset [m]$.*

Proof. The proof of the statement is straightforward observation of the fact that

$$\alpha_1 x^1 + ... + \alpha_n x^n = 0$$

implies

$$\alpha_1 x^1_\Omega + ... + \alpha_n x^n_\Omega = 0_\Omega = 0$$

□

Merging the idea of the algorithm $\mathbf{KS2013}$ with the lemma 1 we get an exact completion algorithm. More concretely, setting $d = \psi(\mathbf{M})+1$ is enough to ensure the underlying matrix will be recovered always as it is enough to decide whether partially observed column is contained in the subspace or not. Details of the algorithm $\mathbf{ERCS}$ (Exact recovery with column sparsity) is provided below.

Under the condition that the column space $\mathbb{U}$ of the underlying matrix satisfies $r - 1 = \overline{\psi}(\mathbb{U})$, the observation complexity of the algorithm $\mathbf{ERCS}$ becomes $m \times r + (n - r) \times r = (m + n - r)r$ which is the degree of freedom of the set of $m \times n$ sized rank-r matrices. Therefore, under this condition $\mathbf{ERCS}$ is absolutely optimal as its sample complexity is equal to degree of

freedom of rank-r matrices. We talk about this case further after technical details of the algorithm.

ERCS: Exact recovery with column sparsity

Input: $d = \overline{\psi}(\mathbb{U}) + 1$

Initialize: $k = 0, \widehat{\mathbf{U}}^0 = \emptyset$

1: Draw uniformly random entries $\Omega \subset [m]$ of size d
2: Observe entire $\mathbf{M}_{\Omega:}$
3: **for** i from 1 to n **do**
4: **if** $\|\mathbf{M}_{\Omega:i} - \mathcal{P}_{\mathcal{U}\Omega}\mathbf{M}_{\Omega:i}\| > 0$
5: Fully observe $\mathbf{M}_{:i}$
6: $\widehat{\mathbf{U}}^{k+1} \leftarrow \widehat{\mathbf{U}}^k \cup \mathbf{M}_{:\mathbf{i}}$
7: Orthogonalize $\widehat{\mathbf{U}}^{k+1}$
8: $k = k + 1$
9: **otherwise:**
10: $\widehat{\mathbf{M}}_{:i} = \widehat{\mathbf{U}}^k \widehat{\mathbf{U}}_{\Omega:}^{k+} \widehat{\mathbf{M}}_{\Omega:i}$

Output: Underlying matrix $\widehat{\mathbf{M}}$

Theorem 1. *Let $\mathbb{U}$ represent the column space of the $m \times n$ sized matrix $\mathbf{M}$ of rank r. Then,* **ERCS** *exactly recovers $\mathbf{M}$ by*

$$m \times r + (n - r)(\overline{\psi}(\mathbb{U}) + 1)$$

observations.

Proof. We start by showing **ERCS** recovers $\mathbf{M}$ exactly and later we focus on observation count. To prove correctness of exact recovery we use mathematical induction as follow:

$Hypothesis$: after i-th iteration **ERCS** already correctly recovered first i columns.

$Base\ case$: $i = 1$ is trivial as if at least one of the observed entries is nonzero we completely observe the column, which guarantees correctness.

On the other hand, if it happens all of $\overline{\psi}(\mathbb{U}) + 1$ entries are zero, then the first column is indeed completely zero because the definition of the *space sparsity number* implies there can be at most $\psi(\mathbb{U})$ many zero coordinates in a nonzero vector in the column space.

$Hypothesis\ proof$: Let assume after step $i-1$, **ERCS** recovered first $i-1$ columns correctly and we want to show the algorithm exactly recovers i-th column too.

From the design of the algorithm $\mathbf{M}_{\Omega:i}$ is already observed. Then, if in the line 4, **ERCS** decides the column is linearly independent with previous columns, as in the next line we completely observe the column there is no chance that the algorithm can do mistake under this case. Therefore, the only remaining case is, if in the line 4 the algorithm decides the column i is linearly dependent.

From the statement of lemma 1, if a set of vectors from a subspace $\mathbb{U}$ are linearly dependent on a given subset of coordinates, then they are indeed linearly dependent. We conclude that the algorithm's decision is correct and by just back projection method, the algorithm recovers remaining entries of the partially observed column. Therefore, column i also recovered correctly and we are done with the proof of induction hypothesis.

Our next goal is to show the observation complexity is $m \times r + (n - r)(\overline{\psi}(\mathbb{U})+1)$. From the lemma 1, we conclude that whenever current column is indeed linearly independent with previous columns, the **ERCS** also de-

cides it is linearly independent. Moreover, from the lemma 10, we conclude that if the current column is linearly dependent with previous columns, then in this case **ERCS** decides it is linearly dependent. As there are r many linearly independent columns in the underlying matrix $\mathbf{M}$, the algorithm decides independence exactly r times and in each of them it does complete observations. However, in remaining $n - r$ columns, number of observations is exactly $\overline{\psi}(\mathbb{U}) + 1$. As a conclusion, number of total observations is: $rm + (n - r)(\overline{\psi}(\mathbb{U}) + 1)$

□

Corollary 1. **ERCS** *still performs correctly under the case* d *set to be any number larger than* $\overline{\psi}(\mathbb{U}) + 1$. *The only difference is that the updated observation complexity will be* $m \times r + (n - r)d$

Proof. The proof is exactly proceeds as proof of the theorem. The key point is to notice, lemma 1 and lemma 2 are still satisfying.

□

Notice that both of algorithms **KS2013** and **ERCS** have two stages of observations.

i : Select subset of rows and observe them completely.

ii : Detect linearly independent columns and observe them completely.

The discussion for tightness of the lemma 1 above implies that **ERCS** is optimal deterministic two stage observation algorithm. Moreover, in the

corollary above we discussed for any $d \geq \overline{\psi}(\mathbb{U}) + 1$ the algorithm **ERCS** algorithm would still perform correctly. Therefore, having constant factor approximation of the *space sparsity number* of the column space would lead asymptotically optimal algorithm:

$$\begin{aligned} \widehat{d} \leq K\overline{\psi}(\mathbb{U}) &\implies \widehat{d}+1 \leq K(\overline{\psi}(\mathbb{U})+1) \\ &\implies (n-r)(\widehat{d}+1) \leq (n-r)K(\overline{\psi}(\mathbb{U})+1) \end{aligned}$$

adding mr to both side leads to

$$\begin{aligned} rm + (n-r)(\widehat{d}+1) &\leq rm + (n-r)K(\overline{\psi}(\mathbb{U})+1) \\ &\leq Krm + K(n-r)(\overline{\psi}(\mathbb{U})+1) \\ &= K\big(rm + (n-r)(\overline{\psi}(\mathbb{U})+1)\big) \end{aligned}$$

Notice that $rm+(n-r)(\widehat{d}+1)$ is the observation complexity we have once we have $\widehat{d}$ as K-approximation of $\overline{\psi}(\mathbb{U})$ and $rm + (n-r)(\overline{\psi}(\mathbb{U})+1)$ is the complexity for optimal two stage as we discussed here. All together, the inequality above implies constanct approximation to *space sparsity number* gives as constant approximation to optimal solution.

In the following lemma we study possible values of the *sparsity number*:

Lemma 3. *For the column space* $\mathbb{U}$ *of* $m \times n$ *sized matrix* $\mathbf{M}$ *with rank-*r*, the following inequality is satisfied*

$$r - 1 \leq \overline{\psi}(\mathbb{U}) \leq m - 1.$$

Proof. $\overline{\psi}(\mathbb{U}) < m$ is straightforward because any nonzero vector in $\mathbb{R}^m$ has at most $m-1$ coordinates equal to zero. Then, it follows from the definition of the *space sparsity number*, $\overline{\psi}(\mathbb{U}) \leq m-1$.

In the rest of the proof we prove $r-1 \leq \overline{\psi}(\mathbf{M})$. $\mathbf{M}$ having rank r implies that we can choose r rows from it that are the basis for the row space of it. Technically, we may find $R \subset [m]$ such that $|R| = r$ and $\mathbf{M}_{R:}$ is rank r. Similarly, we can find $C \subset [n]$ such that $|C| = r$ and $\mathbf{M}_{R:C}$ is an $r \times r$-sized matrix of rank r. It follows that there exists $\alpha \in \mathbb{R}^r$ such that

$$\mathbf{M}_{R:C}\alpha = e_1 = \big(1, 0, \ldots, 0\big).$$

Consequently, $\mathbf{M}_{:C}\alpha \neq 0$ but has zero components in $r-1$ of the indices given by R. Thus , we have:

$$\overline{\psi}(\mathbf{M}_{:C}) \geq r-1$$

which together with upper bound concludes the statement of the lemma. □

It is easy to construct examples to show both sides of the inequality is tight (i.e. equality satisfied). For any matrix that $e_1 = \big(1, 0, 0, \ldots, 0\big)$ is contained in the column space, $\overline{\psi}(\mathbb{U}) = m-1$ is trivially correct. Moreover, column space of the general rank-r matrices creasted as $\mathbf{M} = \mathbf{XY}$ where $\mathbf{X} \in \mathbb{R}^{m\times r}$ and $\mathbf{X}_{i,j} \sim \mathcal{N}(0,1)$, $\mathbf{Y} \in \mathbb{R}^{r\times n}$ and $\mathbf{Y}_{i,j} \sim \mathcal{N}(0,1)$ left side of the inequality is tight (i.e. $r-1 = \overline{\psi}(\mathbb{U})$).

2.1.1 Heterogeneous Observation Cost Exact Recovery

In this section, we discuss the completion problem where entries of the matrix has non uniform cost to observe. We study two types of heterogeneous cost model:

- Each column has its own/different observation cost and entries of the same column has the same cost.
- Each entry of the matrix has different cost.

Uniform Cost Across Columns

Problem: For any fixed j, the cost of observing $\mathbf{M}_{i:j}$ is equal to χ_j for any $1 \leq i \leq m$, and $\chi_1, \chi_2, \ldots, \chi_n$ are arbitrary positive numbers and we target to recover the matrix $\mathbf{M}$ as cheap as possible.

Solution: We propose a slight modification of the $\mathbf{ERCS}$ to solve optimally among the two staged methods as we discussed before. Lets remind that in the algorithm we show that selecting any $d = \overline{\psi}(\mathbb{U}) + 1$ many rows is enough to guarantee exact recovery deterministically. In the next stage, we iteratively go through columns one by one starting with the first column, and if we detect a column is linearly independent with previous ones, we completely observe it. If not, we recover it using the pre-determined subspace.

To adapt the solution for this problem, we just need to change the order of the columns we start to check. Basically, instead of starting with the first

column, we should start with the cheapest one. If we decide its not contained in the current subspace, we completely observe all entries and if it is contained then we just recover with the current subspace. Then, we move to second cheapest column and so on so forth with the increasing order of cost.

Correctness: We can see that the proof of the correctness of **ERCS** is independent of the order of the columns. Therefore, selecting columns with increasing order of the cost would not change the correctness of the algorithm.

Optimality: The set of two stage algorithm can be parametrized by two numbers. First one is - d- the number of rows fully observed and the second is the subset of indices of columns to observe fully. We analyse the optimal algorithm for three cases of values of d:

1. $d \leq \overline{\psi}(\mathbb{U})$. It is obvious that optimal algorithm cannot have $d \leq \overline{\psi}(\mathbb{U})$, because from the discussion for tightness of the lemma 1 and optimality of **ERCS**, there are matrices that selection of $d = \overline{\psi}(\mathbb{U})$ rows is not enough to guarantee the existence of r linearly independent rows.

2. $d = \overline{\psi}(\mathbb{U})+1$ It is a well known fact that the set of column basises are matroids and Greedy algorithms gives the optimal solution for matroids [26]. Note that the algorithm designed above is efficient way of giving greedy solution.

3. $d > \overline{\psi}(\mathbb{U}) + 1$. Lets assume that the optimal algorithm takes $\tilde{d} > d$ rows

in the first phases and columns: $i_1, i_2, \ldots, i_{\tilde{r}}$. We first note that, $\tilde{r} = r$, it is because if $\tilde{r} < r$ then selected columns are not enough to learn the column space and if $\tilde{r} > r$ we can pick subset of these columns that is basis for column space and selecting this basis has less cost which contradicts to optimality. Therefore, $\tilde{r} = r$ for optimal case. Moreover, we can use the same subset selection argument to pick $\overline{\psi}(\mathbb{U}) + 1$ sized subset of rows then select the same set of columns and it will be cheaper. Therefore, for optimality we should select exactly $d = \overline{\psi}(\mathbb{U}) + 1$ rows.

Exact recovery with full heterogeneity: For any given i, j, the cost of observing the entry $\mathbf{M}_{i:j}$ is equal to χ_{ij} and $\chi_{11}, \chi_{12}, \ldots, \chi_{mn}$ are arbitrary positive numbers and similar to the previous problem we target to recover the matrix $\mathbf{M}$ as cheap as possible.

ERHC: Exact recovery with heterogeneous cost

Input: $d = \overline{\psi}(\mathbb{U}) + 1$ here $\mathbb{U}$ is the column space of the underlying matrix
Initialize: $\widehat{\mathbf{M}}$ set to $m \times n$ sized null matrix, $\widehat{\mathbf{U}}^0 = \emptyset$, $k = 0$

for i from 1 to m, **do**
  $\chi^i = \sum_{j=1}^{n} \chi_{ij}$
Sort χ^is with increasing order and select first d and denote their index set by$-R$
Observe entire $\mathbf{M}_{R:}$
for i from 1 to m, **do**
  $\overline{\chi}^i = \sum_{j \in [n] \setminus R} \chi_{ij}$
Sort $\overline{\chi}^i$s with increasing order and lets denote $\{i_1, i_2, \ldots, i_n\}$ as $\overline{\chi}^{i_1} \leq \overline{\chi}^{i_2} \leq \ldots \leq \overline{\chi}^{i_n}$
for h from 1 to m, **do**
    If $\|\mathbf{M}_{R:i_h} - \mathcal{P}_{\widehat{\mathbf{U}}_R^k} \mathbf{M}_{R:i_h}\|^2 > 0$
      Fully observe $\mathbf{M}_{:i_h}$ add it to the basis $\widehat{\mathbf{U}}^k$
      Orthogonalize $\widehat{\mathbf{U}}^k$
      $k = k + 1$
    Otherwise: $\widehat{\mathbf{M}}_{:i_h} = \widehat{\mathbf{U}}^k \widehat{\mathbf{U}}_{R:}^{k+} \widehat{\mathbf{M}}_{R:i_h}$
return $\widehat{\mathbf{M}}$

Output: Underlying matrix $\widehat{\mathbf{M}}$

Solution: We describe the solution in the algorithm $\mathbf{ERHC}$:

Correctness: We can see the correctness of $\mathbf{ERHC}$ is due to the correctness of $\mathbf{ERCS}$ as selecting cheapest $\overline{\psi}(\mathbb{U}) + 1$ is special case of selecting any $\overline{\psi}(\mathbb{U}) + 1$ many columns and iteration order over the columns doesn't matter similarly for this case too.

Optimality: Unlike to the previous case, greedy algorithm doesn't give us the cheapest combination of columns and rows. Following example provides a matrix and entry costs that shows that greedy algorithm is not optimal.

$$\mathbf{M} = \begin{bmatrix} 1 & 1 & 2 & 3 \\ 1 & 2 & 3 & 4 \\ 1 & 3 & 4 & 5 \\ 1 & 4 & 5 & 6 \end{bmatrix} \qquad \chi = \begin{bmatrix} 1 & 1 & 4 & 1 \\ 1 & 5 & 3 & 4 \\ 4 & 3 & 4 & 4 \\ 1 & 4 & 4 & 8 \end{bmatrix}$$

The greedy algorithm for this case observes rows $R = \{1, 2\}$ and columns $C = \{1, 2\}$ which has overall cost of:

$$(1 + 1 + 4 + 1) + (1 + 5 + 3 + 4) + (4 + 1) + (3 + 4) = 32$$

However, observing $R = \{1, 3\}$ and columns $C = \{1, 3\}$ would give us overall cost of:

$$(1 + 1 + 4 + 1) + (4 + 3 + 4 + 4) + (1 + 1) + (3 + 4) = 31$$

which is cheaper than greedy algorithm.

However, with the same cost matrix, there are other matrices that shares the same column space as $\mathbf{M}$ (therefore the same column space sparsity

number) but greedy algorithm is still optimal. For the same cost matrix with a slightly modified underlying matrix, we can give an example:

$$\overline{\mathbf{M}} = \begin{bmatrix} 1 & 1 & 2 & 2 \\ 1 & 2 & 2 & 3 \\ 1 & 3 & 2 & 4 \\ 1 & 4 & 2 & 5 \end{bmatrix} \qquad \chi = \begin{bmatrix} 1 & 1 & 4 & 1 \\ 1 & 5 & 3 & 4 \\ 4 & 3 & 4 & 4 \\ 1 & 4 & 4 & 8 \end{bmatrix}$$

This gives us the conclusion, with just information of the observation cost matrix and column space sparsity number, we cannot pick theoretical optimal set of rows and columns that is guaranteed carrying all of information of the underlying matrix.

2-Optimality: Even though greedy algorithm cannot return the optimal set of rows and columns, here we show that the overall cost of the cost oof the algorithm is at most twice expensive than optimal.

We denote the row set and column set parameter of optimal 2-stage algorithm $\tilde{R}$ and $\tilde{C}$ and cost of it by σ_{OPT}. Then, we can decompose optimal soluton into its parts as following:

$$\sigma_{OPT} = \chi(\mathbf{M}_{\tilde{R}:}) + \chi(\mathbf{M}_{:\tilde{C}}) - \chi(\mathbf{M}_{\tilde{R}:\tilde{C}})$$

Trivially, both of the following inequalities satisfied

$$\chi(\mathbf{M}_{\tilde{R}:\tilde{C}}) \leq \chi(\mathbf{M}_{:\tilde{C}})$$
$$\chi(\mathbf{M}_{\tilde{R}:\tilde{C}}) \leq \chi(\mathbf{M}_{\tilde{R}:})$$

which these inequalities implies that

$$\sigma_{OPT} \geq \max(\chi(\mathbf{M}_{:\tilde{C}}), \chi(\mathbf{M}_{\tilde{R}:})).$$

Now lets decompose cost of greedy algorithm to its pieces:

$$\sigma_G = \chi(\mathbf{M}_{R:}) + \chi(\mathbf{M}_{:C}) - \chi(\mathbf{M}_{R:C})$$

Note that the greedy algorithm doesn't necessarily selects cheapest basis columns, however selected columns minimizes the overall cost after rows selected. Therefore, we conclude that if we denote the set of cheapest columns by C^B, then the following inequality satisfied:

$$\begin{aligned}\sigma_G = \chi(\mathbf{M}_{R:}) + \chi(\mathbf{M}_{:C}) - \chi(\mathbf{M}_{R:C}) &\leq \chi(\mathbf{M}_{R:}) + \chi(\mathbf{M}_{:C^B}) - \chi(\mathbf{M}_{R:C^C}) \\ &\leq 2\max\big(\chi(\mathbf{M}_{R:}), \chi(\mathbf{M}_{:C^B})\big)\end{aligned}$$

As we discussed before in order to have guarantee that we will be able to have full information to detect linearly independent columns we need to observe at least $\psi(\mathbb{U}) + 1$ many rows.

Moreover, as greedy algorithm observe exactly $\overline{\psi}(\mathbb{U}) + 1$ many rows by choosing cheapest columns we are guaranteed to have:

$$\chi(\mathbf{M}_{R:}) \leq \chi(\mathbf{M}_{\tilde{R}:})$$

Similarly as C^B represents the set of cheapest columns, we have:

$$\chi(\mathbf{M}_{:C^B}) \leq \chi(\mathbf{M}_{:\tilde{C}})$$

which together implies

$$\max\big(\chi(\mathbf{M}_{R:}), \chi(\mathbf{M}_{:C^B})\big) \leq \max\big(\chi(\mathbf{M}_{\tilde{R}:}), \chi(\mathbf{M}_{:\tilde{C}})\big).$$

Putting all inequalities together we conclude:

$$\sigma_G \leq 2\max\big(\chi(\mathbf{M}_{R:}), \chi(\mathbf{M}_{:C^B})\big) \leq 2\max\big(\chi(\mathbf{M}_{\tilde{R}:}), \chi(\mathbf{M}_{:\tilde{C}})\big) \leq 2\sigma_{OPT}$$

Therefore, we conclude that greedy algorithm gives us 2-optimal algorithm.

