ON THE SAMPLE COMPLEXITY OF SUBSPACE CLUSTERING WITH MISSING DATA

D. Pimentel, R. Nowak

University of Wisconsin
Electrical and Computer Engineering
Madison, WI, 53706, USA

L. Balzano

University of Michigan
Electrical Engineering and Computer Science
Ann Arbor, MI, 48109, USA

ABSTRACT

Subspace clustering is a useful tool for analyzing large complex data, but in many relevant applications missing data are common. Existing theoretical analysis of this problem employs standard incoherence assumptions and assumes the data are missing at random to show that subspace clustering from incomplete data is possible. However, that analysis requires the number of samples (i.e., partially observed vectors) to be *super-polynomial in the dimension d*. Such huge sample sizes are unnecessary when no data are missing and uncommon in applications. With missing data, huge sample sizes are sufficient, but are they necessary? This important open question is addressed here. There are two main contributions in this paper. First, it is shown that if subspaces have rank at most r and the number of partially observed vectors greater than d^{r+1} (times a poly-logarithmic factor), then with high probability the true subspaces are the only subspaces that agree with the observed data. This implies that subspace clustering may be possible without impractically large sample sizes. Moreover, it tells us that we can certify the output of *any* subspace clustering algorithm by checking its fit to the observed data. The second main contribution is a novel EM-type algorithm for subspace clustering with missing data. We demonstrate and compare it to several other algorithms. Experiments with simulated and real data show that such algorithms work well in practice.

Index Terms— matrix completion, subspace clustering

1. INTRODUCTION

Let $\mathbf{X}$ be a $d \times N$ data matrix whose columns lie in the union of several unknown low-dimensional subspaces of $\mathbb{R}^d$. The goal of subspace clustering is to infer the underlying subspaces from $\mathbf{X}$ and to cluster the columns of $\mathbf{X}$ according to the subspaces. In this paper, we suppose that $\mathbf{X}$ is partially observed with entries missing at random and aim at the same goal.

This problem arises in applications ranging from computer vision [1, 2] to network inference [3]. Existing theoretical analysis of this problem employs standard subspace incoherence assumptions to show that subspace clustering from incomplete data is possible provided the number of samples N is super-polynomial in the dimension of the subspaces [4]. In practice, it is rare to have such huge numbers of samples. Several heuristic algorithms have been proposed for subspace clustering with missing data [5, 6, 7]. These methods sometimes work reasonably well in practice, but lack theoretical justification.

The true sample complexity of subspace clustering with missing data is an important open question addressed here. The main theoretical contribution of the paper shows that if the number of partially observed data vectors per subspace is a rank-degree power of the dimension, then with high probability the true subspaces are the only subspaces that agree with the observed data. This implies that subspace clustering may be possible in practice without large numbers of samples. This is a significant improvement over the large requirements in [4], specially in the cases where r is small, as happens in many real applications, e.g. the ones approached in Section 6, where $r = \mathscr{O}(1)$. The main algorithmic contribution of the paper is a new, computationally-efficient EM algorithm for subspace clustering with missing data. Experiments with real and synthetic data show that the EM algorithm performs better than existing methods even with $\mathscr{O}(Kd)$ samples - fewer than the theoretically derived sample complexity bound.

Subspace clustering with missing data shares aspects of *subspace clustering* [8, 9, 10] and *low-rank matrix completion* [11], and the combination of these areas is the central perspective of prior theoretical analysis of the problem (also referred to as *high-rank matrix completion*) [4]. The main assumptions used in our theoretical analysis are fundamentally different from those arising from traditional Low-Rank Matrix Completion, e.g. [4], but in some sense, more natural and much less restrictive. Specifically, the columns of $\mathbf{X}$ are drawn from a non-atomic distribution supported on a union of low-dimensional generic subspaces, and entries in the columns are missing uniformly at random (precise statements of our assumptions are given in the next section). The main theoretical result of this paper shows that, under these assumptions, subspace clustering with missing data is possible from far fewer observations than the (often impractically) large number required by the algorithm in [4].

In addition to the algorithm in [4], other heuristic procedures have been proposed without theoretical support [5, 7]. For problems with significant fractions of missing data and practically relevant sample sizes, the *k-subspaces* algorithm in [5], which takes the algorithm of [12] and generalizes it to handle missing data, has been shown to be particularly effective. This paper proposes a new EM algorithm for subspace clustering with missing data, which can be viewed as a generalization of [13] to handle missing data and/or a generalization of [14] to low-rank covariances. The EM algorithm is computationally-efficient (just slightly more intensive than the greedy method of [5]) and outperforms the algorithms in [5, 4] in experiments with real and synthetic data. Furthermore, our main theoretical result enables one to certify the result of *any* subspace clustering algorithm by checking the residual error on the observed data.

2. KEY ASSUMPTIONS AND MAIN RESULTS

We state and prove our result in order to give the fundamental ideas behind our approach.

Definition 1. *We denote the set of $d \times N$ matrices with rank r by $\mathcal{M}(r, d \times N)$. A* generic *$(d \times n)$-matrix of rank r is a continuous $\mathcal{M}(r, d, n)$-valued random variable. We say a subspace S is* generic *if a matrix whose columns are drawn i.i.d. according to a non-atomic distribution with support on S is generic a.s.*

A1. The columns of our $d \times N$ data matrix $\mathbf{X}$ are drawn according to a non-atomic distribution with support on the union of at most K generic subspaces. The subspaces, denoted by $\mathcal{S} = \{S_k\}$, each has rank exactly $r < d$.

A2. The probability that a column is drawn from subspace k is ρ_k. Let ρ_* be a bound on $\min_k\{\rho_k\}$.

A3. We observe $\mathbf{X}$ only on a set of entries Ω and denote the observation $\mathbf{X}_\Omega$. Each entry in $\mathbf{X}_\Omega$ is sampled independently with probability p.

