A Converse to Low-Rank Matrix Completion

Daniel L. Pimentel-Alarcón, Robert D. Nowak
University of Wisconsin-Madison

***Abstract*—In many practical applications, one is given a subset $\mathbf{\Omega}$ of the entries in a $d \times N$ data matrix $\mathbf{X}$, and aims to infer all the missing entries. Existing theory in low-rank matrix completion (LRMC) provides conditions on $\mathbf{X}$ (e.g., bounded coherence or genericity) and $\mathbf{\Omega}$ (e.g., uniform random sampling or deterministic combinatorial conditions) to guarantee that *if* $\mathbf{X}$ *is rank-r, then* $\mathbf{X}$ *is the only rank-r matrix that agrees with the observed entries*, and hence $\mathbf{X}$ can be uniquely recovered by some method (e.g., nuclear norm or alternating minimization).**

In many situations, though, one does not know beforehand the rank of $\mathbf{X}$, and depending on $\mathbf{X}$ and $\mathbf{\Omega}$, there may be rank-r matrices that agree with the observed entries, even if $\mathbf{X}$ is not rank-r. Hence one can be deceived into thinking that $\mathbf{X}$ is rank-r when it really is not. In this paper we give conditions on $\mathbf{X}$ (genericity) and a deterministic condition on $\mathbf{\Omega}$ to guarantee that *if there is a rank-r matrix that agrees with the observed entries, then* $\mathbf{X}$ *is indeed rank-r*. While our condition on $\mathbf{\Omega}$ is combinatorial, we provide a deterministic efficient algorithm to verify whether the condition is satisfied. Furthermore, this condition is satisfied with high probability under uniform random sampling schemes with only $\mathcal{O}(\max\{r, \log d\})$ samples per column. This strengthens existing results in LRMC, allowing to drop the assumption that $\mathbf{X}$ is known a priori to be low-rank.

I. Introduction

In many modern applications, one has a large multivariate dataset that has been severely corrupted by missing values. Fortunately, in many situations, the underlying dataset is of intrinsic low dimension, so the missing values may be inferred from the observed ones. Hence the growing interest on low-rank matrix completion (LRMC), which, as the name suggests, aims to infer the missing entries in a partially observed low-rank data matrix [1]. Applications of this problem arise in a wide variety of practical scenarios, such as face recognition [2], recommender systems and collaborative filtering [3] and network topology estimation [4].

Given a $d \times N$ data matrix $\mathbf{X}$, and a matrix $\mathbf{\Omega}$ indicating the locations of its observed entries, existing theory in LRMC has mainly focused on the following problem:

Problem 1. ($\Rightarrow$) *Determine conditions on* $\mathbf{X}$ *and* $\mathbf{\Omega}$ *to guarantee that* if $\mathbf{X}$ *is a rank-r matrix, then* $\mathbf{X}$ *is* the only *rank-r matrix that agrees with* $\mathbf{X}$ *on* $\mathbf{\Omega}$.

Examples of the conditions on $\mathbf{X}$ include bounded coherence [1, 5–12] and genericity [13–15]. Coherence is a parameter indicating how aligned are the columns in a matrix with respect to the canonical axes; typically the lower coherence the better. Genericity essentially asks that the columns of $\mathbf{X}$ are drawn independently according to an absolutely continuous distribution with respect to the Lebesgue measure on an r-dimensional subspace in general position (see Figure 1 for some intuition).

Examples of the conditions on $\mathbf{\Omega}$ include uniform random sampling [1, 5–11], biased random sampling according to the coherence of $\mathbf{X}$ [12], and deterministic combinatorial conditions [15].

There even exist a wide variety of practical methods that will provably complete subsampled low-rank matrices with high probability. Examples include nuclear norm minimization [1, 5–9, 12], alternating minimization [11], and methods based on singular value decomposition [9, 10, 16–18].

In many situations, though, one does not know a priori whether the given dataset is low-dimensional. To build some intuition, consider the full-data case. Imagine we are given a data matrix $\mathbf{X}$, and we want to determine whether it is low-rank. One thing we can do is compute its singular values. If only a few of them are nonzero, then we can be sure that $\mathbf{X}$ is indeed low-rank.

But if data is missing, we can no longer compute singular values. One thing we can do is *suppose* that $\mathbf{X}$ is rank-1, and try to find a rank-1 matrix that agrees with the observed data. Of course, if there exists no such matrix, then $\mathbf{X}$ cannot be rank-1, and we know that $\operatorname{rank}(\mathbf{X}) \geq 2$. We can iteratively repeat this process until we find a rank-r matrix that agrees with the observed entries. At this point we know $\operatorname{rank}(\mathbf{X}) \geq r$. Nonetheless, depending on $\mathbf{X}$ and $\mathbf{\Omega}$, it is possible to find a rank-r matrix that agrees with $\mathbf{X}$ on $\mathbf{\Omega}$ even if $\mathbf{X}$ is not rank-r. In other words, we could be deceived into thinking that $\mathbf{X}$ is rank-r, when it truly is of higher rank. This raises the following question: can we determine whether $\mathbf{X}$ is truly rank-r, based on a proper subset of its entries?

