The Information-Theoretic Requirements of Subspace Clustering with Missing Data

Daniel L. Pimentel-Alarcón PIMENTELALAR@WISC.EDU
Robert D. Nowak NOWAK@ECE.WISC.EDU
University of Wisconsin - Madison, 53706 USA

Abstract

Subspace clustering with missing data (SCMD) is a useful tool for analyzing incomplete datasets. Let d be the ambient dimension, and r the dimension of the subspaces. Existing theory shows that $N_k = \mathcal{O}(rd)$ columns per subspace are necessary for SCMD, and $N_k = \mathcal{O}(\min\{d^{\log d}, d^{r+1}\})$ are sufficient. We close this gap, showing that $N_k = \mathcal{O}(rd)$ is also sufficient. To do this we derive deterministic sampling conditions for SCMD, which give precise information-theoretic requirements and determine sampling regimes. These results explain the performance of SCMD algorithms from the literature. Finally, we give a practical algorithm to certify the output of *any* SCMD method deterministically.

1. Introduction

Let $\mathcal{U}^\star$ be a collection of r-dimensional subspaces of $\mathbb{R}^d$, and let $\mathbf{X}$ be a $d \times N$ data matrix whose columns lie in the union of the subspaces in $\mathcal{U}^\star$. The goal of subspace clustering is to infer $\mathcal{U}^\star$ and cluster the columns of $\mathbf{X}$ according to the subspaces (Vidal, 2011; Elhamifar & Vidal, 2009; 2013; Liu et al., 2010; 2013; Wang & Xu, 2013; Soltanolkotabi et al., 2014; Hu et al., 2015; Qu & Xu, 2015; Peng et al., 2015; Wang et al., 2015). There is growing interest in subspace clustering with missing data (SCMD), where one aims at the same goal, but only observes a subset of the entries in $\mathbf{X}$. This scenario arises in many practical applications, such as computer vision (Kanatani, 2001), network inference and monitoring (Eriksson et al., 2012; Mateos & Rajawat, 2013), and recommender systems (Rennie & Srebro, 2005; Zhang et al., 2012).

There is a tradeoff between the number of samples per column ℓ, and the number of columns per subspace N_k, required for SCMD. If all entries are observed, $N_k = r + 1$ is necessary and sufficient (assuming generic columns). If $\ell = r + 1$ (the minimum required), it is easy to see that $N_k = \mathcal{O}(rd)$ is necessary for SCMD, as $\mathcal{O}(rd)$ columns are necessary for low-rank matrix completion (LRMC) (Candès & Recht, 2009), which is the particular case of SCMD with only one subspace. Under standard random sampling schemes, i.e., with $\ell = \mathcal{O}(\max\{r, \log d\})$, it is known that $N_k = \mathcal{O}(\min\{d^{\log d}, d^{r+1}\})$ is sufficient (Eriksson et al., 2012; Pimentel-Alarcón et al., 2014). This number of samples can be very large, and it is unusual to encounter such huge datasets in practice. Recent work has produced several heuristic algorithms that tend to work reasonably well in practice without these strong requirements. Yet the sample complexity of SCMD remained an important open question until now (Soltanolkotabi, 2014).

Organization and Main Contributions

In Section 2 we formally state the problem and our main result, showing that $N_k = \mathcal{O}(rd)$ is the true sample complexity of SCMD. In Section 3 we present deterministic sampling conditions for SCMD, similar to those in (Pimentel-Alarcón et al., 2015a) for LRMC. These specify precise information-theoretic requirements and determine sampling regimes of SCMD. In Section 3 we also present experiments showing that our theory accurately predicts the performance of SCMD algorithms from the literature.

The main difficulty of SCMD is that the pattern of missing data can cause that $\mathcal{U}^\star$ is not the only collection of subspaces that agrees with the observed data. This implies that in general, even with unlimited computational power, one could try all possible clusterings, and still be unable to determine the right one. Existing theory circumvents this problem by requiring a large number of columns, so that $\mathcal{U}^\star$ is be the only collection of subspaces that agrees with the vast number of observations. In Section 4 we discuss these issues and present an efficient criteria to determine whether a subspace is indeed one of the subspaces in $\mathcal{U}^\star$.

In Section 5 we present the main practical implication of our theory: an efficient algorithm to certify the output of

Proceedings of the 33^{rd} International Conference on Machine Learning, New York, NY, USA, 2016. JMLR: W&CP volume 48.

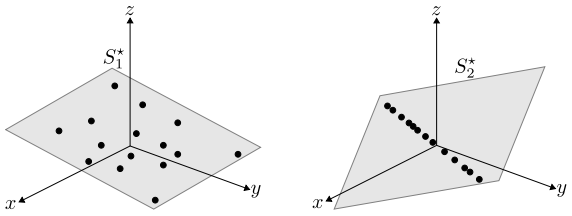

Figure 1. **A1** requires that $\mathcal{U}^\star$ is a generic set of subspaces. Here, $S_1^\star, S_2^\star \in \mathcal{U}^\star$ are 2-dimensional subspaces (planes) in general position. **A2** requires that the columns in $\mathbf{X}^k$ are in general position on $S_k^\star$, as in the left. If we had columns as in the right, all lying in a line (when $S_k^\star$ is a plane), we would be unable to identify $S_k^\star$. Fortunately, these pathological cases have Lebesgue measure 0.

any SCMD method. Our approach is based on a simple idea: one way to verify the output of an algorithm is by splitting the given data into a training set, and a hold-out set. We can use the training set to obtain an estimate $\hat{\mathcal{U}}$ of $\mathcal{U}^\star$. If $\hat{\mathcal{U}}$ does not agree with the hold-out set, we know $\hat{\mathcal{U}}$ is incorrect. What makes SCMD challenging is that depending on the pattern of missing data, $\hat{\mathcal{U}}$ may agree with the hold-out set even if $\hat{\mathcal{U}} \neq \mathcal{U}^\star$. Our validation algorithm uses our results in Section 4 to show that if $\hat{\mathcal{U}}$ agrees with the hold-out set, and the hold-out set satisfies suitable sampling conditions, then $\hat{\mathcal{U}}$ must indeed be equal to $\mathcal{U}^\star$. We prove all our statements in Section 6.

