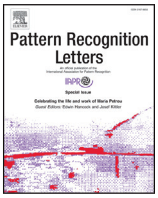
Pattern Recognition Letters 120 (2019) 31–35

Contents lists available at ScienceDirect

Pattern Recognition Letters

journal homepage: www.elsevier.com/locate/patrec



Clustering a union of low-rank subspaces of different dimensions with missing data[☆]

Morteza Ashraphijuo, Xiaodong Wang[*]

Columbia University, New York, NY 10027, USA

ARTICLE INFO

Article history:
Received 8 March 2018
Available online 28 December 2018

Keywords:
Subspace clustering
Low-rank matrix completion
Union of subspaces

ABSTRACT

We derive fundamental conditions for clustering a union of low-rank subspaces with missing data. In particular, given an incomplete matrix, assuming its columns are drawn from K different subspaces with different dimensions, the subspace clustering problem is to cluster the columns that belong to the same subspace. We derive a lower bound on the number of columns from each subspace such that the columns can be clustered correctly with high probability. The analysis focuses on the subspace with the lowest dimension and is a generalization of the corresponding results in [18] that assumes the subspaces are independent and with the same dimension.

1. Introduction

In many practical applications, we need to analyze a collection of datasets like images, text documents, etc. To model such data structures, we can consider a matrix $\mathbf{U} \in \mathbb{R}^{n_1 \times n_2}$ whose columns are chosen from one of K unknown subspaces. The problem of subspace clustering aims to cluster the columns of this matrix to K groups such that the columns in each group belong to the same subspace. Subspace clustering is an important pre-processing step of data analysis when the data lies in a union of subspaces and is well studied [10,11,13]. The problem is much more challenging with missing data, i.e., when the matrix $\mathbf{U}$ is incomplete, which is an important problem in subspace learning for real-world scenarios and is studied broadly [8,15,17–19,21]. Subspace clustering can be also used as a prepossessing step for data completion problem. Retrieving the missing entries, i.e., data completion, has many applications and there are various works in this area [2,5,6,14]. Subspace clustering has many applications in various fields including image processing [12], recommender systems [20], etc.

In [18], it is assumed that all the K unknown subspaces have the same dimension and are chosen independently from the Grassmannian manifold $\text{Gr}(n_1, r)$ (set of all r-dimensional subspaces of the n_1-dimensional space). It is shown that if the number of samples per column is above a threshold, and assuming that there exists an r-dimensional subspace that fits enough number of columns of the sampled matrix, then it is ensured that this subspace is one of the K unknown subspaces and all the covered columns belong to that subspace. The key condition for this to hold is that the number of columns of $\mathbf{U}$ drawn from each of the K subspaces should be more than $(r+1)(n_1-r+1) = \mathcal{O}(rn_1)$. This bound is interesting since before it was only known to be necessary when each column includes $r+1$ sampled entries [9]. The similar algebraic geometry approaches as in [18] have been studied in [1,3,4,7] for data completion and sensing problem.

In this paper, we consider the general scenario that the K unknown subspaces are chosen (not necessarily independently) from K different Grassmannian manifolds with different dimensions $\text{Gr}(n_1, r_1), \dots, \text{Gr}(n_1, r_K)$. Our main result states that if at least $K(r_{\max}+1)(n_1 - r_{\max}+1)$ columns are drawn from each subspace, where $r_{\max} = \max_{1 \leq k \leq K} r_k$, then the columns can be correctly clustered with high probability. The key approach in our analysis is to cluster the subspaces from the lowest dimension to the highest.

2. Background

Given positive integers $r_1, r_2, \dots, r_K$, we consider K different subspaces $\mathcal{S}_1, \dots, \mathcal{S}_K$ chosen from the Grassmannian manifolds $\text{Gr}(n_1, r_k)$, $k = 1, \dots, K$. Let $\mathcal{I}_k$ be a set of c_k columns chosen generically from the mentioned r_k-dimensional subspace (drawn independently according to a continuous distribution with respect to the Lebesgue measure on the mentioned r_k-dimensional subspace), $k = 1, \dots, K$. Assume that $\mathbf{U} \in \mathbb{R}^{n_1 \times n_2}$ is a matrix such that its $n_2 = \Sigma_{k=1}^{K} c_k$ columns are the union of all columns in $\{\mathcal{I}_k, k = 1, \dots, K\}$. However, these n_2 columns are blended so that we do not know the source subspace of each column.

[☆] **Conflict of interest.** The authors, Morteza Ashraphijuo and Xiaodong Wang, are both employed at Columbia University and both of them are supported financially only by Columbia University in New York City.

[*] Corresponding author.

E-mail addresses: ashraphijuo@ee.columbia.edu (M. Ashraphijuo), wangx@ee.columbia.edu (X. Wang).

https://doi.org/10.1016/j.patrec.2018.12.016
0167-8655/© 2019 Elsevier B.V. All rights reserved.