In general, the answer to this question is no. For instance, suppose we observe all but the top-left entry of $\mathbf{X}$, which will be denoted by x_{11}. Further suppose that for every $j = 1, \dots, N$, all the observed entries of the j^{th} column of $\mathbf{X}$ are equal to some constant $c_j \neq 0$. This *suggests* that $\mathbf{X}$ is rank-1.

Without any assumption on $\mathbf{X}$, x_{11} could take any value. The rank of $\mathbf{X}$ will be 1 if and only if $x_{11} = c_1$, and 2 otherwise. But since x_{11} is unknown, we cannot know which is the case.

On the other hand, suppose in addition that the columns of $\mathbf{X}$ are drawn independently according to an absolutely continuous distribution with respect to the Lebesgue measure on *some* underlying subspace (maybe 1-dimensional, but we do not know). This condition essentially asks that the columns of $\mathbf{X}$ are drawn generically from some subspace (see Figure 1 for some intuition). If the underlying subspace is of dimension > 1, then the probability that any two columns of $\mathbf{X}$ are linearly dependent is zero. Since all the observed entries of the j^{th} column are equal to c_j, it follows that columns $2, \dots, N$ are linearly dependent. Then with probability 1, the underlying subspace is 1-dimensional, and $\mathbf{X}$ is rank-1.

Of course, establishing conditions on $\mathbf{X}$ is not enough. For instance, consider the same scenario as before, but suppose instead that we observe none of the entries in the first row of $\mathbf{X}$. Then there exist infinitely many rank-1 matrices that agree with the observed entries. It is possible that $\mathbf{X}$ is one of these matrices, but it is also possible that $\mathbf{X}$ is really rank 2. In this case, because of the observed locations, we cannot know whether $\mathbf{X}$ is rank-1, even if we assume that its columns are generic or ideally coherent. We thus have the following converse of Problem 1:

Problem 2. $(\Leftarrow)$ *Determine conditions on* $\mathbf{X}$ *and* $\boldsymbol{\Omega}$ *to guarantee that if there is a rank-r matrix that agrees with* $\mathbf{X}$ *on* $\boldsymbol{\Omega}$*,* then $\mathbf{X}$ *is indeed rank-r.*

In this paper we study Problem 2. Our main result shows that if $\mathbf{X}$ is a generic matrix observed on $\boldsymbol{\Omega}$ satisfying a deterministic combinatorial condition, and there is a rank-r matrix that agrees with $\mathbf{X}$ on $\boldsymbol{\Omega}$, then $\mathbf{X}$ is indeed rank-r with probability 1. While our condition on $\boldsymbol{\Omega}$ is combinatorial, we provide a deterministic efficient algorithm to verify whether this condition is satisfied. Furthermore, we show that this condition is satisfied with high probability if $\mathbf{X}$ is observed on as little as $\mathcal{O}(\max\{r, \log d\})$ entries per column, selected uniformly at random. This strengthens existing results in LRMC, allowing to drop the assumption that $\mathbf{X}$ is known a priori to be low-rank.

Organization of the paper

In Section II we formally state the problem and our main results. In Section III we discuss the importance of Problem 2, and why we should care. In Section IV we present our efficient algorithm to verify whether our combinatorial conditions on $\boldsymbol{\Omega}$ are satisfied, and we prove our statements in Section V.

II. Model and main results

Let $\mathbf{X}$ be a $d \times N$ data matrix, and $\boldsymbol{\Omega}$ be the $d \times N$ matrix with binary entries indicating the observed locations of $\mathbf{X}$: the $(i,j)^{th}$ entry of $\boldsymbol{\Omega}$ will be 1 if the $(i,j)^{th}$ entry of $\mathbf{X}$ is observed, and zero otherwise. We say that a matrix *agrees* with $\mathbf{X}$ on $\boldsymbol{\Omega}$ if it is equal to $\mathbf{X}$ on all the nonzero locations of $\boldsymbol{\Omega}$.

As mentioned in Section I, since data is missing, we can no longer compute the singular values of $\mathbf{X}$ to determine its rank. Instead, we can try to find a rank-1 matrix that agrees with $\mathbf{X}$ on $\boldsymbol{\Omega}$. If there exists no such matrix, then $\mathbf{X}$ cannot be rank-1, and we know that $\operatorname{rank}(\mathbf{X}) \geq 2$. We can iteratively repeat this process until we find a rank-r matrix that agrees with $\mathbf{X}$ on $\boldsymbol{\Omega}$. At this point we know $\operatorname{rank}(\mathbf{X}) \geq r$, and we want to determine whether $\operatorname{rank}(\mathbf{X}) = r$. Depending on $\mathbf{X}$ and $\boldsymbol{\Omega}$, it is possible to find a rank-r matrix that agrees with $\mathbf{X}$ on $\boldsymbol{\Omega}$ even if $\operatorname{rank}(\mathbf{X}) > r$.