2. Model and Main Result

Let $\mathrm{Gr}(r, \mathbb{R}^d)$ denote the Grassmann manifold of r-dimensional subspaces in $\mathbb{R}^d$. Let $\mathcal{U}^\star := \{S_k^\star\}_{k=1}^K$ be a set of K subspaces in $\mathrm{Gr}(r, \mathbb{R}^d)$. Let $\mathbf{X}$ be a $d \times N$ data matrix whose columns lie in the union of the subspaces in $\mathcal{U}^\star$. Let $\mathbf{X}^k$ denote the matrix with all the columns of $\mathbf{X}$ corresponding to $S_k^\star$. Assume:

- **A1** The subspaces in $\mathcal{U}^\star$ are drawn in independently with respect to the uniform measure over $\mathrm{Gr}(r, \mathbb{R}^d)$.
- **A2** The columns of $\mathbf{X}^k$ are drawn independently according to an absolutely continuous distribution with respect to the Lebesgue measure on $S_k^\star$.
- **A3** $\mathbf{X}^k$ has at least $(r+1)(d-r+1)$ columns.

Assumption **A1** essentially requires that $\mathcal{U}^\star$ is a generic collection of K subspaces in general position (see Figure 1 to build some intuition). Similarly, **A2** requires that the columns in $\mathbf{X}^k$ are in general position on $S_k^\star$. **A1**-**A2** simply discard pathological cases with Lebesgue measure zero, like subspaces perfectly aligned with the canonical axes, or identical columns. Our statements hold with probability (w.p.) 1 with respect to (w.r.t.) the measures in **A1**-**A2**. In contrast, typical results assume bounded coherence, which essentially discards the set of subspaces (with positive measure) that are somewhat aligned to the canonical axes. Finally, **A3** requires that $N_k = \mathcal{O}(rd)$.

Let $\mathbf{\Omega}$ be a $d \times N$ matrix with binary entries, and $\mathbf{X}_{\mathbf{\Omega}}$ be the incomplete version of $\mathbf{X}$, observed only in the nonzero locations of $\mathbf{\Omega}$. Our main result is presented in the following theorem. It states that if $\mathbf{X}$ has at least $\mathcal{O}(rd)$ columns per subspace, and is observed on at least $\mathcal{O}(\max\{r, \log d\})$ entries per column, then $\mathcal{U}^\star$ can be identified with large probability. This shows $N_k = \mathcal{O}(rd)$ to be the true sample complexity of SCMD. The proof is given in Section 6.

Theorem 1. *Assume **A1**-**A3** hold $\forall k$, with $r \leq \frac{d}{6}$. Let $\epsilon > 0$ be given. Suppose that each column of $\mathbf{X}$ is observed on at least ℓ locations, distributed uniformly at random, and independently across columns, with*

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

Then $\mathcal{U}^\star$ can be uniquely identified from $\mathbf{X}_{\mathbf{\Omega}}$ with probability at least $1 - K(r+1)\epsilon$.

Theorem 1 follows by showing that our deterministic sampling conditions (presented in Theorem 2, below) are satisfied with high probability (w.h.p.).

3. Deterministic Conditions for SCMD

In this section we present our deterministic sampling conditions for SCMD. To build some intuition, consider the complete data case. Under **A2**, subspace clustering (with complete data) is possible as long as we have $r+1$ *complete* columns per subspace. To see this, notice that if columns are in general position on $\mathcal{U}^\star$, then any r columns will be linearly independent w.p. 1, and will thus define an r-dimensional *candidate* subspace S. This subspace may or may not be one of the subspaces in $\mathcal{U}^\star$. For example, if the r selected columns come from different subspaces, then S will be none of the subspaces in $\mathcal{U}^\star$ w.p. 1. Fortunately, an $(r+1)^{\text{th}}$ column will lie in S if and only if the $r+1$ selected columns come from the same subspace in $\mathcal{U}^\star$, whence S is such subspace. Therefore, we can identify $\mathcal{U}^\star$ by trying all combinations of $r+1$ columns, using the first r to define a candidate subspace S and the last one to verify whether S is one of the subspaces in $\mathcal{U}^\star$.

When handling incomplete data, we may not have complete columns. Theorem 2 states that we can use a set of $d-r+1$ incomplete columns *observed in the right places* in lieu of *one* complete column, such that, with the same approach as before, we can use r sets of incomplete columns to define a candidate subspace S, and an additional set of incomplete columns to verify whether $S \in \mathcal{U}^\star$. To state precisely what we mean by *observed in the right places*, we introduce the constraint matrix $\breve{\mathbf{\Omega}}$ that encodes the information of the sampling matrix $\mathbf{\Omega}$ in a way that allows us to easily express our results.

Definition 1 (Constraint Matrix)**.** *Let $m_{1,i}, m_{2,i} \ldots, m_{\ell_i,i}$ denote the indices of the ℓ_i observed entries in the i^{th} column of $\mathbf{X}$. Define $\mathbf{\Omega}_i$ as the $d \times (\ell_i - r)$ matrix, whose j^{th} column has the value 1 in rows $m_{j,i}, m_{j+1,i}, \ldots, m_{j+r,i}$, and zeros elsewhere. Define the* constraint matrix *as $\breve{\mathbf{\Omega}} := [\mathbf{\Omega}_1 \ \cdots \ \mathbf{\Omega}_N]$.*

Example 1. *Suppose $m_{1,i} = 1, m_{2,i} = 2, \ldots, m_{\ell_i,i} = \ell_i$, then*

$$\mathbf{\Omega}_i = \begin{bmatrix} \mathbf{0} & \\ \mathbf{1} & \\ \mathbf{0} & \\ \hline & \\ \mathbf{0} & \end{bmatrix} \begin{matrix} \left.\vphantom{\begin{matrix}\mathbf{0}\\\mathbf{1}\\\mathbf{0}\end{matrix}}\right\} \ell_i - r - 1 \\ \left.\vphantom{\begin{matrix}\mathbf{0}\\\mathbf{1}\\\mathbf{0}\end{matrix}}\right\} r+1 \\ \left.\vphantom{\begin{matrix}\mathbf{0}\\\mathbf{1}\\\mathbf{0}\end{matrix}}\right\} d - \ell_i, \end{matrix}$$
$$\underbrace{\qquad\qquad}_{\ell_i - r}$$

where $\mathbf{1}$ and $\mathbf{0}$ denote blocks of all 1's and all 0's.

