High-Rank Matrix Completion

Brian Eriksson
Boston University

Laura Balzano
University of Wisconsin-Madison

Robert Nowak
University of Wisconsin-Madison

Abstract

This paper considers the problem of completing a matrix with many missing entries under the assumption that the columns of the matrix belong to a union of multiple low-rank subspaces. This generalizes the standard *low-rank matrix completion* problem to situations in which the matrix rank can be quite high or even full rank. Since the columns belong to a union of subspaces, this problem may also be viewed as a missing-data version of the *subspace clustering* problem. Let $\mathbf{X}$ be an $n \times N$ matrix whose (complete) columns lie in a union of at most k subspaces, each of rank $\leq r < n$, and assume $N \gg kn$. The main result of the paper shows that under mild assumptions each column of $\mathbf{X}$ can be perfectly recovered with high probability from an incomplete version so long as at least $CrN \log^2(n)$ entries of $\mathbf{X}$ are observed uniformly at random, with $C > 1$ a constant depending on the usual incoherence conditions, the geometrical arrangement of subspaces, and the distribution of columns over the subspaces. The result is illustrated with numerical experiments and an application to Internet distance matrix completion and topology identification.

1 Introduction

Consider a real-valued $n \times N$ dimensional matrix $\mathbf{X}$. Assume that the columns of $\mathbf{X}$ lie in the union of at most k subspaces of $\mathbb{R}^n$, each having dimension at most $r < n$ and assume that $N > kn$. We are especially interested in "high-rank" situations in which the total rank (the rank of the union of the subspaces) may be n. Our goal is to complete $\mathbf{X}$ based on an observation of a small random subset of its entries. We propose a novel method for this matrix completion problem. In the applications we have in mind N may be arbitrarily large, and so we will focus on quantifying the probability that a given column is perfectly completed, rather than the probability that whole matrix is perfectly completed (*i.e.,* every column is perfectly completed). Of course it is possible to translate between these two quantifications using a union bound, but that bound becomes meaningless if N is extremely large.

Suppose the entries of $\mathbf{X}$ are observed uniformly at random with probability p_0. Let Ω denote the set of indices of observed entries and let $\mathbf{X}_\Omega$ denote the observations of $\mathbf{X}$. Our main result shows that under a mild set of assumptions each column of $\mathbf{X}$ can be perfectly recovered from $\mathbf{X}_\Omega$ with high probability using a computationally efficient procedure if

$$p_0 \geq C \frac{r}{n} \log^2(n) \tag{1}$$

where $C > 1$ is a constant depending on the usual incoherence conditions as well as the geometrical arrangement of subspaces and the distribution of the columns in the subspaces.

1.1 Connections to Low-Rank Completion

Low-rank matrix completion theory [1] shows that an $n \times N$ matrix of rank r can be recovered from incomplete observations, as long as the number of entries observed (with locations sampled uniformly at random) exceeds $rN \log^2 N$ (within a constant factor and assuming $n \leq N$). It is also known that, in the same setting, completion is impossible if the number of observed entries is less than a constant times $rN \log N$ [2]. These results imply that if the rank of $\mathbf{X}$ is close to n, then all of the entries are needed in order to determine the matrix.

Here we consider a matrix whose columns lie in the union of at most k subspaces of $\mathbb{R}^n$. Restricting the rank of each subspace to at most r, then the rank of the full matrix in our situation could be as large as kr, yielding the requirement $krN \log^2 N$ using current matrix completion theory. In contrast, the bound in (1) implies that the completion of

Appearing in Proceedings of the 15^{th} International Conference on Artificial Intelligence and Statistics (AISTATS) 2012, La Palma, Canary Islands. Volume XX of JMLR: W&CP XX.

*Brian Eriksson and Laura Balzano contributed equally to this work.

each column is possible from a constant times $rN \log^2 n$ entries sampled uniformly at random. Exact completion of every column can be guaranteed by replacing $\log^2 n$ with $\log^2 N$ is this bound, but since we allow N to be very large we prefer to state our result in terms of per-column completion. Our method, therefore, improves significantly upon conventional low-rank matrix completion, especially when k is large. This does not contradict the lower bound in [2], because the matrices we consider are not arbitrary high-rank matrices, rather the columns must belong to a union of rank $\leq r$ subspaces.

1.2 Connections to Subspace Clustering

Let $x_1, \ldots, x_N \in \mathbb{R}^n$ and assume each x_i lies in one of at most k subspaces of $\mathbb{R}^n$. Subspace clustering is the problem of learning the subspaces from $\{x_i\}_{i=1}^N$ and assigning each vector to its proper subspace; cf. [3] for a overview. This is a challenging problem, both in terms of computation and inference, but provably probably correct subspace clustering algorithms now exist [4, 5, 6]. Here we consider the problem of *high rank matrix completion*, which is essentially equivalent to subspace clustering with missing data. This problem has been looked at in previous works [7, 8], but to the best of our knowledge our method and theoretical bounds are novel. Note that our sampling probability bound (1) requires that only slightly more than r out of n entries are observed in each column, so the matrix may be highly incomplete.

1.3 A Motivating Application

There are many applications of subspace clustering, and it is reasonable to suppose that data may often be missing in high-dimensional problems. One such application is the Internet distance matrix completion and topology identification problem. Infrastructures exist that record the number of routers (*i.e.,* distances) from N end host computers to a set of n monitoring points throughout the Internet. The complete set of distances determines the network topology between the computers and the monitoring points [9]. Some infrastructures are based entirely on passively monitoring of normal traffic, which allows for the ability to monitor a very large portion of the Internet, but with the disadvantage that a subset of the distances may not be observed. This poses a matrix completion problem, with the incomplete distance matrix being potentially full-rank in this application. However, computers tend to be clustered within subnets having a small number of access points to the Internet at large, limiting the rank of the submatrix of distances from subnet computers. Therefore the columns of the $n \times N$ distance matrix lie in the union of k low-rank subspaces, where k is the number of subnets [10].