> *We thus want to establish conditions on* $\mathbf{X}$ *and* $\boldsymbol{\Omega}$ *to guarantee that if* $\operatorname{rank}(\mathbf{X}) \geq r$*, and there is a rank-$r$ matrix that agrees with* $\mathbf{X}$ *on* $\boldsymbol{\Omega}$*, then* $\operatorname{rank}(\mathbf{X}) = r$.

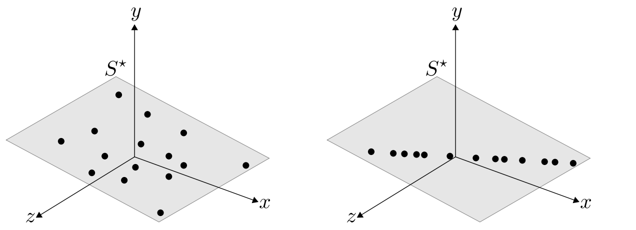

Fig. 1. Each column in a rank-r matrix $\mathbf{X}$ corresponds to a point in an r-dimensional subspace $S^\star$. In these figures, $S^\star$ is a 2-dimensional subspace (plane) in general position. In the **left**, the columns of $\mathbf{X}$ are drawn *generically* from $S^\star$, that is, independently according to an absolutely continuous distribution with respect to the Lebesgue measure on $S^\star$, for example, according to a gaussian distribution on $S^\star$. In this case, the probability of observing a sample as in the **right**, where all columns lie in a line inside $S^\star$, is zero.

We will show that this will be the case if $\boldsymbol{\Omega}$ satisfies condition **C1** below, and $\mathbf{X}$ satisfies the following assumption:

(A1) The columns of $\mathbf{X}$ are drawn independently according to an absolutely continuous distribution with respect to the Lebesgue measure on an $r^\star$-dimensional subspace in general position.

A1 essentially asks that $\mathbf{X}$ is a generic rank-$r^\star$ matrix. To better understand this, let $\operatorname{Gr}(r^\star, \mathbb{R}^d)$ denote the Grassmannian manifold of $r^\star$-dimensional subspaces in $\mathbb{R}^d$. Observe that each $d \times N$ rank-$r^\star$ matrix $\mathbf{X}$ can be uniquely represented in terms of a subspace $S^\star \in \operatorname{Gr}(r^\star, \mathbb{R}^d)$ (spanning the columns of $\mathbf{X}$) and an $r^\star \times N$ coefficient matrix $\boldsymbol{\Theta}^\star$. Let ν_{G} denote the uniform measure on $\operatorname{Gr}(r^\star, \mathbb{R}^d)$, and let ν_Θ denote the Lebesgue measure on $\mathbb{R}^{r^\star \times N}$. Equivalent to **A1**, our statements hold for *almost every* (a.e.) rank-$r^\star$ matrix $\mathbf{X}$, with respect to the product measure $\nu_{\mathrm{G}} \times \nu_\Theta$.

Our condition on $\boldsymbol{\Omega}$ builds on the results in [15], which show that a set of entries in $\mathbf{X}$ observed in the right locations, will determine, up to finite choice, the r-dimensional subspaces that may explain the columns in $\mathbf{X}$. The key insight of our paper is that any additional column observed on $r+1$ entries can be used to verify consistency: if $\operatorname{rank}(\mathbf{X}) > r$, such additional column will agree with none of the candidate r-dimensional subspaces, and equivalently, no rank-r matrix can agree with $\mathbf{X}$ on $\boldsymbol{\Omega}$. This is precisely the contrapositive of the statement we are looking for.

Let us now introduce the constraint matrix $\breve{\boldsymbol{\Omega}}$, as defined in [15], that will allow us to easily express our condition on $\boldsymbol{\Omega}$.

Definition 1 (Constraint Matrix)**.** *Let* $k_{1,j}, k_{2,j}, \ldots, k_{\ell_j,j}$ *denote the indices of the* ℓ_j *observed entries in the* j^{th} *column of* $\mathbf{X}$*. If* $\ell_j \leq r$*, define* $\boldsymbol{\Omega}_j$ *as the empty matrix. Otherwise, define* $\boldsymbol{\Omega}_j$ *as the* $d \times (\ell_j - r)$ *matrix, whose* i^{th} *column has the value* 1 *in rows* $k_{1,j}, k_{2,j}, \ldots, k_{r,j}$ *and* $k_{r+i,j}$*, and zeros elsewhere. Define the* constraint matrix $\breve{\boldsymbol{\Omega}}$ *as* $\breve{\boldsymbol{\Omega}} := [\boldsymbol{\Omega}_1 \cdots \boldsymbol{\Omega}_N]$.