Notice that each column of $\breve{\mathbf{\Omega}}$ has exactly $r+1$ nonzero entries. The key insight behind this construction is that observing more than r entries in a column of $\mathbf{X}$ places constraints on what $\mathcal{U}^\star$ may be. 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. Each column of $\breve{\mathbf{\Omega}}$ encodes one of these constraints.

Our next main contribution is Theorem 2. It gives a deterministic condition on $\mathbf{\Omega}$ to guarantee that $\mathcal{U}^\star$ can be identified from the constraints produced by the observed entries. Define $\mathbf{\Omega}^k$ as the matrix formed with the columns of $\breve{\mathbf{\Omega}}$ corresponding to the k^{th} subspace. Given a matrix, let $n(\cdot)$ denote its number of columns, and $m(\cdot)$ the number of its *nonzero* rows.

Theorem 2. *For every k, assume **A1**-**A2** hold, and that $\mathbf{\Omega}^k$ contains disjoint matrices $\{\mathbf{\Omega}_\tau\}_{\tau=1}^{r+1}$, each of size $d \times (d-r+1)$, such that for every τ,*

(i) *Every matrix $\mathbf{\Omega}'_\tau$ formed with a* proper *subset of the columns in $\mathbf{\Omega}_\tau$ satisfies*

$$m(\mathbf{\Omega}'_\tau) \geq n(\mathbf{\Omega}'_\tau) + r. \tag{2}$$

Then $\mathcal{U}^\star$ can be uniquely identified from $\mathbf{X}_{\mathbf{\Omega}}$ w.p. 1.

The proof of Theorem 2 is given in Section 6. Analogous to the complete data case, Theorem 2 essentially requires that there are at least $r+1$ sets of incomplete columns *observed in the right places* per subspace. Each set of *observations in the right places* corresponds to a matrix $\mathbf{\Omega}_\tau$ satisfying the conditions of Theorem 2. The first r sets can be used to define a candidate subspace S, and the additional one can be used to verify whether $S \in \mathcal{U}^\star$.

In words, (i) requires that every *proper* subset of n columns of $\mathbf{\Omega}_\tau$ has at least $n + r$ nonzero rows. This condition is tightly related to subspace identifiability (Pimentel-Alarcón et al., 2015b). The main difference is the *proper* subset clause, and the condition that $\mathbf{\Omega}_\tau$ has $d - r + 1$ columns, as opposed to $d - r$. More about this is discussed below. In particular, see condition (ii) and Question (Qa).

Example 2. *The next sampling satisfies* (i).

$$\mathbf{\Omega}_\tau = \begin{matrix} r+1 \\ \\ \\ \\ \end{matrix} \overbrace{\begin{bmatrix} & & \boxed{1} \\ & \mathbf{0} & \\ \mathbf{1} & & \\ \mathbf{0} & & \boxed{1} \end{bmatrix}}^{1} \begin{matrix} \left.\vphantom{\begin{matrix}\mathbf{0}\\\mathbf{1}\\\mathbf{0}\end{matrix}}\right\} 1 \\ \left.\vphantom{\begin{matrix}\mathbf{0}\\\mathbf{1}\\\mathbf{0}\end{matrix}}\right\} d-r-1 \\ \left.\vphantom{\begin{matrix}\mathbf{0}\\\mathbf{1}\\\mathbf{0}\end{matrix}}\right\} r. \end{matrix}$$
$$\underbrace{\qquad\qquad}_{d-r+1}$$

To see that the sufficient condition in Theorem 2 is tight, notice that if columns are only observed on $r+1$ entries (the minimum required), then $r(d-r)$ columns per subspace are necessary for SCMD, as a column with $r+1$ observations eliminates at most one of the $r(d-r)$ degrees of freedom in a subspace (Pimentel-Alarcón et al., 2015a). The sufficient condition in Theorem 2 only requires the slightly larger $(r+1)(d-r+1)$ columns per subspace.

Remark 1. *The constraint matrix $\breve{\mathbf{\Omega}}$ explains the interesting tradeoff between ℓ and N_k. The larger ℓ, the more constraints per column we obtain, and the fewer columns are required. This tradeoff, depicted in Figure 2, determines whether SCMD is possible, and can be appreciated in the experiments of Figure 3. This explains why practical algorithms (such as k-GROUSE (Balzano et al., 2012), EM (Pimentel-Alarcón et al., 2014), SSC-EWZF and MC+SSC (Yang et al., 2015), among others), tend to work with only $N_k = \mathcal{O}(rd)$, as opposed to the strong conditions that existing theory required.*

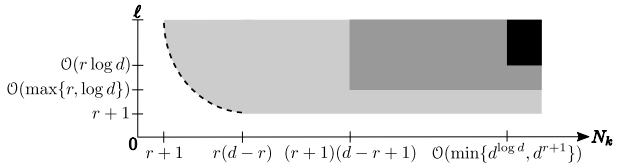

Figure 2. Theoretical sampling regimes of SCMD. In the white region, where the dashed line is given by $\ell = \frac{r(d-r)}{N_k} + r$, it is easy to see that SCMD is impossible by a simple count of the degrees of freedom in a subspace (see Section 3 in (Pimentel-Alarcón et al., 2015b)). In the light-gray region, SCMD is possible provided the entries are *observed in the right places*, e.g., satisfying the conditions of Theorem 2. By Theorem 1, random samplings will satisfy these conditions w.h.p. as long as $N_k \geq (r+1)(d-r+1)$ and $\ell \geq \max\{12(\log(\frac{d}{\epsilon}) + 1),\ 2r\}$, hence w.h.p. SCMD is possible in the dark-grey region. Previous analyses showed that SCMD is possible in the black region (Eriksson et al., 2012; Pimentel-Alarcón et al., 2014), but the rest remained unclear, until now.