1.4 Related Work

The proof of the main result draws on ideas from matrix completion theory, subspace learning and detection with missing data, and subspace clustering. One key ingredient in our approach is the celebrated results on low-rank Matrix Completion [1, 2, 11]. Unfortunately, in many real-world problems where data are missing, particularly when the data are generated from a union of subspaces, these matrices can have high rank (*e.g.,* networking data in [10]). Thus, these prior results will require effectively all the elements be observed to accurately reconstruct the matrix.

Our work builds upon the results of [12], which quantifies the deviation of an incomplete vector norm with respect to the incoherence of the sampling pattern. While this work also examines subspace detection using incomplete data, it assumes complete knowledge of the subspaces.

While research that examines subspace learning has been presented in [13], the work in this paper differs by the concentration on learning from incomplete observations (*i.e.,* when there are missing elements in the matrix), and by the methodological focus (*i.e.,* nearest neighbor clustering versus a multiscale Singular Value Decomposition approach).

1.5 Sketch of Methodology

The algorithm proposed in this paper involves several relatively intuitive steps, outlined below. We go into detail for each of these steps in following sections.

Local Neighborhoods. A subset of columns of $\mathbf{X}_\Omega$ are selected uniformly at random. These are called *seeds*. A set of nearest neighbors is identified for each seed from the remainder of $\mathbf{X}_\Omega$. In Section 3, we show that nearest neighbors can be reliably identified, even though a large portion of the data are missing, under the usual incoherence assumptions.

Local Subspaces. The subspace spanned by each seed and its neighborhood is identified using matrix completion. If matrix completion fails (*i.e.,* if the resulting matrix does not agree with the observed entries and/or the rank of the result is greater than r), then the seed and its neighborhood are discarded. In Section 4 we show that when the number of seeds and the neighborhood sizes are large enough, then with high probability all k subspaces are identified. We may also identify additional subspaces which are unions of the true subspaces, which leads us to the next step. An example of these neighborhoods is shown in Figure 1.

Subspace Refinement. The set of subspaces obtained from the matrix completions is pruned to remove all but k subspaces. The pruning is accomplished by simply discarding all subspaces that are spanned by the union of two or more other subspaces. This can be done efficiently, as is shown in Section 5.

Full Matrix Completion. Each column in $\mathbf{X}_\Omega$ is assigned to its proper subspace and completed by projection onto that subspace, as described in Section 6. Even when many observations are missing, it is possible to find the correct subspace and the projection using results from subspace detection with missing data [12]. The result of this step is a completed matrix $\widehat{\mathbf{X}}$ such that each column is correctly completed with high probability.

The mathematical analysis will be presented organized according to these steps. After the supporting lemmas, whose complete proofs can be found in [14], experimental results are presented in the final section.

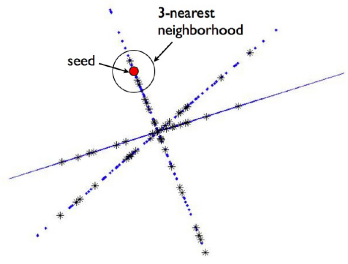

Figure 1: Example of nearest-neighborhood selecting points on from a single subspace.

2 Key Assumptions and Main Result

The notion of incoherence plays a key role in matrix completion and subspace recovery from incomplete observations.

Definition 1. *The* coherence *of an r-dimensional subspace $\mathcal{S} \subseteq \mathbb{R}^n$ is*

$$\mu(\mathcal{S}) := \frac{n}{r} \max_j \|P_{\mathcal{S}} e_j\|_2^2$$

where $P_{\mathcal{S}}$ is the projection operator onto $\mathcal{S}$ and $\{e_j\}$ are the canonical unit vectors for $\mathbb{R}^n$.

Note that $1 \leq \mu(\mathcal{S}) \leq n/r$. The coherence of single vector $x \in \mathbb{R}^n$ is $\mu(x) = \frac{n\|x\|_\infty^2}{\|x\|_2^2}$, which is precisely the coherence of the one-dimensional subspace spanned by x. With this definition, we can state the main assumptions we make about the matrix $\mathbf{X}$.

A1. The columns of $\mathbf{X}$ lie in the union of at most k subspaces, with $k = o(n^d)$ for some $d > 0$. The subspaces are denoted by $\mathcal{S}_1, \ldots, \mathcal{S}_k$ and each has rank at most $r < n$. The ℓ_2-norm of each column is ≤ 1.

A2. The coherence of each subspace is bounded above by μ_0. The coherence of each column is bounded above by μ_1 and for any pair of columns, x_1 and x_2, the coherence of $x_1 - x_2$ is also bounded above by μ_1.

A3. The columns of $\mathbf{X}$ do not lie in the intersection(s) of the subspaces with probability 1, and if $\text{rank}(\mathcal{S}_i) = r_i$, then any subset of r_i columns from $\mathcal{S}_i$ spans $\mathcal{S}_i$ with probability 1. Let $0 < \epsilon_0 < 1$ and $\mathcal{S}_{i,\epsilon_0}$ denote the subset of points in $\mathcal{S}_i$ at least ϵ_0 distance away from any other subspace. There exists a constant $0 < \nu_0 \leq 1$, depending on ϵ_0, such that

- **(i)** The probability that a column selected uniformly at random belongs to $\mathcal{S}_{i,\epsilon_0}$ is at least ν_0/k.
- **(ii)** If $x \in \mathcal{S}_{i,\epsilon_0}$, then the probability that a column selected uniformly at random belongs to the ball of radius ϵ_0 centered at x is at least $\nu_0 \epsilon_0^r / k$.