Throughout the paper we use $\mathbf{X}_\Omega$ to refer indistinctly to the matrix $\mathbf{X}_\Omega$ as well as the set of columns of the matrix $\mathbf{X}_\Omega$. We split $\mathbf{X}_\Omega$ in two sets: the *search* set, $\widetilde{\mathbf{X}}_\Omega$, with $\widetilde{N} := |\widetilde{\mathbf{X}}_\Omega|$ columns, and the *test* set, $\bar{\mathbf{X}}_\Omega$, with $\bar{N} := |\bar{\mathbf{X}}_\Omega|$ columns, s.t. $N := |\mathbf{X}_\Omega| = \widetilde{N} + \bar{N}$. We use $\widetilde{\mathbf{X}}_\Omega^{[k]}$ to denote the columns of $\widetilde{\mathbf{X}}_\Omega$ corresponding to the k^{th} subspace, and equivalently for $\bar{\mathbf{X}}_\Omega^{[k]}$.

We now present our main result, which we prove in Section 3. The theorem below shows that if we observe at least order $d^{r+1}(\log d/r + \log K)$ columns and at least order $r \log^2 d$ entries in each column, then identification of the subspaces is possible with large probability. This result can be easily generalized to a relaxed version of assumption **A1** to the case where the dimensions of the subspaces are upper bounded by r. Note, that the total number of columns needed is only polynomial in d, in contrast to the super-polynomial requirement of $d^{\log d}$ of the best previously existing bounds [4].

Theorem 1. *Suppose* **A1-A3** *hold. Let $\epsilon > 0$ be given. Assume the number of subspaces $K \le \frac{\epsilon}{6} e^{d/4}$, the total number of columns $N = \widetilde{N} + \bar{N} \ge (2d + 4M)/\rho_*$, and*

$$p \ge \frac{1}{d} 128 \mu_1^2 r \beta_0 \log^2(2d),$$

$$\beta_0 = \sqrt{1 + \frac{\log\left(\frac{6K}{\epsilon} 12 \log(d)\right)}{2\log(2d)}},$$

$$M = \left(\frac{de}{r+1}\right)^{r+1} \left((r+1)\log\left(\frac{de}{r+1}\right) + \log\left(\frac{8K}{\epsilon}\right)\right),$$

where $\mu_1^2 := \max_k \frac{d^2}{r} \|U_k V_k^*\|_\infty^2$ and $U_k \Sigma_k V_k^*$ is the singular value decomposition of $\widetilde{\mathbf{X}}^{[k]}$. Then with probability at least $1 - \epsilon$, $\mathcal{S}$ can be uniquely determined from $\mathbf{X}_\Omega$.

3. PROOFS OF MAIN RESULTS

The core of the main result lies in Lemmas 8 and 10 below.

First we make some notational remarks. When we write $x_\omega \in \mathbf{X}_\Omega$, we are referring to one column in the set of columns $\mathbf{X}_\Omega$. Additionally if we write $x_\omega \in S_k$, we mean that the subspace S_k fits x_ω, or by $\mathbf{X}_\Omega \subset S_k$ we mean that S_k fits all the columns of $\mathbf{X}_\Omega$ in the sense that there exists a completion $\hat{x}$ of x_ω such that $\hat{x} \in S_k$.

We also introduce the definition of a *validating set*:

Definition 2. *(Validating set) Consider a collection of columns $\{x_{i_{\omega_i}}\}_{i=1}^m$. Consider a graph $\mathcal{G}$ with m nodes representing these m columns, where edge (i, j) exists if $|\omega_i \cap \omega_j| > r$. We say $\{x_{i_{\omega_i}}\}_{i=1}^m$ is a validating set if $\mathcal{G}$ is connected and $\bigcup_{i=1}^m \omega_i = \{1, ..., d\}$.*

3.1. Intuition

We are now ready to describe the intuition behind our approach. We consider an exhaustive search over every set of d columns in the search set. We use $\widetilde{\mathbf{X}}_\Omega^{(\ell)}$ to denote the ℓ^{th} combination of d columns of $\widetilde{\mathbf{X}}_\Omega$, where ℓ ranges from 1 to $\binom{\widetilde{N}}{d}$. For each of these combinations, if there is a subspace that uniquely fits all d columns, we validate it by finding a subset of the test set that fits the subspace and is also a validating set. We use $\hat{\mathcal{S}}$ to denote the collection of all subspaces satisfying these conditions. This procedure is detailed in Algorithm 1.

By Lemma 8 below, every combination of d columns from a single subspace will have a unique completion and fit a validating subspace with high probability. By Lemma 10 only true subspaces can fit a validating set. Putting together these two results, we get that with high probability $\hat{\mathcal{S}} = \mathcal{S}$. That is, we can identify $\mathcal{S}$ from $\mathbf{X}_\Omega$.

Algorithm 1 Subspaces Identification

```
Input: $\mathbf{X}_\Omega$, $r$, $\rho_*$
Output: $\hat{\mathcal{S}}$
Set $\widetilde{\mathbf{X}}_\Omega$ as the 1st $2d/\rho_*$ columns of $\mathbf{X}_\Omega$, and $\bar{\mathbf{X}}_\Omega$ as the remaining ones.
Initialize $\hat{\mathcal{S}} = 0$.
for $\ell = 1$ to $\binom{N}{d}$ do
    if $\widetilde{\mathbf{X}}^{(\ell)}_\Omega$ is uniquely $r$-completable. then
        $\widetilde{S}_\ell = \text{span}\{\widetilde{\mathbf{X}}^\ell\}$
        if $\widetilde{S}_\ell$ fits a validating set from $\bar{\mathbf{X}}_\Omega$. then
            $\hat{\mathcal{S}} = \hat{\mathcal{S}} \cup \widetilde{S}_\ell$.
        end if
    end if
end for
```

Algorithm 2 High-Rank Matrix Completion

```
Input: $\mathbf{X}_\Omega$, $\hat{\mathcal{S}}$
for $i = 1$ to $N$ do
    $k$ = argument s.t. the residual of $x_\omega$ onto $\hat{S}_k = 0$.
    Complete $x_\omega$ according to $S_k$.
end for
```

Once $\hat{\mathcal{S}}$ is known, completing $\mathbf{X}_\Omega$ becomes a trivial task (Algorithm 2), since with enough observations ($> r$) in each column vector and the assumption of genericity, one can easily determine which subspace each column belongs to by simply projecting the observed coordinates of the column onto each of the subspaces in $\hat{\mathcal{S}}$, and then completing the missing entries according to that subspace.