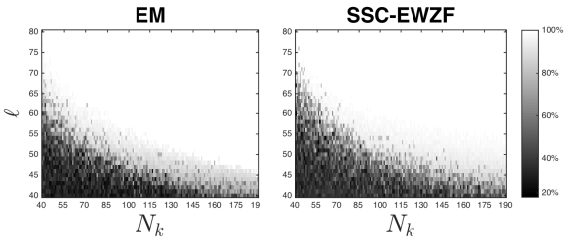

Figure 3. Proportion of correctly classified columns (average over 10 trials) using EM (Pimentel-Alarcón et al., 2014) and SSC-EWZF (Yang et al., 2015) as a function of the number of columns per subspace N_k and the number of observations per column ℓ, with $K = 5$ subspaces of dimension $r = 25$, in ambient dimension $d = 100$. Notice the tradeoff between ℓ and N_k: the smaller N_k, the larger ℓ is required, and vice versa. This tradeoff determines whether SCMD is possible (see Figure 2).

4. What Makes SCMD Hard?

The main difficulty in showing the results above is that depending on the pattern of missing data, there could exist *false* subspaces, that is, subspaces not in $\mathcal{U}^\star$ that agree with arbitrarily many incomplete columns (even if they are observed on identifiable patterns for LRMC). This section gives some insight on this phenomenon, and presents our key result: a deterministic criteria to determine whether a subspace is indeed one of the subspaces in $\mathcal{U}^\star$.

We begin with some terminology. Let $\mathbf{x}$ and $\boldsymbol{\omega}$ denote arbitrary columns of $\mathbf{X}$ and $\boldsymbol{\Omega}$. For any subspace, matrix or vector that is compatible with a binary vector $\boldsymbol{\omega}$, we will use the subscript $\boldsymbol{\omega}$ to denote its restriction to the nonzero coordinates/rows in $\boldsymbol{\omega}$. Let $\mathbf{X}'_{\boldsymbol{\Omega}'}$ be a matrix formed with a subset of the columns in $\mathbf{X}_{\boldsymbol{\Omega}}$. We say that a subspace S *fits* $\mathbf{X}'_{\boldsymbol{\Omega}'}$ if $\mathbf{x}_{\boldsymbol{\omega}} \in S_{\boldsymbol{\omega}}$ for every column $\mathbf{x}_{\boldsymbol{\omega}}$ in $\mathbf{X}'_{\boldsymbol{\Omega}'}$.

The following example shows how false subspaces may fit arbitrarily many incomplete columns.

Example 3. *Let* $\mathcal{U}^\star = \{S_1^\star, S_2^\star\}$*, and*

$$S_1^\star = \operatorname{span}\begin{bmatrix}1\\1\\1\\1\end{bmatrix},\ S_2^\star = \operatorname{span}\begin{bmatrix}1\\2\\3\\4\end{bmatrix},\ \mathbf{X}'_{\boldsymbol{\Omega}'} = \begin{bmatrix}1 & \cdot & 3 & 1\\ 1 & 2 & \cdot & \cdot\\ \cdot & 2 & 3 & \cdot\\ \cdot & \cdot & \cdot & 4\end{bmatrix},$$

such that the first three columns of $\mathbf{X}'_{\boldsymbol{\Omega}'}$ *lie in* $S_1^\star$*, and the last one lies in* $S_2^\star$*. It is easy to see that* $\mathbf{X}'_{\boldsymbol{\Omega}'}$ *fits in a single* 1*-dimensional subspace, namely* $S = \operatorname{span}[1\ \ 1\ \ 1\ \ 4]^\mathsf{T}$*, even though* $S \notin \mathcal{U}^\star$*.*

Moreover, S *is the only* 1*-dimensional subspace that fits* $\mathbf{X}'_{\boldsymbol{\Omega}'}$*. Equivalently, there is only one rank-*1 *matrix that agrees with* $\mathbf{X}'_{\boldsymbol{\Omega}'}$*. In other words, the sampling* $\boldsymbol{\Omega}'$ *is identifiable for LRMC (the particular case of SCMD with just one subspace). This shows that even with unlimited computational power, if we exhaustively find all the identifiable patterns for LRMC, and collect their resulting subspaces, we can end up with false subspaces. Hence the importance of characterizing the identifiable patterns for SCMD.*

Example 3 shows how false subspaces may arise. The core of our main results lies in Theorem 3, which gives a deterministic condition to identify false subspaces, or equivalently, to determine whether a subspace indeed lies in $\mathcal{U}^\star$.

To build some intuition, imagine we *suspect* that S is one of the subspaces in $\mathcal{U}^\star$. For example, S may be a candidate subspace identified using a subset of the data. We want to know whether S is indeed one of the subspaces in $\mathcal{U}^\star$. Suppose first that we have an additional *complete* column $\mathbf{x}$ in general position on $S_k^\star$. Then w.p. 1, $\mathbf{x} \in S$ if and only if $S = S_k^\star$. We can thus verify whether $\mathbf{x} \in S$, and this will determine whether $S = S_k^\star$. It follows that if we had an additional complete column per subspace, we would be able to determine whether $S \in \mathcal{U}^\star$.

When handling incomplete data one cannot count on having complete columns. Instead, suppose we have a collection $\mathbf{X}'_{\boldsymbol{\Omega}'}$ of incomplete columns in general position on $\mathcal{U}^\star$. For example, $\mathbf{X}'_{\boldsymbol{\Omega}'}$ may be a subset of the data, not used to identify S. As we mentioned before, a complete column in general position on $S_k^\star$ will lie in S if and only if $S = S_k^\star$. We emphasize this because here lies the crucial difference with the missing data case: w.p. 1, an incomplete column $\mathbf{x}_{\boldsymbol{\omega}}$ in general position on $S_k^\star$ will fit in S if and only if the projections of S and $S_k^\star$ onto the canonical coordinates indicated by $\boldsymbol{\omega}$ are the same, i.e, if and only if $S_{\boldsymbol{\omega}} = S^\star_{k\boldsymbol{\omega}}$. More generally, a set $\mathbf{X}'_{\boldsymbol{\Omega}'}$ of incomplete columns in general position on $S_k^\star$ will fit in S if and only if $S_{\boldsymbol{\omega}} = S^\star_{k\boldsymbol{\omega}}$ for every column $\boldsymbol{\omega}$ in $\boldsymbol{\Omega}'$. Depending on $\boldsymbol{\Omega}'$, this may or may not imply that $S = S_k^\star$. It is possible that two different subspaces agree on almost all combinations of coordinates. This means S may fit arbitrarily many incomplete columns of $S_k^\star$, even if it is not $S_k^\star$.