Tightness: In the following example, we see that greedy algorithm cannot guarantee better than 2-optimality:

$$\chi = \begin{bmatrix} \frac{\epsilon}{100} & \frac{\epsilon}{100} & \frac{\epsilon}{100} & \frac{\epsilon}{100} & 10-\epsilon & 10-\epsilon \\ \frac{\epsilon}{100} & \frac{\epsilon}{100} & \frac{\epsilon}{100} & \frac{\epsilon}{100} & 10-\epsilon & 10-\epsilon \\ 10 & 10 & \frac{\epsilon}{100} & \frac{\epsilon}{100} & \frac{\epsilon}{100} & \frac{\epsilon}{100} \\ 10 & 10 & \frac{\epsilon}{100} & \frac{\epsilon}{100} & \frac{\epsilon}{100} & \frac{\epsilon}{100} \\ \frac{\epsilon}{100} & \frac{\epsilon}{100} & 10 & 10 & 10-\epsilon & 10-\epsilon \\ \frac{\epsilon}{100} & \frac{\epsilon}{100} & 10 & 10 & 10-\epsilon & 10-\epsilon \end{bmatrix}$$

It is clear that optimal choice is $C = \{1, 2\}$ and $R = \{3, 4\}$ which gives the cost of :

$$\sigma_{OPT} = 10 + 10 + 10 + 10 + 16 \times \frac{\epsilon}{100} = 40 + \frac{\epsilon}{6.25}$$

However, greedy algortihm will pick $R = \{1, 2\}$ in the first stage which has overall cost of

$$(10-\epsilon)+(10-\epsilon)+(10-\epsilon)+(10-\epsilon)+8\times\frac{\epsilon}{100}=40-4\epsilon+\frac{\epsilon}{12.5}$$

Then in the next stage it choose columns $C=\{5,6\}$ which also has cost of

$$(10-\epsilon)+(10-\epsilon)+(10-\epsilon)+(10-\epsilon)+4\times\frac{\epsilon}{100}=40-4\epsilon+\frac{\epsilon}{25}$$

which all together cumulative cost is

$$(40-4\epsilon)+\frac{\epsilon}{12.5}+(40-4\epsilon)+\frac{\epsilon}{25}=80-8\epsilon+\frac{3}{25}\epsilon.$$

To find the fraction of this cost to optimal cost we get

$$\frac{\sigma_G}{\sigma_{OPT}}=\frac{80-8\epsilon+\frac{2\epsilon}{25}}{40+\frac{\epsilon}{6.25}}\approx 2-\frac{\epsilon}{5}.$$

Therefore for any number smaller than 2, we can choose an ϵ which ratio of the cost of greedy algorithm to optimal set is larger than that number. This implies that, 2-optimality of the algorithm $\mathbf{ERHC}$ is tight.

2.2 Fixed Rank Case

Here, we provide an algorithm that exactly recovers a target matrix under the active setting. As we discussed before, one of the strengths of the results of [15, 16, 2] is that the algorithm is its robustness to highly coherent row space compared to previous results as [23]. However, we notice that these algorithms are independent of row space and treat any row space equally. This phenomenon arises a natural question, whether there is an algorithm which enjoys properties of row space to optimize these algorithms further. For example, we can see the following matrices having the same rank $r = 2$ and column space coherence $\mu_0 = 2$. The only difference is due to the coherence of row space, which is 3 for $\mathbf{A}$ and near to 1 for matrix $\mathbf{B}$. Similarly, row space *sparsity-number* is 1 for the matrix $\mathbf{A}$ and 4 for the matrix $\mathbf{B}$.

$$\mathbf{A} = \begin{bmatrix} 1 & 2 & 2 & 2 & 2 & 2 \\ 0 & 2 & 2 & 2 & 2 & 2 \\ 0 & 2 & 2 & 2 & 2 & 2 \\ 0 & 2 & 2 & 2 & 2 & 2 \end{bmatrix} \qquad \mathbf{B} = \begin{bmatrix} 1 & 0 & 1 & 2 & 3 & 4 \\ 0 & 1 & 2 & 3 & 4 & 5 \\ 0 & 1 & 2 & 3 & 4 & 5 \\ 0 & 1 & 2 & 3 & 4 & 5 \end{bmatrix}$$

As previous methods are mainly based on the value of the size of the matrix, r, and μ_0, these methods treat both these matrices equally. Indeed the first column of the matrix $\mathbf{A}$ is crucial to study the column space. That's why we don't want to take the risk of missing the necessary information on this column. That's why we end up observing many entries in each column to make sure we will not miss this column. However, it is entirely different for

the matrix $\mathbf{B}$; any missed column can be replaced by any other one in the study of the column space. That's why it should give us the flexibility of observing less number of entries in each column.

In the following algorithm, we propose a method that exactly recovers the $m \times n$ sized rank r underlying matrix, using just exact information of rank r. The idea of the algorithm is to find r-many linearly independent rows and columns and recover the remaining entries based on them. Finding the linearly independent columns is rely on simple linear algebra fact that if columns of $\mathbf{M}_{R:C}$ are linearly independent, then so are columns of $\mathbf{M}_{:C}$. The statement is valid for rows as well symmetrically. Then, all we need to do is to wait for detecting r many independent columns and rows. Indeed the algorithm does not require an estimate or exact information of coherence as opposed to $\mathbf{KS2013}$, [2, 16] (coherence is crucial to compute d in input phase). The improvement for observation complexity can be observed in the following theorem and corollary.

ERR: Exact recovery for rank r matrices.

Input: Rank of the target matrix - r
Initialize: $R = \emptyset$, $C = \emptyset$, $\widehat{r} = 0$

while $\widehat{r} < r$ **do**
  for j from 1 to n **do**
    Uniformly pick an unobserved entry i from $\mathbf{M}_{:j}$
    $\widehat{R} = R \cup \{i\}, \widehat{C} = C \cup \{j\}$
    If $\mathbf{M}_{\widehat{R}:\widehat{C}}$ is nonsingular
      Fully observe $\mathbf{M}_{:j}$ and $\mathbf{M}_{i:}$
      Set: $R = \widehat{R}$, $C = \widehat{C}$, $\widehat{r} = \widehat{r} + 1$
Orthogonalize column vectors in C and assign to $\widehat{\mathbf{U}}$
for each column $j \in [n] \setminus C$ **do**
  $\widehat{\mathbf{M}}_{:j} = \widehat{\mathbf{U}}\widehat{\mathbf{U}}^{+}_{R:}\widehat{\mathbf{M}}_{R:j}$

Output: $\widehat{\mathbf{M}}$

Theorem 2. *Let r be the rank of underlying $m \times n$ sized matrix $\mathbf{M}$ with column space $\mathbb{U}$ and row space $\mathbb{V}$. Then, $\mathbf{ERR}$ exactly recovers the underlying matrix $\mathbf{M}$ with probability at least $1-\epsilon$ using number of observations at most:*

$$(m+n-r)r + \min\Big(2\frac{mn}{\psi(\mathbb{U})}\log\big(\frac{r}{\epsilon}\big), \frac{\frac{2m}{\psi(\mathbb{U})}(r+2+\log\frac{1}{\epsilon})}{\psi(\mathbb{V})}n\Big).$$

Corollary 2. *Observation complexity of $\mathbf{ERR}$ studied for three different case below:*

- *if $\psi(\mathbb{V}) = \mathcal{O}(1)$ satisfies, then observation complexity is upper bounded by $(m+n-r)r + 2\frac{mn}{\psi(\mathbb{U})}\log\big(\frac{r}{\epsilon}\big) = (m+n-r)r + \mathcal{O}\big(nr\mu_0\log\big(\frac{r}{\epsilon}\big)\big)$. (this bound matches with [2], however in many cases it is much smaller as discussed in the next section)*

- *if $\psi(\mathbb{V}) = \Theta(r)$ satisfies, then observation count is upper bounded by $(m+n-r)r + \mathcal{O}\Big(\frac{\frac{2m}{\psi(\mathbb{U})}(r+2+\log\frac{1}{\epsilon})}{r}n\Big) = (m+n-r)r + \mathcal{O}\big(n\mu_0(r+\log\frac{1}{\epsilon})\big)$. Selecting $\epsilon = \frac{1}{2^{\mathcal{O}(r)}}$ gives the bound of: $mr + \mathcal{O}(n\mu_0 r)$*

- *if $\psi(\mathbb{V}) = \Theta(n)$ satisfies, then observation count is upper bounded by $(m+n-r)r + \mathcal{O}\big(\frac{2m(r+2+\log\frac{1}{\epsilon})}{\psi(\mathbb{U})}\big) = (m+n-r)r + \mathcal{O}\big(\mu_0 r(r+\log\frac{1}{\epsilon})\big)$ Selecting $\epsilon = \frac{1}{2^{\mathcal{O}(r)}}$ gives bound: $\mathcal{O}((m+n-r)r)$*

Note that, for the last case, observation complexity is bounded by $\mathcal{O}((m+n-r)r)$, which is absolute lower bound for any algorithm.

We split the statement of the theorem above into two and prove each of them

separately. First, we show that observation complexity is upper bounded by

$$(m+n-r)r + 2\frac{mn}{\psi(\mathbb{U})}\log(\frac{r}{\epsilon}).$$

Then, later we show that the observation complexity is bounded by

$$(m+n-r)r + \frac{\frac{2m}{\psi(\mathbb{U})}(r+2+\log\frac{1}{\epsilon})}{\psi(\mathbb{V})}$$

and the statement of theorem follows from these two results.

2.2.1 Matrices with low row space sparsity-number

In this section, we prove that the observation complexity of the algorithm **ERR** is always upper bounded by

$$(m+n-r)r + 2\frac{mn}{\psi(\mathbb{U})}\log(\frac{r}{\epsilon})$$

Proof. The proof is consisting following steps:

- *step 1.* Give terminology will be used throughout the proof. Identifying type of observations to two classes : informative and non-informative.
- *step 2.* Provide a bound to number of informative observations.
- *step 3.* In remaining steps, we try to give bound to non-informative observations. We start by giving upper bound to the unsuccessful observations in line 5.
- *step 4.* We model the execution of **ERR** with a stochastic process and design another process which terminates faster than this.

- *step 5.* We relate the problem to basic combinatorial counting problem and analyse
- *step 6.* Conclude that total number of observations is

$$(m+n-r)r + 2\frac{mn}{\psi(\mathbb{U})}\log(\frac{r}{\epsilon}).$$

Step 1: For ease of readability we denote $\psi(\mathbb{U})$ by k during the proof. Lets start a process in the beginning of the algorithm for each column. We call process of the column $\mathbf{M}_{:j}$ dies in one of the following cases happens:

a. $\mathbf{M}_{:j}$ is fully observed in line 6 in some intermediate step of $\mathbf{ERR}$

b. $\mathbf{M}_{:j}$ is contained in the column space of the already fully observed columns in underlying matrix $\mathbf{M}$ (i.e. columns in C).

c. Algorithm already learns entire column space : $\widehat{r} = r$

If a column/process is not dead then we call it is active. We call an observation is informative if it is observed at line 6 (i.e. it contributes to the studied column/row space learned by $\mathbf{ERR}$ and uninformative if it observed at line 3. Obviously some entries are observed both at line 3 and 6, so they count in both non-informative and informative observations.

Step 2: We can simply observe that the number of informative observations is exactly $mr + nr - r^2$. Because, at the end of the algorithm set of informative observations is just set of r many linearly independent columns (we have mr observations here) and r many linearly independent rows (we have nr observations here). By observing that entries in $r \times r$ sub-matrix is

counted twice, we conclude that overall observations is just : $mr + nr - r^2$

Step 3: In order to give upper bound to the number of non-informative observations, we see it is enough to bound the number of phases the algorithm **ERR** passes through. Specifically, if the number of phases is bounded by T then overall number of non-informative observations is bounded by Tn_2. In order to give upper bound to T, we first explore the probability of an detecting independence of an observation in line 3 for an active column:

Lemma 4. *The probability of detecting independence of an active column in the j'th phase of the algorithm* **ERR** *is lower bounded by* $\frac{k}{m-j}$

Proof. In an intermediate step of **ERR** we have $|C| = |R| = \widehat{r}$ and $M_{R:C}$ is $\widehat{r} \times \widehat{r}$ matrix of rank $\widehat{r}$. Then, for any $i \in [n]$, $\mathbf{M}_{R:i}$ is in the column space of $\mathbf{M}_{R:C}$ as the matrix is full rank and therefore its column space is entire $\mathbb{R}^{\widehat{r}}$. Then there exists unique coefficients $\alpha_1, \alpha_2, ..., \alpha_{\widehat{r}}$ for columns $C = \{c_1, ..., c_{\widehat{r}}\}$ that following equality satisfied.

$$\begin{aligned} &\alpha_1 \mathbf{M}_{R:c_1} + \ldots + \alpha_{\widehat{r}} \mathbf{M}_{R:c_{\widehat{r}}} = \mathbf{M}_{R:i} \\ \implies &\alpha_1 \mathbf{M}_{R:c_1} + \ldots + \alpha_{\widehat{r}} \mathbf{M}_{R:c_{\widehat{r}}} - \mathbf{M}_{R:i} = 0 \end{aligned}$$

Now, let observe the vector

$$y = \alpha_1 \mathbf{M}_{:c_1} + \alpha_2 \mathbf{M}_{:c_2} + ... + \alpha_{\widehat{r}} \mathbf{M}_{:c_{\widehat{r}}} - \mathbf{M}_{:i}$$

We know that $y \neq 0$ because we know column i is linearly independent with

previous observed columns - $\mathbf{M}_{:c_1}, \mathbf{M}_{:c_2}, ..., \mathbf{M}_{:c_{\hat{r}}}$. Moreover for any row index $a \in R$, $y_a = 0$ because $y_R = 0$ from the definition of α_j's. As y is in the column space of M it has at most $m - k$-many zero coordinates. Moreover, for any $a \notin R$ but $\mathbf{M}_{i_a:i}$ observed in line 6 $y_{i_a} = 0$ should satisfy, because otherwise in one of previous iterations we would already decide $\mathbf{M}_{:i}$ is linearly independent and we would add index i to C, but here we know $i \notin C$.

Basically we conclude that all known coordinates of y is 0 and number of known coordinates is represented by $observed$. We know at least k many coordinates of y is nonzero and we already have $m - observed$ many coordinates of y is zero, then with probability at least: $\frac{k}{m-observed}$ uniformly selected next observation will be zero. Being nonzero of y_a implies nonsingularity of the matrix $\widehat{\mathbf{M}}_{\widehat{R}:\widehat{C}}$ where $\widehat{R} = R \cup \{a\}$ and $\widehat{C} = C \cup \{i\}$. It is because if this matrix was not invertible then there would be coefficients $\beta_1, ..., \beta_{r'+1}$ (not all of them are zero) such that

$$\beta_1 \mathbf{M}_{\widehat{R}:c_1} + ... + \beta_{\hat{r}} \mathbf{M}_{\widehat{R}:c_{\hat{r}}} + \beta_{\hat{r}+1} \mathbf{M}_{\widehat{R}:i} = 0.$$

From lemma 2, linear independence of $\mathbf{M}_{R:c_1}, \mathbf{M}_{R:c_{\hat{r}}}$ implies linear independence of $\mathbf{M}_{\widehat{R}:c_1}, \mathbf{M}_{\widehat{R}:c_{\hat{r}}}$. which concludes $\beta_{\hat{r}+1}$ is nonzero, so we can simply assume it is -1. Then

$$\begin{aligned} &\beta_1 \mathbf{M}_{\widehat{R}:c_1} + ... + \beta_{\hat{r}} \mathbf{M}_{\widehat{R}:c_{\hat{r}}} = \mathbf{M}_{\widehat{R}:i} \\ \implies &\beta_1 \mathbf{M}_{R:c_1} + ... + \beta_{\hat{r}} \mathbf{M}_{R:c_{\hat{r}}} = \mathbf{M}_{R:i} \end{aligned}$$

Due to uniqueness of α_j's above, we can tell that

$$\alpha_1 = \beta_1 \qquad \alpha_2 = \beta_2 \qquad \dots \qquad \alpha_{\widehat{r}} = \beta_{\widehat{r}} \qquad \alpha_{\widehat{r}} = \beta_{\widehat{r}}$$

Then the vector $y_{\widehat{R}} = \alpha_1 \mathbf{M}_{\widehat{R}:c_1} + \ldots + \alpha_{\widehat{r}} \mathbf{M}_{\widehat{R}:c_{\widehat{r}}} - \mathbf{M}_{\widehat{R}:i} = 0$ is a zero vector. However it is a contradiction because if $y_{\widehat{R}}$ is zero vector then $y_a = 0$ due to $a \in \widehat{R}$ which we already know $y_a \neq 0$.

Therefore, being nonzero of y_a implies non-singularity of $\mathbf{M}_{\widehat{R}:\widehat{C}}$ which is equivalent to the detection of the independence of column $\mathbf{M}_{:i}$ due to lemma 2. As a conclusion, probability of detection of independence of an active column is at least $\frac{k}{m-observed}$ and considering the fact that $observed \geq j$ it follows that $\frac{k}{m-observed} > \frac{k}{m-j}$ and it give the final conclusion of the desired probability is lower bounded by: $\frac{k}{m-j}$. As desired. □

Step 4 : We can model execution of $\mathbf{ERR}$ as following stochastic process:

$$S_0 = X_{0,1} + X_{0,2} + ... + X_{0,n}$$

where each of the $X_{0,j}$ corresponds to the indicator variable of the activeness of the column $\mathbf{M}_{:j}$. Obviously, initially at least r of these random variables are equal to 1. We define S_1 similarly:

$$S_1 = X_{1,1} + X_{1,2} + ... + X_{1,n}$$

and for any j that $X_{0,j} = 1$ satisfied, at this phase $X_{1,j}$ will be equal to 0 with probability at least $\frac{k}{m-0}$ from lemma 4. For remaining j's that $X_{0,j} = 0$

satisfied then $X_{1,j} = 0$ also to be satisfied. For the next step S_2 defined as:

$$S_2 = S_1 + X_{2,1} + X_{2,2} + ... + X_{2,n_2}$$

where again for any j that $X_{0,j} = 1$ satisfied, at this phase $X_{1,j}$ will be equal to 0 with probability at least $\frac{k}{m-1}$ from lemma 4. Remaining j's will stay as $X_{2,j}$ to be equal to 0. In general

$$S_{iter} = X_{iter,1} + X_{iter,2} + ... + X_{iter,n_2}$$

where again for any j that $X_{iter-1,j} = 1$ satisfied, at this phase $X_{iter,j}$ will be equal to 0 with probability at least $\frac{k}{m-(iter-1)}$ from lemma 4. The termination of algorithm is equivalent to the point $S_p = 0$ in this model. One can see termination of this process is upper bounded by termination of the following process :

$$S'_0 = X'_{0,1} + X'_{0,2} + ... + X'_{0,r}$$

where each of the $X'_{0,j}$ is set to be equal to 1. we define S'_1 in a similar way:

$$S'_1 = X'_{1,1} + X'_{1,2} + ... + X'_{1,r}$$

where each of $X'_{1,j}$ is equal to 0 with probability $\frac{k}{m}$. Then:

$$S'_2 = X'_{2,1} + X'_{2,2} + ... + X'_{2,r}.$$

Similarly $X'_{2,j}$ is set to be 0 if $X'_{1,j} = 0$ and $X'_{2,j}$ is equal to 0 with probability

$\frac{k}{m-1}$ otherwise. In general:

$$S'_{iter} = X'_{iter,1} + X'_{iter,2} + ... + X'_{iter,r}$$

again $X'_{iter,j} = 0$ if $X'_{iter-1,j} = 0$ and, $X'_{iter,j} = 0$ with probability $\frac{k}{m-(iter-1)}$ otherwise.

Step 5: Here we use a combinatorial argument to bound number of observation in each column.

Lemma 5. *Let X' be a process that is zero initially: $X'_0 = 1$ and remaining entries defined as*

$$X'_{i+1} = \begin{cases} \begin{cases} 0 & \textit{with probability } \frac{k}{m-i} \\ 1 & \textit{otherwise} \end{cases} & \textit{if} \quad X'_i = 1 \\ 0 & \textit{if} \quad X'_i = 0 \end{cases}$$

Then expected point that X' to switch to 0 is $\frac{m+1}{k+1}$.

Proof. Lets denote the switch time with st and write the expression for it:

$$\begin{aligned} \mathbb{E}[st] &= \sum iP(st = i) \\ &= 1\frac{k}{m} + 2\frac{k}{m-1}\Big(1 - \frac{k}{m}\Big) + 3\frac{k}{m-2}\Big(1 - \frac{k}{m}\Big)\Big(1 - \frac{k}{m-1}\Big) + \dots \end{aligned}$$

We claim that this sum is equal to the expected position of the first 1 in a random binary string with k many 1 and $m - k$ many 0. To observe truth of the claim we notice followings:

- First 1 being in the first position is obviously $\frac{k}{m}$ as there are k many 1's out of m many characters.
- The probability of the first 1 being in the second place is $(1-\frac{k}{m})\frac{k}{m-1}$. The first entry being zero has probability: $1-\frac{k}{m}$ and the second entry being one is $\frac{k}{m-1}$
- The probability of the first 1 being in the i-th place is

$$\Big(1-\frac{k}{m}\Big)\Big(1-\frac{k}{m-1}\Big)\dots\Big(1-\frac{k}{m-(i-1)}\Big)\Big(\frac{k}{m-(i-1)}\Big).$$

  The first entry being zero has probability: $1-\frac{k}{m}$, the second entry being zero has probability $1-\frac{k}{m-1}$ and so on so forth. Finally out of remaining $m-i+1$ entries the next one being 1 is equal to $\frac{k}{m-(i-1)}$.