For example, if $k_1 = 1, k_2 = 2, \ldots, k_{\ell_j} = \ell_j$, then

$$\boldsymbol{\Omega}_j = \underbrace{\left[\begin{array}{c} \mathbf{1} \\ \hline \mathbf{I} \\ \hline \mathbf{0} \end{array}\right]}_{\ell_j - r} \begin{array}{l} \} \, r \\ \} \, \ell_j - r \\ \} \, d - \ell_j, \end{array}$$

where $\mathbf{1}$ denotes a block of all 1's and $\mathbf{I}$ the identity matrix.

The key insight behind this construction is that observing more than r entries in a column of $\mathbf{X}$ places constraints on the r-dimensional subspaces that can explain it. For example, if we observe $r+1$ entries of a particular column, then not all r-dimensional subspaces will be consistent with the entries. If we observe more entries, then even fewer subspaces will be consistent with them. In effect, each observed entry, in addition to the first r observations, places one constraint that an r-dimensional subspace must satisfy in order to be consistent with the observations. The matrix $\breve{\boldsymbol{\Omega}}$ encodes all these constraints. **C1** below is a simple, combinatorial condition on $\breve{\boldsymbol{\Omega}}$ that guarantees that if $\operatorname{rank}(\mathbf{X}) > r$, then the observed entries will produce inconsistent constraints, implying that no r-dimensional subspace can explain $\mathbf{X}$, or equivalently, that no rank-r matrix can agree with $\mathbf{X}$ on $\boldsymbol{\Omega}$.

Given a matrix, let $n(\cdot)$ denote its number of columns and $m(\cdot)$ the number of its *nonzero* rows. With this, we are ready to present our condition on $\boldsymbol{\Omega}$:

(**C1**) The constraint matrix $\breve{\boldsymbol{\Omega}}$ contains a column $\boldsymbol{\omega}$, in addition to r disjoint matrices $\{\breve{\boldsymbol{\Omega}}_\tau\}_{\tau=1}^r$, each of size $d \times (d-r)$, such that for every τ:

(**C2**) Every matrix $\boldsymbol{\Omega}'$ formed with a subset of the columns in $\breve{\boldsymbol{\Omega}}_\tau$ satisfies

$$m(\boldsymbol{\Omega}') \geq n(\boldsymbol{\Omega}') + r. \tag{1}$$

In words, **C2** asks that every subset of n columns of $\breve{\boldsymbol{\Omega}}_\tau$ has at least $n+r$ nonzero rows.

Example 1. *The following sampling satisfies* **C2***.*

$$\breve{\boldsymbol{\Omega}}_\tau = \left[\begin{array}{c} \mathbf{1} \\ \hline \mathbf{I} \end{array}\right] \begin{array}{l} \} \, r \\ \} \, d - r. \end{array}$$

While condition **C2** is combinatorial, we show in Section IV that one can easily verify whether this condition is satisfied by checking the dimension of the null-space of a sparse matrix. This is summarized in Algorithm 1, which in turn provides a practical criteria to verify whether $\mathbf{X}$ is indeed rank-r.

The paper's main result is the following theorem, which gives an answer to Problem 2. It states that for almost every matrix $\mathbf{X}$ with $\operatorname{rank}(\mathbf{X}) \geq r$ (thus establishing conditions on $\mathbf{X}$, namely genericity), if $\boldsymbol{\Omega}$ satisfies condition **C1** and there is a rank-r matrix that agrees with $\mathbf{X}$ on $\boldsymbol{\Omega}$, then $\mathbf{X}$ is indeed rank-r. The proof is given in Section V.

Theorem 1. *Let* **A1** *hold. Suppose* $\operatorname{rank}(\mathbf{X}) \geq r$ *and* $\boldsymbol{\Omega}$ *satisfies* **C1***. If there exists a rank-r matrix that agrees with* $\mathbf{X}$ *on* $\boldsymbol{\Omega}$*, then* $\operatorname{rank}(\mathbf{X}) = r$ *with probability* 1.

In a nutshell, Theorem 1 states that if $\mathbf{X}$ is generic, and there is a rank-r matrix that agrees with $\mathbf{X}$ in the right places, then $\mathbf{X}$ must be rank-r with probability 1.

Furthermore, the next theorem states that sampling patterns satisfying **C1** appear with high probability under uniform random sampling schemes with only $\mathcal{O}(\max\{r, \log d\})$ samples per column. The proof is given in Section V.