The conditions of **A3** are met if, for example, the columns are drawn from a mixture of continuous distributions on each of the subspaces. The value of ν_0 depends on the geometrical arrangement of the subspaces and the distribution of the columns within the subspaces. If the subspaces are not too close to each other, and the distributions within the subspaces are fairly uniform, then typically ν_0 will be not too close to 0. We define three key quantities, the confidence parameter δ_0, the required number of "seed" columns s_0, and a quantity ℓ_0 related to the neighborhood formation process (see Algorithm 1 in Section 3):

$$\begin{aligned} \delta_0 &:= n^{2-2\beta^{1/2}} \log n \,, \text{ for some } \beta > 1 \,, \qquad (2) \\ s_0 &:= \left\lceil \frac{k(\log k + \log 1/\delta_0)}{(1-e^{-4})\nu_0} \right\rceil \,, \\ \ell_0 &:= \left\lceil \max\left\{ \frac{2k}{\nu_0(\frac{\epsilon_0}{\sqrt{3}})^r} \,, \frac{8k\log(s_0/\delta_0)}{n\nu_0(\frac{\epsilon_0}{\sqrt{3}})^r} \right\} \right\rceil \,. \end{aligned}$$

We can now state the main result of the paper.

Theorem 2.1. *Let $\mathbf{X}$ be an $n \times N$ matrix satisfying* **A1**-**A3**. *Suppose that each entry of $\mathbf{X}$ is observed independently with probability p_0. If*

$$p_0 \;\geq\; \frac{128\,\beta \max\{\mu_1^2, \mu_0\}}{\nu_0} \, \frac{r \, \log^2(n)}{n}$$

and

$$N \;\geq\; \ell_0 n (2\delta_0^{-1} s_0 \ell_0 n)^{\mu_0^2 \log p_0^{-1}}$$

then each column of $\mathbf{X}$ can be perfectly recovered with probability at least $1 - (6 + 15 s_0)\,\delta_0$, using the methodology sketched above (and detailed later in the paper).

The requirements on sampling are essentially the same as those for standard low-rank matrix completion, apart from requirement that the total number of columns N is sufficiently large. This is needed to ensure that each of the subspaces is sufficiently represented in the matrix. The requirement on N is polynomial in n for fixed p_0, which is easy to see based on the definitions of δ_0, s_0, and ℓ_0 (see further discussion at the end of Section 3).

Perfect recovery of each column is guaranteed with probability that decreases linearly in s_0, which itself is linear in k (ignoring log factors). This is expected since this problem is more difficult than k individual low-rank matrix completions. We state our results in terms of a per-column (rather than full matrix) recovery guarantee. A full matrix recovery guarantee can be given by replacing $\log^2 n$ with $\log^2 N$. This is evident from the final completion step discussed in Lemma 8, below. However, since N may be quite large (perhaps arbitrarily) in the applications we envision, we state our results in terms of a per-column guarantee.

The details of the methodology and lemmas leading to the theorem above are developed in the subsequent sections. Proof sketches are shown and complete proofs can be found in [14]. In certain cases it will be convenient to consider sampling the locations of observed entries uniformly at random *with replacement* rather than without replacement, as assumed above. The following lemma will be useful for translating bounds derived assuming sampling with replacement to our situation (the same sort of relation is noted in Proposition 3.1 in [1]). Sampling with replacement could give duplicates in a set; sampling without replacement gives more information and therefore better probability performance. For the proof see [1, 14].

Lemma 1. *Draw m samples independently and uniformly from $\{1,\ldots,n\}$ and let Ω' denote the resulting subset of unique values. Let Ω_m be a subset of size m selected uniformly at random from $\{1,\ldots,n\}$. Let E denote an event (e.g., a failure or error event) depending on a random subset of $\{1,\ldots,n\}$. If $\mathbb{P}(E(\Omega_m))$ is a non-increasing function of m, then $\mathbb{P}(E(\Omega')) \geq \mathbb{P}(E(\Omega_m))$.*

3 Local Neighborhoods

In this first step, s columns of $\mathbf{X}_\Omega$, called "seeds," are selected uniformly at random and a set of "nearby" columns are identified for each, constituting a local neighborhood of size n. All bounds that hold are designed with probability at least $1-\delta_0$, where δ_0 is defined in (2) above. The required size of s is determined with the following lemma, which is a slight generalization of the classic coupon collector problem.

Lemma 2. *Assume* **A3** *holds. If the number of chosen seeds,*

$$s \;\geq\; \frac{k(\log k + \log 1/\delta_0)}{(1-e^{-4})\nu_0}\,,$$

then with probability greater than $1-\delta_0$ for each $i = 1,\ldots,k$, at least one seed is in $\mathcal{S}_{i,\epsilon_0}$ and each seed column has at least

$$\eta_0 := \frac{64\,\beta\max\{\mu_1^2,\mu_0\}}{\nu_0}\; r\,\log^2(n) \qquad (3)$$

observed entries.

Proof Sketch: From Theorem 2.1, the expected number of observed entries per column is at least

$$\eta = \frac{128\,\beta\max\{\mu_1^2,\mu_0\}}{\nu_0}\; r\,\log^2(n)$$

Combining this with **A3**, the probability that the set does not contain at least one column from $\mathcal{S}_{i,\epsilon_0}$ with $\eta/2$ or more observed entries, for $i=1,\ldots,k$ is less than $\delta_0 = k(1-\nu_0'/k)^s$. Solving for s in terms of δ_0 yields the result. □

Next, for each seed we must find a set of n columns from the same subspace as the seed. This will be accomplished by identifying columns that are ϵ_0-close to the seed, so that if the seed belongs to $\mathcal{S}_{i,\epsilon_0}$, the columns must belong to the same subspace. Clearly the total number of columns N must be sufficiently large so that n or more such columns can be found. We will return to the requirement on N a bit later, after first dealing with the following challenge.