Then expected position of the first 1 is equal to

$$1\frac{k}{m}+2\frac{k}{m-1}\Big(1-\frac{k}{m}\Big)+3\frac{k}{m-2}\Big(1-\frac{k}{m}\Big)\Big(1-\frac{k}{m-1}\Big)+\dots$$

which is equal to $\mathbf{E}[st]$. Lets find the position of the first 1 by double counting. A word with k number of 1 and $m-k$ number of 0 can be represented as $a_0 1 a_1 1 a_2 ... 1 a_k$ where a_i represents number of zeros between two 1's. Now, lets find number of first 1 in the $k+1$ sized set of following words

$$a_0 1 a_1 1 a_2 ... 1 a_k,$$

$$a_1 1 a_2, ..., a_k 1 a_0,$$

$$\vdots$$

$$a_k 1 a_0, ..., a_k 1 a_0$$

Expected number of first 1 here is simply

$$\frac{a_0+1}{k+1}+\frac{a_1+1}{k+1}+\ldots+\frac{a_k+1}{k+1}=\frac{a_0+a_1+...+a_k+k+1}{k+1}=\frac{m+1}{k+1}.$$

So, we can divide set of all words with k many 1's and $m-k$ many 0's into $k+1$-sized sets. For each group the average position of the first 1 will be $\frac{m+1}{k+1}$. Therefore, in overall the average position of the first 1 is $\frac{m+1}{k+1}$. □

A simple followup of this lemma is to notice:

$$\mathbb{E}[st]=\frac{m+1}{k+1}<\frac{m}{k}$$

due to $m>k$. Then, we can use the Markov inequality to get:

$$P\Big(st>2\frac{m}{k}\Big)<\frac{1}{2}.$$

Moreover, from the combinatorial counting argument we can imply that the probability of $st>a$ will be given as

$$P(st>a)=\frac{\binom{m-a}{k}}{\binom{m}{k}}$$

using the previous inequality we can observe that:

$$P\Big(st>\frac{2m}{k}\Big)=\frac{\binom{m-2m/k}{k}}{\binom{m}{k}}<\frac{1}{2}$$

Considering the fact

$$f(x)=\frac{\binom{x-2m/k}{k}}{\binom{x}{k}}$$

is an increasing function and

$$P\Big(st > \frac{\alpha m}{k}\Big) = \frac{\binom{m-2m/k}{k}}{\binom{m}{k}} \frac{\binom{m-4m/k}{k}}{\binom{m-2m/k}{k}} \cdots \frac{\binom{m-2\alpha m/k}{k}}{\binom{m-2(\alpha-1)m/k}{k}} < (\frac{1}{2})^{\alpha}$$

For a given ϵ, if we set $\alpha = \log \frac{1}{\epsilon}$ we conclude that with probability at least $1 - \epsilon$ the following inequality satisfied:

$$st > 2\frac{m}{k} \log \frac{1}{\epsilon}.$$

Step 6: So, we can tell

$$P\Big(X' \geq 2 \log (\frac{1}{\epsilon})\frac{m}{k}\Big) \leq \epsilon.$$

Which means for a given j, with probability more than $1-\epsilon$, $X'_{i,j}$ will switch to zero before $2\log(\frac{1}{\epsilon})\frac{m}{k}$ for any $j \in [r]$. Using union bound argument, after $2\log(\frac{1}{\epsilon})\frac{m}{k}$ iteration with probability more than $1 - \epsilon r$, for any $i \in [r]$, $X'_{i,j}$ will switch to zero.

Therefore, the process S will stop before $2\log \frac{r}{\epsilon}\frac{m}{k}$ iteration with probability $1 - \epsilon$. Remind that, termination time of S corresponds to the value of T and number of total red points is bounded by Tn. Then number of total red observations is bounded by:

$$2\frac{mn}{k} \log \frac{r}{\epsilon}.$$

Finaly, total number of observations is equal to the number of red observa-

tions plus number of blue observations which gives the bound:

$$(m+n-r)r + 2\frac{mn}{k}\log\frac{r}{\epsilon}.$$

□

To translate this result to coherence number rather than *space sparsity number*, we use the following lemma:

Lemma 6. *Let $\mathbb{U}$ be an r-dimensional subspace of $\mathbb{R}^m$. Then the below relation between $\psi(\mathbb{U})$ and $\mu(U)$ holds:*

$$\mu(U) \geq \frac{m}{r}\frac{1}{\psi(\mathbb{U})}.$$

Proof. We again denote $\psi(\mathbb{U})$ with k for ease of reading. By the definition of the *space sparsity number*, we see that there exists a vector $v \in U$ and k different indices $i_1, i_2, ..., i_k$ such that the only nonzero components of v are $v_{i_1}, v_{i_2}, ..., v_{i_k}$. Up to scaling, we may assume that v is a unit vector. This is equivalent to

$${v_{i_1}}^2 + ... + {v_{i_k}}^2 = 1$$

Therefore, we observe that there is an index i_a satisfies ${v_{i_a}}^2 \geq \frac{1}{k}$. If this was not the case, then for all j with $1 \leq j \leq k$, ${v_{i_j}}^2 < \frac{1}{k}$ should satisfy, and this implies

$$1 = {v_{i_1}}^2 + ... + {v_{i_k}}^2 < k\frac{1}{k} = 1$$

and this is a contradiction. Using these facts, we can see that

$$||P_U e_{i_a}||^2 \geq ||v \cdot e_{i_a}||^2 = |v \cdot e_{i_a}|^2 = {v_{i_a}}^2 \geq \frac{1}{k}$$

where e_{i_a} is i_a'th standard basis of $\mathbb{R}^m$. The first inequality follows from the fact that the length of projection of any vector to the subspace $\mathbb{U}$ is always greater or equal than the length of the projection onto a vector of that subspace. Thus, we have

$$\mu(U) = \frac{m}{r} \max_{1 \leq j \leq m} ||P_U e_j||^2 \geq \frac{m}{r} ||P_U e_{i_a}||^2 \geq \frac{m}{r} \frac{1}{k} = \frac{m}{r} \frac{1}{\psi(\mathbb{U})}$$

□

Comparison of Sparsity Number with Coherence: In the lemma 6 we show that

$$\mu(\mathbb{U}) \geq \frac{m}{r} \frac{1}{\psi(\mathbb{U})}.$$

and in the theorem above we prove that the observation complexity of $\mathbf{ERR}$ is upper bounded by $(m+n-r)r + 2\frac{mn}{\psi(\mathbb{U})} \log \frac{r}{\epsilon}$ where $\mathbb{U}$ is column space of the matrix $\mathbf{M}$. Lets denote the fraction

$$\gamma = \frac{m}{\psi(\mathbb{U})} \frac{1}{\mu(\mathbb{U})r}$$

then lemma 6 is equivalent to $\gamma \leq 1$. Lets transfer observation complexity of $\mathbf{ERR}$ with respect to $\mu(\mathbb{U})$ using γ. Then the observation complexity is

$$(m+n-r)r + 2\gamma\mu(\mathbb{U})r \log r/\epsilon$$

and using the fact that $\gamma \leq 1$ this number is smaller than bound due to [2]:

$$(m + n - r) + 2\mu(\mathbb{U})r \log r/\epsilon$$

In many cases γ can be very small. For any matrix that has high value of-$\psi(\mathbb{U})$ or low value of $\mu(\mathbb{U})$, γ is guaranteed to be very small. Specifically, if $\psi(\mathbb{U})$ is $\Theta(m)$ or $\mu(\mathbb{U})$ is $\Theta\frac{m}{r}$ then γ is $\mathcal{O}(\frac{1}{r})$. Proofs for each case provided below:

$\psi(\mathbb{U})$ *is* $\Theta(m)$: Assigning $\psi(\mathbb{U})$ being $\Theta(m)$ in the definition of γ, we conclude that γ is $\Theta(\frac{1}{\mu(U)r})$. Remember from the definition of the coherence, $\mu(\mathbb{U}) \geq 1$ for any subspace, which gives the final conclusion of γ is $\mathcal{O}(\frac{1}{r})$.

$\mu(\mathbb{U})$ *is* $\Theta\frac{m}{r}$: Assigning $\mu(\mathbb{U})$ being $\Theta(\frac{m}{r})$ in the definition of γ, we conclude that γ is $\Theta(\frac{1}{\psi(\mathbb{U})})$. Moreover, remember that from lemma 3, we know that $\psi(\mathbb{U})$ is $\Omega(r)$ which gives final conclusion of $\mathcal{O}(\frac{1}{r})$.

2.2.2 Matrices with high row space sparsity-number

In this section, we show that the observation complexity of the algorithm **ERR** is upper bounded by

$$(m + n - r)r + \frac{\frac{2m}{\psi(\mathbb{U})}\left(r + 2 + \log\frac{1}{\epsilon}\right)}{\psi(\mathbb{V})}n$$

Proof. We use the same terminology as previous theorem and k and t stands for $\psi(\mathbb{U})$ and $\psi(\mathbb{V})$ correspondingly. So, if a column is not in the column

space of C then we call it active.

From lemma 4, we know that at any step if a column is still active, then probability of its detection is at least $\frac{k}{m}$ where k is the *space non-sparsity number* for column space. Let's just focus on active observations, and estimate the number of required active observations to detect r-th linearly independent column. We can see that, under the condition of each observation being active observation and the probability of detection being exactly $\frac{k}{m}$ the process of the detection of r-th independent column can be modelled as negative binomial distribution.

Lets remind the formula of the probability mass function negative binomial distribution as getting a-th success in the $a+b$'th step while success probability being p :

$$f(a, b, p) = \binom{a+b-1}{a} p^a (1-p)^b$$

For this problem, we are interested to find the probability for finding r-th success at N-th trial which corresponds to :

$$f(r, N-r, \frac{k}{m}) = \binom{N-1}{r} \left(\frac{k}{m}\right)^r \left(1-\frac{k}{m}\right)^{N-r-1}$$

As the number of observations is the focus of this theorem, we fix parameters k, m, r and investigate the behaviour of the function while N being variable. Intuitively, we use the following notation:

$$\tau_{k,m,r}(N) = f(r, N-r, \frac{k}{m})$$

In lemma 7 and 8 we investigate properties of this function to have better understanding of failure probability of **ERR**:

Lemma 7. *$\tau_{k,m,r}(N)$ is a decreasing function after N being larger than $(\frac{2m}{k}+1)r$. Specifically, we can give the following bound for the decreasing rate:*

$$1-\frac{k}{m}<\frac{\tau_{k,m,r}(N+1)}{\tau_{k,m,r}(N)}<1-\frac{k}{2m}$$

Proof. To show the decreasing we analyse the fraction :

$$\begin{aligned}\frac{\tau_{k,m,r}(N+1)}{\tau_{k,m,r}(N)} &= \frac{\binom{N}{r}\left(\frac{k}{m}\right)^r\left(1-\frac{k}{m}\right)^{N-r}}{\binom{N-1}{r}\left(\frac{k}{m}\right)^r\left(1-\frac{k}{m}\right)^{N-r-1}}\\ &= \frac{\frac{N!}{r!(N-r)!}\left(\frac{k}{m}\right)^r\left(1-\frac{k}{m}\right)^{N-r}}{\frac{(N-1)!}{r!(N-1-r)!}\left(\frac{k}{m}\right)^r\left(1-\frac{k}{m}\right)^{N-r-1}}\\ &= \frac{N}{N-r}(1-\frac{k}{m})\end{aligned}$$

So, we get following recursive formula:

$$\tau_{k,m,r}(N+1)=\frac{N}{N-r}\Big(1-\frac{k}{m}\Big)\tau_{k,m,r}(N).$$

Left side of the the target inequality is easy to prove as $\frac{N}{N-r}>1$ implies

$$\frac{\tau_{k,m,r}(N+1)}{\tau_{k,m,r}(N)}>1-\frac{k}{m}.$$

Then, we only need to prove the right side of the inequality. Lets make

the following observations

$$\frac{N}{N-r} = 1 + \frac{r}{N-r}$$

and from the hypothesis of the lemma we have

$$N > \Big(\frac{2m}{k} + 1\Big)r \implies N - r > \frac{2m}{k}r \implies \frac{r}{N-r} < \frac{k}{2m}.$$

Now, we are ready to prove rigth side:

$$\begin{aligned}
\frac{\tau_{k,m,r}(N+1)}{\tau_{k,m,r}(N)} = \frac{N}{N-r}\Big(1 - \frac{k}{m}\Big) &= \Big(1 + \frac{r}{N-r}\Big)\Big(1 - \frac{k}{m}\Big) \\
&< \Big(1 + \frac{k}{2m}\Big)\Big(1 - \frac{k}{m}\Big) \\
&= 1 - \frac{k}{2m} - \frac{k^2}{2m^2} \\
&< 1 - \frac{k}{2m}.
\end{aligned}$$

Therefore:

$$1 - \frac{k}{m} < \frac{\tau_{k,m,r}(N+1)}{\tau_{k,m,r}(N)} < 1 - \frac{k}{2m}.$$

Note that we can claim decreasing of $\tau_{k,m,r}$ just follows from the right side of the inequality.

□

To explore more properties of the function $\tau_{k,m,r}$ we prove the following lemma.

Lemma 8. *Lets assume n is a positive integer. Then $\tau_{k,m,r}$ satisfies the following inequality*

$$\tau_{k,m,r}\Big(\frac{2m}{k}(r+1)+n\Big) \leq \frac{1}{n}.$$

Proof. It is clear that $\tau_{k,m,r}\Big(\frac{2m}{k}(r+1)\Big) < 1$ as it is value of a probability mass function. In lemma 7 we proved that the functions $\tau_{k,m,r}$ is decreasing after $\frac{2m}{k}(r+1)$. Therefore, for any positive integer n the following inequalities satisfied:

$$\begin{aligned}
\tau_{k,m,r}\Big(\frac{2m}{k}(r+1)+n\Big) &< \tau_{k,m,r}\Big(\frac{2m}{k}(r+1)+n-1\Big)\\
\tau_{k,m,r}\Big(\frac{2m}{k}(r+1)+n\Big) &< \tau_{k,m,r}\Big(\frac{2m}{k}(r+1)+n-2\Big)\\
&\vdots\\
\tau_{k,m,r}\Big(\frac{2m}{k}(r+1)+n\Big) &< \tau_{k,m,r}\Big(\frac{2m}{k}(r+1)\Big)
\end{aligned}$$

By summing all these inequalities we conclude:

$$n\tau_{k,m,r}\Big(\frac{2m}{k}(r+1)+n\Big) < \sum_{i=1}^{n}\tau_{k,m,r}\Big(\frac{2m}{k}(r+1)+i\Big)$$

To bound the second term, we can use:

$$\sum_{i=1}^{n}\tau_{k,m,r}\Big(\frac{2m}{k}(r+1)+i\Big) \leq \sum_{i=r}^{\infty}\tau_{k,m,r}(i) = 1$$

and dividing left and rigth side of the inequality above concludes:

$$n\tau_{k,m,r}\Big(\frac{2m}{k}(r+1)+n\Big) < 1.$$

□

To apply the lemma above for $n = \frac{2m}{k}$ we get,

$$\tau_{k,m,r}\Big(\frac{2m}{k}(r+1) + \frac{2m}{k}\Big) < \frac{k}{2m}$$

Using right side of the lemma 7 we notice :

$$\tau_{k,m,r}\Big(\frac{2m}{k}(r+1) + \frac{2m}{k} + i\Big) < \frac{k}{2m}\Big(1 - \frac{k}{2m}\Big)^i$$

for any positive integer i. Picking $i = \frac{2m}{k}\log\frac{1}{\epsilon}$ follows as :

$$\begin{aligned}\tau_{k,m,r}\Big(\frac{2m}{k}(r+1) + \frac{2m}{k} + \frac{2m}{k}\log\frac{1}{\epsilon}\Big) &< \frac{k}{2m}\Big(1 - \frac{k}{2m}\Big)^{\frac{2m}{k}\log\frac{1}{\epsilon}} \\ &< \frac{k}{2m}e^{-\log\frac{1}{\epsilon}} = \frac{k}{2m}\epsilon\end{aligned}$$

second inequality here is application of the $(1 - \frac{1}{\alpha})^\alpha < \frac{1}{e}$ for $\alpha > 0$. Therefore we currently have :

$$\tau_{k,m,r}\Big(\frac{2m}{k}(r+1) + \frac{2m}{k} + \frac{2m}{k}\log\frac{1}{\epsilon}\Big) < \frac{k}{2m}\epsilon$$

and we target to bound :

$$\sum_{i=0}^{\infty}\tau_{k,m,r}\Big(\frac{2m}{k}(r+1) + \frac{2m}{k} + \frac{2m}{k}\log\frac{1}{\epsilon} + i\Big).$$

To apply right side of lemma 7, i times we conclude :

$$\tau_{k,m,r}\Big(\frac{2m}{k}(r+1)+\frac{2m}{k}+\frac{2m}{k}\log\frac{1}{\epsilon}+i\Big)<$$
$$\tau_{k,m,r}\Big(\frac{2m}{k}(r+1)+\frac{2m}{k}+\frac{2m}{k}\log\frac{1}{\epsilon}\Big)\Big(1-\frac{k}{2m}\Big)^i$$

Therefore the summation above can be upper bounded as:

$$\sum_{i=0}^{\infty}\tau_{k,m,r}\Big(\frac{2m}{k}(r+1)+\frac{2m}{k}+\frac{2m}{k}\log\frac{1}{\epsilon}+i\Big)<$$
$$<\sum_{i=0}^{\infty}\tau_{k,m,r}\Big(\frac{2m}{k}(r+1)+\frac{2m}{k}+\frac{2m}{k}\log\frac{1}{\epsilon}\Big)\Big(1-\frac{k}{2m}\Big)^i$$
$$=\tau_{k,m,r}\Big(\frac{2m}{k}(r+1)+\frac{2m}{k}+\frac{2m}{k}\log\frac{1}{\epsilon}\Big)\sum_{i=0}^{\infty}\Big(1-\frac{k}{2m}\Big)^i$$
$$=\tau_{k,m,r}\Big(\frac{2m}{k}(r+1)+\frac{2m}{k}+\frac{2m}{k}\log\frac{1}{\epsilon}\Big)\frac{2m}{k}$$
$$<\epsilon\frac{k}{2m}\frac{2m}{k}=\epsilon.$$

Therefore, we can conclude that the probability of $\mathbf{ERR}$ terminating after $\frac{2m}{k}\big(r+2+\log\frac{1}{\epsilon}\big)$ is smaller than ϵ.

At this point we have number upper bound for number of active observations in order to have $1-\epsilon$ probability of termination. However, we need to give the bound with respect to number of overall observations. Following lemma will help us for that purpose

Lemma 9. *At every phase of the algorith-*$\mathbf{ERR}$*, if there is at least one active observation, then there is at least* t *many active observations.*

Proof. The first step is to observe that, any column in C is already inactive as they are already in temporary column space. The second observation is any column that is linear combination of columns in C also already inactive. We prove the lemma by assuming the hypothesis of the lemma is not correct and we will deduce contradiction from that. Therefore, we assume that there is a step that the number of active columns is less than t, under the condition not all columns are inactive.

Number of active columns being smaller than t implies that the number of inactive columns is larger than $n - t$. Which implies there is a subset of columns- Ω' that satisfies $|\Omega'| > n - t$ and $\mathbf{M}_{:\Omega'}$ has rank of at most $r - 1$ (as there are still some active columns).

We know that the rank of $\mathbf{M}$ being r implies there is at least one set of r many linearly independent rows. Lets denote one of these sets by $R = \{j_1, j_2, ..., j_r\}$ and naturally, the set of row vectors $\mathbf{M}_{j_1:}, \mathbf{M}_{j_2:}, ..., \mathbf{M}_{j_r:}$ are linearly dependent.

Returning back to the argument $M_{:\Omega'}$ having a rank of at most $r - 1$, implies the rank of $M_{R:\Omega'}$ is also at most $r - 1$. Therefore, there is a linear dependence relation among the vectors $\mathbf{M}_{j_1:\Omega'}$, $\mathbf{M}_{j_2:\Omega'}$, . . . ,$\mathbf{M}_{j_r:\Omega'}$. As we already have $\Omega' > n - t$ then using lemma 1 we conclude that there is linear dependence relation among $\mathbf{M}_{j_1:}, \mathbf{M}_{j_2:}, ..., \mathbf{M}_{j_r:}$ which is a contradiction. Therefore, if there is one active column we can conclude there is at least t many active columns. □

Rest of the proof is simple counting argument. We know that if we have $\frac{2m}{k}\big(r+2+\log\frac{1}{\epsilon}\big)$ many observations then with probability larger than $1-\epsilon$ our algorithm succeeds. Moreover, from the lemma above, as at each phase we have at least $\psi(\mathbb{V})$ many observations,

$$\frac{\frac{2m}{k}\big(r+2+\log\frac{1}{\epsilon}\big)}{t}$$

many phase is enough to have desired number of active observations. Note that, at each step we have at most n many observation, which concludes the statement

$$\frac{\frac{2m}{k}\big(r+2+\log\frac{1}{\epsilon}\big)}{t}n$$

many observation is enough to guarantee with probability $1-\epsilon$ □

Proof of corollary 2

Proof. Case : $\psi(\mathbb{V})=\mathcal{O}(1)$. From the theorem 2 the observation complexity is upper bounded by

$$(m+n-r)r+\min\Big(2\frac{mn}{\psi(\mathbb{U})}\log(\frac{r}{\epsilon}),\frac{\frac{2m}{\psi(\mathbb{U})}(r+2+\log\frac{1}{\epsilon})}{\psi(\mathbb{V})}n\Big)$$

therefore it is upper bounded by

$$(m+n-r)r+2\frac{mn}{\psi(\mathbb{U})}\log(\frac{r}{\epsilon}).$$

Moreover, in lemma 6 we show that

$$\mu(\mathbb{U})\geq\frac{m}{r}\frac{1}{\psi(\mathbb{U})}$$

which upper bounds the last quantity by

$$(m+n-r)r + \mathcal{O}\big(nr\mu_0 \log(\frac{r}{\epsilon})\big).$$

Case $\psi(\mathbb{V}) = \Theta(r)$:For this case we are choosing the second term in the $\min$ operator of the theorem 2. We already know that $\frac{m}{\psi(\mathbb{U})}$ can be upper bounded by $\mu(\mathbb{U})r$ and plugging it together with $\psi(\mathbb{V}) = \Theta(r)$ gives us upper bound of

$$(m+n-r)r + \mathcal{O}\big(n\mu_0(r + \log\frac{1}{\epsilon})\big).$$

Moreover, if $\epsilon = \frac{1}{2^{\mathcal{O}(r)}}$ then $\log\frac{1}{\epsilon}$ is $\mathcal{O}(r)$, therefore right summand is bounded by $\mathcal{O}(n\mu_0 r)$. Considering the fact $\mu_0 \geq 1$ always, then overall expression is upper bounded by $mr + \mathcal{O}(n\mu_0 r)$

Case $\psi(\mathbb{V}) = \Theta(n)$: This case is just similar too previous case with the difference of plugging $\psi(\mathbb{V}) = \Theta(n)$ gives us the bound of

$$(m+n-r)r + \mathcal{O}\big(\mu_0 r(r + \log\frac{1}{\epsilon})\big).$$

Using the similar bound to ϵ makes the right summand to be $\mathcal{O}(\mu_0 r^2)$. Moreover from the definition of coherence we have $\mu_0 \leq \frac{m}{r}$ which upper bounds this term by $\mathcal{O}(mr)$ therefore overall sum is upper bounded by $\mathcal{O}((m+n-r)r)$. □

2.3 Exact Recovery While Rank Estimation

In this section, we solve the exact completion problem in a slightly different setup. **ERR** assumes that we know the exact rank of the underlying matrix. However, here we assume that we don't have this information. Therefore we don't know precisely at which point the process of searching a new independent row/column should stop. In what follows, we show that if at a given state new independent column/row not detected for long enough time, then it means that it is likely no more one exists. We formalize this statement in the following theorem.