Moreover, we do not know a priori whether the columns in $\mathbf{X}'_{\boldsymbol{\Omega}'}$ come from the same subspace. So if S fits $\mathbf{X}'_{\boldsymbol{\Omega}'}$, all we know is that S agrees with *some* subspaces of $\mathcal{U}^\star$ (the subspaces corresponding to the columns in $\mathbf{X}'_{\boldsymbol{\Omega}'}$) on *some* combinations of coordinates (indicated by the columns in $\boldsymbol{\Omega}'$). For example, if the columns in $\mathbf{X}'_{\boldsymbol{\Omega}'}$ come from two different subspaces of $\mathcal{U}^\star$, and S fits $\mathbf{X}'_{\boldsymbol{\Omega}'}$, then all we know is that S agrees with one subspace of $\mathcal{U}^\star$ in some coordinates, and with an other subspace of $\mathcal{U}^\star$ in other coordinates. This is what happened in Example 3: S agrees with $S_1^\star$ on the first three coordinates (rows), and with $S_2^\star$ on the first and fourth coordinates. Here lies the importance of our next main contribution: Theorem 3.

Theorem 3 states that if S fits $\mathbf{X}'_{\boldsymbol{\Omega}'}$, and $\mathbf{X}'_{\boldsymbol{\Omega}'}$ is *observed in the right places* (indicated by the matrix $\boldsymbol{\Omega}_\tau$ satisfying (i), defined in the lemma), then we can be sure that all the columns in $\mathbf{X}'_{\boldsymbol{\Omega}'}$ come from the same subspace in $\mathcal{U}^\star$, and that S is such subspace. The proof is given in Section 6.

Algorithm 1 Subspace Clustering Certification.

Input: Matrix $\mathbf{X}_{\mathbf{\Omega}}$.
· Split $\mathbf{X}_{\mathbf{\Omega}}$ into $\mathbf{X}_{\mathbf{\Omega}_1}$ and $\mathbf{X}_{\mathbf{\Omega}_2}$.
· Subspace cluster $\mathbf{X}_{\mathbf{\Omega}_1}$ to obtain $\hat{\mathcal{U}} = \{\hat{S}_k\}_{k=1}^K$.
for $k = 1$ **to** K **do**
  · $\mathbf{\Omega}_2^k$ = columns of $\breve{\mathbf{\Omega}}$ corresponding to the columns of $\mathbf{X}_{\mathbf{\Omega}_2}$ that fit in $\hat{S}_k$.
  if $\mathbf{\Omega}_2^k$ contains a $d \times (d-r+1)$ submatrix $\mathbf{\Omega}_\tau$ satisfying condition (i) (see Algorithm 2) **then**
    · **Output:** $\hat{S}_k$ is one of the subspaces in $\mathcal{U}^\star$.
  end if
end for

Theorem 3. *Let **A1**-**A2** hold. Let $\mathbf{X}'_{\mathbf{\Omega}'}$ be a matrix formed with a subset of the columns in $\mathbf{X}_{\mathbf{\Omega}}$. Let $\breve{\mathbf{\Omega}}'$ be the matrix containing the columns of $\breve{\mathbf{\Omega}}$ corresponding to $\mathbf{\Omega}'$. Suppose there is a matrix $\mathbf{\Omega}_\tau$ formed with $d-r+1$ columns of $\breve{\mathbf{\Omega}}'$ that satisfies* (i). *Let $S \in \mathrm{Gr}(r, \mathbb{R}^d)$ be a subspace identified without using $\mathbf{X}'_{\mathbf{\Omega}'}$. If S fits $\mathbf{X}'_{\mathbf{\Omega}'}$, then $S \in \mathcal{U}^\star$ w.p.* 1.

5. Practical Implications

In this section we present the main practical implication of our theoretical results: an efficient algorithm to certify the output of *any* SCMD method deterministically, in lieu of sampling and coherence assumptions.

Example 3 shows that finding a set of subspaces that agrees with the observed data does not guarantee that it is the correct set. It is possible that there exist false subspaces, that is, subspaces not in $\mathcal{U}^\star$ that agree with the observed data. Under certain assumptions on the subset of observed entries (e.g., random sampling) and $\mathcal{U}^\star$ (e.g., incoherence and distance between the subspaces), existing analyses have produced conditions to guarantee that this will not be the case (Eriksson et al., 2012; Pimentel-Alarcón et al., 2014). These assumptions and conditions are sometimes unverifiable, unjustifiable, or hardly met. For instance, (Eriksson et al., 2012) require $\mathcal{O}(d^{\log d})$ columns per subspace, and (Pimentel-Alarcón et al., 2014) require $\mathcal{O}(d^{r+1})$, and it is unusual to encounter such huge datasets. So in practice, the result of a SCMD algorithm can be suspect.

Hence using previous theory, even when we obtained a solution that *appeared* to be correct, we were be unable to tell whether it truly was. With our results, we now can. Theorem 3 implies that if one runs a SCMD algorithm on a subset of the data, then the uniqueness and correctness of the resulting clustering can be verified by testing whether it agrees with an other portion of the data that is *observed in the right places*. This is summarized in Algorithm 1.

Algorithm 2 Determine whether $\mathbf{\Omega}_\tau$ satisfies (i).

Input: Matrix $\mathbf{\Omega}_\tau$ with $d-r+1$ columns of $\breve{\mathbf{\Omega}}$
· Draw $\mathbf{U} \in \mathbb{R}^{d \times r}$ with i.i.d. Gaussian entries.
for $i = 1$ **to** $d-r+1$ **do**
  · $\mathbf{a}_{\boldsymbol{\omega}_i}$ = nonzero vector in $\ker \mathbf{U}_{\boldsymbol{\omega}_i}^\mathsf{T}$.
  · $\mathbf{a}_i$ = vector in $\mathbb{R}^d$ with entries of $\mathbf{a}_{\boldsymbol{\omega}_i}$ in the nonzero locations of $\boldsymbol{\omega}_i$ and zeros elsewhere.
end for
· $\mathbf{A}_{\tau-i}$ = matrix formed with all but the i^{th} column in $\{\mathbf{a}_i\}_{i=1}^{d-r+1}$.
if $\dim \ker \mathbf{A}_{\tau-i}^\mathsf{T} = r \ \ \forall\, i$ **then**
  · **Output:** $\mathbf{\Omega}_\tau$ satisfies (i).
else
  · **Output:** $\mathbf{\Omega}_\tau$ does not satisfy (i).
end if