Since the columns are only partially observed, it may not be possible to determine how close each is to the seed. We address this by showing that if a column and the seed are both observed on enough common indices, then the incoherence assumption **A2** allows us reliably estimate the distance.

Lemma 3. *Assume* **A2** *and let $y = x_1 - x_2$, where x_1 and x_2 are two columns of $\mathbf{X}$. Assume there is a common set of indices of size $q \leq n$ where both x_1 and x_2 are observed. Let ω denote this common set of indices and let y_ω denote the corresponding subset of y. Then for any $\delta_0 > 0$, if the number of commonly observed elements*

$$q \;\geq\; 8\mu_1^2\log(2/\delta_0)\,,$$

then with probability at least $1-\delta_0$

$$\frac{1}{2}\|y\|_2^2 \;\leq\; \frac{n}{q}\|y_\omega\|_2^2 \;\leq\; \frac{3}{2}\|y\|_2^2\,.$$

Proof Sketch: Assumption **A2** implies that $n^2\|y\|_\infty^4 \leq \mu_1^2\|y\|_2^4$. Using McDiramid's Inequality we then have

$$\mathbb{P}\left(\left|\frac{n}{q}\|y_\omega\|_2^2 - \|y\|_2^2\right| \geq t\right) \;\leq\; 2\,\exp\left(\frac{-qt^2}{2\mu_1^2\|y\|_2^4}\right)\,.$$

Taking $t = \frac{1}{2}\|y\|_2^2$ yields the result. □

Suppose that $x_1 \in \mathcal{S}_{i,\epsilon_0}$ (for some i) and that $x_2 \notin \mathcal{S}_i$, and that both x_1, x_2 observe $q \geq 2\mu_0^2\log(2/\delta_0)$ common indices. Let y_ω denote the difference between x_1 and x_2 on the common support set. If the *partial distance* $\frac{n}{q}\|y_\omega\|_2^2 \leq \epsilon_0^2/2$, then the result above implies that with probability at least $1-\delta_0$

$$\|x_1 - x_2\|_2^2 \;\leq\; 2\,\frac{n}{q}\,\|y_\omega\|_2^2 \;\leq\; \epsilon_0^2.$$

On the other hand if $x_2 \in \mathcal{S}_i$ and $\|x_1 - x_2\|_2^2 \leq \epsilon_0^2/3$, then with probability at least $1-\delta_0$

$$\frac{n}{q}\,\|y_\omega\|_2^2 \;\leq\; \frac{3}{2}\|x_1 - x_2\|_2^2 \;\leq\; \epsilon_0^2/2\,.$$

Using these results we will proceed as follows. For each seed we find all columns that have at least $t_0 > 2\mu_0^2 \log(2/\delta_0)$ observations at indices in common with the seed (the precise value of t_0 will be specified in a moment). Assuming that this set is sufficiently large, we will select ℓn these columns uniformly at random, for some integer $\ell \geq 1$. In particular, ℓ will be chosen so that with high probability at least n of the columns will be within $\epsilon_0/\sqrt{3}$ of the seed, ensuring that with probability at least δ_0 the corresponding partial distance of each will be within $\epsilon_0/\sqrt{2}$. That is enough to guarantee with the same probability that the columns are within ϵ_0 of the seed. Of course, a union bound will be needed so that the distance bounds above hold uniformly over the $s\ell n$ columns under consideration, which means that we will need each to have at least $t_0 := 2\mu_0^2 \log(2s\ell n/\delta_0)$ observations at indices in common with the corresponding seed. All this is predicated on N being large enough so that such columns exist in $\mathbf{X}_\Omega$. We will return to this issue later, after determining the requirement for ℓ. For now we will simply assume that $N \geq \ell n$.

Lemma 4. *Assume* **A3** *and for each seed x let T_{x,ϵ_0} denote the number of columns of $\mathbf{X}$ in the ball of radius $\epsilon_0/\sqrt{3}$ about x. If the number of columns selected for each seed, ℓn, such that,*

$$\ell \;\geq\; \max\left\{ \frac{2k}{\nu_0(\frac{\epsilon_0}{\sqrt{3}})^r} \,,\, \frac{8k\log(s/\delta_0)}{n\nu_0(\frac{\epsilon_0}{\sqrt{3}})^r} \right\} \,,$$

then $\mathbb{P}\left(T_{x,\epsilon_0} \leq n\right) \;\leq\; \delta_0$ for all s seeds.

Proof Sketch: The probability that a column chosen uniformly at random from $\mathbf{X}$ belongs to this ball is at least $\nu_0(\epsilon_0/\sqrt{3})^r/k$, by Assumption **A3**. Therefore the expected number of points is

$$\mathbb{E}[T_{x,\epsilon_0}] \geq \frac{\ell n \nu_0(\frac{\epsilon_0}{\sqrt{3}})^r}{k} \,.$$

By Chernoff's bound for any $0 < \gamma < 1$

$$\mathbb{P}\left(T_{x,\epsilon_0} \leq (1-\gamma)\frac{\ell n \nu_0(\frac{\epsilon_0}{\sqrt{3}})^r}{k} \right) \leq \exp\left(\frac{-\gamma^2}{2} \frac{\ell n \nu_0(\frac{\epsilon_0}{\sqrt{3}})^r}{k} \right)$$

Take $\gamma = 1/2$. Requiring ℓ such that $\frac{\ell n \nu_0(\frac{\epsilon_0}{\sqrt{3}})^r}{2k} \geq n$ and $\exp\left(-\frac{\ell n \nu_0(\frac{\epsilon_0}{\sqrt{3}})^r}{8k}\right) \leq \delta_0/s$ yields the results. □

We can now formally state the procedure for finding local neighborhoods in Algorithm 1. Recall that the number of observed entries in each seed is at least η_0, per Lemma 2.