3.2. Low-Rank Matrix Completion

We begin the proof of our main theorem using Theorems 2 of [11] and 2.6 of [15] with some adjustments to our context. We state our versions here as Lemmas 1 and 2.

Lemma 1 (Low-Rank Matrix Completion [11])**.** *Consider a $d \times d$ rank-r matrix $\mathbf{Y}$ with singular value decomposition $\mathbf{Y} = U\Sigma V^*$. Let the row and column spaces of $\mathbf{Y}$ have coherences (as in Definition 1 of [11]) bounded above by some positive μ_0. Suppose the matrix UV^* has a maximum entry bounded by $\mu_1 \sqrt{r/Cd^2}$ in absolute value for some positive μ_1. Suppose that every entry of $\mathbf{Y}$ has been observed independently with probability p to yield Y_Ω, with*

$$p \geq \frac{1}{d} 128 \max\{\mu_1^2, \mu_0\} r \beta_0 \log^2(2d)$$

and β_0 as in Theorem 1. Then $\mathbf{Y}^*$, the minimizer to the nuclear norm minimization problem (Equation 2 of [11]) is unique and equal to $\mathbf{Y}$ with probability at least $1 - \frac{\epsilon}{3K}$.

Proof. $\mathsf{P}((i,j) \in \Omega) = p$ by definition, so $\mathsf{E}[|\Omega|] = pd^2$. Also, $|\Omega| = \sum_{i,j=1}^{d,d} \mathbb{1}_{\{(i,j)\in\Omega\}}$, so using the multiplicative form of the Chernoff bound we get

$$\mathsf{P}\left(|\Omega| \leq \frac{pd^2}{2}\right) = \mathsf{P}\left(|\Omega| \leq (1-\beta)\mathsf{E}\left[|\Omega|\right]\right) \leq e^{-\frac{\beta^2}{2}\mathsf{E}[|\Omega|]}$$
$$\leq e^{-\frac{pd^2}{8}} \leq \frac{\epsilon}{6K}$$

by taking $\beta = 1/2$ and since $pd^2 > 8\log(\frac{6K}{\epsilon})$.

Given $|\Omega| \geq pd^2/2$, Equation 2 of [11] is satisfied for β_0 as in Theorem 1, and so all assumptions of Theorem 2 of [11] hold, and therefore $\mathbf{Y}^* = \mathbf{Y}$, i.e. $\mathbf{Y}$ can be recovered from $\mathbf{Y}_\Omega$, with probability at least $1 - \frac{\epsilon}{6K}$.

Using the Law of Total Probability on the event that $\mathbf{Y}^* \neq \mathbf{Y}$ conditioning on the event that $|\Omega| > pd^2/2$ and remembering that $\mathsf{P}(\cdot) \leq 1$ we have the Lemma:

$$\mathsf{P}(\mathbf{Y}^* \neq \mathbf{Y}) \leq \mathsf{P}\left(\mathbf{Y}^* \neq \mathbf{Y} | \, |\Omega| > m\right) + \mathsf{P}(|\Omega| \leq m)$$
$$\leq \frac{\epsilon}{6K} + \frac{\epsilon}{6K} = \frac{\epsilon}{3K}$$

□

Lemma 2 (Completion Identifiability [15])**.** *Let Ω be given. Let $\mathbf{X}$ and $\mathbf{Y}$ be two different generic rank-r matrices. Then $\mathbf{X}_\Omega$ is completable (i.e. $\mathbf{X}$ can be recovered from $\mathbf{X}_\Omega$) if and only if $\mathbf{Y}_\Omega$ is completable.*

Proof. See Theorem 2.6 of [15]. □

These two lemmas are used to prove Lemma 3, which is a version of Lemma 1 that gives us a probability of Low-Rank completion of generic matrices.

Lemma 3 (Generic Low-Rank Matrix Completion)**.** *Consider a $d \times d$ generic matrix $\mathbf{X}$. Suppose that every entry of $\mathbf{X}$ has been observed independently with probability p to yield $\mathbf{X}_\Omega$, with p and β_0 as in Theorem 1. Then $\mathbf{X}$ can be recovered with probability at least $1 - \frac{\epsilon}{3K}$.*

Proof. In Lemma 1, μ_0 and μ_1 satisfy $1 \leq \mu_0 \leq d/r$ and $\mu_1 \geq 1$. So we can take a generic matrix $\mathbf{Y}$ that satisfies all assumptions of Lemma 1 with $\mu_0 = 1$. We know we can do this, as there exist matrices that are both incoherent and generic. Then Ω satisfies the sample assumptions of Lemma 1 with p as in Theorem 1, so if $\mathbf{Y}$ were sampled in Ω, all assumptions of Lemma 1 would be satisfied. Whence, with probability at least $1 - \frac{\epsilon}{3K}$, $\mathbf{Y}_\Omega$ is uniquely completable, and so is $\mathbf{X}_\Omega$ by Lemma 2, as both $\mathbf{X}$ and $\mathbf{Y}$ are generic. □

3.3. Probability of a Validating Set

The following Lemmas bound the probability of having a validating set in $\bar{\mathbf{X}}^{[k]}_\Omega$.

Lemma 4. *If $|\bar{\mathbf{X}}^{[k]}_\Omega| \geq 2M$, with M as in Theorem 1, then $\bar{\mathbf{X}}^{[k]}_\Omega$ has at least M columns with more than r entries each, with probability at least $1 - \frac{\epsilon}{8K}$.*

Proof. First, notice that for any $x_\omega \in \bar{\mathbf{X}}^{[k]}_\Omega$, $\mathsf{E}[|\omega|] = dp$ and $|\omega| = \sum_{j=1}^{d} \mathbb{1}_{\{j\in\omega\}}$, so using again the multiplicative form of the Chernoff bound we have

$$\begin{aligned}\mathsf{P}(|\omega| \le r) &\le \mathsf{P}(|\omega| \le (1-\beta)\mathsf{E}[|\omega|]) \\ &\le e^{-\frac{\beta^2}{2}\mathsf{E}[|\omega|]} \le e^{-\frac{1}{8}dp} \le 1 - 4\left(\frac{r+1}{de}\right)^{r+1}.\end{aligned}$$

by taking $\beta = 1/2$ and substituting dp.