ERRE: Exact recovery while rank estimation.

Input: T-delay parameter at the end of algorithm
Initialize: $R = \emptyset, C = \emptyset, \widehat{r} = 0, delay = 0$

while $delay < T$ **do**
  $delay = delay + 1$
  for j from 1 to n **do**
    Uniformly pick an unobserved entry i from $\mathbf{M}_{:j}$
    $\widehat{R} = R \cup \{i\}, \widehat{C} = C \cup \{j\}$
    If $\mathbf{M}_{\widehat{R}:\widehat{C}}$ is nonsingular :
      Fully observe $\mathbf{M}_{:j}$ and $\mathbf{M}_{i:}$
      Set $R = \widehat{R}$, $C = \widehat{C}$, $\widehat{r} = \widehat{r} + 1, delay = 0$
Orthogonalize column vectors in C and assign to $\mathbf{U}$
for each column $j \in [n] \setminus C$ **do**
  $\widehat{\mathbf{M}}_{:j} = \widehat{\mathbf{U}}\widehat{\mathbf{U}}^{+}_{R:}\widehat{\mathbf{M}}_{R:j}$

Output: $\widehat{\mathbf{M}}, \widehat{r}$

Theorem 3. *Let r be the rank of underlying $m \times n$ sized matrix $\mathbf{M}$ that has column space $\mathbb{U}$ and row space $\mathbb{V}$. Then, the algorithm* **ERRE** *exactly recovers the underlying matrix $\mathbf{M}$ while estimating rank with probability at least $1 - (\epsilon + e^{-T\frac{\psi(\mathbb{U})\psi(\mathbb{V})}{m}})$ using number of observations at most:*

$$(m + n - r)r + Tn + \min\Big(2\frac{mn}{\psi(\mathbb{U})}\log(\frac{r}{\epsilon}), \frac{\frac{2m}{\psi(\mathbb{U})}(r + 2 + \log\frac{1}{\epsilon})}{\psi(\mathbb{V})}n)\Big)$$

Corollary 3. *Lets assume that either $\psi(\mathbb{U})$ or $\psi(\mathbb{V})$ is big enough i.e. $\psi(\mathbb{U})\psi(\mathbb{V}) \geq m$. For a given ϵ set $T = \log \frac{1}{\epsilon}$, then* **ERRE** *recovers underlying matrix with probability $1 - 2\epsilon$ using just*

$$(m+n-r)r + n\log\frac{1}{\epsilon} + \min\Big(2\frac{mn}{\psi(\mathbb{U})}\log(\frac{r}{\epsilon}), \frac{\frac{2m}{\psi(\mathbb{U})}(r+2+\log\frac{1}{\epsilon})}{\psi(\mathbb{V})}n\Big)$$

We refer to the analysis of corollary 2 to understand this expression better.

Proof. We again use k and t for $\psi(\mathbb{U})$ and $\psi(\mathbb{V})$ correspondingly and use all the terminology from the previous proofs. Then, we start by proving that under the scenario there is still active column, then with probability $1 - e^{-T\frac{kt}{m}}$, it will be detected in T phases. We prove the following key lemma in order to accomplish the proof of the theorem.

Lemma 10. *Lets assume the underlying matrix* $\mathbf{M}$ *has row space non-sparsity number k and column space non-sparsity number t. Then, if at an intermediate step of* **ERRE** *still column space not recovered completely, then with probability $1 - e^{-T\frac{kt}{m}}$ new independent column will be detected within T phases.*

Proof. For every active column observation, the probability of detecting independence is at least $1 - \frac{k}{m}$ from the lemma 4. From the lemma 9 , if there is one active column, then there is at least t many active column in that phase. Therefore, the probability of detection of an active column is at least

$$\Big(1 - \frac{k}{m}\Big)^t.$$

Then, we conclude that after T many phase, detection probability is at least

$$\Big(1-\frac{k}{m}\Big)^{tT}.$$

Using the inequality $1+x \leq e^x$ for $\forall x \in \mathbb{R}$ the quantity above can be bounded by:

$$\Big(1-\frac{k}{m}\Big)^{tT} < e^{-T\frac{kt}{m}}.$$

□

Now, we show that with probability at least $1-e^{-T\frac{kt}{m}}$, estimated rank $\widehat{r}$ is equal to r. We have :$P(r=\widehat{r}) = 1-\big(P(r<\widehat{r})+P(r>\widehat{r})\big)$

$P(r<\widehat{r})=0$ trivially satisfied, $\widehat{r}$ represents number of detected linearly independent columns of $\mathbf{M}$ which is always bounded by r. Now, all we need to do is to bound $P(r<\widehat{r})$. We denote the event of existence of active column by ACE. Then, trivially:

$$P(\widehat{r}<r) = P\big(\widehat{r}<r \text{ and } ACE\big)$$

Moreover, we can write

$$P\big(\widehat{r}<r \text{ and } ACE\big) = \sum_{i=0}^{r-1} P\big(\widehat{r}=i \text{ and } ACE\big) = P\big(ACE \mid \widehat{r}=i\big)P(\widehat{r}=i)$$

To finish the proof we just need to observe following equality / inequality :

$$P(\widehat{r}<r) = \sum_{i=0}^{r-1} P\big(\widehat{r}=i \wedge ACE\big) = \sum_{i=0}^{r-1} P\big(ACE|\widehat{r}=i\big)P(\widehat{r}=i)$$

From the lemma above, we can imply that

$$P\big(ACE|\widehat{r}=i\big) \leq e^{-T\frac{kt}{m}}$$

and as probability of $P(\widehat{r}=r) \neq 0$, we conclude $P(\widehat{r}<r)<1$. Equivalently,

$$\sum_{i=0}^{r-1} P\big(\widehat{r}=i\big) < 1$$

which all together these two inequalities concludes

$$P(\widehat{r}<r) = \sum_{i=0}^{r-1} P\big(ACE|\widehat{r}=i\big)P(\widehat{r}=i) \leq e^{-T\frac{kt}{m}} \sum_{i=0}^{r-1} P(\widehat{r}=i) \leq e^{-T\frac{kt}{m}}$$

To finalize the proof, we divide the algorithm $\mathbf{ERRE}$ into two parts. First part, is the detection point of the last independent column by algorithm, and second part is waiting T many rounds to check if there is any independent column left. Moreover, $\mathbf{ERRE}$ would fail generating correct matrix only if failure in the second part happens (there is still independent column not detected, but checking tells us that there is no left) i.e. $\widehat{r}<r$ which we just show $P(\widehat{r}<r)<e^{-T\frac{kt}{m}}$. This concludes that with probability at least $1-e^{-T\frac{kt}{m}}$ recovered matrix is correct.

Therefore, with probability $1-e^{-T\frac{kt}{m}}$ the first part of the algorithm is just equivalent to the algorithm $\mathbf{ERR}$, which with probability more than $1-\epsilon$ observation complexity is bounded by

$$(m+n-r)r + \min\Big(2\frac{mn}{k}\log(\frac{4r}{\epsilon}), \frac{\frac{2m}{k}(r+2+\log\frac{1}{\epsilon})}{t}n\Big)$$

using the union bound we conclude that with probability at least $1 - \epsilon + e^{-T\frac{kt}{m}}$, the algorithm recovers underlying matrix correctly and observation complexity is bounded by

$$(m+n-r)r + Tn + \min\Big(2\frac{mn}{k}\log(\frac{4r}{\epsilon}), \frac{\frac{2m}{k}(r+2+\log\frac{1}{\epsilon})}{t}n\Big)$$

□

Proof of Corollary 3

Proof. First observe the following inequalitys:

$$kt \geq m \implies \frac{kt}{m} \geq 1 \implies e^{\frac{kt}{m}} \geq e \implies e^{-\frac{kt}{m}} \leq \frac{1}{e}$$

The rest of the proof is just straightforward application of the theorem. Setting $T = \log\frac{1}{\epsilon}$ to the statement of theorem tells with probability at least $1 - (\epsilon + e^{-\frac{kt}{m}\log\frac{1}{\epsilon}})$ using

$$(m+n-r)r + n\log\frac{1}{\epsilon} + \min\Big(2\frac{mn}{k}\log(\frac{4r}{\epsilon}), \frac{\frac{2m}{k}(r+2+\log\frac{1}{\epsilon})}{t}n\Big)$$

observations. Considering the fact that $e^{-\frac{kt}{m}} \leq \frac{1}{e}$ we conclude that

$$e^{-\frac{kt}{m}\log\frac{1}{\epsilon}} \leq \Big(\frac{1}{e}\Big)^{\log\frac{1}{\epsilon}} = e^{-\log\frac{1}{\epsilon}} = e^{\log\epsilon} = \epsilon$$

Which concludes that with probability at least $1 - 2\epsilon$ the observation complexity is bounded by

$$(m+n-r)r + n\log\frac{1}{\epsilon} + \min\Big(2\frac{mn}{k}\log(\frac{4r}{\epsilon}), \frac{\frac{2m}{k}(r+2+\log\frac{1}{\epsilon})}{t}n\Big)$$

□

2.4 Exact Recovery with Estimated Information

In this section, we approach the exact completion problem under the scenario we have estimated pre-information, using ideas from previous sections. This setting is similar to **KS2013**, [2] and [16] that estimated values of rank r and μ_0 - coherence of column space got used. Methods proposed in these papers work efficiently under the condition row spaces are highly coherent. An in-depth analysis of these methods tells us that these algorithms designed for the highest value of row space coherence, and they also work perfectly well for remaining cases. In the following algorithm, we somehow extend these algorithms to a method that can enjoy the properties of row space; meanwhile, performing similarly good for highly coherent row space matrices.

Note that the algorithm below designed and analyzed under the condition that estimated r, $\psi(\mathbb{U})$, and $\psi(\mathbb{V})$ provided. However, under the condition just estimated r and μ_0 provided, we can set $\psi(\mathbb{V})$ to be equal to 1 and use the inequality $\mu_0 r > \frac{m}{k}$ to transfer the information of μ_0 to $\psi(\mathbb{U})$. Then the algorithm below will perform as good as [2] with observation complexity of $mr + \mathcal{O}(nr\mu_0 \log \frac{r}{\epsilon})$.

We demonstrate the performance of the algorithm in an example matrix below after providing the technical details. The idea of the algorithm is very similar to **ERR** with the only difference here, instead of observing just one entry from each column each time through many steps, we observe just large amount once and to show it also works with the same probability.

EREI: Exact recovery with estimated information

Input: $r, \psi(\mathbb{U}), \psi(\mathbb{V})$

Initialize: $R = \emptyset, k = 0, \widehat{\mathbf{U}}^0 = \emptyset, d = \min\Big(2\frac{m}{\psi(\mathbb{U})}\log\left(\frac{r}{\epsilon}\right), \frac{\frac{2m}{\psi(\mathbb{U}}(r+2+\log\frac{1}{\epsilon})}{\psi(\mathbb{V})}\Big)$

Draw uniformly random entries $\Omega \subset [m]$ of size d
for i from 1 to n **do**
    if $\|\mathbf{M}_{\Omega:i} - \mathcal{P}_{\widehat{\mathbf{U}}^{\mathbf{k}}_{\mathbf{\Omega}}}\mathbf{M}_{\Omega:i}\| > 0$
        Fully observe $\mathbf{M}_{:i}$, $\widehat{\mathbf{U}}^{k+1} \leftarrow \widehat{\mathbf{U}}^k \cup \mathbf{M}_{:\mathbf{i}}$ and orthogonalize $\widehat{\mathbf{U}}^{k+1}$, $k = k+1$
        Select a row $a \in \Omega \setminus R$ that, $\widehat{\mathbf{U}}^{\mathbf{k+1}}_{\mathbf{R}\cup\{\mathbf{a}\}}$ is rank $k+1$ then $R \leftarrow R \cup \{a\}$
    Draw uniformly random entries $\Delta \subset [m] \setminus R$ of size d and $\Omega = \Delta \cup R$
Observe entire $\mathbf{M}_{R:}$
for i from 1 to n **do**
    if $\mathbf{M}_{:i}$ not fully observed **then** : $\widehat{\mathbf{M}}_{:i} = \widehat{\mathbf{U}}^k\widehat{\mathbf{U}}^{k^+}_{R:}\widehat{\mathbf{M}}_{R:i}$

Output: $\widehat{\mathbf{M}}$

We show the execution of the algorithm for $d = 2$ and $r = 1$ below. $\times$ stands for entries that is observed randomly, $\times$ stands for deterministically observed and $\times$ stands for recovered entries. Note that $\Omega_1 = \{1, 5\}, \Omega_2 = \{2, 5\}, \Omega_3 = \{1, 3\}$ and after 3-rd iteration R becomes $\{3\}$ therefore, in fourth column $\mathbf{M}_{3:4}$ observed deterministically, besides together random observations $\Omega_4 = \{5, 6\}$. After all of the iterations completed, we observe the entire $\mathbf{M}_{3:}$, and in the next step, we recover remaining entries.

$$\begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 1 & 3 & 2 & 3 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 2 & 6 & 4 & 6 \end{bmatrix} \quad \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 1 & 3 & 2 & 3 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 2 & 6 & 4 & 6 \end{bmatrix} \quad \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 1 & 3 & 2 & 3 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 2 & 6 & 4 & 6 \end{bmatrix}$$

$$\begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 1 & 3 & 2 & 3 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 2 & 6 & 4 & 6 \end{bmatrix} \quad \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 1 & 3 & 2 & 3 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 2 & 6 & 4 & 6 \end{bmatrix} \quad \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 1 & 3 & 2 & 3 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 2 & 6 & 4 & 6 \end{bmatrix}$$

Theorem 4. *Let r be the rank of underlying $m \times n$ sized matrix $\mathbf{M}$ with column space $\mathbb{U}$ and row space $\mathbb{V}$. Then, with probability $1-\epsilon$ the algorithm* **EREI** *exactly recovers the underlying matrix $\mathbf{M}$ using number of observations at most*

$$(m+n-r)r + \min\Big(2\frac{mn}{\psi(\mathbb{U})}\log\big(\frac{r}{\epsilon}\big), \frac{\frac{2m}{\psi(\mathbb{U})}(r+2+\log\frac{1}{\epsilon})}{\psi(\mathbb{V})}n\Big)$$

Proof for low row space sparsity-number : $d = 2\frac{m}{\psi(\mathbb{U})}\log\big(\frac{r}{\epsilon}\big)$

Proof. We first start with the case that if minimum of these two quantities is $2\frac{m}{k}\log\big(\frac{r}{\epsilon}\big)$. As the matrix has rank r, there exists at least one set of linearly independent columns with r columns. We select the set of linear independent columns$-C$ that has lexicographically smallest indices. We show sampling $2\frac{m}{k}\log\big(\frac{1}{\epsilon}\big)$ entries from each column will give us the probability of at least $1-r\epsilon$ correctly recovery.

From the Step 5 of proof of the theorem 3 we can see sampling $2\frac{m}{k}\log\big(\frac{1}{\epsilon}\big)$ entries from an active column, would give guarantee of probability of at least $1-\epsilon$ detection of independence. As C is lexicographically smallest, each column is active on the time entries sampled from it, and each of r columns will succeed with probability at least $1-\epsilon$. Therefore, using union bound, with probability $1-r\epsilon$ all of the columns in C will succeed, which guarantees the exact recovery.

Replacing ϵ by $\frac{\epsilon}{r}$ will conclude the result that sampling $2\frac{m}{k}\log\big(\frac{r}{\epsilon}\big)$ from each column will guarantees the correctness of the algorithm with probability at least $1-\epsilon$. □

Proof for high row space sparsity-number : $d = \frac{\frac{2m}{\psi(\mathbb{U})}(r+2+\log\frac{1}{\epsilon})}{\psi(\mathbb{V})}$

Proof. From the follow up of lemma 4, we conclude that in a process of $\frac{k}{m}$ probability success and $1-\frac{k}{m}$ probability of failure, having $\frac{2m}{k}\big(r+2+\log\frac{1}{\epsilon}\big)$ trial is enough to guarantee getting r many success with probability $1-\epsilon$.

Failure probability of the algorithm **EREI** is equal to failing finding r linearly independent columns. Consider following equation:

$$\begin{aligned}\mathbb{P}\big(\mathbf{EREI}fails\big) = \mathbb{P}\Big(\mathbf{EREI}fails \text{ and TNAO} \geq \frac{2m}{k}\big(r+2+\log\frac{1}{\epsilon}\big)\Big)\\ + \mathbb{P}\Big(\mathbf{EREI}fails \text{ and TNAO} < \frac{2m}{k}\big(r+2+\log\frac{1}{\epsilon}\big)\Big)\end{aligned}$$

where we denote TNAO as total number of active observations. Recall that we call an observation active, if it is active in the execution time (the column is still not contained in the current column space). Intuitively we represent NAO by number of active observations executed by the algorithm **EREI** in the given specific time.

From lemma 9, if there exists an active column, then there exists at least t many active columns. Therefore, failure of the algorithm is equivalent to existence of an active column when algorithm terminates. Moreover each of our observations in those columns were active observations and considering the fact that we observed $\frac{\frac{2m}{k}\left(r+2+\log\frac{1}{\epsilon}\right)}{t}$ many entries in each of them, total number of active observations is at least

$$t\frac{\frac{2m}{k}\big(r+2+\log\frac{1}{\epsilon}\big)}{t} = \frac{2m}{k}\Big(r+2+\log\frac{1}{\epsilon}\Big)$$

Therefore $\mathbb{P}\Big(\text{NAO} < \frac{2m}{k}\big(r+2+\log\frac{1}{\epsilon}\big)|\mathbf{EREI}fails\Big) = 0$ and using Bayesian rule we conclude

$$\mathbb{P}\Big(\mathbf{EREI}fails \text{ and TNAO} \leq \frac{2m}{k}\big(r+2+\log\frac{1}{\epsilon}\big)\Big) = 0$$

Then, following equation simplly satisfied:

$$\mathbb{P}\big(\mathbf{EREI}fails\big) = P\Big(\mathbf{EREI}fails \text{ and TNAO} \geq \frac{2m}{k}\big(r+2+\log\frac{1}{\epsilon}\big)\Big)$$

We can observe the following inequality as **EREI** may tamporarily fail at the point that the number of active observations is $\frac{2m}{k}\big(r+2+\log\frac{1}{\epsilon}\big)$ but it can succeed finding remaining independent columns later during the execution:

$$\mathbb{P}\Big(\mathbf{EREI} \text{ currently fail when NAO} = \frac{2m}{k}\big(r+2+\log\frac{1}{\epsilon}\big)\Big) \geq \mathbb{P}\Big(\mathbf{EREI}fails \text{ and TNAO} \geq \frac{2m}{k}\big(r+2+\log\frac{1}{\epsilon}\big)\Big)$$

Therefore we conclude that:

$$\mathbb{P}\big(\mathbf{EREI}fails\big) \leq \mathbb{P}\Big(\mathbf{EREI} \text{ currently fail when NAO} = \frac{2m}{k}\big(r+2+\log\frac{1}{\epsilon}\big)\Big)$$

Remember the fact that at each active observation probability of **EREI** detecting linear independence of is larger or equal than $\frac{k}{m}$. From the previous discussion if the probability is exactly equal to $\frac{k}{m}$ then still not finding r linearly independent column at $\frac{2m}{k}\big(r+2+\log\frac{1}{\epsilon}\big)$ observations is less than ϵ. Therefore, **EREI** not detecting r linearly independent column after

$\frac{2m}{k}\big(r+2+\log\frac{1}{\epsilon}\big)$ observations is smaller than ϵ, which is equivalent to

$$\mathbb{P}\big(\mathbf{EREI}fails\big) \leq \epsilon$$

as desired. □

Straightforward conclusion of the theorem 4 is as following:

Corollary 4. *Lets assume that we have estimated values of rank r, column space coherence number μ_0 and estimated row space sparsity-number is $\psi(\mathbb{V})$. Then, if $\psi(\mathbb{V})$ is $\mathcal{O}(1)$ then observation complexity is buounded by $(m+n-r)r+\mathcal{O}(nr\mu_0\log\frac{r}{\epsilon})$ and if $\psi(\mathbb{V})$ is $\mathcal{O}(n)$ then with probabiliy $\frac{1}{2^{\mathcal{O}(r)}}$ the bound is $\mathcal{O}((m+n-r)r)$ which is theoretical lower bound for exact completion problem.*

Proof. Proof of this corollary is the same as the proof of the corollary 3. □

Chapter 3

Noisy Matrix Completion

In this chapter, we analyze the completion problem with the condition that entries of the underlying matrix can be noisy. Similar to [2], we focus on two types of noise model: sparse random noise and bounded noise. First, we assume that several columns of the matrix are completely noisy, and we target to recover clean entries using as little as possible observations. We show how to extend exact completion algorithms proposed here to handle this type of noise. Second, we assume that each entry of the underlying matrix can contain some small noise.