Lemma 5. *If N is sufficiently large and $\eta_0 > t_0$, then the Local Neighborhood Procedure in Algorithm 1 produces at least n columns within ϵ_0 of each seed, and at least one seed will belong to each of $\mathcal{S}_{i,\epsilon_0}$, for $i = 1, \ldots, k$, with probability at least $1 - 3\delta_0$.*

Algorithm 1 - Local Neighborhood Procedure

Input: n, k, μ_0, ϵ_0, ν_0, η_0, $\delta_0 > 0$.

$$\begin{aligned} s_0 &:= \left\lceil \frac{k(\log k + \log 1/\delta_0)}{(1-e^{-4})\nu_0} \right\rceil \\ \ell_0 &:= \left\lceil \max\left\{ \frac{2k}{\nu_0(\frac{\epsilon_0}{\sqrt{3}})^r} \,,\, \frac{8k\log(s_0/\delta_0)}{n\nu_0(\frac{\epsilon_0}{\sqrt{3}})^r} \right\} \right\rceil \\ t_0 &:= \lceil 2\mu_0^2 \log(2s_0\ell_0 n/\delta_0) \rceil \end{aligned}$$

Steps:

1. Select s_0 "seed" columns uniformly at random and discard all with less than η_0 observations
2. For each seed, find all columns with t_0 observations at locations observed in the seed
3. Randomly select $\ell_0 n$ columns from each such set
4. Form local neighborhood for each seed by randomly selecting n columns with partial distance less than $\epsilon_0/\sqrt{2}$ from the seed

Proof. Lemma 2 states that if we select s_0 seeds, then with probability at least $1 - \delta_0$ there is a seed in each $\mathcal{S}_{i,\epsilon_0}$, $i = 1, \ldots, k$, with at least η_0 observed entries, where η_0 is defined in (3). Lemma 4 implies that if $\ell_0 n$ columns are selected uniformly at random for each seed, then with probability at least $1 - \delta_0$ for each seed at least n of the columns will be within a distance $\epsilon_0/\sqrt{3}$ of the seed. Each seed has at least η_0 observed entries and we need to find $\ell_0 n$ other columns with at least t_0 observations at indices where the seed was observed. Provided that $\eta_0 \geq t_0$, this is certainly possible if N is large enough. It follows from Lemma 3 that $\ell_0 n$ columns have at least t_0 observations at indices where the seed was also observed, then with probability at least $1 - \delta_0$ the partial distances will be within $\epsilon_0/\sqrt{2}$, which implies the true distances are within ϵ_0. The result follows by the union bound. □

Finally, we quantify just how large N needs to be. Lemma 4 also shows that we require at least

$$N \;\geq\; \ell n \;\geq\; \max\left\{ \frac{2kn}{\nu_0(\frac{\epsilon_0}{\sqrt{3}})^r} \,,\, \frac{8k\log(s/\delta_0)}{\nu_0(\frac{\epsilon_0}{\sqrt{3}})^r} \right\} \,.$$

However, we must also determine a lower bound on the probability that a column selected uniformly at random has at least t_0 observed indices in common with a seed. Let γ_0 denote this probability, and let p_0 denote the probability of observing each entry in $\mathbf{X}$. Note that our main result, Theorem 2.1, assumes that

$$p_0 \geq \frac{128\,\beta \max\{\mu_1^2, \mu_0\}}{\nu_0} \frac{r\,\log^2(n)}{n} \,.$$

Since each seed has at least η_0 entries observed, γ_0 is greater than or equal to the probability that a Binomial(η_0, p_0) random variable is at least t_0. Thus,

$$\gamma_0 \;\geq\; \sum_{j=t_0}^{\eta_0} \binom{\eta_0}{j} p_0^j (1-p_0)^{\eta_0 - j} \;.$$

This implies that the expected number of columns with t_0 or more observed indices in common with a seed is at least $\gamma_0 N$. If $\widetilde{n}$ is the actual number with this property, then by Chernoff's bound, $\mathbb{P}(\widetilde{n} \leq \gamma_0 N/2) \leq \exp(-\gamma_0 N/8)$. So $N \geq 2\ell_0 \gamma_0^{-1} n$ will suffice to guarantee that enough columns can be found for each seed with probability at least $1 - s_0 \exp(-\ell_0 n/4) \geq 1 - \delta_0$ since this will be far larger than $1 - \delta_0$, since δ_0 is polynomial in n.

To take this a step further, a simple lower bound on γ_0 is obtained as follows. Suppose we consider only a t_0-sized subset of the indices where the seed is observed. The probability that another column selected at random is observed at all t_0 indices in this subset is $p_0^{t_0}$. Clearly $\gamma_0 \geq p_0^{t_0} = \exp(t_0 \log p_0) \geq (2 s_0 \ell_0 n)^{2\mu_0^2 \log p_0}$. This yields the following sufficient condition on the size of N:

$$N \;\geq\; \ell_0 n (2 s_0 \ell_0 n / \delta_0)^{2\mu_0^2 \log p_0^{-1}} \;.$$

From the definitions of s_0 and ℓ_0, this implies that if $2\mu_0^2 \log p_0$ is a fixed constant, then a sufficient number of columns will exist if $N = O(\text{poly}(kn/\delta_0))$. For example, if $\mu_0^2 = 1$ and $p_0 = 1/2$, then $N = O((kn)/\delta_0)^{2.4})$ will suffice; i.e., N need only grow polynomially in n. On the other hand, in the extremely undersampled case p_0 scales like $\log^2(n)/n$ (as n grows and r and k stay constant) and N will need to grow almost exponentially in n, like $n^{\log n - 2 \log \log n}$.