Theorem 2. *Let* $0 < \epsilon \leq 1$ *be given. Suppose* $r \leq \frac{d}{6}$, $N > r(d-r)$*, and that each column of* $\boldsymbol{\Omega}$ *has at least* ℓ *nonzero entries, distributed uniformly at random and independently across columns, with*

$$\ell \geq \max\left\{12\left(\log(\tfrac{d}{\epsilon}) + 1\right), \ 2r\right\}. \tag{2}$$

Then with probability at least $1-\epsilon$*,* $\boldsymbol{\Omega}$ *will satisfy* **C1***.*

In a nutshell, Theorem 2 implies that if $\mathbf{X}$ is generic, and there is a rank-r matrix that agrees with $\mathbf{X}$ on enough entries selected uniformly at random, then $\mathbf{X}$ must be rank-r.

III. Why should we care?

In many modern applications one is given an incomplete data matrix, and aims to infer its missing entries. If the intrinsic dimension of the complete (yet unknown) dataset is too large for the number of observed entries, nothing can be done to infer the missing entries. Fortunately, in many situations the rank of the complete data matrix is very low, whence the whole matrix can be inferred from a small fraction of its entries.

The importance of Problem 2 is that in many situations we do not know a priori the rank of the complete data matrix $\mathbf{X}$. In such case, all we can do is *suppose* that the matrix is rank-r, and try to find a rank-r matrix that agrees with the observed data. If we find such a matrix, our *hope* is that it *is* $\mathbf{X}$. But it is possible that it is not, even if it is *the only* rank-r matrix that agrees with the observed data. Theorem 1 states that if $\mathbf{X}$ is generic, and $\boldsymbol{\Omega}$ satisfies **C1**, then this will not be the case. Furthermore, Theorem 2 implies that if $\mathbf{X}$ is generic, and there is a rank-r matrix that agrees with $\mathbf{X}$ on $\mathcal{O}(\max\{r, \log d\})$ entries per column, selected uniformly at random, then $\mathbf{X}$ must be rank-r.

These are incredibly good news! This implies that if our data matrix is known to be generic (as is often the case), and $\boldsymbol{\Omega}$ satisfies **C1** (which will happen with high probability according to Theorem 2), then we do not have to worry about being deceived into thinking that our data is rank-r, when it truly is not.

Theorem 2 is particularly relevant because a large portion of the theory and methods for LRMC operate under uniform random sampling schemes with $\mathcal{O}(r \log d)$ observations per column. We often used these methods without knowing whether $\mathbf{X}$ is truly low-rank, *hoping* (but not knowing) that it truly is. Theorem 2 strengthens these results for data matrices that are both generic *and* of bounded coherence (as is often the case). In such case, *now* we know that if we run any of these methods, and find a rank-r matrix that agrees with the observed data, then the underlying matrix is truly rank-r.

IV. Solving Problem 2 in practice

Theorem 2 states that samplings with $\mathcal{O}(\max\{r, \log d\})$ observations per column drawn uniformly at random will satisfy **C1** with high probability.

In many situations, though, sampling is not uniform. For instance, in vision, occlusion of objects can produce missing data in very non-uniform random patterns. In cases like this, one would still like to verify whether a given matrix is rank-r. We can do this using Theorem 1 directly. For example, we can split $\breve{\mathbf{\Omega}}$ (e.g., randomly) into disjoint matrices $\{\breve{\mathbf{\Omega}}_\tau\}_{\tau=1}^r$, and verify whether each $\breve{\mathbf{\Omega}}_\tau$ satisfies **C2**. Of course, **C2** is a combinatorial condition, hence verifying it directly may be computationally prohibitive, especially for large d. Fortunately, we can easily verify whether a matrix $\breve{\mathbf{\Omega}}_\tau$ satisfies **C2** by checking the dimension of the null-space of a sparse matrix. This is summarized in Algorithm 1, which in turn provides a practical criteria to verify whether $\mathbf{X}$ is indeed rank-r.

To present this algorithm, let us introduce the matrix $\mathbf{A}$ that will allow us to determine efficiently whether a sampling $\breve{\mathbf{\Omega}}_\tau$ satisfies **C2**. To this end, let $\boldsymbol{\omega}_j$ denote the j^{th} column of $\breve{\mathbf{\Omega}}_\tau$, and let $\mathbf{U}$ be a $d \times r$ matrix drawn according to $\nu_{\mathbf{U}}$, an absolutely continuous distribution with respect to the Lebesgue measure on $\mathbb{R}^{d \times r}$. Let $\mathbf{U}_{\boldsymbol{\omega}_j}$ denote the restriction of $\mathbf{U}$ to the nonzero rows in $\boldsymbol{\omega}_j$. Let $\mathbf{a}_{\boldsymbol{\omega}_j} \in \mathbb{R}^{r+1}$ be a nonzero vector in $\ker \mathbf{U}_{\boldsymbol{\omega}_j}^\mathsf{T}$, and $\mathbf{a}_j$ be the vector in $\mathbb{R}^d$ with the entries of $\mathbf{a}_{\boldsymbol{\omega}_j}$ in the nonzero locations of $\boldsymbol{\omega}_j$ and zeros elsewhere. Finally, let $\mathbf{A}$ denote the $d \times (d-r)$ matrix with $\{\mathbf{a}_j\}_{j=1}^{d-r}$ as columns.