3.1 Random Noise

In this section, we discuss how to extend exact completion algorithms proposed in here to the case that some of the columns have random noise that is coming from the non-degenerate distribution $\mathcal{U}$. Here, we extend the algorithm **EREI** to make it robust to this type of noise. As it is discussed in [2], if a column is an entirely random noise, then it should be linearly dependent

with the rest of the columns. It is trivial to notice that a non-degenerate random noise will not be in a predefined subspace as much as the subspace is not full-dimensional.

Therefore, intuitively to clean noisy columns, we should detect columns that without them the rank of the matrix decrements. Surprisingly, this phenomenon can be related to space *sparsity-number*. Given that *sparsity-number* of row space is higher than one, none of the columns of the underlying matrix satisfies this condition.

We propose a method that whenever the underlying matrix having coherent row space (or low value of $\psi(\mathbb{V})$). However, it is given that the column space is recoverable, we run **EREI** with $d = 2\frac{m}{\psi(\mathbb{U})}\log(\frac{r}{\epsilon})$ and at the end, all of the columns that their deletion causes rank decrement are detected as noise. For matrices that have highly incoherent row space or high value of *nonsparsity-number*, we approach little differently because noise can immensely effect $\psi(\mathbb{V})$. However, this quantity can be affected at most by the number of noisy columns which is proved in one of the following lemmas.

Therefore, it is safe to condition if the number of noisy columns: $\xi < \frac{\psi(\mathbf{V})}{2}$, we can safely use the same method as **EREI** however, here d will be twice larger d than noise-free case: $d = \frac{4m}{\psi(\mathbb{U}\psi(\mathbb{V})}(r + 2 + \log\frac{1}{\epsilon})$. However, if the number of noisy columns-ξ get closer to $\psi(\mathbf{V})$, using $d = 2\frac{m}{\psi(\mathbb{U})}\log(\frac{r}{\epsilon})$ is more safe as much as no further information provided. We detect noise columns similar to the previous case. The advantage of this method is due to the efficiency of the exact completion method.

EEREI: Extended Exact recovery with estimated information

Input: $r, \psi(\mathbb{U}), \psi(\mathbb{V})$

Initialize: $k = 0, \widehat{\mathbf{U}}^0 = \emptyset, d = 2\frac{m}{\psi(\mathbb{U})}\log\left(\frac{r}{\epsilon}\right), R = \emptyset$

if $\xi \leq \frac{\psi(\mathbb{V})}{2}$

  $d = \min\Big(2\frac{m}{\psi(\mathbb{U})}\log\left(\frac{r}{\epsilon}\right), \frac{\frac{4m}{\psi(\mathbb{U})}(r+2+\log\frac{1}{\epsilon})}{\psi(\mathbb{V})}\Big)$

Draw uniformly random entries $\Omega \subset [m]$ of size d

for i from 1 to n **do**

  if $\|\mathbf{M}_{\Omega:i} - \mathcal{P}_{\widehat{\mathbf{U}}^{\mathbf{k}}_{\mathbf{\Omega}}}\mathbf{M}_{\Omega:i}\| > 0$

    Fully observe $\mathbf{M}_{:i}$

    $\widehat{\mathbf{U}}^{k+1} \leftarrow \widehat{\mathbf{U}}^k \cup \mathbf{M}_{:\mathbf{i}}$, Orthogonalize $\widehat{\mathbf{U}}^{k+1}$, $k = k+1$

    Select a row $a \in \Omega \setminus R$ that, $\widehat{\mathbf{U}}^{\mathbf{k+1}}_{\mathbf{R}\cup\{\mathbf{a}\}}$ is rank $k+1$ then $R \leftarrow R \cup \{a\}$

  Draw uniformly random entries $\Delta \subset [m] \setminus R$ of size d and $\Omega = \Delta \cup R$

Observe unobserved entries in $\mathbf{M}_{R:}$

for i from 1 to n **do**

  if $\mathbf{M}_{:i}$ not fully observed : $\widehat{\mathbf{M}}_{:i} = \widehat{\mathbf{U}}^k\widehat{\mathbf{U}}^{k+}_{R:}\widehat{\mathbf{M}}_{R:i}$

Detect all the columns that their deletion decrements rank and collect them in Σ

Output: Noisy Columns - Σ, recovered underlying matrix $\widehat{\mathbf{M}}_{:[n]\setminus\Sigma}$

The proof of the algorithm **EEREI** correctly detects noisy columns is the same as the proof provided in [2], therefore we don't see necessity to provide it here. Moreover, the correct recovery of the remaining entries is the same as the proof of **EREI** with only difference, if $\xi \leq \frac{\psi(\mathbb{V})}{2}$ happens and $2\frac{m}{\psi(\mathbb{U})}\log\left(\frac{r}{\epsilon}\right) > \frac{\frac{4m}{\psi(\mathbb{U})}(r+2+\log\frac{1}{\epsilon})}{\psi(\mathbb{V})}$. However, the crucial step here will be to notice number of active columns will be at least $\frac{\psi(\mathbb{V})}{2}$ and at each active column we are doing $\frac{\frac{4m}{\psi(\mathbb{U})}(r+2+\log\frac{1}{\epsilon})}{\psi(\mathbb{V})}$ many observations which gives overall

$$\frac{\frac{4m}{\psi(\mathbb{U})}(r+2+\log\frac{1}{\epsilon})}{\psi(\mathbb{V})} \times \frac{\psi(\mathbb{V})}{2} = 2\frac{m}{\psi(\mathbb{U})}(r+2+\log\frac{1}{\epsilon})$$

active observations, which is the required number of active observations in order to detect r complete column space with probability $1-\epsilon$.

Lemma 11. *Lets assume that for a given $m \times n$ sized rank$-r$ matrix $\widetilde{\mathbf{M}}$, the nonsparsity-number of it row space $\widetilde{\mathbb{V}}$ is equal to one: $\psi(\widetilde{\mathbb{V}}) = 1$. Then there is a column$-c$, that deletion of it decrements the rank: $\mathrm{rank}(\widetilde{\mathbf{M}}_{:[n]\setminus\{c\}}) = r - 1$.*

Proof. From the definition of *nonsparsity-number*, we notice that, it is always positive and moreover, only case that it is equal to 1 is if some of the standard basis vector e_i is contained in the space. Then lets assume that the row space contains e_{i_0}, then we claim that the deletion of the column i_0 decrements the rank.

Lets assume that $\mathrm{rank}(\widetilde{\mathbf{M}}_{:[n]\setminus\{i_0\}}) = r$. Moreover, lets denote the $\widetilde{\mathbf{M}}^0$ by the matrix that is i_0-th column replaced by zero vector. It is trivial to see that that $\mathrm{rank}(\widetilde{\mathbf{M}}_{:[n]\setminus\{i_0\}}) = \mathrm{rank}(\widetilde{\mathbf{M}}^0)$. Therefore $\mathrm{rank}(\widetilde{\mathbf{M}}^0) = r$ satisfies. Moreover, row space of $\widetilde{\mathbf{M}}^0$ is subset of row space of $\widetilde{\mathbf{M}}$, however, as both of them has the same rank r, these space are the same. But, this is a contradiction to the fact that, e_{i_0} is contained in the row space of $\widetilde{\mathbf{M}}$ but not in the row space of $\widetilde{\mathbf{M}}^0$ (due to the fact that, entire row space has 0 in its i_0-th coordinate). Then, our assumption is wrong, $\mathrm{rank}(\widetilde{\mathbf{M}}_{:[n]\setminus\{i_0\}})$ cannot be r and only other possible option is $r - 1$. Therefore $\mathrm{rank}(\widetilde{\mathbf{M}}_{:[n]\setminus\{c\}}) = r - 1$ satisfies.

□

Lemma 12. *Lets assume that for a given $m \times n$ sized rank$-r$ matrix $\widetilde{\mathbf{M}}$ with row space $\widetilde{\mathbb{V}}$, several columns are deleted and number of deleted columns is smaller than the non-sparsity number of row space. Then, rank of the matrix*

is preserved, and its non-sparsity number is decreased by at most $a-$where a stands for the number of columns those are deleted.

Proof. The proof is just inductively application of following claims:

- Deletion of one column, can reduce *nonsparsity-number* of the row space at most by 1.
- if $\psi(\widetilde{\mathbb{V}}) > 1$, deletion of a column cannot decrement the rank.

First we prove the first claim: Lets assume for contradiction that, there is a column c that its deletion causes drop of row space *nonsparsity-number* by at least 2, i.e. $\psi(\widetilde{\mathbb{V}}) - \psi(\mathbb{V}') \geq 2$, where $\mathbb{V}'$ is the row space of $\widetilde{\mathbf{M}}_{:[n]\setminus\{c\}}$.

Remind the definition: $\psi(\mathbb{V}') = \min\{\psi(x) | x \in \mathbb{V}' \text{ and } x \neq 0\}$. Lets pick the nonzero vector $x_0 \in \mathbb{V}'$ which satisfies the minimality in this definition. As $\mathbb{V}'$ is restriction of $\widetilde{\mathbb{V}}$ to the index set $[n] \setminus \{c\}$, there is a vector $y \in \mathbb{V}$ such that $y_{[n]\setminus\{c\}} = x_0$. Therefore

$$\|y\|_0 \leq \|x_0\|_0 + 1 = \psi(\mathbb{V}') + 1.$$

Considering the inequality

$$\psi(\widetilde{\mathbb{V}}) \leq \|y\|_0 \leq \|x_0\|_0 + 1 = \psi(\mathbb{V}') + 1$$

which implies $\psi(\widetilde{\mathbb{V}}) - \psi(\mathbb{V}') \leq 1$ and this contradicts to the assumption. Therefore, first claim indeed always satisfied.

Now, we prove the second claim: This is a simple statement, once we notice that if deletion of the column c results decrement in the rank of the matrix,

then e_c (i.e. $c-$th standard vector) is contained in the row space of $\widetilde{\mathbf{M}}$. The proof is simple as following:

Lets assume that without lost of generality columns that has the index set $\mathcal{B} = \{c_1, c_2, \ldots, c_{r-1}, c\}$ is a basis for column space. Note that, c should be part of any basis, because without column$-c$ the column space has dimension $r-1$. Moreover, we know that for any $i \notin \mathcal{B}$, the column $\widetilde{\mathbf{M}}_{:i}$ can be written as linear combination of columns $\{c_1, c_2, \ldots, c_{r-1}\}$ as:

$$\widetilde{\mathbf{M}}_{:i} = \sum_{j=1}^{r-1} \alpha_j \widetilde{\mathbf{M}}_{:c_j} \text{ for scalar } \alpha_j' s.$$

Moreover, considering the fact that, $\widetilde{\mathbf{M}}_{:\mathcal{B}}$ has rank equal to r, we can find the set of rows $R = \{R_1, R_2, \ldots, R_r\}$, that $\mathbf{M}_{R:\mathcal{B}}$ also has rank r. Considering the fact that, $\widetilde{\mathbf{M}}_{R:\mathcal{B}}$ is $r \times r$ sized matrix, then its row space contains the standard basis vector: $(0, 0, \ldots, 0, 1)$. This follows that there is a linear combination of rows $R = \{R_1, R_2, \ldots, R_r\}$ with some scalars $\beta_1, \beta_2, \ldots, \beta_r$ that

$$\beta_1 \widetilde{\mathbf{M}}_{R_1:\mathcal{B}} + \beta_2 \widetilde{\mathbf{M}}_{R_2:\mathcal{B}} + \ldots + \beta_r \widetilde{\mathbf{M}}_{R_r:\mathcal{B}} = (0, 0, \ldots, 0, 1).$$

For a given $i \notin \mathcal{B}$ lets check the $i-$th coordinate of the vector

$$\beta_1 \widetilde{\mathbf{M}}_{R_1:i} + \beta_2 \widetilde{\mathbf{M}}_{R_2:i} + \ldots + \beta_r \widetilde{\mathbf{M}}_{R_r:i}.$$

Remember that $\widetilde{\mathbf{M}}_{:i} = \sum_{j=1}^{r-1} \alpha_j \widetilde{\mathbf{M}}_{:c_j}$ as $i \notin \mathcal{B}$, therefore the sum above can

be written as:

$$\begin{aligned}
&\beta_1\widetilde{\mathbf{M}}_{R_1:i} + \beta_2\widetilde{\mathbf{M}}_{R_2:i} + \ldots + \beta_r\widetilde{\mathbf{M}}_{R_r:i} \\
&\qquad = \beta_1 \sum_{j=1}^{r-1} \alpha_j \widetilde{\mathbf{M}}_{:c_j} + \beta_2 \sum_{j=1}^{r-1} \alpha_j \widetilde{\mathbf{M}}_{:c_j} + \ldots + \beta_r \sum_{j=1}^{r-1} \alpha_j \widetilde{\mathbf{M}}_{:c_j} \\
&\qquad = \alpha_1 \sum_{i=1}^{r} \beta_i \widetilde{\mathbf{M}}_{:c_1} + \alpha_2 \sum_{i=1}^{r} \beta_i \widetilde{\mathbf{M}}_{:c_2} + \ldots + \alpha_{r-1} \sum_{i=1}^{r} \beta_i \widetilde{\mathbf{M}}_{:c_{r-1}} \\
&\qquad = \alpha_1 \times 0 + \alpha_2 \times 0 + \ldots + \alpha_{r-1} \times 0 = 0
\end{aligned}$$

and this concludes that

$$\beta_1\widetilde{\mathbf{M}}_{R_1:} + \beta_2\widetilde{\mathbf{M}}_{R_2:} + \ldots + \beta_r\widetilde{\mathbf{M}}_{R_r:} = (0, 0, \ldots, 0, 1, 0, \ldots, 0)$$

Therefore e_c is in the row space, and this implies that $\psi(\widetilde{\mathbb{V}}) = 1$. Therefore,if $\psi(\widetilde{\mathbb{V}}) > 1$ then deletion of a column cannot decrement rank.

□

Number of Observations: The number of observations is simply ξ many additional observed columns and rows, compared to the algorithm **EREI** with the slight difference if $\xi < \frac{\psi(\mathbb{V})}{2}$ then

$$(m+n-r)r + \min\Big(2\frac{mn}{\psi(\mathbb{U})}\log(\frac{r}{\epsilon}), \frac{\frac{4m}{\psi(\mathbb{U})}(r+2+\log\frac{1}{\epsilon})}{\psi(\mathbb{V})}n\Big) + \xi(m+n)$$

and if $\xi > \frac{\psi(\mathbb{V})}{2}$

$$(m+n-r)r + 2\frac{mn}{\psi(\mathbb{U})}\log(\frac{r}{\epsilon}) + \xi(m+n)$$

Recall for properties of non-degenerate subspace: Lets remind the characteristics of non-degenerate random vectors provided in [2]:

Let $\mathbf{E}^s \in \mathbb{R}^{m\times s}$ be matrix consisting of corrupted vectors drawn from any non-degenerate distribution. Let $\mathbb{U}^k \in \mathbb{R}^{m\times k}$ be any fixed matrix with rank k. Then with probability 1, we have

- $\mathrm{rank}(\mathbf{E}^s) = s$ for any $s \leq m$
- $\mathrm{rank}(\mathbf{E}^s, x) = s+1$ holds for $x \in \mathbf{U}^k \subset \mathbb{R}^{>}$ uniformly and $s \leq m-k$, where x can be depend or independent on $\mathbf{E}^s$
- $\mathrm{rank}(\mathbf{E}^s, \mathbb{U}^k) = s + k$ given that $s + k \leq m$
- The marginal of non-degenerate distribution is non-degenerate

3.2 Bounded Deterministic Noise

In this section, we propose an algorithm that gives a low-rank estimation to a matrix with small noise additional to a low-rank structure. Specifically, we assume that the observed matrix $\mathbf{M}$ is created by adding small noise to the underlying low-rank matrix $\mathbf{L}$.

$$\mathbf{M} = \mathbf{L} + \zeta \quad \text{such that} \quad \|\mathbf{L}_{:i}\|_2 = 1 \;\; and \;\; \|\zeta_{:i}\|_2 \leq \epsilon \;\; \forall i \in [n]$$

The main novelty of the algorithm provided here is to decide the number of entries to be observed adaptively depending on the angle between estimated column space and actual column space. This approach to observation complexity opens further space for future improvements. In lemma 25 we show that the angle between estimated space and actual space cannot be too much different using similar argument to [2], and the angle between them is upper bounded by $3\frac{\pi}{2}\sqrt{k\epsilon}$, which gives the worst observation complexity for **LREBN** with $d = \mathcal{O}(\mu(\mathbf{U})r\log^2\frac{1}{\delta} + mk\epsilon\log\frac{1}{\delta})$ which improves the previous rate $\mathcal{O}(\mu(\mathbf{U})r\log^2\frac{1}{\delta} + mk\epsilon\log^2\frac{1}{\delta})$, especially when ϵ is relatively big that the term $mk\epsilon$ is dominating over $\mu(\mathbf{U})r\log^2\frac{1}{\epsilon}$.

Moreover, there are many cases that estimated angle is much smaller than $\sqrt{k\epsilon}$. especially, when the basis vectors of the matrix $\mathbf{L}$ are far enough by each other (the angles between them is big enough) this quantity can be as small as $k\epsilon$, which in this case observation complexity for a given column would be $d = 72\mu(\mathbf{U})r\log^2\frac{1}{\delta} + 8mk^2\epsilon^2\log\frac{r}{\delta}$ which is further smaller.

LREBN: Low-rank estimation for bounded noise.

Input: $d = 72\mu(\mathbf{U})r\log^2\frac{1}{\delta}$

Initialize: $k = 0, \widetilde{\mathbf{U}}^0 = \emptyset, \tilde{\theta}(\widetilde{\mathbf{U}}^0, \mathbf{U}^0) = 0$

1: Draw uniformly random entries $\Omega \subset [m]$ of size d
2: **for** i from 1 to n **do**
3: **if** $\|\mathbf{M}_{\Omega:i} - \mathcal{P}_{\widehat{\mathbf{U}}^{\mathbf{k}}_{\mathbf{\Omega}}}\mathbf{M}_{\Omega:i}\| > (1+\epsilon)\Big(\sqrt{\frac{3d}{2m}}\tilde{\theta}(\mathbf{U}^k, \widetilde{\mathbf{U}}^k) + \sqrt{\frac{3dk\epsilon}{2m}}\Big)$
4: Fully observe $\mathbf{M}_{:i}$
5: $\widetilde{\mathbf{U}}^{k+1} \leftarrow \widetilde{\mathbf{U}}^k \cup \mathbf{M}_{:\mathbf{i}}$, Orthogonalize $\widehat{\mathbf{U}}^{k+1}$
6: Estimate $\tilde{\theta}(\mathbf{U}^k, \widetilde{\mathbf{U}}^k)$ the upper bound for $\theta(\mathbf{U}^k, \widetilde{\mathbf{U}}^k)$
7: $d = 72\mu(\mathbf{U})r\log^2\frac{1}{\delta} + 8m\tilde{\theta}(\widetilde{\mathbf{U}}^k, \mathbf{U}^k)^2\log\frac{r}{\delta}$ and set $k = k+1$
8: Draw uniformly random entries $\Omega \subset [m]$ of size d
9: **otherwise:** $\widetilde{\mathbf{M}}_{:i} = \widehat{\mathbf{U}}^k\widehat{\mathbf{U}}^{k+}_{\Omega}\widetilde{\mathbf{M}}_{\Omega:i}$

Output: $\widetilde{\mathbf{M}}$

To estimate the upper bound $\tilde{\theta}(\mathbf{U}^k, \widetilde{\mathbf{U}}^k)$ for $\theta(\mathbf{U}^k, \widetilde{\mathbf{U}}^k)$ we use the idea due to [17]:

$$\tilde{\theta}(\widetilde{\mathbf{U}}^k, \mathbf{U}^k) = \frac{\pi}{2}\frac{\theta(u_k, \widetilde{u}_k)}{\theta(\widetilde{u}_k, \widetilde{\mathbf{U}}^{k-1}) - \tilde{\theta}(\widetilde{\mathbf{U}}^{k-1}, \mathbf{U}^{k-1})} + \tilde{\theta}(\widetilde{\mathbf{U}}^{k-1}, \mathbf{U}^{k-1})$$

Theorem 5. *Given the* $\mathbf{L}$ *be an* $m \times n$ *sized underlying rank-*r *matrix where each column has* ℓ_2 *norm of 1. Moreover,* $\mathbf{M}$ *is a full rank matrix where each column is created by adding at most* ℓ_2 *norm*$-\epsilon$ *noise to the corresponding column of* $\mathbf{L}$*. Then the algorithm* **LREBN** *estimates underlying matrix with* ℓ_2 *norm of error is* $\Theta(\frac{m}{d}\sqrt{k\epsilon})$ *by sampling* $d = 72\mu(\mathbf{U})r\log^2\frac{1}{\delta} + 8m\tilde{\theta}(\widetilde{\mathbf{U}}^k, \mathbf{U}^k)^2\log\frac{r}{\delta}$ *entries in each column*

Proof. The proof consists the following steps:

- *step 1.* Show that estimated matrix $\widetilde{\mathbf{M}}$ has rank of at most r.
- *step 2.* Decompose the error of recovered matrix
- *step 3.* Bound each of the terms in decomposition.