4 Local Subspace Completion

For each of our local neighbor sets, we will have an incompletely observed $n \times n$ matrix; if all the neighbors belong to a single subspace, the matrix will have rank $\leq r$. First, we recall the following result from low-rank matrix completion theory [1].

Lemma 6. *Consider an $n \times n$ matrix of rank $\leq r$ and row and column spaces with coherences bounded above by some constant μ_0. Then the matrix can be exactly completed if*

$$m' \geq 64 \max\left(\mu_1^2, \mu_0\right) \beta r n \log^2(2n) \tag{4}$$

entries are observed uniformly at random, for constants $\beta > 0$ and with probability $\geq 1 - 6(2n)^{2-2\beta} \log n - n^{2-2\beta^{1/2}}$.

We wish to apply these results to our local neighbor sets, but we have three issues we must address: First, the sampling of the matrices formed by local neighborhood sets is not uniform since the set is selected based on the observed indices of the seed. Second, given Lemma 2 we must complete not one, but s_0 (see Algorithm 1) incomplete matrices simultaneously with high probability. Third, some of the local neighbor sets may have columns from more than one subspace. Let us consider each issue separately.

Firstly, the fact that our incomplete submatrices are not sampled uniformly is due to the fact that the columns in each submatrix are selected to have at least some amount of overlap in observations as the seed column. Therefore, the indices on which the seed is observed are highly sampled on the other columns as well. This can be corrected with a simple thinning procedure which is detailed in the full length version of this paper [14]. The dependence on the seed column cannot be eliminated, and so the seed column must be removed from the submatrix.

Once each neighborhood matrix has been thinned, we have the following matrix completion guarantee for each neighborhood matrix.

Lemma 7. *Assume all s_0 seed neighborhood matrices are thinned according to the discussion above, have rank $\leq r$, and the matrix entries are observed uniformly at random with probability,*

$$p_0 \geq \frac{128\,\beta \max\{\mu_1^2, \mu_0\}}{\nu_0} \; \frac{r \, \log^2(n)}{n} \tag{5}$$

Then with probability $\geq 1 - 12 s_0 n^{2-2\beta^{1/2}} \log n$, all s_0 matrices can be perfectly completed.

Proof Sketch: From Lemma 6 and the union bound, if each matrix has

$$m' \geq 64 \max\left(\mu_1^2, \mu_0\right) \beta r n \log^2(2n)$$

entries observed uniformly at random (with replacement), then with probability $\geq 1 - 12 s_0 n^{2-2\beta^{1/2}} \log n$, all s_0 matrices are perfectly completed.

Under our sampling assumptions, enough entries are observed in each of the s_0 seed matrices with probability at least $1 - \exp(-n^2 p_0/8 + \log s_0)$ if $p_0 \geq \frac{128\,\beta \max\{\mu_1^2, \mu_0\}}{\nu_0} \; \frac{r \, \log^2(n)}{n}$. Since $n^2 p_0 > rn \log^2 n$ and $s_0 = O(k(\log k + \log n))$, this probability tends to zero exponentially in n as long as $k = o(e^n)$, which holds according to Assumption **A1**. Therefore this holds with probability at least $1 - 12 s_0 n^{2-2\beta^{1/2}} \log n$. □

Finally, let us consider the third issue, the possibility that one or more of the points in the neighborhood of a seed lies in a subspace different than the seed subspace. When this occurs, the rank of the submatrix formed by the seed's neighbor columns will be larger than the dimension of the seed subspace. Without loss of generality assume that we have only two subspaces represented in the neighbor set, and assume their dimensions are r' and r''. First, in the case

that $r' + r'' > r$, when a rank $\geq r$ matrix is completed to a rank r matrix, with overwhelming probability there will be errors with respect to the observations as long as the number of samples in each column is $O(r \log r)$, which is assumed in our case; see [12]. Thus we can detect and discard these candidates. Secondly, in the case that $r' + r'' \leq r$, we still have enough samples to complete this matrix successfully with high probability. However, since we have drawn enough seeds to guarantee that every subspace has a seed with a neighborhood entirely in that subspace, we will find that this problem seed is redundant. This is determined in the Subspace Refinement step.

5 Subspace Refinement

Each of the matrix completion steps above yields a low-rank matrix with a corresponding column subspace, which we will call the *candidate* subspaces. While the true number of subspaces will not be known in advance, since $s_0 = O(k(\log k + \log(1/\delta_0))$, the candidate subspaces will contain the true subspaces with high probability (see Lemma 4). We must now deal with the algorithmic issue of determining the true set of subspaces.

We first note that, from Assumption **A3**, with probability 1 a set of points of size $\geq r$ all drawn from a single subspace $\mathcal{S}$ of dimension $\leq r$ will span $\mathcal{S}$. In fact, any $b < r$ points will span a b-dimensional subspace of the r-dimensional subspace $\mathcal{S}$.

Assume that $r < n$, since otherwise it is clearly necessary to observe all entries. Therefore, if a seed's nearest neighborhood set is confined to a single subspace, then the columns in span their subspace. And if the seed's nearest neighborhood contains columns from two or more subspaces, then the matrix will have rank larger than that of any of the constituent subspaces. Thus, if a certain candidate subspace is spanned by the union of two or more smaller candidate subspaces, then it follows that that subspace is not a true subspace (since we assume that none of the true subspaces are contained within another).