Then define $\bar{\mathbf{X}}^{[k]}_{\Omega*}$ as the subset of $\bar{\mathbf{X}}^{[k]}_\Omega$ which columns have more than r entries, i.e.

$$\bar{\mathbf{X}}^{[k]}_{\Omega*} = \{x_\omega \in \bar{\mathbf{X}}^{[k]}_\Omega : |\omega| > r\}.$$

If we let $p^* := 1 - \mathsf{P}(|\omega| \le r)$, by our previous argument, $\mathsf{E}[|\bar{\mathbf{X}}^{[k]}_{\Omega*}|] = p^*|\bar{\mathbf{X}}^{[k]}_\Omega|$, and since $|\bar{\mathbf{X}}^{[k]}_{\Omega*}| = \sum_{i=1}^{2M} \mathbb{1}_{\{|\omega_i|\ge r\}}$, by the multiplicative form of the Chernoff bound we obtain

$$\begin{aligned}\mathsf{P}\left(|\bar{\mathbf{X}}^{[k]}_{\Omega*}| \le M\right) &= \mathsf{P}\left(|\bar{\mathbf{X}}^{[k]}_{\Omega*}| \le (1-\beta)\mathsf{E}[|\bar{\mathbf{X}}^{[k]}_{\Omega*}|]\right) \\ &\le e^{-\frac{\beta^2}{2}\mathsf{E}[|\bar{\mathbf{X}}^{[k]}_{\Omega*}|]} \le e^{-\frac{1}{4}p^*M} \le \frac{\epsilon}{8K}.\end{aligned}$$

by taking $\beta = \frac{1}{2}$ and substituting p^* and our choice of M. □

Lemma 5 (Coupons Collector)**.** *Consider a collection of columns $\mathbf{Y}_\Omega$ with $|\mathbf{Y}_\Omega| = M$ as in Theorem 1, and whose columns (sampled uniformly and independently at random) all have exactly $r+1$ entries. Then with probability at least $1 - \frac{\epsilon}{8K}$, the columns of $\mathbf{Y}_\Omega$ have all the different $\binom{d}{r+1}$ observation sets (of size $r+1$).*

Proof. This is a simple application of the well known Coupons Collector problem. We give a proof for completeness. First notice that there are $\binom{d}{r+1} \le \left(\frac{de}{r+1}\right)^{r+1}$ different observation sets (coupons). Let Z_j be the event that none of the first M columns in $\mathbf{Y}_\Omega$ have the j^{th} observation set. It is easy to see that

$$\mathsf{P}(Z_j) = \left(1 - \frac{1}{\binom{d}{r+1}}\right)^M \le \left(1 - \frac{1}{\left(\frac{de}{r+1}\right)^{r+1}}\right)^M \le e^{-M\left(\frac{r+1}{de}\right)^{r+1}}$$

Union bounding over these events and substituting M we obtain the Lemma:

$$\mathsf{P}\left(\bigcup_{j=1}^{\binom{d}{r+1}} Z_j\right) \le \binom{d}{r+1}\mathsf{P}(Z_j) \le \left(\frac{de}{r+1}\right)^{r+1} e^{-M\left(\frac{r+1}{de}\right)^{r+1}} \le \frac{\epsilon}{8K}.$$

□

Lemma 6. *Assume $|\bar{\mathbf{X}}^{[k]}_\Omega| \ge 2M$, with M as in Theorem 1. Then with probability at least $1 - \frac{\epsilon}{4K}$ $\bar{\mathbf{X}}^{[k]}_\Omega$ has at least one validation set.*

Proof. Consider a matrix $\mathbf{Y}_\Omega$ as in Lemma 5. It is clear that the probability that $\bar{\mathbf{X}}^{[k]}_{\Omega*}$ contains a validating set is larger than the probability that $\mathbf{Y}_\Omega$ does. And the latter is larger than the probability that the columns of $\mathbf{Y}_\Omega$ have all the different $\binom{d}{r+1}$ observation sets (of size $r+1$). By Lemma 5, this probability is at least $1 - \frac{\epsilon}{8K}$. And by Lemma 4 the probability that $\bar{\mathbf{X}}^{[k]}_\Omega$ has at least M columns with more than r entries is at least $1 - \frac{\epsilon}{8K}$. A simple use of the Law of Total Probability gives the desired result. □

3.4. Proof of Main Result

First a Lemma to bound the probability of having sufficient columns of each subspace.

Lemma 7. *Let the conditions of Theorem 1 hold. Then $|\widetilde{\mathbf{X}}^{[k]}_\Omega| > d$ with probability at least $1 - \frac{\epsilon}{6K}$, and $|\bar{\mathbf{X}}^{[k]}_\Omega| > 2M$ with probability at least $1 - \frac{\epsilon}{4K}$.*

Proof. Remember that $\widetilde{N} = 2d/\rho_*$ and $\bar{N} = 4M/\rho_*$.

Notice that $|\widetilde{\mathbf{X}}^{[k]}_\Omega| = \sum_i^{\widetilde{N}} \mathbb{1}_{\{x_i\in S_k\}}$. By **A2** $\mathsf{P}(x \in S_k) = \rho_k$, so $\mathsf{E}[|\widetilde{\mathbf{X}}^{[k]}_\Omega|] = \widetilde{N}\rho_k$. So by the Multiplicative form of the Chernoff bound we have

$$\begin{aligned}\mathsf{P}(|\widetilde{\mathbf{X}}^{[k]}_\Omega| \le d) &\le \mathsf{P}\left(|\widetilde{\mathbf{X}}^{[k]}_\Omega| \le (1-\beta)\mathsf{E}\left[|\widetilde{\mathbf{X}}^{[k]}_\Omega|\right]\right) \\ &\le e^{-\frac{\beta^2}{2}\mathsf{E}\left[|\widetilde{\mathbf{X}}^{[k]}_\Omega|\right]} = e^{-\frac{\widetilde{N}\rho_k}{8}} \le e^{-\frac{\widetilde{N}\rho_*}{8}} \le \frac{\epsilon}{6K},\end{aligned}$$

by taking $\beta = \frac{1}{2}$ and noticing that $\mathsf{E}[|\widetilde{\mathbf{X}}^{[k]}_\Omega|] = \widetilde{N}\rho_k > \widetilde{N}\rho_* = 2d$ and $d \ge 4\log(\frac{6K}{\epsilon})$.