Step 1:

Lemma 13. *Let assume that* $\mathbf{M}$ *is can be decomposed as rank* r *matrix* $\mathbf{L}$ *with additional small noise in each column that, its* ℓ_2 *norm is bounded by* ϵ*. Then, at the end of the termination of the algorithm* **LREBN***, estimated subspace* $\widetilde{\mathbf{U}}^k$ *has dimension at most* r*.*

Proof. We prove that in the execution of the algorithm, we show if a column $\mathbf{M}_{:t}$ has been detected as new column that cannot be contained in preselected $\widetilde{\mathbf{U}}^k$, then $\mathbf{L}_{:t}$ is indeed cannot be contained in the $\mathbf{U}^k$. To use triangle inequality, we notice

$$\theta(\mathbf{L}_{:t}, \mathbf{U}^k) \geq \theta(\mathbf{L}_{:t}, \widetilde{\mathbf{U}}^k) - \theta(\widetilde{\mathbf{U}}^k, \mathbf{U}^k)$$

Using the lemma 20 we can notice that following inequalities are get satisfied:

$$\begin{aligned}
\|\mathbf{M}_{\Omega t} - \mathcal{P}_{\widetilde{\mathbf{U}}^k_\Omega}\mathbf{M}_{\Omega:t}\| &\leq \sqrt{\frac{3d}{2m}}\Big(\|\mathbf{M}_{:t} - \mathcal{P}_{\widetilde{\mathbf{U}}^k}\mathbf{M}_{:t}\|\Big) \\
&\leq \sqrt{\frac{3d}{2m}}\Big(\|\mathbf{M}_{:t} - \mathbf{L}_{:t}\| + \|\mathbf{L}_{:t} - \mathcal{P}_{\widetilde{\mathbf{U}}^k}\mathbf{L}_{:t}\|\Big) \\
&\qquad + \|\mathcal{P}_{\widetilde{\mathbf{U}}^k}(\mathbf{L}_{:t} - \mathbf{M}_{:t})\| \\
&\leq \sqrt{\frac{3d}{2m}}\Big(\epsilon + \theta(\mathbf{L}_{:t}, \widetilde{\mathbf{U}}^k) + \epsilon\Big)
\end{aligned}$$

From the design of the algorithm

$$\|\mathbf{M}_{\Omega t} - \mathcal{P}_{\widetilde{\mathbf{U}}^k_\Omega}\mathbf{M}_{\Omega:t}\| > (1+\epsilon)\Big(\sqrt{\frac{3d}{2m}}\theta(\mathbf{U}^k, \widetilde{\mathbf{U}}^k) + \sqrt{\frac{3dk\epsilon}{2m}}\Big)$$

and using this inequality above, we conclude that

$$\sqrt{\frac{3d}{2m}}\theta(\mathbf{U}^k, \widetilde{\mathbf{U}}^k) + \sqrt{\frac{3dk\epsilon}{2m}} < \sqrt{\frac{3d}{2m}}\Big(\epsilon + \theta(\mathbf{L}_{:t}, \widetilde{\mathbf{U}}^k) + \epsilon\Big)$$

which follows that

$$\theta(\mathbf{U}^k, \widetilde{\mathbf{U}}^k) + \sqrt{k\epsilon} < \Big(\epsilon + \theta(\mathbf{L}_{:t}, \widetilde{\mathbf{U}}^k) + \epsilon\Big)$$

considering the fact that $\epsilon < \frac{1}{4}$ we conclude that

$$\theta(\mathbf{L}_{:t}, \widetilde{\mathbf{U}}^k) \geq \theta(\mathbf{L}_{:t}, \widetilde{\mathbf{U}}^k) + 2\epsilon - \sqrt{k\epsilon} > \theta(\mathbf{U}^k, \widetilde{\mathbf{U}}^k)$$

therefore we conclude that $\theta(\mathbf{U}^k, \widetilde{\mathbf{U}}^k) < \theta(\mathbf{L}_{:t}, \widetilde{\mathbf{U}}^k)$ and it follows that $\theta(\mathbf{L}_{:t}, \mathbf{U}^k) > 0$. Moreover, one can see that after every time this inequality get satisfied, dimension of $\mathbf{U}^k$ increases by one, and considering the fact that $\mathbf{U}^k$'s are subspace of column space of $\mathbf{L}$, its dimension cannot increase more than r times. □

Step 2:

In this section we are formulating an upper bound to recovery error. Note that, if the algorithm decides completely observe the column, then ℓ_2 norm of the error is upper bounded by ϵ. Then, all we need to do is to give upper bound to columns those recovered by estimated subspace.

$$\|\widetilde{\mathbf{M}}_{:t} - \mathbf{L}_{:t}\| = \|\widetilde{\mathbf{U}}^k\widetilde{\mathbf{U}}^{k^+}_{\Omega:}\mathbf{M}_{\Omega t} - \mathbf{L}_{:t}\|$$

$$
\begin{aligned}
&\leq \|\widetilde{\mathbf{U}}^k\widetilde{\mathbf{U}}^{k^+}_{\Omega:}\mathbf{M}_{\Omega t} - \widetilde{\mathbf{U}}^k\widetilde{\mathbf{U}}^{k^+}_{\Omega:}\mathbf{L}_{\Omega:t}\| \\
&\qquad + \|\widetilde{\mathbf{U}}^k\widetilde{\mathbf{U}}^{k^+}_{\Omega}\mathbf{L}_{\Omega:t} - \widetilde{\mathbf{U}}^k\widetilde{\mathbf{U}}^{k^+}\mathbf{L}_{:t}\| + \|\widetilde{\mathbf{U}}^k\widetilde{\mathbf{U}}^{k^+}\mathbf{L}_{:t} - \mathbf{L}_{:t}\| \\
&\leq \|\widetilde{\mathbf{U}}^k\widetilde{\mathbf{U}}^{k^+}_{\Omega:}(\mathbf{M}_{\Omega t} - \mathbf{L}_{\Omega:t})\| \\
&\qquad + \|\widetilde{\mathbf{U}}^k\widetilde{\mathbf{U}}^{k^+}_{\Omega}\mathbf{L}_{\Omega:t} - \widetilde{\mathbf{U}}^k\widetilde{\mathbf{U}}^{k^+}\mathbf{L}_{:t}\| + \sin\theta(\mathbf{L}_{:t}, \widetilde{\mathbf{U}}^k) \\
&\leq \|\widetilde{\mathbf{U}}^k\widetilde{\mathbf{U}}^{k^+}_{\Omega:}\|\|(\mathbf{M}_{\Omega t} - \mathbf{L}_{\Omega:t})\| \\
&\qquad + \|\widetilde{\mathbf{U}}^k\widetilde{\mathbf{U}}^{k^+}_{\Omega}\mathbf{L}_{\Omega:t} - \widetilde{\mathbf{U}}^k\widetilde{\mathbf{U}}^{k^+}\mathbf{L}_{:t}\| + \theta(\mathbf{L}_{:t}, \widetilde{\mathbf{U}}^k)
\end{aligned}
$$

Step 3:

Then all we need to do is to give an upper bound to the final expression. Lets start with the second term here: $\mathbf{L}_{:t} = \widetilde{\mathbf{U}}^k v + e$ where $\widetilde{\mathbf{U}}^k v = \widetilde{\mathbf{U}}^k\widetilde{\mathbf{U}}^{k^+}\mathbf{L}_{:t}$ and note $\|e\| = \sin\theta(\mathbf{L}_{:t}, \widetilde{\mathbf{U}}^k) \leq \theta(\mathbf{L}_{:t}, \widetilde{\mathbf{U}}^k)$. Therefore:

$$
\widetilde{\mathbf{U}}^k\widetilde{\mathbf{U}}^{k^+}_{\Omega}\mathbf{L}_{\Omega:t} - \widetilde{\mathbf{U}}^k\widetilde{\mathbf{U}}^{k^+}\mathbf{L}_{:t} = \widetilde{\mathbf{U}}^k\widetilde{\mathbf{U}}^{k^+}_{\Omega}(\widetilde{\mathbf{U}}^k v + e) - \widetilde{\mathbf{U}}^k v = \widetilde{\mathbf{U}}^k\widetilde{\mathbf{U}}^{k^+}_{\Omega} e
$$

Hence we conclude that:

$$
\begin{aligned}
\|\widetilde{\mathbf{M}}_{:t} - \mathbf{L}_{:t}\| &\leq \|\widetilde{\mathbf{U}}^k\widetilde{\mathbf{U}}^{k^+}_{\Omega:}\|\|(\mathbf{M}_{\Omega t} - \mathbf{L}_{\Omega:t})\| \\
&\qquad + \|\widetilde{\mathbf{U}}^k\widetilde{\mathbf{U}}^{k^+}_{\Omega} e_\Omega\| + \theta(\mathbf{L}_{:t}, \widetilde{\mathbf{U}}^k) \\
&\leq \|\widetilde{\mathbf{U}}^k\widetilde{\mathbf{U}}^{k^+}_{\Omega:}\|\|(\mathbf{M}_{\Omega t} - \mathbf{L}_{\Omega:t})\| \\
&\qquad + \|\widetilde{\mathbf{U}}^k\widetilde{\mathbf{U}}^{k^+}_{\Omega}\|\theta(\mathbf{L}_{:t}, \widetilde{\mathbf{U}}^k) + \theta(\mathbf{L}_{:t}, \widetilde{\mathbf{U}}^k).
\end{aligned}
$$

To give upper bound to this expression, we notice

$$\|\widetilde{\mathbf{U}}^k\widetilde{\mathbf{U}}^{k+}_{\Omega:t}\| \leq \frac{\sigma_1(\widetilde{\mathbf{U}}^k)}{\sigma_k(\widetilde{\mathbf{U}}^k_{\Omega:})} \leq \Theta(\frac{m}{d})$$

given the condition that $d \geq 4\mu(\widetilde{\mathbf{U}}^k)k\log\frac{k}{\delta}$ from the lemma 21. From lemma 17 we know that $\mu(\widetilde{\mathbf{U}}^k) \leq 2\mu(\mathbf{U}^k)+2\frac{m}{k}\theta(\widetilde{\mathbf{U}}^k,\mathbf{U}^k)^2$ and from lemma 15 we notice that $k\mu(\widetilde{\mathbf{U}}^k) \leq r\mu(\mathbf{U})$. Then all together these facts concludes the selected

$$\begin{aligned} d &= 72\mu(\mathbf{U})r\log^2\frac{1}{\delta} + 8m\theta(\widetilde{\mathbf{U}}^k,\mathbf{U}^k)^2\log\frac{r}{\delta} \\ &\geq 8\mu(\mathbf{U})r\log\frac{r}{\delta} + 8m\theta(\widetilde{\mathbf{U}}^k,\mathbf{U}^k)^2\log\frac{r}{\delta} \end{aligned}$$

satisfies $d \geq 4\mu(\widetilde{\mathbf{U}}^k)k\log\frac{k}{\delta}$ (it is assumed that $\delta \leq \frac{1}{r^{1/8}}$). Therefore, we can bound $\|\widetilde{\mathbf{U}}^k\widetilde{\mathbf{U}}^{k+}_{\Omega:t}\|$ above by $\Theta(\frac{m}{d})$.

Now, only remaining term in the error bound above is $\theta(\mathbf{L}_{:t},\widetilde{\mathbf{U}}^k)$, and we use the following inequality to compare it with quantities provided as input:

$$\begin{aligned} \|\mathcal{P}_{\widetilde{\mathbf{U}}^k}\mathbf{M}_{:t} - \mathbf{L}_{:t}\| &\geq \sin\theta(\mathcal{P}_{\widetilde{\mathbf{U}}^k}\mathbf{M}_{:t},\mathbf{L}_{:t}) \\ &\geq \frac{\theta(\mathcal{P}_{\widetilde{\mathbf{U}}^k}\mathbf{M}_{:t},\mathbf{L}_{:t})}{2} \\ &\geq \frac{\theta(\widetilde{\mathbf{U}}^k,\mathbf{L}_{:t})}{2} \end{aligned}$$

and to relate the term $\|\mathcal{P}_{\widetilde{\mathbf{U}}^k}\mathbf{M}_{:t} - \mathbf{L}_{:t}\|$ with observed entries we again use the inequality 20 and the fact that $\left(1+2\log\frac{1}{\delta}\right)^2 \leq 6\log^2\frac{1}{\delta}$ once $\delta < 0.1$, lemma 15 and lemma 18

$$
\begin{aligned}
&\|\mathbf{M}_{\Omega:t} - \mathcal{P}_{\widetilde{\mathbf{U}}^k}\mathbf{M}_{\Omega:t}\| \geq \\
&\qquad \geq \sqrt{\frac{1}{m}\Big(\frac{d}{2} - \frac{3k\mu(\widetilde{\mathbf{U}}^k)\beta}{2}\Big)}\,\|\mathbf{M}_{:t} - \mathcal{P}_{\widetilde{\mathbf{U}}^k}\mathbf{M}_{:t}\| \\
&\qquad \geq \sqrt{\frac{1}{m}\Big(\frac{d}{2} - \frac{3k\mu(\widetilde{\mathbf{U}}^k)\beta}{2}\Big)}\,\Big(\|\mathbf{L}_{:t} - \mathcal{P}_{\widetilde{\mathbf{U}}^k}\mathbf{M}_{\Omega:t}\| - \|\mathbf{L}_{:t} - \mathbf{M}_{:t}\|\Big) \\
&\qquad \geq \sqrt{\frac{1}{m}\Big(\frac{d}{2} - \frac{3k\mu(\widetilde{\mathbf{U}}^k)\beta}{2}\Big)}\,\Big(\|\mathcal{P}_{\widetilde{\mathbf{U}}^k}\mathbf{M}_{:t} - \mathbf{L}_{:t}\| - \epsilon\Big) \\
&\qquad \geq \sqrt{\frac{1}{m}\Big(\frac{d}{2} - \frac{3k\mu(\widetilde{\mathbf{U}}^k)\beta}{2}\Big)}\,\Big(\frac{\theta(\widetilde{\mathbf{U}}^k, \mathbf{L}_{:t})}{2} - \epsilon\Big) \\
&\qquad \geq \sqrt{\frac{1}{m}\Big(\frac{d}{2} - 9k\mu(\widetilde{\mathbf{U}}^k)\log^2\frac{1}{\delta}\Big)}\,\Big(\frac{\theta(\widetilde{\mathbf{U}}^k, \mathbf{L}_{:t})}{2} - \epsilon\Big) \\
&\qquad \geq \sqrt{\frac{1}{m}\Big(\frac{d}{2} - 18k\mu(\mathbf{U}^k)\log^2\frac{1}{\delta} - 18m\theta(\widetilde{\mathbf{U}}^k, \mathbf{U}^k)^2\log^2\frac{1}{\delta}\Big)}\,\Big(\frac{\theta(\widetilde{\mathbf{U}}^k, \mathbf{L}_{:t})}{2} - \epsilon\Big) \\
&\qquad \geq \sqrt{\frac{d}{2m} - \frac{18k}{m}\mu(\mathbf{U}^k)\log^2\frac{1}{\delta} - 18\theta(\widetilde{\mathbf{U}}^k, \mathbf{U}^k)^2\log^2\frac{1}{\delta}}\Big(\frac{\theta(\widetilde{\mathbf{U}}^k, \mathbf{L}_{:t})}{2} - \epsilon\Big) \\
&\qquad \geq \sqrt{\frac{d}{2m} - \frac{18r}{m}\mu(\mathbf{U})\log^2\frac{1}{\delta} - 18\theta(\widetilde{\mathbf{U}}^k, \mathbf{U}^k)^2\log^2\frac{1}{\delta}}\Big(\frac{\theta(\widetilde{\mathbf{U}}^k, \mathbf{L}_{:t})}{2} - \epsilon\Big) \\
&\qquad \geq \sqrt{\frac{d}{4m}}\Big(\frac{\theta(\widetilde{\mathbf{U}}^k, \mathbf{L}_{:t})}{2} - \epsilon\Big)
\end{aligned}
$$

From here, we conclude that:

$$\begin{aligned}\theta(\widetilde{\mathbf{U}}^k, \mathbf{L}_{:t}) &\leq 4\sqrt{\frac{m}{d}}\|\mathbf{M}_{\Omega:t} - \mathcal{P}_{\widetilde{\mathbf{U}}^k}\mathbf{M}_{\Omega:t}\| + 2\epsilon \\ &\leq 4\sqrt{\frac{m}{d}}(1+\epsilon)\Big(\sqrt{\frac{3d}{2m}}\theta(\mathbf{U}^k, \widetilde{\mathbf{U}}^k) + \sqrt{\frac{3dk\epsilon}{2m}} + 2\epsilon\Big) \\ &\leq (1+\epsilon)\Big(\sqrt{24}\theta(\mathbf{U}^k, \widetilde{\mathbf{U}}^k) + \sqrt{8k\epsilon}\Big)\end{aligned}$$

Finally, returning back to the recovery error:

$$\begin{aligned}\|\widetilde{\mathbf{M}}_{:t} - \mathbf{L}_{:t}\| &\leq \|\widetilde{\mathbf{U}}^k\widetilde{\mathbf{U}}^{k+}_{\Omega:}\|\|(\mathbf{M}_{\Omega t} - \mathbf{L}_{\Omega:t})\| + \|\widetilde{\mathbf{U}}^k\widetilde{\mathbf{U}}^{k+}_{\Omega}\|\theta(\mathbf{L}_{:t}, \widetilde{\mathbf{U}}^k) + \theta(\mathbf{L}_{:t}, \widetilde{\mathbf{U}}^k) \\ &\leq \frac{m}{d}\epsilon + \Big(\frac{m}{d}+1\Big)\Big(\sqrt{24}\theta(\mathbf{U}^k, \widetilde{\mathbf{U}}^k) + \sqrt{8k\epsilon}\Big)(1+\epsilon)\end{aligned}$$

Then all we need to do is to give upper bound to $\theta(\widetilde{\mathbf{U}}^k, \mathbf{U}^k)$. In the proof below, we use similar argument to [17]. Lets assume $\mathbf{U}^k = \{u_1, u_2, \ldots, u_k\}$ and $\widetilde{\mathbf{U}}^k = \{\widetilde{u}_1, \widetilde{u}_2, \ldots, \widetilde{u}_k\}$ where each of $\|u_i - \widetilde{u}_i\| \leq \epsilon$ satisfied. Then using triangle inequality, lemma 19 and lemma 14

$$\begin{aligned}\theta(\widetilde{\mathbf{U}}^k, \mathbf{U}^k) &\leq \theta(\widetilde{\mathbf{U}}^k, \widehat{\mathbf{U}}) + \theta(\widehat{\mathbf{U}}, \mathbf{U}^k) \\ &\leq \frac{\pi}{2}\frac{\theta(u_k, \widetilde{u}_k)}{\theta(\widetilde{u}_k, \mathbf{U}^{k-1})} + \theta(\widetilde{\mathbf{U}}^{k-1}, \mathbf{U}^{k-1}) \\ &\leq \frac{\pi}{2}\frac{\theta(u_k, \widetilde{u}_k)}{\theta(\widetilde{u}_k, \widetilde{\mathbf{U}}^{k-1}) - \theta(\widetilde{\mathbf{U}}^{k-1}, \mathbf{U}^{k-1})} + \theta(\widetilde{\mathbf{U}}^{k-1}, \mathbf{U}^{k-1}) \\ &\leq \frac{\pi}{2}\frac{\theta(u_k, \widetilde{u}_k)}{\sqrt{k\epsilon} + \theta(\widetilde{\mathbf{U}}^{k-1}, \mathbf{U}^{k-1}) - \theta(\widetilde{\mathbf{U}}^{k-1}, \mathbf{U}^{k-1})} + \theta(\widetilde{\mathbf{U}}^{k-1}, \mathbf{U}^{k-1}) \\ &\leq \frac{\pi}{2}\frac{\epsilon}{\sqrt{k\epsilon}} + \theta(\widetilde{\mathbf{U}}^{k-1}, \mathbf{U}^{k-1})\end{aligned}$$

and using lemma 16, we can conclude that $\theta(\widetilde{\mathbf{U}}^k, \mathbf{U}^k) \leq \frac{3\pi}{2}\sqrt{k\epsilon}$, which gives the final bound to $\|\widetilde{\mathbf{M}}_{:t} - \mathbf{L}_{:t}\|$ to be $\Theta(\frac{m}{d}\sqrt{k\epsilon})$.