This observation suggests the following subspace refinement procedure. The s_0 matrix completions yield $s \leq s_0$ candidate column subspaces; s may be less than s_0 since completions that fail are discarded as described above. First sort the estimated subspaces in order of rank from smallest to largest (with arbitrary ordering of subspaces of the same rank), which we write as $\mathcal{S}_{(1)}, \ldots, \mathcal{S}_{(s)}$. We will denote the final set of estimated subspaces as $\widehat{\mathcal{S}}_1, \ldots, \widehat{\mathcal{S}}_k$. The first subspace $\widehat{\mathcal{S}}_1 := \mathcal{S}_{(1)}$, a lowest-rank subspace in the candidate set. Next, $\widehat{\mathcal{S}}_2 = \mathcal{S}_{(2)}$ if and only if $\mathcal{S}_{(2)}$ is not contained in $\widehat{\mathcal{S}}_1$. Following this simple sequential strategy, suppose that when we reach the candidate $\mathcal{S}_{(j)}$ we have so far determined $\widehat{\mathcal{S}}_1, \ldots, \widehat{\mathcal{S}}_i$, $i < j$. If $\mathcal{S}_{(j)}$ is not in the span of $\cup_{\ell=1}^{i} \widehat{\mathcal{S}}_\ell$, then we set $\widehat{\mathcal{S}}_{i+1} = \mathcal{S}_{(j)}$, otherwise we move on to the next candidate. In this way, we can proceed sequentially through the rank-ordered list of candidates, and we will identify all true subspaces.

6 The Full Monty

Now all will be revealed. At this point, we have identified the true subspaces, and all N columns lie in the span of one of those subspaces. For ease of presentation, we assume that the number of subspaces is exactly k. However if columns lie in the span of fewer than k, then the procedure above will produce the correct number. To complete the full matrix, we proceed one column at a time. For each column of $\mathbf{X}_\Omega$, we determine the correct subspace to which this column belongs, and we then complete the column using that subspace. We can do this with high probability due to results from [12, 15].

The first step is that of subspace assignment, determining the correct subspace to which this column belongs. In [15], it is shown that given k subspaces, an incomplete vector can be assigned to its closest subspace with high probability given enough observations. In the situation at hand, we have a special case of the results of [15] because we are considering the more specific situation where our incomplete vector lies exactly in one of the candidate subspaces, and we have an upper bound for both the dimension and coherence of those subspaces.

Lemma 8. *Let $\{\mathcal{S}_1, \ldots, \mathcal{S}_k\}$ be a collection of k subspaces of dimension $\leq r$ and coherence parameter bounded above by μ_0. Consider column vector x with index set $\Omega \in \{1, \ldots, n\}$, and define $P_{\Omega,\mathcal{S}_j} = U^j_\Omega \left(\left(U^j_\Omega\right)^T U^j_\Omega \right)^{-1} \left(U^j_\Omega\right)^T$, where U^j is the orthonormal column span of $\mathcal{S}_j$ and U^j_Ω is the column span of $\mathcal{S}_j$ restricted to the observed rows, Ω. Without loss of generality, suppose the column of interest $x \in \mathcal{S}_1$. If* **A3** *holds, and the probability of observing each entry of x is independent and Bernoulli with parameter*

$$p_0 \geq \frac{128\,\beta \max\{\mu_1^2, \mu_0\}}{\nu_0} \; \frac{r \, \log^2(n)}{n} \; .$$

Then with probability at least $1 - (3(k-1)+2)\delta_0$,

$$\|x_\Omega - P_{\Omega,\mathcal{S}_1} x_\Omega\|_2^2 = 0 \tag{6}$$

and for $j = 2, \ldots, k$

$$\|x_\Omega - P_{\Omega,\mathcal{S}_j} x_\Omega\|_2^2 \; > \; 0 \; . \tag{7}$$

Proof. The full proof of this lemma can be found in [14]; it is a straightforward extension of ideas from [12, 15]. □

Finally, denote the column to be completed by x_Ω. To complete x_Ω we first determine which subspace it belongs to using the results above. For a given column we can use the *incomplete data projection residual* of (6). With

probability at least $1 - (3(k-1)+2)\delta_0$, the residual will be zero for the correct subspace and strictly positive for all other subspaces. Using the span of the chosen subspace, U, we can then complete the column by using $\widehat{x} = U \left(U_\Omega^T U_\Omega\right)^{-1} U_\Omega^T x_\Omega$.

We reiterate that Lemma 8 allows us to complete a single column x with probability $1 - (3(k-1)+2)\delta_0$. If we wish to complete the entire matrix, we will need another union bound over all N columns, leading to a $\log N$ factor in our requirement on p_0. Since N may be quite large in applications, we prefer to state our result in terms of per-column completion bound.

The confidence level stated in Theorem 2.1 is the result of applying the union bound to all the steps required in the Sections 3, 4, and 6. All hold simultaneously with probability at least $1-(6+3(k-1)+12s_0)\delta_0 < 1-(6+15s_0)\delta_0$, which proves the theorem.

7 Experiments

The following experiments evaluate the performance of the proposed high-rank matrix completion procedure and compare results with standard low-rank matrix completion based on nuclear norm minimization.

7.1 Numerical Simulations

We begin by examining a highly synthesized experiment where the data exactly matches the assumptions of our high-rank matrix completion procedure. The key parameters were chosen as follows: $n = 100$, $N = 5000$, $k = 10$, and $r = 5$. The k subspaces were r-dimensional, and each was generated by r vectors drawn from the $\mathcal{N}(0, I_n)$ distribution and taking their span. The resulting subspaces are highly incoherent with the canonical basis for $\mathbb{R}^n$. For each subspace, we generate 500 points drawn from a $\mathcal{N}(0, UU^T)$ distribution, where U is a $n \times r$ matrix whose orthonormal columns span the subspace. Our procedure was implemented using $\lceil 3k \log k \rceil$ seeds. The matrix completion software called GROUSE (available here [16]) was used in our procedure and to implement the standard low-rank matrix completions. We ran 50 independent trials of our procedure and compared it to standard low-rank matrix completion. The results are summarized in the Figure 2.