Similarly, $\mathsf{P}(|\bar{\mathbf{X}}^{[k]}_\Omega| \le 2M) \le e^{-\frac{M}{2}} \le \frac{\epsilon}{4K}$, as $M \ge 2\log(\frac{4K}{\epsilon})$ □

Lemmas 3, 6 and 7 give us Lemma 8, which shows that with high probability the true subspaces will be contained in $\hat{\mathcal{S}}$.

Lemma 8 (True Positive)**.** *Let* **A1-A3** *hold. Suppose N, M, p and β_0 are as in Theorem 1. Then $S_k \in \hat{\mathcal{S}}$ with probability at least $1 - \frac{\epsilon}{K}$ for every k.*

Proof. Given that $\widetilde{\mathbf{X}}_\Omega$ has at least d columns from S_k, there will be at least one ℓ for which $\widetilde{\mathbf{X}}^{(\ell)}_\Omega \subset \widetilde{\mathbf{X}}^{[k]}_\Omega$ deterministically. Whence, all the assumptions of Lemma 3 are satisfied for $\widetilde{\mathbf{X}}^{(\ell)}_\Omega$, and therefore we know it is uniquely completable with probability at least $1 - \frac{\epsilon}{3K}$. Equivalently, $\widetilde{S}_\ell$ will be the unique r-dimensional subspace that will fit $\widetilde{\mathbf{X}}^{(\ell)}_\Omega$, and equal to the true subspace S_k for some k.

Similarly, given $|\bar{\mathbf{X}}^{[k]}_\Omega| \ge M$, by Lemma 6 $\bar{\mathbf{X}}^{[k]}_\Omega$ will contain a validating set with probability at least $1 - \frac{\epsilon}{4K}$ (which $\widetilde{S}_\ell = S_k$ will obviously fit).

Using the Law of Total Probability on $\mathsf{P}(S_k \notin \hat{\mathcal{S}})$ conditioning on the events that $|\widetilde{\mathbf{X}}^{[k]}_\Omega| \ge d$ and $|\bar{\mathbf{X}}^{[k]}_\Omega| \ge M$, which have low probability according to Lemma 7, and remembering that $\mathsf{P}(\cdot) \leq 1$, we obtain the desired result

$$\begin{aligned}\mathsf{P}(S_k \notin \hat{\mathcal{S}}) \leq\ & \mathsf{P}\left(\widetilde{\mathbf{X}}_\Omega^{(\ell)} \subset \widetilde{\mathbf{X}}_\Omega^{[k]} \text{ is uniquely completable} \middle| |\widetilde{\mathbf{X}}_\Omega^{[k]}| \geq d\right) \\ & + \mathsf{P}\left(\bar{\mathbf{X}}_\Omega^{[k]} \text{ contains a validating set} \middle| |\bar{\mathbf{X}}_\Omega^{[k]}| \geq M\right) \\ & + \mathsf{P}\left(|\widetilde{\mathbf{X}}_\Omega^{[k]}| < d\right) + \mathsf{P}\left(|\bar{\mathbf{X}}_\Omega^{[k]}| < k\right) \\ & \leq \frac{\epsilon}{3K} + \frac{\epsilon}{6K} + \frac{\epsilon}{4K} + \frac{\epsilon}{4K} = \frac{\epsilon}{K}.\end{aligned}$$

□

Before we present the proof of the main theorem, we state the following Lemmas that prove that no subspace other than the true ones will be contained in $\hat{\mathcal{S}}$.

Lemma 9. *Let $x_{\omega_x}, y_{\omega_y} \in \bar{\mathbf{X}}_\Omega$, with $|\omega_x \cap \omega_y| > r$. Suppose $\widetilde{S}$ fits x_{ω_x} and y_{ω_y}. Then x_{ω_x} and y_{ω_y} belong to the same subspace, say S_k, a.s. Furthermore, letting $\omega = \omega_x \cup \omega_y$, $\widetilde{S}_\omega = S_{k_\omega}$ a.s.*

Proof. Since $|\omega_x| \geq |\omega_x \cap \omega_y| > r$, and all columns in $\bar{\mathbf{X}}_\Omega$ are generic w.r.t. the columns in $\widetilde{\mathbf{X}}_\Omega$ that produced $\widetilde{S}$, x_{ω_x} can only lie in $\widetilde{S}$ iff $\widetilde{S}_{\omega_x} = S_{k_{\omega_x}}$, where k is the subspace that x belongs to.

Similarly, $|\omega_y| > r$ and $y_{\omega_y} \in \widetilde{S}$ imply $\widetilde{S}_{\omega_y} = S_{k'_{\omega_y}}$, where k' is the subspace y belongs to.

Furthermore, since $|\omega_x \cap \omega_y| > r$ and both x_{ω_x} and y_{ω_y} lie in $\widetilde{S}$, k and k' must be the same and also $\widetilde{S}_\omega = S_{k_\omega}$. □

Lemma 10. *If $\widetilde{S}$ fits a validating set, then $\widetilde{S} = S_k$ for some k a.s.*

Proof. Let $\{x_{i_{\omega_i}}\}_{i=1}^m$ be a validating set. Then by induction on Lemma 9, $\widetilde{S}_\omega = S_{k_\omega}$ for some k, where $\omega = \cup_i^m \omega_i$. But since $\omega = \{1, ..., d\}$ by the very definition of a validating set, we conclude that $\widetilde{S} = S_k$. □

We now present the proof of our main result, Theorem 1.