Lemma 14. *Given that*

$$\|\mathbf{M}_{\Omega t} - \mathcal{P}_{\widetilde{\mathbf{U}}_{\Omega}^{k-1}}\mathbf{M}_{\Omega:t}\| \geq (1+\epsilon)\Big(\sqrt{\frac{3d}{2m}}\theta(\widetilde{\mathbf{U}}^{k-1}, \mathbf{U}^{k-1}) + \sqrt{\frac{3dk\epsilon}{2m}}\Big)$$

satisfies. Then following also satisfies:

$$\theta(\widetilde{u}_k, \widetilde{\mathbf{U}}^{k-1}) \geq \theta(\widetilde{\mathbf{U}}^{k-1}, \mathbf{U}^{k-1}) + \sqrt{k\epsilon}$$

Proof. Note that simple triangle inequality implies $\|\mathbf{M}_{:t}\| \leq 1 + \epsilon$

$$\begin{aligned}
\theta(\widetilde{u}_k, \widetilde{\mathbf{U}}^{k-1}) = \theta(\mathbf{M}_{:t}, \widetilde{\mathbf{U}}^{k-1}) &\geq \sin\theta(\mathbf{M}_{:t}, \widetilde{\mathbf{U}}^{k-1}) \\
&\geq \frac{\mathbf{M}_{:t}}{1+\epsilon}\sin\theta(\mathbf{M}_{:t}, \widetilde{\mathbf{U}}^{k-1}) \\
&= \frac{1}{1+\epsilon}\|\mathbf{M}_{:t} - \mathcal{P}_{\widetilde{\mathbf{U}}^{k-1}}\mathbf{M}_{:t}\| \\
&\geq \frac{1}{1+\epsilon}\sqrt{\frac{2m}{3d}}\|\mathbf{M}_{\Omega t} - \mathcal{P}_{\widetilde{\mathbf{U}}_{\Omega}^{k-1}}\mathbf{M}_{\Omega:t}\|
\end{aligned}$$

Using, the fact that

$$\|\mathbf{M}_{\Omega t} - \mathcal{P}_{\widetilde{\mathbf{U}}_{\Omega}^{k-1}}\mathbf{M}_{\Omega:t}\| \geq (1+\epsilon)\Big(\sqrt{\frac{3d}{2m}}\theta(\widetilde{\mathbf{U}}^{k-1}, \mathbf{U}^{k-1}) + \sqrt{\frac{3dk\epsilon}{2m}}\Big)$$

we conclude

$$\theta(\widetilde{u}_k, \widetilde{\mathbf{U}}^{k-1}) \geq \sqrt{k\epsilon} + \theta(\widetilde{\mathbf{U}}^{k-1}, \mathbf{U}^{k-1}).$$

□

Lemma 15. *Let* $\mathbf{U}^k$ *be a* k*-dimensional subspace of* $\mathbf{U}$ *which is subspace of* $\mathbb{R}^m$ *with dimension* r*. Then following inequality satisfied:*

$$k\mu(\mathbf{U}^k) \leq r\mu(\mathbf{U})$$

Proof.

$$\begin{aligned} k\mu(\mathbf{U}^k) = k\frac{m}{k}\max_{1\leq j\leq m}\|\mathcal{P}_{\mathbf{U}^k}e_i\|^2 &= r\frac{m}{r}\max_{1\leq j\leq m}\|\mathcal{P}_{\mathbf{U}^k}e_i\|^2 \\ &\leq r\frac{m}{r}\max_{1\leq j\leq m}\|\mathcal{P}_{\mathbf{U}}e_i\|^2 \\ &= r\mu(\mathbf{U}) \end{aligned}$$

and the inequality due to $\mathbf{U}^k \subseteq \mathbf{U}$ □

Lemma 16. *Let assume that* $a_0 = 0$ *and* $a_k \leq a_{k-1} + \frac{\pi}{2}\sqrt{\frac{\epsilon}{k}}$*. Then it follows that* $a_k \leq \frac{3\pi}{2}\sqrt{k\epsilon}$

Proof. Its trivial to notice that $a_1 \leq \frac{\pi}{2}\sqrt{\epsilon} \leq 3\frac{\pi}{2}\sqrt{\epsilon}$. Lets assume by induction for a given k any index $i \leq k$ satisfies $a_i \leq 3\frac{\pi}{2}\sqrt{i\epsilon}$ and then we will prove that $a_{k+1} \leq 3\frac{\pi}{2}\sqrt{(k+1)\epsilon}$. We prove it by contradiction, by assuming $a_{k+1} > 3\frac{\pi}{2}\sqrt{(k+1)\epsilon}$ and conclude to a contradiction.

$$\begin{aligned} a_{k+1} &> 3\frac{\pi}{2}\sqrt{(k+1)\epsilon} \\ -a_k &\geq -3\frac{\pi}{2}\sqrt{k\epsilon} \end{aligned}$$

Therefore, $a_{k+1} - a_k \geq 3\frac{\pi}{2}\sqrt{\epsilon}\Big(\sqrt{k+1} - \sqrt{k}\Big) = 3\frac{\pi}{2}\sqrt{\epsilon}\frac{1}{\sqrt{k}+\sqrt{k+1}} \geq 3\frac{\pi}{2}\sqrt{\epsilon}\frac{1}{3\sqrt{k}} = \frac{\pi}{2}\sqrt{\frac{\epsilon}{k}}$ which contradicts to the statement of the lemma. Therefore, assumption cannot be satisfied which follows $a_{k+1} \leq 3\frac{\pi}{2}\sqrt{(k+1)\epsilon}$ □

Lemma 17. *Let $\widetilde{\mathbf{U}}^k$ and $\mathbf{U}^k$ be as defined above then, coherence number of these spaces satisfies the following inequality:*

$$\mu(\widetilde{\mathbf{U}}^k) \leq 2\mu(\mathbf{U}^k) + 2\frac{m}{k}\theta(\widetilde{\mathbf{U}}^k, \mathbf{U}^k)^2$$

Proof. In order to achieve the goal of comparing $\mu(\widetilde{\mathbf{U}}^k)$ and $\mu(\mathbf{U}^k)$, we first need to understand how projection to standard vectors to $\mathbf{U}^k$ differ than projection of them to $\widetilde{\mathbf{U}}^k$:

$$\begin{aligned}\|\mathcal{P}_{\widetilde{\mathbf{U}}^k} e_i\| \leq \|\mathcal{P}_{\mathbf{U}^k} e_i\| + \|\mathcal{P}_{\widetilde{\mathbf{U}}^k} e_i - \mathcal{P}_{\mathbf{U}^k} e_i\| &\leq \|\mathcal{P}_{\mathbf{U}^k} e_i\| + \|\mathcal{P}_{\widetilde{\mathbf{U}}^k} - \mathcal{P}_{\mathbf{U}^k}\|\|e_i\| \\ &= \|\mathcal{P}_{\mathbf{U}^k} e_i\| + \sin\theta(\widetilde{\mathbf{U}}^k, \mathbf{U}^k) \\ &\leq \|\mathcal{P}_{\mathbf{U}^k} e_i\| + \theta(\widetilde{\mathbf{U}}^k, \mathbf{U}^k)\end{aligned}$$

Therefore:

$$\begin{aligned}\mu(\widetilde{\mathbf{U}}^k) = \frac{m}{k}\max_{1\leq j\leq n}\|\mathcal{P}_{\widetilde{\mathbf{U}}^k} e_i\|^2 &\leq \frac{m}{k}\Big(2\max_{1\leq j\leq n}\|\mathcal{P}_{\mathbf{U}^k} e_i\|^2 + 2\theta(\widetilde{\mathbf{U}}^k, \mathbf{U}^k)^2\Big) \\ &= 2\mu(\mathbf{U}^k) + 2\frac{m}{k}\theta(\widetilde{\mathbf{U}}^k, \mathbf{U}^k)^2\end{aligned}$$

□

Lemma 18. *Lets assume the setting as discussed in the proof above. Then,*

$$\frac{d}{4m} > \frac{18r}{m}\mu(\mathbf{U})\log^2\frac{1}{\delta} + 18\theta(\widetilde{\mathbf{U}}^k, \mathbf{U}^k)^2\log^2\frac{1}{\delta}$$

Proof. Remind $d = 72\mu(\mathbf{U})r\log^2\frac{1}{\delta} + 8m\theta(\widetilde{\mathbf{U}}^k, \mathbf{U}^k)^2\log\frac{r}{\delta}$ and it implies $\frac{d}{4m} = 18\mu(\mathbf{U})r\log^2\frac{1}{\delta} + 2m\theta(\widetilde{\mathbf{U}}^k, \mathbf{U}^k)^2\log\frac{r}{\delta}$. Then all we need to show is $2m\log\frac{r}{\delta} > 18\log^2\frac{1}{\delta}$ and this is correct due to $m \geq d \geq 9\log\frac{1}{\delta}$ □

Lemma 19 ([17]). *Let subspaces U, V, and $\widetilde{V}$ defined as $U = span\{a_1, \ldots, a_{k-1}\}$, $V = span\{a_1, \ldots, a_{k-1}, b\}$ and $\widetilde{V} = span\{a_1, \ldots, a_{k-1}, \tilde{b}\}$. Then following inequality satisfied:*

$$\theta(V, \widetilde{V}) \leq \frac{\pi}{2} \frac{\theta(\tilde{b}, b)}{\theta(\tilde{b}, U)}$$

Lemma 20 ([16]). *Let $\widetilde{\mathbf{U}}^k$ be a k-dimensional subspace of $\mathbb{R}^m$, and set $d = \max(\frac{8}{3} k\mu(\widetilde{\mathbf{U}}^k) \log \frac{2k}{\delta}, 4\mu\mathcal{P}_{\widetilde{\mathbf{U}}^k} \log \frac{1}{\delta})$. Given that Ω stands for uniformly selected subset of $[m]$ then following inequality get satisfied:*

$$\frac{d(1-\alpha) - k\mu(\widetilde{\mathbf{U}}^k)\frac{\beta}{1-\zeta}}{m} \|y - \mathcal{P}_{\widetilde{\mathbf{U}}^k} y\| \leq \|y_\Omega - \mathcal{P}_{\widetilde{\mathbf{U}}^k_\Omega} y_\Omega\| \leq (1+\alpha)\frac{d}{m} \|y - \mathcal{P}_{\widetilde{\mathbf{U}}^k} y\|$$

where $\alpha = \sqrt{2\frac{\mu(\mathcal{P}_{\widetilde{\mathbf{U}}^{k\perp}} y)}{d} \log \frac{1}{\delta}} + 2\frac{\mu(\mathcal{P}_{\widetilde{\mathbf{U}}^{k\perp}} y)}{3d} \log \frac{1}{\delta}, \quad \beta = (1 + 2\log \frac{1}{\delta})^2$

$\zeta = \sqrt{\frac{8k\mu(\widetilde{\mathbf{U}}^{k\perp})}{3d} \log \frac{2r}{\delta}}$ *we use $\alpha < 1/2$ and $\gamma < 1/3$ similar to [16]*

Lemma 21 ([8]). *Consider a finite sequence $\{\mathbf{X}_k\} \in \mathbb{R}^{n\times n}$ independent random, Hermitian matrices those satisfies:*

$$0 \leq \lambda_{\min}(\mathbf{X}_k) \leq \lambda_{\max}(\mathbf{X}_k) \leq L.$$

Let $\mathbf{Y} = \sum_k \mathbf{X}_k$ and μ_r be the r-th largest eigenvalue of $\mathbb{E}[\mathbf{Y}]$ ($\mu_r = \lambda_r(\mathbb{E}[\mathbf{Y}])$), then for any $\epsilon \in [0, 1)$ following inequality satisfied:

$$\mathbf{Pr}(\lambda_r(\mathbf{Y}) \geq (1-\epsilon)\mu_r) \geq 1 - r\Big(\frac{e^{-\epsilon}}{(1-\epsilon)^{1-\epsilon}}\Big)^{\frac{\mu_r}{L}} \geq 1 - re^{\frac{\mu_r \epsilon^2}{2L}}$$

Chapter 4

Experimental Results

4.1 Design of Experiments

Exact Completion Designs: To generate $m \times n$ sized underlying low-rank matrix $\mathbf{M}$, we created $m \times r$ and $r \times n$ sized matrices $\mathbf{X}, \mathbf{Y}$, where $\mathbf{X}_{i,j}, \mathbf{Y}_{i,j} \sim \mathcal{N}(0,1)$ and we set $\mathbf{M} = \mathbf{XY}$.

To design an $m \times n$ sized rank r with *space nonsparsity number* equal to 1, we generate $m \times r-1$ and $r-1 \times n$ sized matrices $\mathbf{X}, \mathbf{Y}$, where $X_{i,j}, Y_{i,j} \sim \mathcal{N}(0,1)$. Multiplication of these matrices would gives us a rank $r-1$ matrix. As the column space of $\mathbf{M}$ is column space of $\mathbf{Y}$. Given that $Y_{i,j} \sim \mathcal{N}(0,1)$ implies that coherence of column space of $\mathbf{Y}$ is small. Therefore, $\mathbf{M}$ has small column space coherence number. Then we generate random a vector in $\mathbb{R}^n$, and replace it with one of the rows: row$-i$.

We can guarantee that the resulting matrix will contain i-th standard basis vector e_i in the column space. To observe this phenomenon, lets analyse the restriction of the matrix $\mathbf{M}$ to the first r columns and all the rows but

row i. As this is a submatrix of rank $r-1$ matrix (initial $\mathbf{M}$) this matrix also has rank at most $r-1$. Therefore, we have non-trivial coefficients $\alpha_1, \alpha_2, \ldots, \alpha_r$, that makes linear combination of columns of submatrix to be equal to zero vector. Therefore,

$$\alpha_1 \mathbf{M}_{:1} + \alpha_1 \mathbf{M}_{:2} + \ldots + \alpha_r \mathbf{M}_{:r} = we_i$$

for some w. Then all we need to show is $w \neq 0$, however it is straightforward because $w = \alpha_1 u_1 + \alpha_2 u_2 + \ldots + \alpha_r u_r$ which is nonzero because u is random. In conclusion, e_i is contained in the column space of $\mathbf{M}$ and therefore from the definition of coherence of the columns space of the matrix is equal to $\frac{m}{r}$ which is also maximum value. Moreover, it has nonsparsity-number to be equal to 1 as e_i is in column space.

Note that, we might change the process above and generate several random vectors and replace them with some rows of the matrix, we would get matrix which has *nonsparsity-number* to be equal to a.

1. generate a many random vector $u \in \mathbb{R}^n$
2. randomly select subset of $[m]$ that has size a
3. replace selected rows with randomly generated vectors.

We test our method for fixed $r = 5$ and $r = 10$ with n varying over $\{1000, 2000, \ldots, 12000\}$. Similarly, we fixed $n = 3000$ and varied r over $\{1, 2, \ldots, 12\}$. Moreover, we classified matrices into four types for coherence/incoherence of the column and row space.

Low Rank Estimation Designs: We use the Hoopokins-155 dataset for low-rank estimation algorithms.

4.2 Exact Completion Experiments

We tested all of our exact completion algorithms-**ERR**,**ERRE** and **EREI** for synthetically generated low-rank matrices. From tables below, we can get the comparison of the performance of our algorithms to previous methods ([2, 15, 16, 24]). We refer to Supplementary Material regarding generation method of coherent/incoherent, row/column subspace matrices.