Example 4. *To give a practical example, consider $d = 100$ and $r = 10$. In this case, the previous best guarantees that we are aware of require at least $N_k = \mathcal{O}(\min\{d^{\log d}, d^{r+1}\}) = \mathcal{O}(10^9)$, and that all entries are observed (Eriksson et al., 2012). Experiments show that practical algorithms can cluster perfectly even when fewer than half of the entries are observed, and with as little as $N_k = \mathcal{O}(rd)$. While previous theory for SCMD gives no guarantees in scenarios like this, our new results do.*

To see this, split $\mathbf{X}_{\mathbf{\Omega}}$ into two submatrices $\mathbf{X}_{\mathbf{\Omega}_1}$ and $\mathbf{X}_{\mathbf{\Omega}_2}$. Use any SCMD method to obtain an estimate $\hat{\mathcal{U}}$ of $\mathcal{U}^\star$, using only $\mathbf{X}_{\mathbf{\Omega}_1}$. Next cluster $\mathbf{X}_{\mathbf{\Omega}_2}$ according to $\hat{\mathcal{U}}$. Let $\mathbf{\Omega}_2^k$ denote the columns of $\breve{\mathbf{\Omega}}$ corresponding to the columns of $\mathbf{X}_{\mathbf{\Omega}_2}$ that fit in the k^{th} subspace of $\hat{\mathcal{U}}$. If $\mathbf{\Omega}_2^k$ is observed in the right places*, i.e., if $\mathbf{\Omega}_2^k$ contains a matrix $\mathbf{\Omega}_\tau$ satisfying* (i)*, then by Theorem 3 the k^{th} subspace of $\hat{\mathcal{U}}$ must be equal to one of the subspaces in $\mathcal{U}^\star$. It follows that if the subspaces in $\hat{\mathcal{U}}$ fit the columns in $\mathbf{X}_{\mathbf{\Omega}_2}$, and each $\hat{\mathbf{\Omega}}_2^k$ satisfies* (i)*, then the clustering is unique and correct, i.e., $\hat{\mathcal{U}} = \mathcal{U}^\star$.*

Algorithm 1 states that a clustering will be unique and correct if it is consistent with a hold-out subset of the data that satisfies (i). In Section 6 we show that sampling patterns with as little as $\mathcal{O}(\max\{r, \log d\})$ uniform random samples per column will satisfy (i) w.h.p. If this is the case, a clustering will be correct w.h.p. if it is consistent with enough hold-out columns ($d-r+1$ per subspace).

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, we can use Algorithm 2, which applies the results in (Pimentel-Alarcón et al., 2015b) to efficiently determine whether a matrix $\mathbf{\Omega}_\tau$ satisfies (i). This way, Algorithm 1 together with Algorithm 2 allow one to drop the sampling and incoherence assumptions, and validate the result of *any* SCMD algorithm deterministically and efficiently.

Algorithm 2 checks the dimension of the null-space of $d-r+1$ sparse matrices. To present the algorithm, let $\boldsymbol{\Omega}_\tau$ be a matrix formed with $d-r+1$ columns of $\breve{\boldsymbol{\Omega}}$, and let $\boldsymbol{\Omega}_{\tau-i}$ denote the matrix formed with all but the i^{th} column of $\boldsymbol{\Omega}_\tau$. Consider the following condition:

(ii) Every matrix $\boldsymbol{\Omega}'_\tau$ formed with a subset of the columns in $\boldsymbol{\Omega}_{\tau-i}$ (including $\boldsymbol{\Omega}_{\tau-i}$) satisfies (2).

It is easy to see that $\boldsymbol{\Omega}_\tau$ will satisfy (i) if and only if $\boldsymbol{\Omega}_{\tau-i}$ satisfies (ii) for every $i=1,\dots,d-r+1$. Fortunately, the results in (Pimentel-Alarcón et al., 2015b) provide an efficient algorithm to verify whether (ii) is satisfied.

Next let $\boldsymbol{\omega}_i$ denote the i^{th} column of $\boldsymbol{\Omega}_\tau$, and let $\mathbf{U}$ be a $d\times r$ matrix drawn according to an absolutely continuous distribution w.r.t. the Lebesgue measure on $\mathbb{R}^{d\times r}$ (e.g., with i.i.d. Gaussian entries). Recall that $\mathbf{U}_{\boldsymbol{\omega}_i}$ denotes the restriction of $\mathbf{U}$ to the nonzero rows in $\boldsymbol{\omega}_i$. Let $\mathbf{a}_{\boldsymbol{\omega}_i}\in\mathbb{R}^{r+1}$ be a nonzero vector in $\ker \mathbf{U}_{\boldsymbol{\omega}_i}^{\mathsf{T}}$, and $\mathbf{a}_i$ be the vector in $\mathbb{R}^d$ with the entries of $\mathbf{a}_{\boldsymbol{\omega}_i}$ in the nonzero locations of $\boldsymbol{\omega}_i$ and zeros elsewhere. Finally, let $\mathbf{A}_{\tau-i}$ denote the $d\times(d-r)$ matrix with all but the i^{th} column in $\{\mathbf{a}_i\}_{i=1}^{d-r+1}$.

Section 3 in (Pimentel-Alarcón et al., 2015b) shows that $\boldsymbol{\Omega}_{\tau-i}$ satisfies (ii) if and only if $\dim\ker \mathbf{A}_{\tau-i}^{\mathsf{T}}=r$. Algorithm 2 will verify whether $\dim\ker \mathbf{A}_{\tau-i}^{\mathsf{T}}=r$ for every i, and this will determine whether $\boldsymbol{\Omega}_\tau$ satisfies (i). We thus have the next lemma, which states that w.p. 1, Algorithm 2 will determine whether $\boldsymbol{\Omega}_\tau$ satisfies (i).