We assume that $\mathbf{U}$ is randomly sampled, i.e., each entry of $\mathbf{U}$ is independently sampled with probability $0 < p < 1$. Let $\mathbf{\Omega}$ be an $n_1 \times n_2$ binary sampling matrix such that $\mathbf{\Omega}(i, j) = 1$ if $\mathbf{U}(i, j)$ is sampled and $\mathbf{\Omega}(i, j) = 0$ otherwise. Let $\mathbf{U}_{\mathbf{\Omega}}$ denote the incomplete matrix consisting of only the sampled entries of $\mathbf{U}$. We are interested in clustering the columns of the sampled matrix $\mathbf{U}_{\mathbf{\Omega}}$ into K groups such that the columns in each group belong to one subspace with high probability. The number of columns in $\mathcal{I}_k$, i.e., c_k and the sampling probability p or the number of sampled entries in $\mathbf{U}_{\mathbf{\Omega}}$ are key parameters in this problem.

In [18], the above subspace clustering problem with missing data is studied for the special case that $r_1 = r_2 = \cdots = r_K = r$ and the subspaces $\mathcal{S}_1, \ldots, \mathcal{S}_K$ are independent. Here, we restate the main result of Pimentel-Alarcón and Nowak [18] (i.e., Theorems 1 and 3 in [18]) using our notations in Theorem 2.2, which provides a lower bound on c_k's and the number of samples per column (which can be translated in terms of p) such that the columns of $\mathbf{U}_{\mathbf{\Omega}}$ chosen from the same subspace, can be correctly clustered with high probability. First we restate Theorem 3 in [16] that characterizes a condition on unique completability of $\mathbf{U}_{\mathbf{\Omega}}$, i.e., a condition that ensures there exists a unique way to complete $\mathbf{U}_{\mathbf{\Omega}}$ while satisfying the rank constraint, as the following theorem. This theorem is used to show the main result in [18], i.e., Theorem 2.2 below, as well as our new result, i.e., Theorem 3.4 in Section 3.

Theorem 2.1 ([16]). *Assume that a generic rank-r matrix* $\mathbf{U} \in \mathbb{R}^{n_1 \times n_2}$ *with* $r \leq \frac{n_1}{6}$ *and* $n_2 \geq (r+1)(n_1 - r + 1)$ *is randomly sampled such that each column of* $\mathbf{U}_{\mathbf{\Omega}}$ *includes at least l sampled entries where*

$$l > \max\left\{12\left(\log\left(\frac{n_1(r+1)}{\epsilon}\right) + 1\right), 2r\right\} \tag{1}$$

for some $0 < \epsilon < 1$. *Then, the sampled matrix* $\mathbf{U}_{\mathbf{\Omega}}$ *is uniquely completable with probability at least* $1 - \epsilon$.

Definition 2.1. Consider a subspace $\mathcal{S} \in \text{Gr}(n_1, r)$ and a sampled column $\mathbf{u}_{\mathbf{\Omega}} \in \mathbb{R}^{n_1 \times 1}$. We say that $\mathcal{S}$ fits $\mathbf{u}_{\mathbf{\Omega}}$ or $\mathbf{u}_{\mathbf{\Omega}}$ can be covered (generated) by $\mathcal{S}$ if there exists at least one completion of $\mathbf{u}_{\mathbf{\Omega}}$ that belongs to $\mathcal{S}$.

Theorem 2.2 ([18]). *Assume that the subspaces* $\mathcal{S}_1, \ldots, \mathcal{S}_K$ *are* ***independently*** *chosen from* $\text{Gr}(n_1, r)$, $r = r_1 = r_2 = \cdots = r_K \leq \frac{n_1}{6}$ *and* $c_k \geq (r+1)(n_1 - r + 1)$, $k = 1, \ldots, K$. *Moreover suppose that each column of* $\mathbf{U}_{\mathbf{\Omega}}$ *includes at least l sampled entries such that (1) holds. Let* $\bar{\mathcal{S}}$ *denote an r-dimensional subspace that fits exactly* $\bar{c}$ *columns of* $\mathbf{U}_{\mathbf{\Omega}}$ *(i.e.,* $\bar{c}$ *is the maximum number of columns of* $\mathbf{U}_{\mathbf{\Omega}}$ *that can be covered by* $\bar{\mathcal{S}}$*) and assume that* $\bar{c} \geq (r+1)(n_1 - r + 1)$. *Then, with probability at least* $1 - K\epsilon$, *the following statement holds: All the* $\bar{c}$ *columns of* $\mathbf{U}_{\mathbf{\Omega}}$ *covered by* $\bar{\mathcal{S}}$ *belong to one source* $\mathcal{I}_{k_0}$ *for some* $1 \leq k_0 \leq K$ *and the rest of the columns of* $\mathbf{U}_{\mathbf{\Omega}}$ *do not belong to* $\mathcal{I}_{k_0}$ *and moreover,* $\bar{c} = c_{k_0}$ *and* $\bar{\mathcal{S}} = \mathcal{S}_{k_0}$.

3. Main results

We are interested in generalizing Theorem 2.2 to the general scenario when $r_1, \ldots, r_K$ are not necessarily equal and also, the subspaces $\mathcal{S}_1, \ldots, \mathcal{S}_K$ are not chosen independently.