Proof. (**Theorem 1**) It suffices to show that with high probability $\hat{\mathcal{S}} = \mathcal{S}$. Write

$$\begin{aligned}\mathsf{P}\left(\mathcal{S} \neq \hat{\mathcal{S}}\right) &= \mathsf{P}\left(\mathcal{S} \not\subset \hat{\mathcal{S}} \cup \mathcal{S} \not\supset \hat{\mathcal{S}}\right) \\ &\leq \mathsf{P}\left(\bigcup_k S_k \notin \hat{\mathcal{S}} \cup \bigcup_{\ell=1}^{\binom{\widetilde{N}}{L}} S_\ell \in \hat{\mathcal{S}} \backslash \mathcal{S}\right) \\ &\leq \sum_{k=1}^K \mathsf{P}\left(S_k \notin \hat{\mathcal{S}}\right) + \sum_{\ell=1}^{\binom{\widetilde{N}}{L}} \mathsf{P}\left(S_\ell \in \hat{\mathcal{S}} \backslash \mathcal{S}\right)\end{aligned}$$

By Lemma 8, $\mathsf{P}(S_k \notin \hat{\mathcal{S}}) \leq \frac{\epsilon}{K}\ \forall\ k$. Also, notice that $\mathsf{P}(S_\ell \in \hat{\mathcal{S}} \backslash \mathcal{S})$ is equivalent to the probability that S_ℓ fits a validating set given that $S_\ell \neq S_k\ \forall\ k$, and this probability is zero by Lemma 10. Thus $\mathsf{P}\left(\mathcal{S} \neq \hat{\mathcal{S}}\right) \leq \sum_{k=1}^K \frac{\epsilon}{K} + 0 = \epsilon$, as desired. □

4. EM ALGORITHM FOR SUBSPACE CLUSTERING WITH MISSING DATA

The problem of subspace clustering with missing data can be posed as fitting a mixture of Gaussians with low-rank covariances to incomplete data. This naturally suggests considering extensions of the EM algorithm in [13] to handle missing data, or alternatively, a generalization of the EM algorithm in [14] to low-rank covariance matrices. We propose the following EM algorithm for this task, based largely on [13]. To begin we assume the data are contaminated with additive Gaussian noise.

Consider the usual Gaussian mixture framework, and additionally split every $x_i \in \mathbf{X}_\Omega$ into its observed and missing parts:

$$\begin{bmatrix} x_i^o \\ x_i^m \end{bmatrix} = \sum_{k=1}^K \mathbb{1}_{\{z_i=k\}} \left(\begin{bmatrix} W_k^{o_i} \\ W_k^{m_i} \end{bmatrix} y_i + \begin{bmatrix} \mu_k^{o_i} \\ \mu_k^{m_i} \end{bmatrix} + \eta_i \right), \tag{1}$$

where $\{1, ..., K\} \ni z_i \overset{iid}{\sim} \rho \perp y_i \overset{iid}{\sim} \mathcal{N}(0, I)$, W_k is a $d \times r$ matrix whose span is S_k, and $\eta_i | z_i \overset{iid}{\sim} \mathcal{N}(0, \sigma_{z_i}^2 I)$ models the noise in the z_i^{th} subspace. We are interested on the Maximum Likelihood Estimate (MLE) of $\theta = \{W, \mu, \rho, \sigma^2\}$, where $W := \{W_k\}_{k=1}^K$, $\mu := \{\mu_k\}_{k=1}^K$, ρ and $\sigma^2 := \{\sigma_k^2\}_{k=1}^K$.

Let $\mathbf{X}^o := \{x_i^o\}_{i=1}^N$, $\mathbf{X}^m := \{x_i^m\}_{i=1}^N$, $\mathbf{Y} := \{y_i\}_{i=1}^N$, $\mathbf{Z} := \{z_i\}_{i=1}^N$ s.t. $\mathbf{X}^o$ is our data, θ is the parameter of interest, and $\mathbf{X}^m$, $\mathbf{Y}$ and $\mathbf{Z}$ are the *hidden* variables in the EM algorithm. The iterates of the algorithm are computed as follows, where $\mathsf{E}_k \langle \cdot \rangle$ denotes $\mathsf{E}_{\cdot | x_i^o, z_i = k, \hat{\theta}}[\cdot]$.

$$\begin{aligned}\widetilde{W}_k = & \left[\sum_{i=1}^N \mathsf{p}_{i,k} \mathsf{E}_k \left\langle x_i y_i^T \right\rangle - \frac{\left(\sum_{i=1}^N \mathsf{p}_{i,k} \mathsf{E}_k \langle x_i \rangle\right)\left(\sum_{i=1}^N \mathsf{p}_{i,k} \mathsf{E}_k \langle y_i \rangle^T\right)}{\sum_{i=1}^N \mathsf{p}_{i,k}} \right] \\ & \left[\sum_{i=1}^N \mathsf{p}_{i,k} \mathsf{E}_k \left\langle y_i y_i^T \right\rangle - \frac{\left(\sum_{i=1}^N \mathsf{p}_{i,k} \mathsf{E}_k \langle y_i \rangle\right)\left(\sum_{i=1}^N \mathsf{p}_{i,k} \mathsf{E}_k \langle y_i \rangle^T\right)}{\sum_{i=1}^N \mathsf{p}_{i,k}} \right]^{-1},\end{aligned} \tag{2}$$

$$\widetilde{\mu}_k = \frac{\sum_{i=1}^N \mathsf{p}_{i,k} \left(\mathsf{E}_k \langle x_i \rangle - \widetilde{W}_k \mathsf{E}_k \langle y_i \rangle \right)}{\sum_{i=1}^N \mathsf{p}_{i,k}}, \tag{3}$$

$$\begin{aligned}\widetilde{\sigma}_k^2 = & \frac{1}{d \sum_{i=1}^N \mathsf{p}_{i,k}} \left[\sum_{i=1}^N \mathsf{p}_{i,k} \left(tr\left(\mathsf{E}_k \left\langle x_i x_i^T \right\rangle\right) - 2\widetilde{\mu}_k^T \mathsf{E}_k \langle x_i \rangle + \widetilde{\mu}_k^T \widetilde{\mu}_k \right.\right. \\ & \left.\left. -2tr\left(\mathsf{E}_k \left\langle y_i x_i^T \right\rangle \widetilde{W}_k\right) + 2\widetilde{\mu}_k^T \widetilde{W}_k \mathsf{E}_k \langle y_i \rangle + tr\left(\mathsf{E}_k \left\langle y_i y_i^T \right\rangle \widetilde{W}_k^T \widetilde{W}_k\right) \right) \right],\end{aligned} \tag{4}$$

$$\widetilde{\rho}_k = \frac{1}{N} \sum_{i=1}^N \mathsf{p}_{i,k}, \qquad \mathsf{p}_{i,k} := \mathsf{P}_{z_i | x_i^o, \hat{\theta}}(k) = \frac{\hat{\rho}_k \mathsf{P}_{x_i^o | z_i = k, \hat{\theta}}(x_i^o)}{\sum_{j=1}^K \hat{\rho}_j \mathsf{P}_{x_i^o | z_i = j, \hat{\theta}}(x_i^o)}. \tag{5}$$