Lemma 1. *Let $\boldsymbol{\Omega}_\tau$ be a matrix formed with $d-r+1$ columns of $\breve{\boldsymbol{\Omega}}$. Let $\{\mathbf{A}_{\tau-i}\}_{i=1}^{d-r+1}$ be constructed as in Algorithm 2. Then w.p. 1, $\boldsymbol{\Omega}_\tau$ satisfies* (i) *if and only if* $\dim\ker \mathbf{A}_{\tau-i}^{\mathsf{T}}=r$ *for every* $i=1,\dots,d-r+1$.

6. Proofs

Similar to $\breve{\boldsymbol{\Omega}}$, let us introduce the *expanded matrix* $\breve{\mathbf{X}}_{\breve{\boldsymbol{\Omega}}}$ of $\mathbf{X}_{\boldsymbol{\Omega}}$. Recall that ℓ_i denotes the number of observed entries of the i^{th} column of $\mathbf{X}_{\boldsymbol{\Omega}}$.

Definition 2 (Expanded Matrix)**.** *Define $\breve{\mathbf{X}}_i$ as the matrix with ℓ_i-r columns, all identical to the i^{th} column of $\mathbf{X}$. Let $\breve{\mathbf{X}}:=[\breve{\mathbf{X}}_1\ \cdots\ \breve{\mathbf{X}}_N]$. Define the* expanded *matrix $\breve{\mathbf{X}}_{\breve{\boldsymbol{\Omega}}}$ as the matrix with the values of $\breve{\mathbf{X}}$ in the nonzero locations of $\breve{\boldsymbol{\Omega}}$, and missing values in the zero locations of $\breve{\boldsymbol{\Omega}}$.*

Each column in $\breve{\mathbf{X}}_{\breve{\boldsymbol{\Omega}}}$ specifies one constraint on what $\mathcal{U}^\star$ may be, and the pattern $\breve{\boldsymbol{\Omega}}$ determines whether these constraints are redundant. We will show that if the conditions of Theorem 2 are satisfied, then $\mathcal{U}^\star$ can be uniquely identified from these constraints.

To identify $\mathcal{U}^\star$, we will exhaustively search for subspaces that fit combinations of $(r+1)(d-r+1)$ columns of $\breve{\mathbf{X}}_{\breve{\boldsymbol{\Omega}}}$ that are *observed in the right places*, that is, satisfying the conditions of Theorem 2. More precisely, let $\boldsymbol{\Omega}^\eta$ denote the matrix formed with the η^{th} combination of $(r+1)(d-r+1)$ columns of $\breve{\boldsymbol{\Omega}}$. For each η we will verify whether $\boldsymbol{\Omega}^\eta$ can be partitioned into $r+1$ submatrices $\{\boldsymbol{\Omega}_\tau^\eta\}_{\tau=1}^{r+1}$ satisfying (i). In this case, we will use the first r sets of $d-r+1$ incomplete columns of $\breve{\mathbf{X}}_{\breve{\boldsymbol{\Omega}}}$ corresponding to $\{\boldsymbol{\Omega}_\tau^\eta\}_{\tau=1}^{r}$ to identify a candidate subspace S^η. Next we will verify whether the last set of $d-r+1$ incomplete columns of $\breve{\mathbf{X}}_{\breve{\boldsymbol{\Omega}}}$ corresponding to $\boldsymbol{\Omega}_{r+1}^\eta$ fit in S^η. In this case, we will keep S in the collection of subspaces $\hat{\mathcal{U}}$. We will show that if the assumptions of Theorem 2 are satisfied, then the output of this procedure, $\hat{\mathcal{U}}$, will be equal to $\mathcal{U}^\star$.

Theorem 2 is based on Theorem 3 above, and Lemma 8 in (Pimentel-Alarcón et al., 2015a), which states that if $\{\boldsymbol{\Omega}_\tau^\eta\}_{\tau=1}^{r}$ satisfy (i), there are at most finitely many r-dimensional subspaces that fit $\{\mathbf{X}_{\boldsymbol{\Omega}_\tau}^\eta\}_{\tau=1}^{r}$ (the corresponding submatrices of $\breve{\mathbf{X}}_{\breve{\boldsymbol{\Omega}}}$ observed on $\{\boldsymbol{\Omega}_\tau^\eta\}_{\tau=1}^{r}$). We restate this result as the following lemma, with some adaptations.

Lemma 2. *Let **A1**-**A2** hold $\forall k$. Suppose $\{\boldsymbol{\Omega}_\tau^\eta\}_{\tau=1}^{r}$ satisfy* (i)*. Then w.p.* 1*, there exist at most finitely many r-dimensional subspaces that fit $\{\mathbf{X}_{\boldsymbol{\Omega}_\tau}^\eta\}_{\tau=1}^{r}$.*

Remark 2. *Lemma 8 in (Pimentel-Alarcón et al., 2015a) holds for a different construction of $\boldsymbol{\Omega}_i$. However, both constructions define the same variety, and hence they can be used interchangeably. We prefer this construction because it spreads the nonzero entries in $\boldsymbol{\Omega}_i$ more uniformly.*

Proof of Theorem 2

We will show that $\hat{\mathcal{U}}=\mathcal{U}^\star$.

($\subset$) Suppose $\boldsymbol{\Omega}^\eta$ can be partitioned into $r+1$ submatrices $\{\boldsymbol{\Omega}_\tau^\eta\}_{\tau=1}^{r+1}$ satisfying (i). By Lemma 2, there are at most finitely many r-dimensional subspaces that fit $\{\mathbf{X}_{\boldsymbol{\Omega}_\tau}^\eta\}_{\tau=1}^{r}$. Let S^η be one of these subspaces. Since $\boldsymbol{\Omega}_{r+1}^\eta$ also satisfies (i), it follows by Theorem 3 that S^η will only fit $\mathbf{X}_{\boldsymbol{\Omega}_{r+1}}^\eta$ if $S^\eta\in\mathcal{U}^\star$. Since this argument holds for all of the finitely many r-dimensional subspaces that fit $\{\mathbf{X}_{\boldsymbol{\Omega}_\tau}^\eta\}_{\tau=1}^{r}$, it follows that only subspaces in $\mathcal{U}^\star$ may fit $\{\mathbf{X}_{\boldsymbol{\Omega}_\tau}^\eta\}_{\tau=1}^{r+1}$. Since η was arbitrary, it follows that $\hat{\mathcal{U}}\subset\mathcal{U}^\star$.