We start by stating some basic properties as a consequence of the genericity assumption. Consider a matrix $\mathbf{X} \in \mathbb{R}^{n_1 \times n_2}$ whose columns are drawn generically from a subspace that belongs to $\text{Gr}(n_1, r)$, where $n_1 \geq r$ and $n_2 \geq r$. Then, with probability one, $\mathbf{X}$ is a rank-r matrix. More specifically, with probability one, any r columns (or any r rows) of $\mathbf{X}$ are linearly independent; and any $r \times r$ submatrix of $\mathbf{X}$ is full-rank. Further, given two different subspaces $\mathcal{S}_1 \subset \mathcal{S}_2$ and a column $\mathbf{u}$ that is drawn generically from $\mathcal{S}_2$, we have $\mathbf{u} \notin \mathcal{S}_1$ with probability one.

The following three lemmas are instrumental to the proof of our clustering result, i.e., Theorem 3.4.

Lemma 3.1. *Let* $\mathbf{X}_0$ *be a rank-*$(r-1)$ *matrix and* $\mathbf{X}_0(i, j) = x$ *be an entry of this matrix. Assume that changing the value of entry* $\mathbf{X}_0(i, j)$ *from x to y results in* $\mathbf{X}_0'$, *which is a rank-r matrix. Then, there are infinitely many scalars z such that changing the value of entry* $\mathbf{X}_0(i, j)$ *from x to z results in a rank-r matrix.*

Proof. Since $\mathbf{X}_0$ is a rank-$(r-1)$ matrix, the determinant of any $r \times r$ submatrix of $\mathbf{X}_0$ is zero. Moreover, $\mathbf{X}_0'$ is a rank-r matrix and hence, there exists an $r \times r$ submatrix $\mathbf{X}_r$ of $\mathbf{X}_0$ such that changing the value of the corresponding entry of $\mathbf{X}_r$ from x to y results in a non-zero determinant. Since changing the order of the rows and columns does not affect the values of rank and determinant, we can assume that $\mathbf{X}_r(1, 1) = \mathbf{X}_0(i, j) = x$. Hence, $\det(\mathbf{X}_r) = 0$ and changing the value of $\mathbf{X}_r(1, 1)$ from x to y makes the determinant of $\mathbf{X}_r$ non-zero. On the other hand, we have

$$\begin{aligned} 0 = \det(\mathbf{X}_r) &= \mathbf{X}_r(1,1)\det(\mathbf{X}_r(2:r, 2:r)) \\ &- \Sigma_{i=2}^{r}(-1)^i \mathbf{X}_r(1,i)\det(\mathbf{X}_r(2:r, \{1,\ldots,r\} \backslash \{i\})), \end{aligned} \tag{2}$$

or equivalently,

$$\begin{aligned} x&\det(\mathbf{X}_r(2:r, 2:r)) \\ &= \Sigma_{i=2}^{r}(-1)^i \mathbf{X}_r(1,i)\det(\mathbf{X}_r(2:r, \{1,\ldots,r\} \backslash \{i\})). \end{aligned} \tag{3}$$

Moreover, we have

$$\begin{aligned} y&\det(\mathbf{X}_r(2:r, 2:r)) \\ &\neq \Sigma_{i=2}^{r}(-1)^i \mathbf{X}_r(1,i)\det(\mathbf{X}_r(2:r, \{1,\ldots,r\} \backslash \{i\})). \end{aligned} \tag{4}$$

Note that if $\det(\mathbf{X}_r(2:r, 2:r)) = 0$, then both sides of (3) are zero; and hence both sides of (4) are zero; which contradicts the inequality in (4). Hence, $\det(\mathbf{X}_r(2: r, 2: r)) \neq 0$, then (3) and (4) can be written as

$$y \neq x = \frac{\Sigma_{i=2}^{r}(-1)^i \mathbf{X}_r(1,i)\det(\mathbf{X}_r(2:r, \{1,\ldots,r\} \backslash \{i\}))}{\det(\mathbf{X}_r(2:r, 2:r))}. \tag{5}$$

Therefore, changing the value of $\mathbf{X}_r(1, 1)$ (i.e., $\mathbf{X}_0(i, j)$) from x to any $z \neq x$ leads $\mathbf{X}_r$ to an $r \times r$ full-rank matrix. As a result, there are infinitely many scalars z such that changing the value of $\mathbf{X}_0(i, j)$ from x to z results in the existence of a full-rank $r \times r$ submatrix, i.e., results in a matrix with rank at least r.

On the other hand, changing the value of only one entry of a matrix can affect the rank of the matrix by at most one, i.e., the rank can decrease or increase by one or stay the same. This is because changing one single entry of the matrix affects only one column of the matrix. Hence, the rank cannot decrease or increase by more than one. Therefore, there are infinitely many scalars z such that changing the value of $\mathbf{X}_0(i, j)$ from x to z results in a rank-r matrix. □

Lemma 3.2. *Consider a sampled matrix* $\mathbf{X}_{\mathbf{\Omega}}$ *such that there exist at least one rank-*$(r-1)$ *and one rank-r completion for some* $r > 1$. *Then, there exists infinitely many rank-r completions of* $\mathbf{X}_{\mathbf{\Omega}}$.