Algorithm 1 will verify whether $\dim \ker \mathbf{A}^\mathsf{T} = r$, and this will determine whether $\breve{\mathbf{\Omega}}_\tau$ satisfies **C2**. The key insight behind Algorithm 1 is that $\mathbf{A}$ encodes the information of the projections of $S = \operatorname{span}\{\mathbf{U}\}$ onto the canonical coordinates indicated by $\breve{\mathbf{\Omega}}_\tau$. We know from Theorem 1 in [19] that $\nu_{\mathbf{U}}$-a.s., these projections will uniquely determine S if and only if $\dim \ker \mathbf{A}^\mathsf{T} = r$, which will be the case if and only if $\breve{\mathbf{\Omega}}_\tau$ has $d-r$ columns and $\breve{\mathbf{\Omega}}_\tau$ satisfies **C2**.

We have thus shown the following lemma, which states that with probability 1, Algorithm 1 will determine whether $\breve{\mathbf{\Omega}}_\tau$ satisfies **C2**.

Lemma 1. *Let $\breve{\mathbf{\Omega}}_\tau$ be a matrix formed with $d-r$ columns of $\breve{\mathbf{\Omega}}$. Then $\nu_{\mathbf{U}}$-a.s., $\breve{\mathbf{\Omega}}_\tau$ satisfies* **C2** *if and only if* $\dim \ker \mathbf{A}^\mathsf{T} = r$.

V. Proofs

The proof of Theorem 1 is largely based on Theorem 1 and Lemma 8 in [15], which together give a combinatorial condition of sampling patterns that can only be completed in finitely many ways. We combine these results in the following lemma. Recall that $\breve{\mathbf{\Omega}}$ denotes the matrix encoding all the constraints imposed by $\mathbf{\Omega}$, as defined in Section II.

Algorithm 1: Determine whether $\breve{\mathbf{\Omega}}_\tau$ satisfies **C2**.

```
Input: Matrix $\breve{\mathbf{\Omega}}_\tau$ with $d-r$ columns of $\breve{\mathbf{\Omega}}$.
- Draw $\mathbf{U} \in \mathbb{R}^{d \times r}$ drawn according to $\nu_{\mathbf{U}}$.
- for $j = 1$ to $d-r$ do
    - $\mathbf{a}_{\boldsymbol{\omega}_j}$ = nonzero vector in $\ker \mathbf{U}_{\boldsymbol{\omega}_j}^\mathsf{T}$.
    - $\mathbf{a}_j$ = vector in $\mathbb{R}^d$ with entries of $\mathbf{a}_{\boldsymbol{\omega}_j}$ in the
        nonzero locations of $\boldsymbol{\omega}_j$ and zeros elsewhere.
- $\mathbf{A}$ = matrix formed with $\{\mathbf{a}_j\}_{j=1}^{d-r}$ as columns.
- if $\dim \ker \mathbf{A}^\mathsf{T} = r$ then
    - Output: $\breve{\mathbf{\Omega}}_\tau$ satisfies C2.
- else
    - Output: $\breve{\mathbf{\Omega}}_\tau$ does not satisfy C2.
```

Lemma 2. *Let* **A1** *hold, and suppose* $\operatorname{rank}(\mathbf{X}) = r$. *If* $\mathbf{\Omega}$ *satisfies* **C3** *below, then with probability* 1 *there exist at most finitely many rank-r matrices that agree with* $\mathbf{X}$ *on* $\mathbf{\Omega}$.

(C3) The constraint matrix $\breve{\mathbf{\Omega}}$ contains disjoint matrices $\{\breve{\mathbf{\Omega}}_\tau\}_{\tau=1}^r$, each of size $d \times (d-r)$, such that each $\breve{\mathbf{\Omega}}_\tau$ satisfies **C2**.

Lemma 2 implies that if $\operatorname{rank}(\mathbf{X}) = r$ and $\mathbf{\Omega}$ satisfies **C3**, then there exist at most finitely many r-dimensional subspaces that may explain the columns of $\mathbf{X}$. We now use Lemma 2 to show that this will also be the case if $\operatorname{rank}(\mathbf{X}) > r$.