The expectations in (2) - (5) are easily derived from the following conditional distribution, where $M_k := \sigma^2 I +$

$W_k^{o_iT} W_k^{o_i}$,

$$\begin{bmatrix} x_i^m \\ y_i \end{bmatrix} \Bigg\| x_i^o, z_i = k, \theta \sim \mathcal{N}\left(\begin{bmatrix} \mu_k^{m_i} + W_k^{m_i} M_k^{-1} W_k^{o_iT}(x_i^o - \mu_k^{o_i}) \\ M_k^{-1} W_k^{o_iT}(x_i^o - \mu_k^{o_i}) \end{bmatrix}, \right.$$
$$\left. \sigma^2 \begin{bmatrix} I + W_k^{m_i} M_k^{-1} W_k^{m_iT} & W_k^{m_i} M_k^{-1} \\ M_k^{-1} W_k^{m_iT} & M_k^{-1} \end{bmatrix} \right). \tag{6}$$

Notice that in the noiseless case we can no longer compute (5), as $(\widetilde{W}_k^{o_i} \widetilde{W}_k^{o_iT})$ is not invertible. Nevertheless, viewing the noiseless case as the limit as $\sigma^2 \to 0$, we can compute $\mathsf{p}_{i,k}$ for a fixed arbitrarily small σ^2 and with it we can find estimates under such σ^2, say $\hat{W}_{\sigma^2}$ and $\hat{\rho}_{\sigma^2}$, s.t. the noiseless estimates are given by $\hat{W} = \lim_{\sigma^2 \to 0} \hat{W}_{\sigma^2}$ and $\hat{\rho} = \lim_{\sigma^2 \to 0} \hat{\rho}_{\sigma^2}$. In other words, we can estimate W and ρ in the noiseless case with arbitrary precision by letting σ^2 be arbitrarily small. This issue arises only when computing $\mathsf{p}_{i,k}$; all the other desired expectations can be evaluated directly by substituting $\sigma^2 = 0$ in (6), and since $\mathsf{p}_{i,k}$ converges to $\mathbb{1}_{\{z_i = k\}} \mathbb{1}_{\{\hat{W}_k = W_k\}}$ as $\sigma^2 \to 0$, we can also do a hard assignment at the end of EM to improve precision.

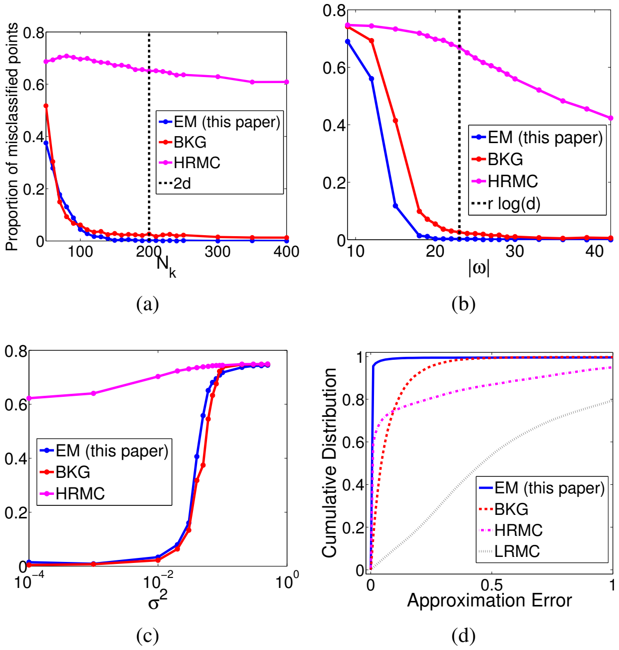
The computations of the expected means and covariances have the highest computational complexity in the noiseless and noisy case respectively, with $|\omega_i^c| r$ and $|\omega_i^c|^2 r$ operations per column per subspace per iteration. Since $|\omega_i^c|$ is close to and upper bounded by d, the computational complexity of the EM algorithm per iteration will be in the order of $NKdr$ and NKd^2r in the noiseless and noisy cases, respectively. In conclusion, noise in our measurements increases the computational complexity of the algorithm by one order of magnitude.

5. SIMULATIONS

The first experiment we present is a set of simulations of the EM setup above with $d = 100$, $K = 4$, $r = 5$. For each simulation we generated K subspaces and K initial estimates, each spanned by an orthonormal basis generated from r i.i.d. standard gaussian d-dimensional vectors — known to be highly incoherent — and N_k columns from each subspace with $|\omega|$ observed entries each. We evaluated the performance of the EM algorithm derived before, batch k-GROUSE (BKG) [5] and the HRMC algorithm from [4]. We ran 450 independent trials of this experiment as a function of N_k, $|\omega|$ and σ^2. The results are summarized in Figure 1 (a)-(c).

For a second simulation, we consider an application in which unions of subspaces are indeed a good model for data. Distances in a network measured in number of hop counts between passive monitors and computers determine the network's topology [16]. As measurements in such monitors is not controlled, not all distances can be observed. Fortunately, these distances lie in a union of K 2-dimensional subspaces with K being the number of subnets [3].

With this in mind we simulated a network and measured hop counts based on shortest-path routing using a Heuristically Optimal Topology from [17] with $d = 75$ passive monitors randomly located and $N_k = 225$ IP addresses from each of the $K = 12$ subnets. In Figure 1(d) we compare the results of the hop count matrix estimation from only 40% of the total hop counts using EM, BKG, HRMC and LRMC.