Proof. Note that changing the value of only one entry of a matrix results in changing the rank of the matrix by at most one. Let $\mathbf{X}_1$ and $\mathbf{X}_2$ denote the rank-$(r-1)$ and rank-r completions, respectively. $\mathbf{X}_1$ and $\mathbf{X}_2$ are the same over the sampled entries, i.e., $(\mathbf{X}_1)_{\mathbf{\Omega}} = (\mathbf{X}_2)_{\mathbf{\Omega}}$, and their difference is only over some of the non-sampled entries. We start changing the value of non-sampled entries of $\mathbf{X}_1$ one by one to the value of the corresponding non-sampled entries of $\mathbf{X}_2$, which will eventually result in $\mathbf{X}_2$ if we continue this for all non-sampled entries. While performing this simple process, we simply increase the rank from $r-1$ to r at some step by changing a non-sampled entry. This is because at the beginning the rank of the matrix is $r-1$ and at the end the rank is r and also at each step the rank changes by at most one.

Hence, there exists a rank-$(r-1)$ completion $\mathbf{X}_3$ of the sampled matrix $\mathbf{X}_{\mathbf{\Omega}}$ such that changing the value of some entry $\mathbf{X}_3(i, j)$ from

v to v' increases the rank to r for some scalars v and v'. The rest of the proof is straight-forward due to Lemma 3.1. □

Lemma 3.3. *Consider a sampled matrix $\mathbf{X}_{\mathbf{\Omega}}$ such that there exist at least one rank-$(r-i)$ and one rank-r completion for some $i<r$. Then, there exist infinitely many rank-r completions of $\mathbf{X}_{\mathbf{\Omega}}$.*

Proof. Using the same process described in the proof of Lemma 3.2, i.e., changing the values of the non-sampled entries of the rank-$(r-i)$ completion to reach to the rank-r completion, it is trivial to see that there exists at least one rank-$(r-1)$ completion as well. Hence, according to Lemma 3.2, there exists infinitely many rank-r completions of $\mathbf{X}_{\mathbf{\Omega}}$. □

The following theorem extends Theorem 2.2 to the general case and provides the conditions to correctly cluster the columns chosen from one of the subspaces that is of the lowest dimension, with high probability.

Theorem 3.4. *Without loss of generality, assume that $r_1 \leq r_2 \leq \cdots \leq r_K$ and denote $r_{\max} = \max_{1\leq k\leq K} r_k = r_K$. Assume further that $r_{\max} \leq \frac{n_1}{6}$, $c_k \geq K(r_{\max}+1)(n_1 - r_{\max}+1)$, $k=1,\dots,K$, and also, each column of $\mathbf{U}_{\mathbf{\Omega}}$ includes at least l sampled entries such that $l > \max\{12(\log(\frac{n_1(r_{\max}+1)}{\epsilon})+1), 2r_{\max}\}$. Let $\bar{\mathcal{S}}$ denote an r_1-dimensional subspace that fits exactly $\bar{c}$ columns of $\mathbf{U}_{\mathbf{\Omega}}$ (i.e., $\bar{c}$ is the maximum number of columns of $\mathbf{U}_{\mathbf{\Omega}}$ that can be covered by $\bar{\mathcal{S}}$) and assume that $\bar{c} \geq K(r_{\max}+1)(n_1-r_{\max}+1)$. Then, with probability at least $1-\epsilon$ the following statement holds: All the $\bar{c}$ columns of $\mathbf{U}_{\mathbf{\Omega}}$ covered by $\bar{\mathcal{S}}$ belong to one source $\mathcal{I}_{k_0}$ for some $1 \leq k_0 \leq K$ that $r_{k_0} = r_1$ (if $r_1 < r_2$ then $k_0 = 1$ and otherwise there are more options for k_0) and the rest of the columns of $\mathbf{U}_{\mathbf{\Omega}}$ do not belong to $\mathcal{I}_{k_0}$ and moreover, $\bar{c} = c_{k_0}$ and $\bar{\mathcal{S}} = \mathcal{S}_{k_0}$.*

Proof. According to pigeonhole principle, at least $\lceil \frac{\bar{c}}{K} \rceil \geq (r_{\max}+1)(n_1 - r_{\max}+1)$ columns of the $\bar{c}$ covered columns by $\bar{\mathcal{S}}$ are chosen from one source $\mathcal{I}_{k_0}$. Note that due to the assumptions $r_{\max} \geq r_{k_0}$ and $r_{\max} \leq \frac{n_1}{6}$ we have $(r_{\max}+1)(n_1-r_{\max}+1) \geq (r_{k_0}+1)(n_1-r_{k_0}+1)$ and hence, there are at least $(r_{k_0}+1)(n_1-r_{k_0}+1)$ columns covered by $\bar{\mathcal{S}}$ that are chosen from one source $\mathcal{I}_{k_0}$. Then, according to Theorem 2.1, there exists a unique rank-r_{k_0} completion for the mentioned $(r_{k_0}+1)(n_1-r_{k_0}+1)$ columns with probability at least $1-\epsilon$. In the rest of the proof, we assume the mentioned unique completability holds and show the mentioned statement holds with probability one.