Corollary 1. *Let* **A1** *hold, and suppose* $\operatorname{rank}(\mathbf{X}) \geq r$. *If* $\breve{\mathbf{\Omega}}$ *satisfies* **C3**, *then with probability* 1 *there exist at most finitely many rank-r matrices that agree with* $\mathbf{X}$ *on* $\mathbf{\Omega}$.

Proof. If $\operatorname{rank}(\mathbf{X}) = r$, the corollary follows directly from Lemma 2. Now suppose $\operatorname{rank}(\mathbf{X}) > r$. If there is no rank-$r$ matrix that agrees with $\mathbf{X}$ on $\mathbf{\Omega}$, then the corollary is trivially true. Now suppose that there is at least one rank-r matrix $\mathbf{Y}$ that agrees with $\mathbf{X}$ on $\mathbf{\Omega}$. By Lemma 2, with probability 1 there exist at most finitely many rank-r matrices that agree with $\mathbf{Y}$ on $\mathbf{\Omega}$. It follows that there exist at most finitely many rank-r matrices that agree with $\mathbf{X}$ on $\mathbf{\Omega}$. □

Corollary 1 shows that if $\operatorname{rank}(\mathbf{X}) \geq r$ and $\mathbf{\Omega}$ satisfies **C3**, then there exist at most finitely many r-dimensional subspaces that may explain the columns of $\mathbf{X}$. Now we will show that any additional column observed on $r+1$ entries can be used to verify whether $\operatorname{rank}(\mathbf{X}) = r$ or $\operatorname{rank}(\mathbf{X}) > r$. The main intuition is that if $\operatorname{rank}(\mathbf{X}) > r$, the additional column will agree with none of the candidate r-dimensional subspaces. Equivalently, no rank-r matrix can agree with $\mathbf{X}$ on $\mathbf{\Omega}$. This will be the contrapositive of the statement in Theorem 1.

Proof. (Theorem 1) Suppose $\mathrm{rank}(\mathbf{X}) \geq r$ and that $\boldsymbol{\Omega}$ satisfies **C1**. Then $\boldsymbol{\Omega}$ also satisfies **C3**. By Corollary 1, there are at most finitely many r-dimensional subspaces that may explain the columns in $\mathbf{X}$. Let S be one of these subspaces. In addition, let $\boldsymbol{\omega}$ denote a column in $\breve{\boldsymbol{\Omega}}$ that is not in $\{\breve{\boldsymbol{\Omega}}_\tau\}_{\tau=1}^r$. Recall that each column in $\breve{\boldsymbol{\Omega}}$ corresponds to a column in $\boldsymbol{\Omega}$, which in turn corresponds to a column in $\mathbf{X}$. Let $\mathbf{x}$ denote the column in $\mathbf{X}$ corresponding to $\boldsymbol{\omega}$.

Next suppose for contrapositive that $\mathrm{rank}(\mathbf{X}) = r^\star > r$. This means that the columns of $\mathbf{X}$ lie in an $r^\star$-dimensional subspace $S^\star$. Observe that for ν_{G}-a.e. $r^\star$-dimensional subspace $S^\star$, the restriction of $S^\star$ to $\ell \leq r^\star$ coordinates is $\mathbb{R}^\ell$. Let $S^\star_{\boldsymbol{\omega}}, S_{\boldsymbol{\omega}}$ and $\mathbf{x}_{\boldsymbol{\omega}}$ denote the restrictions of $S^\star, S$ and $\mathbf{x}$ to the nonzero rows in $\boldsymbol{\omega}$. Since $\boldsymbol{\omega}$ has exactly $r+1$ nonzero entries by construction, it follows that $S^\star_{\boldsymbol{\omega}} = \mathbb{R}^{r+1}$. In contrast, $S_{\boldsymbol{\omega}}$ is an r-dimensional subspace of $\mathbb{R}^{r+1}$. Recall that $\mathbf{x}$ is drawn according to an absolutely continuous with respect to the Lebesgue measure on $S^\star$. Equivalently, $\mathbf{x}_{\boldsymbol{\omega}}$ is drawn according to an absolutely continuous distribution with respect to the Lebesgue measure on $S^\star_{\boldsymbol{\omega}} = \mathbb{R}^{r+1}$. Intuitively, this means that that $\mathbf{x}_{\boldsymbol{\omega}}$ could take any value in $\mathbb{R}^{r+1}$. Since $S_{\boldsymbol{\omega}}$ is an r-dimensional subspace of $\mathbb{R}^{r+1}$, it has measure zero. It follows that almost surely, $\mathbf{x}_{\boldsymbol{\omega}} \notin S_{\boldsymbol{\omega}}$. This is true for all of the finitely many r-dimensional subspaces that could explain the columns in $\mathbf{X}$. It follows that no r-dimensional subspace can explain the observed entries of $\mathbf{X}$. Equivalently, there exists no rank-r matrix that agrees with $\mathbf{X}$ on $\boldsymbol{\Omega}$. This is the contrapositive of the statement in Theorem 1. □

The proof of Theorem 2 is based on Lemma 9 in [15], which shows that sampling patterns satisfying **C2** appear with high probability under uniform random sampling schemes with only $\mathcal{O}(\max\{r, \log d\})$ samples per column. We restate this result as the following lemma.