7.2 Network Topology Inference Experiments

As a complement to the heavy network load of standard active probing methods (*e.g.,* [17]), which scale poorly for Internet-scale networks, recent research has focused on the ability to recover Internet connectivity from passively observed measurements [9]. An example of this measurement infrastructure can be seen in Figure 3-(Left). As detailed in the introduction, this passive measurements infrastructure results in a massively incomplete $n \times N$ matrix, where n is the number of passive monitors, N is the total unique IP addresses observed, and the matrix lies in the union of k low-rank subspaces, where k is the number of subnets.

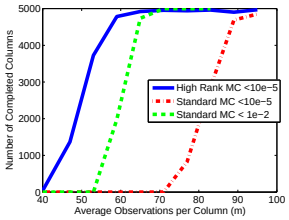

Figure 2: The number of correctly completed columns (with tolerances shown above, 10e-5 or 0.01), versus the average number of observations per column. The high rank MC method provides accurate completion with only about 50 samples per column, meanwhile the standard low rank MC method requires almost all samples in each column.

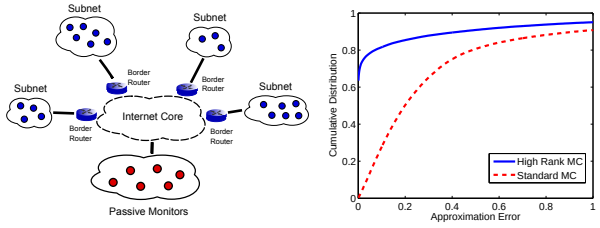

Figure 3: Passive Internet measurements. (Left) - Internet topology example of subnets sending traffic to passive monitors through the Internet core and common border routers, (Right) - Hop count imputation results: Cumulative distribution of estimation error is shown with respect to observing 40% of the total elements.

Using a Heuristically Optimal Topology from [18], we simulated a network topology and measurement infrastructure consisting of $N = 2700$ total IP addresses uniformly distributed over $k = 12$ different subnets. The hop counts are generated on the topology using shortest-path routing from $n = 75$ passive monitors. Observing only 40% of the total hop counts, in Figure 3-(Right) we compare the performance of the high-rank procedure with standard low-rank matrix completion. The experiment shows dramatic improvements, as over 70% of the missing hop counts can be imputed exactly using the high-rank matrix completion methodology, while approximately no missing elements are imputed exactly using low-rank matrix completion.

8 Acknowledgements

This work was supported in part by AFOSR grant number FA9550-09-1-0140 and FA9550-09-1-0643. Any opinions, findings, conclusions or other recommendations expressed in this material are those of the authors and do not necessarily reflect the AFOSR.

References

[1] B. Recht, "A Simpler Approach to Matrix Completion," in *To appear in Journal of Machine Learning Research, arXiv:0910.0651v2*.

[2] E. J. Candès and T. Tao, "The Power of Convex Relaxation: Near-Optimal Matrix Completion." in *IEEE Transactions on Information Theory*, vol. 56, May 2010, pp. 2053–2080.

[3] R. Vidal, "A Tutorial on Subspace Clustering," in *Johns Hopkins Technical Report*, 2010.

[4] K. Kanatani, "Motion Segmentation by Subspace Separation and Model Selection," in *Computer Vision, 2001. ICCV 2001. Proceedings. Eighth IEEE International Conference on*, vol. 2, 2001, pp. 586–591.

[5] R. Vidal, Y. Ma, and S. Sastry, "Generalized Principal Component Analysis (GPCA)," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, December 2005.

[6] G. Lerman and T. Zhang, "Robust Recovery of Multiple Subspaces by L_p Minimization," 2011, Preprint at http://arxiv.org/abs/1104.3770.

[7] A. Gruber and Y. Weiss, "Multibody Factorization with Uncertainty and Missing Data using the EM Algorithm," in *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, vol. 1, June 2004.

[8] R. Vidal, R. Tron, and R. Hartley, "Multiframe Motion Segmentation with Missing Data Using Power Factorization and GPCA," *International Journal of Computer Vision*, vol. 79, pp. 85–105, 2008.

[9] B. Eriksson, P. Barford, R. Nowak, and M. Crovella, "Learning Network Structure from Passive Measurements," in *Proceedings of ACM Internet Measurement Conference*, October 2007.

[10] B. Eriksson, P. Barford, and R. Nowak, "Network Discovery from Passive Measurements," in *Proceedings of ACM SIGCOMM Conference*, August 2008.

[11] E. Candès and B. Recht, "Exact Matrix Completion Via Convex Optimization." in *Foundations of Computational Mathematics*, vol. 9, 2009, pp. 717–772.

[12] L. Balzano, B. Recht, and R. Nowak, "High-Dimensional Matched Subspace Detection When Data are Missing," in *Proceedings of the International Conference on Information Theory*, June 2010, available at http://arxiv.org/abs/1002.0852.

[13] G. Chen and M. Maggioni, "Multiscale Geometric and Spectral Analysis of Plane Arrangements," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2011.

[14] B. Eriksson, L. Balzano, and R. Nowak, "High-Rank Matrix Completion and Subspace Clustering with Missing Data," in *Technical Report*, December 2011. [Online]. Available: http://arxiv.org/abs/1112.5629

[15] L. Balzano, R. Nowak, A. Szlam, and B. Recht, "k-Subspaces with missing data," University of Wisconsin, Madison, Tech. Rep. ECE-11-02, February 2011.

[16] L. Balzano and B. Recht, 2010, http://sunbeam.ece.wisc.edu/grouse/.

[17] N. Spring, R. Mahajan, and D. Wetherall, "Measuring ISP Topologies with Rocketfuel," in *Proceedings of ACM SIGCOMM*, August 2002.

[18] L. Li, D. Alderson, W. Willinger, and J. Doyle, "A First-Principles Approach to Understanding the Internet's Router-Level Topology," in *Proceedings of ACM SIGCOMM Conference*, August 2004.