First, we show that $r_{k_0} = r_1$. By contradiction, assume otherwise that $r_1 < r_{k_0}$. Recall that $\bar{\mathcal{S}}$ is an r_1-dimensional subspace that fits the mentioned $(r_{k_0}+1)(n_1-r_{k_0}+1)$ columns and hence, there exists a rank-r_1 completion of these columns. Hence, according to Lemma 3.3, there exist infinitely many rank-r_{k_0} completions of these columns, which contradicts the earlier uniqueness assumption. As a result, we have $r_{k_0} = r_1$ with probability one.

Therefore, again according to the uniqueness of rank-r_{k_0} completion assumption, and due to the fact that both subspaces $\bar{\mathcal{S}}$ and $\mathcal{S}_{k_0}$ are r_1-dimensional (since $r_{k_0} = r_1$) and they both fit the mentioned $(r_{k_0}+1)(n_1-r_{k_0}+1)$ columns, we simply conclude $\bar{\mathcal{S}} = \mathcal{S}_{k_0}$. Consequently, $\bar{\mathcal{S}}$ covers all c_{k_0} columns of $\mathbf{U}_{\mathbf{\Omega}}$ that belong to $\mathcal{I}_{k_0}$. In order to complete the proof, it suffices to show that $\bar{c} = c_{k_0}$, i.e., $\bar{\mathcal{S}}$ does not cover any other column of $\mathbf{U}_{\mathbf{\Omega}}$ that belongs to other sources $\mathcal{I}_k$ for $k \neq k_0$, with probability one.

Note that any column chosen from sources other than I_{k_0} does not belong to $\mathcal{S}_{k_0}$ with probability one (this statement is not valid if $r_{k_0} \neq \min\{r_1, r_2, \dots, r_K\}$ as will be discussed in Remark 3.1). This is because none of the other subspaces can be a subspace of $\mathcal{S}_{k_0}$ as $r_{k_0} = r_1 = \min\{r_1, r_2, \dots, r_K\}$. By contradiction, assume that a column $\mathbf{u}_{\mathbf{\Omega}}$ of $\mathbf{U}_{\mathbf{\Omega}}$ is chosen from I_{k_1} (for some $k_1 \neq k_0$) and it can be covered by $\bar{\mathcal{S}}$. Recall that $l > \max\{12(\log(\frac{n_1(r_{\max}+1)}{\epsilon})+1), 2r_{\max}\}$ holds and therefore, $\mathbf{u}_{\mathbf{\Omega}}$ includes at least $2r_{\max} \geq 2r_1$ sampled entries. Now, consider r_1 random columns of $\mathbf{U}_{\mathbf{\Omega}}$ that belong to I_{k_0} and denote it by $\mathbf{U}_{0_{\mathbf{\Omega}}}$. Also, let the unique completion of $\mathbf{U}_{0_{\mathbf{\Omega}}}$ be $\mathbf{U}_0$. Then, define $\mathbf{U}_{1_{\mathbf{\Omega}}} = [\mathbf{U}_0 | \mathbf{u}_{\mathbf{\Omega}}] \in \mathbb{R}^{n_1 \times (r_1+1)}$ (where $\mathbf{U}_0$ denotes the corresponding r_1 columns of the unique completion that is not given to us and $\mathbf{u}_{\mathbf{\Omega}}$ is an incomplete column; so only the last column of $\mathbf{U}_{1_{\mathbf{\Omega}}}$ is incomplete) and consider an $(r_1+1)\times(r_1+1)$ submatrix of $\mathbf{U}_{1_{\mathbf{\Omega}}}$ that includes r_1+1 of the sampled entries of $\mathbf{u}_{\mathbf{\Omega}}$ and denote it by $\mathbf{U}'_1$. Since $r_{k_1} \geq r_{k_0} = r_1$ and $\mathcal{S}_{k_1} \neq \mathcal{S}_{k_0}$ (because $k_1 \neq k_0$), we conclude that $\text{rank}(\mathbf{U}_1) = r_1+1$ and hence, $\text{rank}(\mathbf{U}'_1) = r_1+1$ with probability one, where $\mathbf{U}_1$ denotes the original (before sampling) matrix corresponding to $\mathbf{U}_{1_{\mathbf{\Omega}}}$. Hence, for any completion of $\mathbf{U}_{1_{\mathbf{\Omega}}}$, there exists a "fixed" and full-rank $(r_1+1)\times(r_1+1)$ submatrix. Therefore, $\bar{\mathcal{S}}$ cannot fit $\mathbf{u}_{\mathbf{\Omega}}$ with probability one (since $\bar{\mathcal{S}}$ is an r_1-dimensional subspace) and the proof is complete due to this contradiction. □