Lemma 3. *Let the assumptions of Theorem 2 hold, and let $\boldsymbol{\Omega}_\tau$ be a matrix formed with $d-r$ columns of $\boldsymbol{\Omega}$. With probability at least $1 - \frac{\epsilon}{d}$, $\boldsymbol{\Omega}_\tau$ will satisfy* **C2**.

Theorem 2 follows directly from Lemma 3 by applying a union bound.

Proof. (Theorem 2) If $N > r(d-r)$, randomly select disjoint matrices $\{\boldsymbol{\Omega}_\tau\}_{\tau=1}^r$, each formed with $d-r$ columns of $\boldsymbol{\Omega}$. Let $\mathcal{E}_\tau$ denote the even that $\boldsymbol{\Omega}_\tau$ fails to satisfy **C2**.

Union bounding over τ, we may upper bound the probability that $\boldsymbol{\Omega}$ fails to satisfy **C1** by

$$\sum_{\tau=1}^{r} \mathsf{P}(\mathcal{E}_\tau) < \sum_{\tau=1}^{r} \frac{\epsilon}{d} < \sum_{\tau=1}^{r} \frac{\epsilon}{r} = \epsilon,$$

where the first inequality follows by Lemma 3.

□

VI. Conclusions

In this paper we show that if a generic data matrix $\mathbf{X}$ is observed on the locations indicated by $\boldsymbol{\Omega}$ satisfying a deterministic combinatorial condition, and there is a rank-r matrix that agrees with the observed data, then $\mathbf{X}$ is indeed rank-r with probability 1. Our condition on $\boldsymbol{\Omega}$ is combinatorial, yet we provide a deterministic efficient criteria to verify whether this condition is satisfied. Furthermore, we show that this condition is satisfied with high probability if $\mathbf{X}$ is observed on as little as $\mathcal{O}(\max\{r, \log d\})$ entries per column, selected uniformly at random. This strengthens existing results in LRMC, allowing to drop the assumption that $\mathbf{X}$ is known a priori to be low-rank.

References

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

[2] R. Basri and D. Jacobs, *Lambertian reflectance and linear subspaces*, IEEE Transactions on Pattern Analysis and Machine Intelligence, 2003.

[3] J. Rennie and N. Srebro, *Fast maximum margin matrix factorization for collaborative prediction*, International Conference on Machine Learning, 2005.

[4] B. Eriksson, P. Barford, J. Sommers and R. Nowak, *DomainImpute: Inferring Unseen Components in the Internet*, IEEE INFOCOM Mini-Conference, 2011.

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

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

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

[8] S. Bhojanapalli and P. Jain, *Universal matrix completion*, International Conference on Machine Learning, 2014.

[9] Y. Chen, *Incoherence-optimal matrix completion*, IEEE Transactions on Information Theory, 2013.

[10] R. Keshavan, A. Montanari and S. Oh, *Matrix completion from a few entries*, IEEE Transactions on Information Theory, 2010.

[11] P. Jain, P. Netrapalli and S. Sanghavi, *Low-rank matrix completion using alternating minimization*, ACM Symposium on Theory Of Computing, 2013.

[12] Y. Chen, S. Bhojanapalli, S. Sanghavi and R. Ward, *Coherent matrix completion*, International Conference on Machine Learning, 2014.

[13] A. Singer and M. Cucuringu, *Uniqueness of low-rank matrix completion by rigidity theory*, SIAM Journal on Matrix Analysis and Applications, 2010.

[14] F. Király and R. Tomioka, *A combinatorial algebraic approach for the identifiability of low-rank matrix completion*, International Conference on Machine Learning, 2012.

[15] D. Pimentel-Alarcón, N. Boston and R. Nowak, *A characterization of deterministic sampling patterns for low-rank matrix completion*, Allerton, 2015.

[16] J. Cai, E. Candès and and Z. Shen, *A singular value thresholding algorithm for matrix completion*, SIAM Journal on Optimization, 2010.

[17] S. Ma, D. Goldfarb, L. Chen, *Fixed point and Bregman iterative methods for matrix rank minimization*, Mathematical Programming, 2011.

[18] E. Chunikhina, R. Raich and T. Nguyen, *Performance analysis for matrix completion via iterative hard-thresholded SVD*, IEEE Statistical Signal Processing Workshop, 2014.

[19] D. Pimentel-Alarcón, N. Boston and R. Nowak, *Deterministic conditions for subspace identifiability from incomplete sampling*, IEEE International Symposium on Information Theory, 2015.