(a)

(b)

(c)

(d)

Fig. 1. (a) - (c): Proportion of misclassified points, (a) as a function of N_k with $|\omega| = \lceil r \log d \rceil = 24$ fixed; (b) as a function of $|\omega|$ with $N_k = 2.1d$ fixed; (c) as a function of σ^2 with $N_k = 3d$ and $|\omega| = \lceil r \log d \rceil$ fixed. (d) Cumulative Distribution of hop count estimation error for $K = 12$ subnets, $d = 75$ passive monitors and $N = 2700$ IP addresses from 40% of total observations.

6. REAL DATA - COMPUTER VISION: HOPKINS DATASET

Finally, we tested our EM algorithm using real data from the Hopkins 155 Motion Segmentation Dataset [18]. In each video of this dataset, a collection of points are identified over the frames. Each point belongs to an unknown cluster, e.g. a car, a person, background, etc., and the positions of these points are known to lie in a union of subspaces [8]. However, in real life it is unusual to be able to identify *every* point over *all* the frames of a video, due to occlusion, objects leaving the video window, objects rotation, miss detection, etc. Therefore missing data arises naturally.

Table 1 shows a summary of EM's performance on this dataset, where we synthetically removed data uniformly at random.

Comparing to the related table in [8] we see that our EM algorithm performs about average of the algorithms in mean,

Table 1. Classification Errors (in %) of EM on the Hopkins 155 Motion Segmentation Dataset

Two Motions

$\lvert\omega\rvert(\%)$	Check.(78)		Traffic (31)		Articul. (11)	
	Mean	Median	Mean	Median	Mean	Median
100	4.3	0	0.1	0	0.5	0
70	3.6	0.5	0.9	0	4.6	0
50	3.2	0.3	1.3	0	2.4	0
30	5.8	0.9	3.4	0.4	2.4	0

Three Motions

$\lvert\omega\rvert(\%)$	Check.(78)		Traffic (31)		Articul. (11)	
	Mean	Median	Mean	Median	Mean	Median
100	16.9	17.9	1.3	0.4	0	0
70	16.2	17.9	10.5	6.1	9.0	9.0
50	17.4	17.9	10.7	8.9	21.7	21.7
30	25.4	25.6	19.6	13.0	22.4	22.4

but its performance is nearly as good in median, even with missing data, as the best algorithms with full data. We can interpret that to mean that there are a few datasets where EM does very poorly. In these datasets, there may be overlapping subspaces or ill conditioned data, which would be problematic for any algorithm.

7. CONCLUSION

In this paper we showed that only $\mathscr{O}(Kd^{r+1})$ columns are sufficient to guarantee a unique solution for subspace clustering with missing data, as opposed to $\mathscr{O}(d^{\log d})$ from previous existing bounds. A powerful conclusion of our theory is that *if* there is an algorithm that finds a set of K low-dimensional subspaces that fit independent generic validation sets, then that solution is the true $\mathcal{S}$ a.s. Furthermore, we presented a novel EM-type algorithm that in practice performs very well even with fewer columns than theoretically derived, suggesting that our bound is over sufficient. The true sample complexity of this problem, conjectured to be $\mathscr{O}(Kd)$, and a practical algorithm that provably solves it under such conditions, remain important open questions that our immediate future work will aim to answer.

8. REFERENCES

[1] Joao Paulo Costeira and Takeo Kanade, "A multibody factorization method for independently moving objects," *International Journal of Computer Vision*, vol. 29, 1998.

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

[3] Brian Eriksson, Paul Barford, Joel Sommers, and Robert Nowak, "DomainImpute: Inferring Unseen Components in the Internet," in *Proceedings of IEEE INFOCOM Mini-Conference*, April 2011, pp. 171–175.

[4] Brian Eriksson, Laura Balzano, and Robert Nowak, "High-Rank Matrix Completion and Subspace Clustering with Missing Data," in *Proceedings of the Conference on Artificial Intelligence and Statistics (AI Stats)*, 2012.

[5] Laura Balzano, Robert Nowak, Arthur Szlam, and Benjamin Recht, "*k*-Subspaces with missing data," in *Proceedings of the Statistical Signal Processing Workshop*, 2012.

[6] Ehsan Elhamifar and René Vidal, "Sparse subspace clustering: Algorithm, theory, and applications," *CoRR*, vol. abs/1203.1005, 2012.

[7] S. Gunnemann, E. Muller, S. Raubach, and T. Seidl, "Flexible fault tolerant subspace clustering for data with missing values," in *Data Mining (ICDM), 2011 IEEE 11th International Conference on*, 2011, pp. 231–240.

[8] René Vidal, "Subspace clustering," *IEEE Signal Processing Magazine*, 2010.

[9] Teng Zhang, Arthur Szlam, Yi Wang, and Gilad Lerman, "Randomized hybrid linear modeling by local best fit flats," in *The Twenty-Third IEEE Conference on Computer Vision and Pattern Recognition*, San Francisco, CA, June 2010.

[10] Mahdi Soltanolkotabi, Ehsan Elhamifar, and Emmanuel Candès, "Robust subspace clustering," 2013, Available at `http://arxiv.org/abs/1002.0852`.

[11] Benjamin Recht, "A simpler approach to matrix completion," *Jrnl. of Machine Learning Rsrch.*, vol. 12, pp. 3413–3430, 2011.

[12] Paul S. Bradley and Olvi L. Mangasarian, "*k*-Plane clustering," *Journal of Global Optimization*, vol. 16, pp. 23–32, 2000.

[13] Michael E. Tipping and Christopher M. Bishop, "Mixtures of probabilistic principal component analysers," *Neural Computation*, vol. 11, no. 2, pp. 443–482, 1999.

[14] Zoubin Ghahramani and Michael I. Jordan, "Supervised learning from incomplete data via an em approach," in *Advances in Neural Information Processing Systems 6*. 1994, pp. 120–127, Morgan Kaufmann.

[15] Franz Király and Ryota Tomioka, "A combinatorial algebraic approach for the identifiability of low-rank matrix completion," in *Proceedings of the 29th International Conference on Machine Learning*, 2012.

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

[17] 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.

[18] Johns Hopkins Vision Lab, ," Hopkins 155 video dataset available at `http://www.vision.jhu.edu/data/hopkins155/`.