Remark 3.1. Note that the above proof is valid since $r_{k_0} = r_1 = \min\{r_1, r_2, \dots, r_K\}$, as mentioned in the last part of the proof. Moreover, we can show that if $r_{k_0} \neq \min\{r_1, r_2, \dots, r_K\}$, the statement of the theorem does not hold. For example, consider the scenario when $r_1 < r_2 < \ldots < r_K$ and $\mathcal{S}_k$ is a subspace of $\mathcal{S}_{k+1}$ (this can happen as the subspaces are not necessarily independent), $k = 1, \dots, K-1$. Now, assume that $\bar{\mathcal{S}}$ in the statement of the above theorem is r_2-dimensional instead of r_1-dimensional. Then if $\bar{\mathcal{S}} = \mathcal{S}_2$, $\bar{\mathcal{S}}$ also fits the columns drawn from $\mathcal{S}_1$ (recall that $\mathcal{S}_1$ is a subspace of $\mathcal{S}_2$) and hence, we cannot distinguish the columns drawn from $\mathcal{S}_1$ and $\mathcal{S}_2$.

Remark 3.2. Theorem 3.4 requires K times more columns from each unknown subspace in comparison with Theorem 2.2 to identify the columns of one subspace. However, Theorem 3.4 does not require all ranks to be the same or the independent subspace assumption. Moreover, the probability of clustering failure in Theorem 3.4 is K times less than that in Theorem 2.2.

After identifying all columns chosen from an r_1-dimensional subspace correctly, we can exclude the identified columns from the sampled matrix. Then, the problem reduces to the similar problem with $K-1$ subspaces of ranks $r_2 \leq \cdots \leq r_K$ and a smaller number of columns for the sampled matrix. Hence, the same analysis is applicable again.

Specifically, let $\bar{\mathcal{S}}_1, \dots, \bar{\mathcal{S}}_{K'}$ (for some $1 \leq K' < K$) denote different $r_1-, \dots, r_{K'}-$dimensional subspaces that fit exactly $\bar{c}_1, \dots, \bar{c}_{K'}$ columns of $\mathbf{U}_{\mathbf{\Omega}}$ (i.e., $\bar{c}_k$ is the maximum number of columns of $\mathbf{U}_{\mathbf{\Omega}}$ that can be covered by $\bar{\mathcal{S}}_k$), respectively, and assume that $\bar{c}_k \geq K(r_{\max}+1)(n_1-r_{\max}+1)$, $k=1,\dots,K'$. Moreover, assume that there exist $K(r_{\max}+1)(n_1-r_{\max}+1)$ columns covered by $\bar{\mathcal{S}}_k$ that cannot be covered by any of $\bar{\mathcal{S}}_1, \dots, \bar{\mathcal{S}}_{k-1}$, $k=1,\dots,K'$.

Then, according to Theorem 3.4, we have $\bar{\mathcal{S}}_1 = \mathcal{S}_{k_1}$ and $\bar{c}_1 = c_{k_1}$ with probability at least $1-\epsilon$. Moreover, we can exclude all the c_{k_1} columns from the sampled matrix and the identified subspace $\mathcal{S}_{k_1}$. Then, the new sampled matrix is $n_1 \times (n_2 - c_{k_1})$ and the columns of this matrix are chosen from the $K-1$ remaining subspaces. Then, similarly, we apply Theorem 3.4 for the r_2-dimensional subspace $\bar{\mathcal{S}}_2$ that has the lowest dimension now (because one r_1-dimensional subspace has been excluded).

Now, assuming that the clustering of the c_{k_1} columns in the previous step was correct, we can cluster the columns of the next subspace correctly with probability at least $1-\epsilon$. This can be done because due to the assumption, after excluding the columns of the first cluster, there exist $K(r_{\max}+1)(n_1-r_{\max}+1)$ columns covered by $\bar{\mathcal{S}}_2$ that cannot be covered by $\bar{\mathcal{S}}_1$. Hence, we apply Theorem 3.4 again and therefore, with probability at least $(1-\epsilon)^2$ the following statement holds: All the $\bar{c}_k$ columns of $\mathbf{U}_{\mathbf{\Omega}}$ covered by $\bar{\mathcal{S}}_k$ belong to one source $\mathcal{I}_{k'}$ such that $r_{k'} = r_k$ and the rest of the columns of $\mathbf{U}_{\mathbf{\Omega}}$ do not belong to $\mathcal{I}_{k'}$ and moreover, $\bar{c}_k = c_{k'}$ and

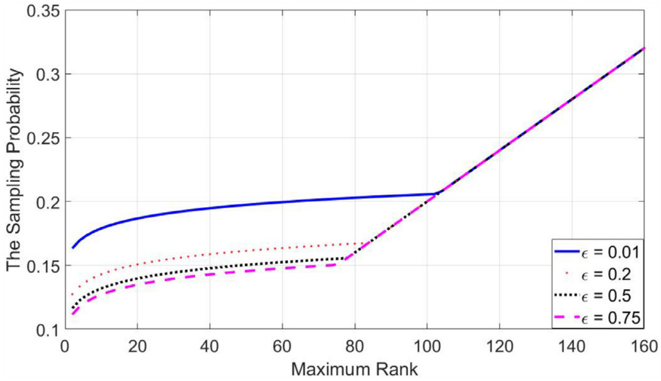


Fig. 1. The required sampling probability for correctly clustering with probability at least 0.99, where $n_1 = 1000$ and $c_1 = c_2 = c_3 = 600000$.