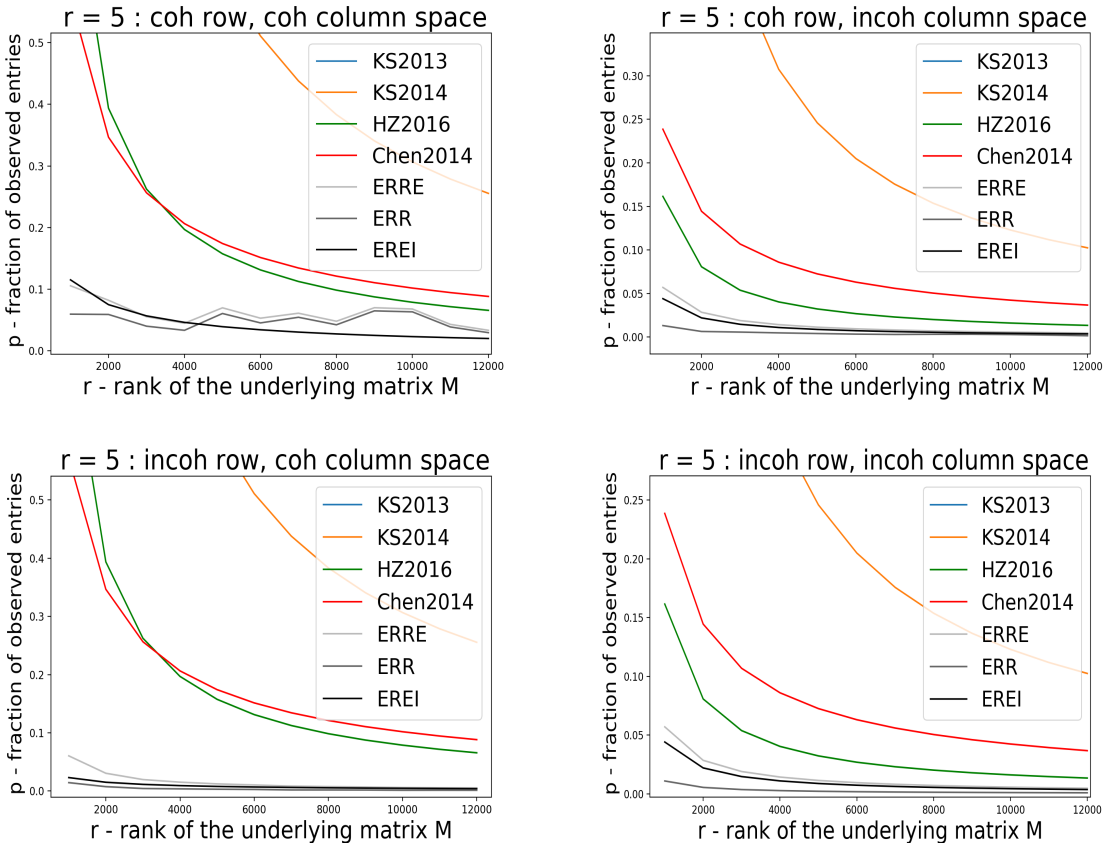


Table 4.1: Experiments for rank 5 matrices

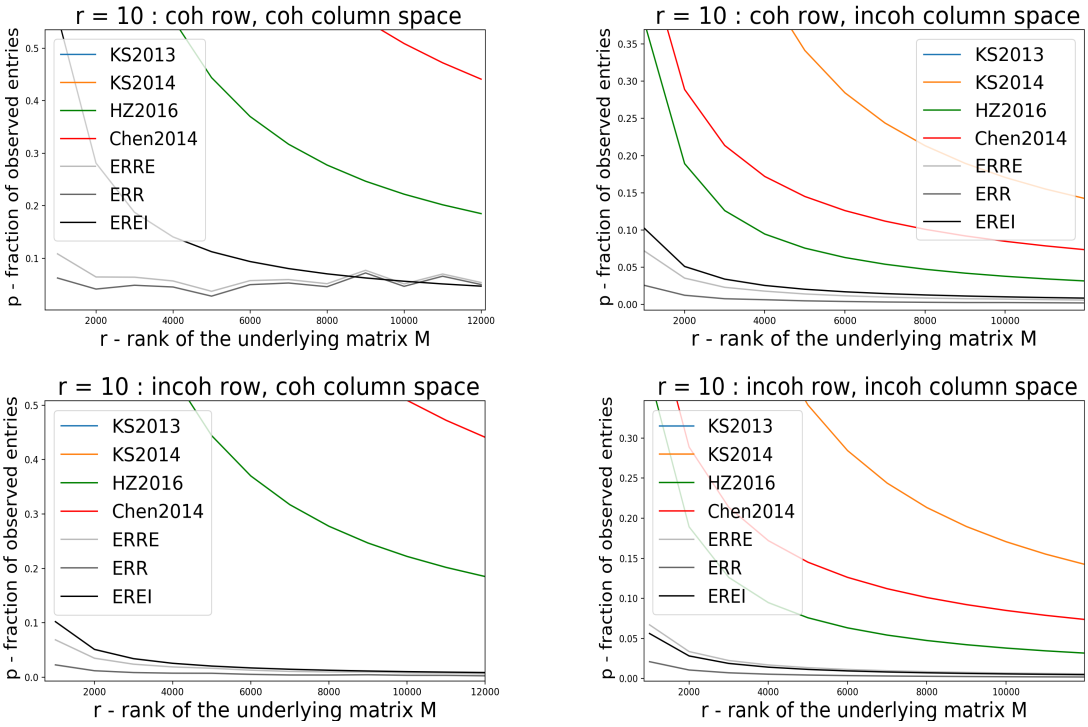


Table 4.2: Experiments for rank 10 matrices

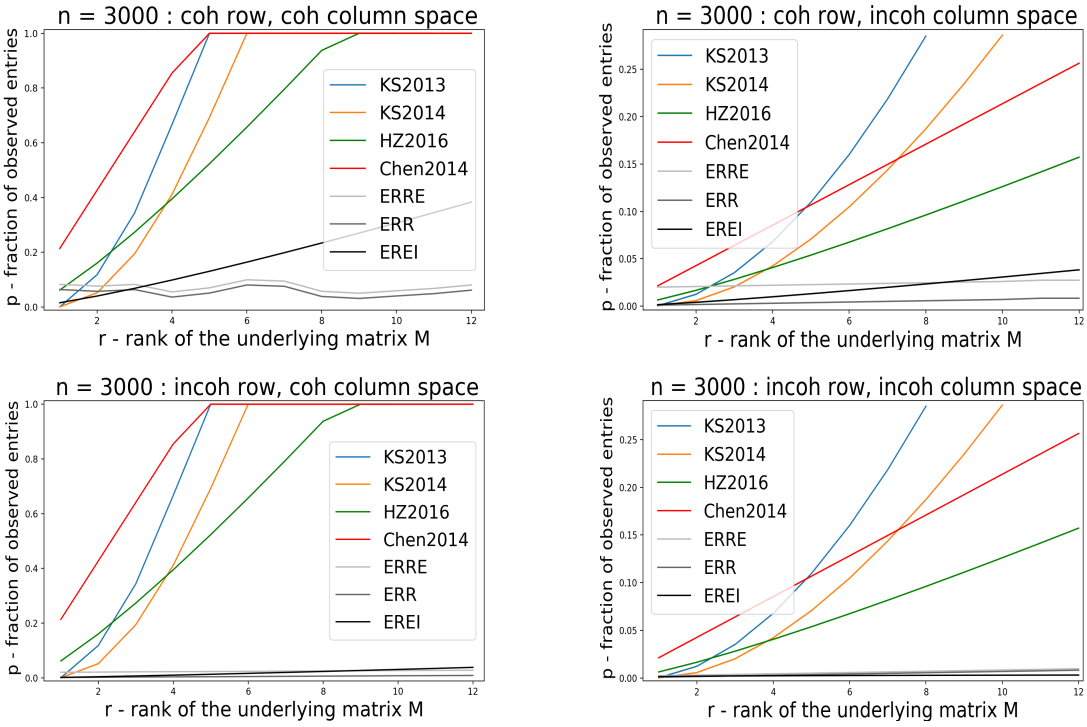


Table 4.3: Experiments for rank 3000×3000 sized matrices

4.3 Low-Rank Estimation Experiments

Experiments for real data: We tested the low-rank estimation method using adaptive sampling with respect to the angle between estimated and underlying subspace. Experiments run on Hopkins 155 dataset. Hopkins dataset contains 155 images, and each of them represents two or three motion objects. The trajectory of motion of each object lies in a low-rank subspace. We reshape 3-dimensional tensor data to 2-dimensional matrix. Almost all reshaped matrices are smaller than the size 500×50. As the constants are high in the algorithm LREBN and works of previous authors together with the multiplicative term of $\log(1/\epsilon)$, applying for these numbers here concludes observing the entire column. Therefore, instead of sampling precisely these numbers, we use other numbers that generated using a similar idea. If a newly added independent column has a high angle with the estimated subspace, then we don't increase the number of samples for the next columns. Otherwise, we increase the value of the d-observation count parameter. We compare this method to the method that no matter what is the angle, we always increase d once a new independent column detected. ϵ set to 0.0003, and the ℓ_2 norm of the column normalized using ℓ_2 norm of observed entries. We run each of the experiments 20 times and take the average of them to compute corresponding error and sample size. On the left side, we see plots for ℓ_∞-norm of estimation error in a given column. In the right, we see the number of observations used during the execution. Each row below corresponds to different images. Computing average error over columns and average sample sizes conclude that the adaptive method

roughly gives $8 - 12\%$ benefit on reducing observation complexity. However, the average ℓ_∞ norm of error doesn't change significantly.

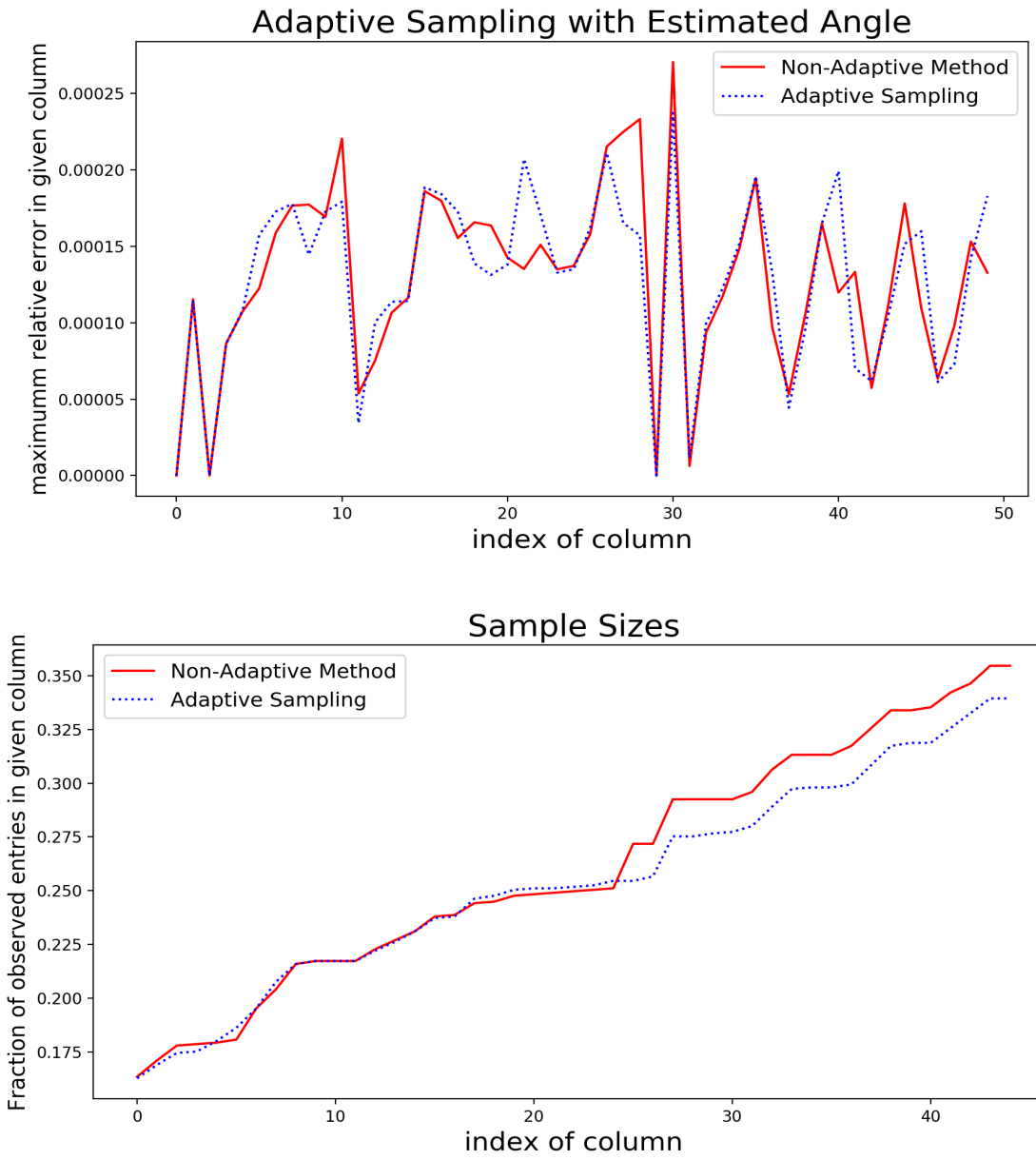
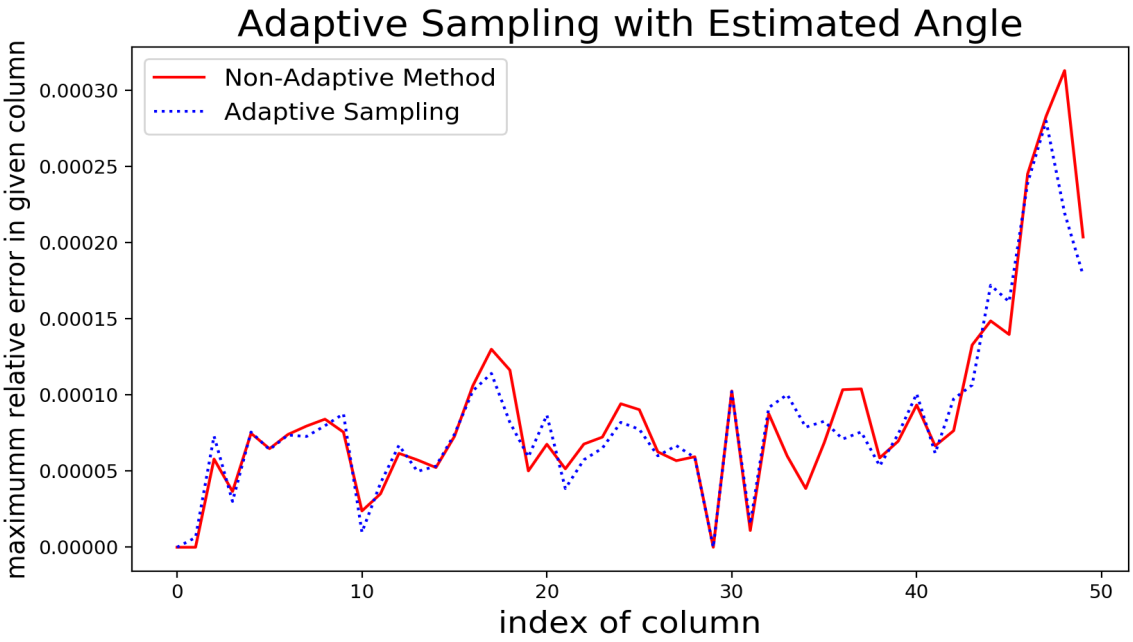


Table 4.4: Experiments for one of Hopkins 155 dataset



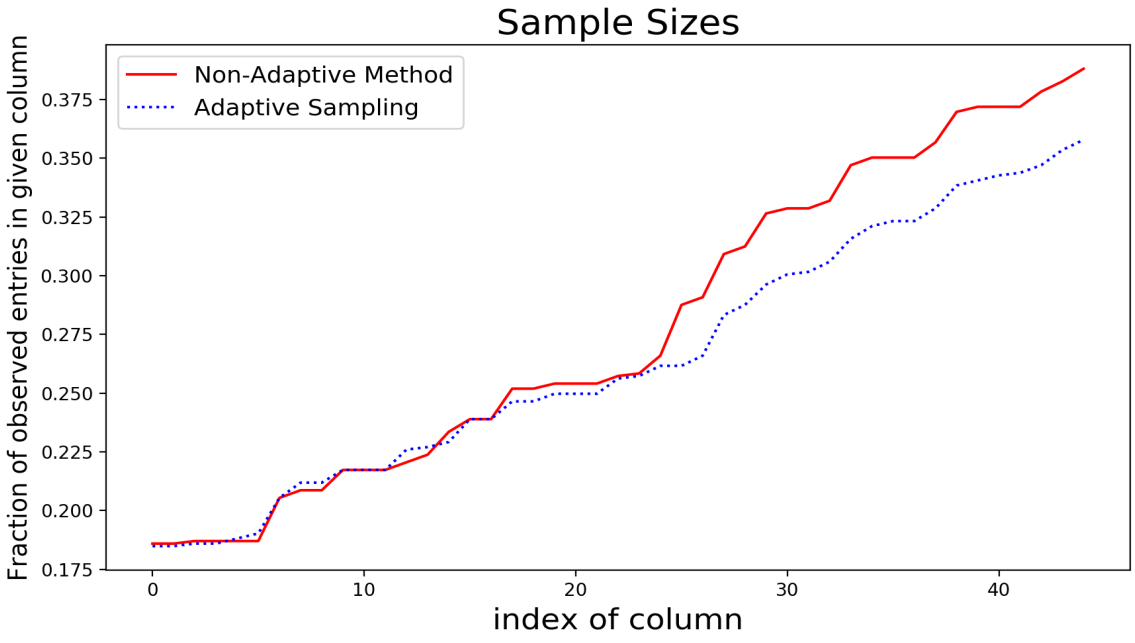
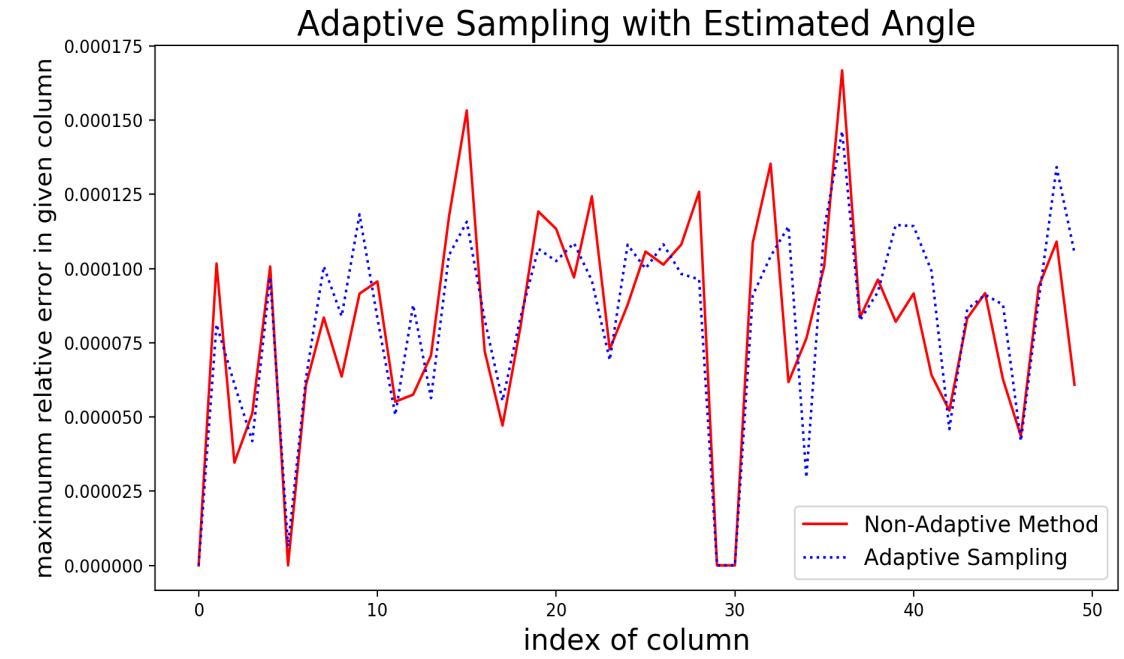


Table 4.5: Experiments for one of Hopkins 155 dataset



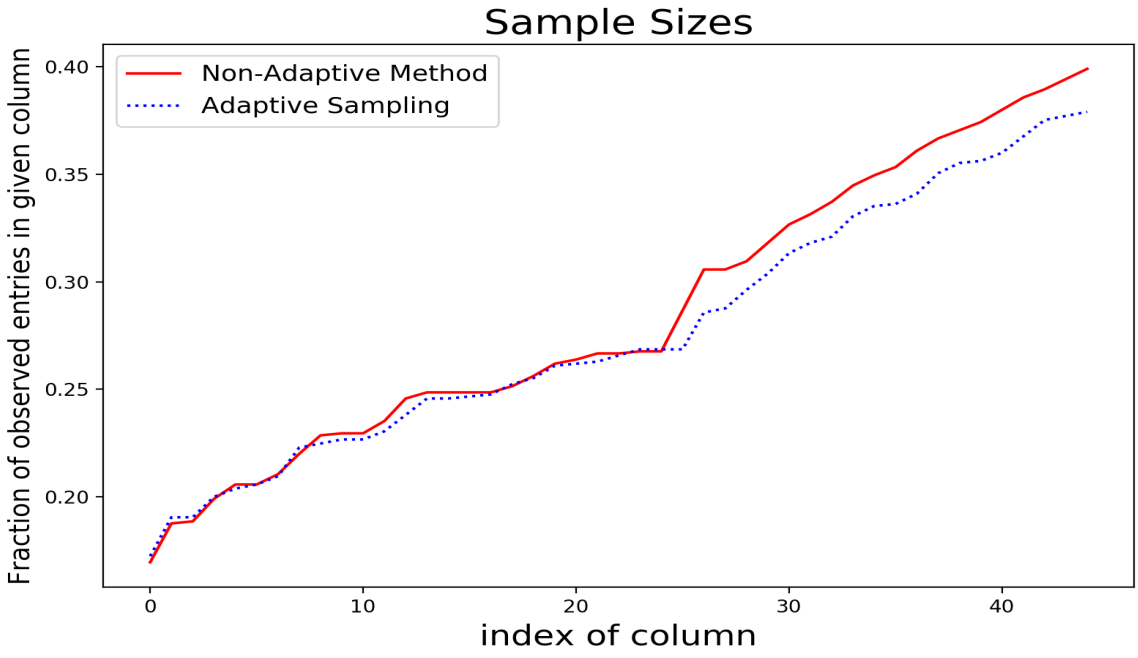


Table 4.6: Experiments for one of Hopkins 155 dataset

Bibliography

[1] Kolar M. Rinaldo A. Balakrishnan, S. and A. Singh. Recovering block-structured activations using compressive measurements. *arXiv:1209.3431*, 2012.

[2] M.F Balcan and H. Zhang. Noise-tolerant life-long matrix completion via adaptive sampling. *Advances in Neural Information Processing Systems*, 2016.

[3] E. J. Cand‘es and Y. Plan. Matrix completion with noise. *Proceedings of the IEEE.*, 2010.

[4] E. J. Candes and B. Recht. Exact matrix completion via convex optimization. *Foundations of Computational Mathematics*, 2009.

[5] E. J. Candés and T. Tao. The power of convex relaxation: Near-optimal matrix completion. *IEEE Transactions on Information Theory*, 2010.

[6] Y. Chen, S. Bhojanapalli, S. Sanghavi, and R. Ward. Coherent matrix completion. In *In Proceedings of The 31st International Conference on Machine Learning*, 2014.

[7] T. F. Coleman and A. Pothen. The null space problem i. complexity. *SIAM Journal on Algebraic Discrete Methods*, 1986.

[8] Alex Gittens and Joel A Tropp. Tail bounds for all eigenvalues of a sum of random matrices. *arXiv preprint: 1104.4513*, 2011.

[9] D. Gross. Recovering low-rank matrices from few coefficients in any basis. *Information Theory, IEEE Transactions on.*, 2011.

[10] J. Haupt, R. Castro, and R. Nowak. Adaptive sampling for sparse detection and estimation. *IEEE Transactions on Information Theory*, 2011.

[11] C. a. E. H.Zhang, Z.Lin. Exact recoverability of robust pca via outlier pursuit with tight recovery bounds. *In AAAI Conference on Artificial Intelligence.*, 2015.

[12] Ramazanli Ilqar and Barnabas Poczos. Optimal adaptive matrix completion. In *https://arxiv.org/pdf/2002.02431.pdf*, 2013.

[13] AndreaMontanari Keshavan, R. H. and S. Oh. Matrix completion from noisy entries. *The Journal of Machine Learning Research.*, 2010.

[14] V. Koltchinskii, K. Lounici, and A. B. Tsybakov. Nuclear-norm penalization and optimal rates for noisy low-rank matrix completion. *The Annals of Statistics*, 2011.

[15] A. Krishnamurthy and A. Singh. Low-rank matrix and tensor completion via adaptive sampling. In *Advances in Neural Information Processing Systems*, 2013.

[16] A. Krishnamurthy and A. Singh. On the power of adaptivity in matrix completion and approximation. *arXiv preprint*, 2014.

[17] A. Blum M.-F. Balcan and S. Vempala. Efficient representations for life-long learning and autoencoding. *In Annual Conference on Learning Theory*, 2015.

[18] T. T. Mai and P. Alquier. A bayesian approach for noisy matrix completion: Optimal rate under general sampling distribution. *Electronic Journal of Statistics*, 2015.

[19] M. Malloy and R. Nowak. Sequential analysis in high-dimensional multiple testing and sparse recovery. *IEEE International Symposium on Information Theory.*, 2011.

[20] S. T. McCormick. A combinatorial approach to some sparse matrix problems. *DTIC Document*, 1983.

[21] S. Negahban and M. J. Wainwright. Restricted strong convexity and weighted matrix completion: Optimal bounds with noise. *The Journal of Machine Learning Research.*, 2012.

[22] Ju Sun Qing Qu and John Wright. Finding a sparse vector in a subspace: linear sparsity using alternating directions. *IEEE Transaction-RANSACTION onON information theoryINFORMATION THEORY*, 2015.

[23] B. Recht. A simpler approach to matrix completion. *The Journal of Machine Learning Research*, 2011.

[24] Srinadh Bhojanapalli. Sujay Sanghavi Rachel Ward. Yudong Chen. Coherent matrix completion. *International Conference on Machine Learning*, 2014.

[25] E. T‘anczos and R. Castro. Adaptive sensing for estimation of structured sparse signals. *arXiv:1311.7118*, 2013.

[26] A. Vince. A framework for the greedy algorithm. *Discrete Applied Mathematics*, 2002.

[27] M. K. Warmuth and D. Kuzmin. Randomized online pca algorithms with regret bounds that are logarithmic in the dimension. *Journal of Machine Learning Research*, 2008.