$\bar{\mathcal{S}}_k = \mathcal{S}_{k'}$, $k = 1, 2$. By simply repeating this procedure, we conclude the following corollary.

Corollary 3.5. *Without loss of generality, assume that* $r_1 \le r_2 \le \cdots \le r_K$ *and denote* $r_{\max} = \max_{1 \le k \le K} r_k = r_K$. *Assume further that* $r_{\max} \le \frac{n_1}{6}$, $c_k \ge K(r_{\max}+1)(n_1 - r_{\max}+1)$, $k = 1, \ldots, K$, *and also, each column of* $\mathbf{U_\Omega}$ *includes at least* l *sampled entries such that* $l > \max\{12(\log(\frac{n_1(r_{\max}+1)}{\epsilon})+1), 2r_{\max}\}$. *Let* $\bar{\mathcal{S}}_1, \ldots, \bar{\mathcal{S}}_{K'}$ *(for some* $1 \le K' < K$*) denote different* $r_1, \ldots, r_{K'}$ *dimensional subspaces that fits exactly* $\bar{c}_1, \ldots, \bar{c}_{K'}$ *columns of* $\mathbf{U_\Omega}$ *(i.e.,* $\bar{c}_k$ *is the maximum number of columns of* $\mathbf{U_\Omega}$ *that can be covered by* $\bar{\mathcal{S}}_k$*), respectively, and assume that* $\bar{c}_k \ge K(r_{\max}+1)(n_1 - r_{\max}+1)$, $k = 1, \ldots, K'$. *Moreover, assume that there exist* $K(r_{\max}+1)(n_1 - r_{\max}+1)$ *columns covered by* $\bar{\mathcal{S}}_k$ *that cannot be covered by any of* $\bar{\mathcal{S}}_1, \ldots, \bar{\mathcal{S}}_{k-1}$, $k = 1, \ldots, K'$. *Then, with probability at least* $(1-\epsilon)^{K'}$ *the following statement holds: All the* $\bar{c}_k$ *columns of* $\mathbf{U_\Omega}$ *covered by* $\bar{\mathcal{S}}_k$ *belong to one source* $\mathcal{I}_{k'}$ *such that* $r_{k'} = r_k$ *and the rest of the columns of* $\mathbf{U_\Omega}$ *do not belong to* $\mathcal{I}_{k'}$ *and moreover,* $\bar{c}_k = c_{k'}$ *and* $\bar{\mathcal{S}}_k = \mathcal{S}_{k'}$, $k = 1, \ldots, K'$.

We would like to emphaszie the advantage of our results when the number of sampled entries are as low as $\mathcal{O}(n_1 r_{\max})$ per column. Please refer to [18] to see the discussion on how tight our information-theoretic bounds on the number of samples are in comparison with the theoretical bounds in the existing works on subspace clustering with missing data. Moreover, our results in this paper not only improved the bound on the number of sampled entries in [18], but also removed the strong restrictions such as independency of the subspaces or subspaces being of the same size.

4. Numerical experiments

Assume that $n_1 = 1000$ and $c_1 = c_2 = c_3 = 600000$. We construct $K = 3$ matrices of rank r_i by multiplying a random $n_1 \times r_i$ matrix by a random $r_i \times c_i$ matrix. We assume each entry is sampled uniformly and independently with some sampling probability p. Since in our probabilistic analysis, only the maximum rank $r_{\max}$ matters (our bounds and analyses are based on the maximum rank), the x-axis in Fig. 1 represents the maximum rank. Also, the y-axis represents the required sampling probability. Then, using Corollary 3.5, the average number of required samples to guarantee the correct clustering with probability at least $1-\epsilon$ is $\frac{((\max\{12(\log(\frac{n_1(r_{\max}+1)}{\epsilon})+1), 2r_{\max}\})}{n_1}$.

Hence, in Fig. 1, we have provided several curves to represent the value of sampling probability and certainty value using our analysis. Each curve represents the probability of sampling (for different rank value) such that according to Corollary 3.5, we can guarantee the correct clustering with probability at least $1-\epsilon$, for different values of ϵ.

Note that our analysis is more efficient for relatively low-rank scenarios. This is because as long as $2r_{\max} < 12(\log(\frac{n_1(r_{\max}+1)}{\epsilon})+1)$, we basically provide a very tight bound on the number of samples to for correctly clustering with probability $1-\epsilon$. However, as we need $2r_{\max}$ samples as well in Corollary 3.5 (since we used Theorem 2.1), we can observe that by increasing the value of rank to a very large number (high-rank scenarios) the bound can be slightly weak and ϵ disappears in the curves as it means we can guarantee the correct clustering with probability almost 1.

5. Conclusions

We have developed a generalization to the low-rank subspace clustering conditions in [18]. In particular, given an incomplete matrix whose columns are drawn from K independent subspaces with the same dimension, a lower bound on the number of columns from each subspace is given in [18], such that, with high probability, the columns are clustered correctly. In order to treat the general case that the subspaces are not independently chosen, and their dimensions can be different, we have provided a new analysis that leads to the lower bound on the number of columns from each subspace, for the general case, which is K times that in [18]; however, the probability of clustering failure is reduced by a factor of K compared with that in [18]. The key approach in our analysis is to focus on the subspace of the lowest dimension.

Acknowledgments

This work was supported in part by the U.S. National Science Foundation (NSF) under grant CIF1064575, and in part by the U.S. Office of Naval Research (ONR) under grant N000141410667.

References

[1] M. Ashraphijuo, V. Aggarwal, X. Wang, On deterministic sampling patterns for robust low-rank matrix completion, IEEE Signal Process. Lett. 25 (3) (2018) 343–347.

[2] M. Ashraphijuo, R. Madani, J. Lavaei, Characterization of rank-constrained feasibility problems via a finite number of convex programs, in: IEEE 55th Conference on Decision and Control, IEEE, 2016, pp. 6544–6550.

[3] M. Ashraphijuo, X. Wang, Fundamental conditions for low-cp-rank tensor completion, J. Mach. Learn. Res. 18 (63) (2017) 1–29.

[4] M. Ashraphijuo, X. Wang, A characterization of sampling patterns for union of low-rank subspaces retrieval problem, Int. Symp. Artif. Intell. Math. (2018) 1–8.

[5] M. Ashraphijuo, X. Wang, V. Aggarwal, An approximation of the CP-rank of a partially sampled tensor, in: Annual Allerton Conference on Communication, Control, and Computing, IEEE, 2017, pp. 604–611.

[6] M. Ashraphijuo, X. Wang, V. Aggarwal, A characterization of sampling patterns for low-rank multi-view data completion problem, in: IEEE International Symposium on Information Theory, IEEE, 2017, pp. 1147–1151.

[7] M. Ashraphijuo, X. Wang, V. Aggarwal, Rank determination for low-rank data completion, J. Mach. Learn. Res. 18 (98) (2017) 1–29.

[8] L. Balzano, B. Eriksson, R. Nowak, High rank matrix completion and subspace clustering with missing data, in: Conference on Artificial Intelligence and Statistics, 2012.

[9] E.J. Candès, B. Recht, Exact matrix completion via convex optimization, Found. Comput. Math. 9 (6) (2009) 717.

[10] E. Elhamifar, R. Vidal, Sparse subspace clustering, in: IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2009, pp. 2790–2797.

[11] E. Elhamifar, R. Vidal, Sparse subspace clustering: algorithm, theory, and applications, IEEE Trans. Pattern Anal. Mach. Intell. 35 (11) (2013) 2765–2781.

[12] W. Hong, J. Wright, K. Huang, Y. Ma, Multiscale hybrid linear models for lossy image representation, IEEE Trans. Image Process. 15 (12) (2006) 3655–3671.

[13] G. Liu, Z. Lin, S. Yan, J. Sun, Y. Yu, Y. Ma, Robust recovery of subspace structures by low-rank representation, IEEE Trans. Pattern Anal. Mach. Intell. 35 (1) (2013) 171–184.

[14] X.-Y. Liu, S. Aeron, V. Aggarwal, X. Wang, M.-Y. Wu, Tensor completion via adaptive sampling of tensor fibers: Application to efficient indoor rf fingerprinting, in: IEEE International Conference on Acoustics, Speech and Signal Processing, IEEE, 2016, pp. 2529–2533.

[15] D. Pimentel-Alarcón, L. Balzano, R. Nowak, Necessary and sufficient conditions for sketched subspace clustering, in: Annual Allerton Conference on Communication, Control, and Computing, IEEE, 2016, pp. 1335–1343.

[16] D. Pimentel-Alarcón, N. Boston, R.D. Nowak, A characterization of deterministic sampling patterns for low-rank matrix completion, IEEE J. Sel. Top. Signal Process. 10 (4) (2016) 623–636.

[17] D. Pimentel-Alarcón, R. Nowak, L. Balzano, On the sample complexity of subspace clustering with missing data, in: IEEE Workshop on Statistical Signal Processing, IEEE, 2014, pp. 280–283.

[18] D. Pimentel-Alarcón, R. Nowak, The information-theoretic requirements of subspace clustering with missing data, in: International Conference on Machine Learning, 2016, pp. 802–810.

[19] D.L. Pimentel-Alarcón, N. Boston, R.D. Nowak, Deterministic conditions for subspace identifiability from incomplete sampling, in: IEEE International Symposium on Information Theory, IEEE, 2015, pp. 2191–2195.

[20] J.D. Rennie, N. Srebro, Fast maximum margin matrix factorization for collaborative prediction, in: International Conference on Machine Learning, ACM, 2005, pp. 713–719.

[21] C. Yang, D. Robinson, R. Vidal, Sparse subspace clustering with missing entries, in: International Conference on Machine Learning, 2015, pp. 2463–2472.