($\supset$) By **A3**, $\mathbf{X}_{\boldsymbol{\Omega}}$ has at least $(r+1)(d-r+1)$ columns from each of the K subspaces in $\mathcal{U}^\star$. By assumption, there is some η such that all the columns in the η^{th} combination belong to $S_k^\star$ and $\boldsymbol{\Omega}^\eta$ can be partitioned into $r+1$ submatrices $\{\boldsymbol{\Omega}_\tau^\eta\}_{\tau=1}^{r+1}$ satisfying (i). Take such η. Then $S_k^\star\in\hat{\mathcal{U}}$, as $S_k^\star$ trivially fits this combination of columns. Since this is true for every k, it follows that $\mathcal{U}^\star\subset\hat{\mathcal{U}}$. □

Proof of Theorem 1

Theorem 1 follows as a consequence of Theorem 2 and Lemma 9 in (Pimentel-Alarcón et al., 2015a), which we restate here with some adaptations as Lemma 3.

Lemma 3. *Let the sampling assumptions of Theorem 1 hold. Let $\boldsymbol{\Omega}_{\tau-i}$ be a matrix formed with $d-r$ columns of $\boldsymbol{\Omega}$. Then $\boldsymbol{\Omega}_{\tau-i}$ satisfies* (ii) *w.p. at least* $1-\frac{\epsilon}{d}$.

By **A3**, $\mathbf{X}^k$ has at least $(r+1)(d-r+1)$ columns. Randomly partition the matrix indicating the observed entries of $\mathbf{X}^k$ into matrices $\{\mathbf{\Omega}_\tau\}_{\tau=1}^{r+1}$, each of size $d \times (d-r+1)$. Let $\mathbf{\Omega}_{\tau-i}$ denote the $d \times (d-r)$ matrix formed with all but the i^{th} column in $\mathbf{\Omega}_\tau$. It is easy to see that $\mathbf{\Omega}_\tau$ will satisfy (i) if every $\mathbf{\Omega}_{\tau-i}$ satisfies (ii).

By Lemma 3 and a union bound, $\mathbf{\Omega}_\tau$ will satisfy (i) w.p. at least $1-\epsilon$. Using two more union bounds we conclude that the conditions of Theorem 2 are satisfied w.p. at least $1-K(r+1)\epsilon$, whence, by Theorem 2, $\mathcal{U}^\star$ can be uniquely identified from $\mathbf{X}_{\mathbf{\Omega}}$.

Proof of Theorem 3

As mentioned in Section 4, the main difficulty in showing our main results is that there could exist false subspaces, that is, subspaces not in $\mathcal{U}^\star$ that could fit arbitrarily many incomplete columns. Theorem 3 provides a deterministic condition to determine whether a subspace lies in $\mathcal{U}^\star$. We use this section to give the proof of this statement and expose the main ideas used to derive it.

To build some intuition, imagine we *suspect* that S is one of the subspaces in $\mathcal{U}^\star$. We want to determine whether it truly is one of the subspaces in $\mathcal{U}^\star$. From the discussion in Section 4 it follows that if we had a *complete* column in general position from each of the subspaces in $\mathcal{U}^\star$, we could check whether S fits any of these columns, knowing that almost surely, it will do so if and only if $S \in \mathcal{U}^\star$.

When handling incomplete data one cannot count on having a complete column. But what if we had several incomplete ones instead? Could a set of incomplete columns *behave* just as a complete column in the sense that S will only fit such set if $S \in \mathcal{U}^\star$? The answer to this question is yes, and is given by Theorem 3, which, in a way, is telling us that a set of incomplete columns will *behave* as a single but complete one if it is *observed in the right places*.

We are thus interested in knowing when will a set of N_τ incomplete columns have the property that only a subspace from $\mathcal{U}^\star$ can fit them. As discussed in Section 4, a complete column in general position on $S_k^\star$ will fit in S if and only if $S = S_k^\star$. Similarly, an incomplete column $\mathbf{x}_{\boldsymbol{\omega}}$ in general position on $S_k^\star$ will fit in S if and only if the projections of S and $S_k^\star$ on $\boldsymbol{\omega}$ are the same, i.e., if $S_{\boldsymbol{\omega}} = S_{k\boldsymbol{\omega}}^\star$.

Therefore, we can restate the problem of interest as follows: suppose we are given N_τ projections of some of the subspaces in $\mathcal{U}^\star$ onto small subsets of the canonical coordinates. When will only one of the subspaces in $\mathcal{U}^\star$ agree with this set of projections?

Question. (Qa) *Can we guarantee that only a subspace in $\mathcal{U}^\star$ will agree with the given projections if there is only one r-dimensional subspace that agrees with these projections? The answer is no.*

Example 5. *Let $\mathcal{U}^\star$ be as in Example 3, and assume that we only observe the following set of projections:*

$$\left\{ \operatorname{span}\begin{bmatrix}1\\1\\0\\0\end{bmatrix}, \operatorname{span}\begin{bmatrix}0\\1\\1\\0\end{bmatrix}, \operatorname{span}\begin{bmatrix}1\\0\\1\\0\end{bmatrix}, \operatorname{span}\begin{bmatrix}1\\0\\0\\4\end{bmatrix} \right\},$$

corresponding to subspaces $\{1,1,1,2\}$ in $\mathcal{U}^\star$. It is not hard to see that $S = \operatorname{span}[1 \;\; 1 \;\; 1 \;\; 4]^\mathsf{T}$ is the only 1-dimensional subspace that agrees with these projections. But $S \notin \mathcal{U}^\star$.

Question. (Qb) *Can we guarantee that only a subspace in $\mathcal{U}^\star$ will agree with the set of given projections if all the projections correspond to the same subspace in $\mathcal{U}^\star$? Again, the answer is no.*

Example 6. *With $\mathcal{U}^\star$ as in Example 3, suppose that we only observe the following projections: $\{\operatorname{span}[1 \;\; 1 \;\; 0 \;\; 0]^\mathsf{T}, \operatorname{span}[0 \;\; 0 \;\; 1 \;\; 1]^\mathsf{T}\}$, both corresponding to $S_1^\star$. It is easy to see that for $\alpha \in \mathbb{R}\backslash\{0,1\}$, $S = \operatorname{span}[1 \;\; 1 \;\; \alpha \;\; \alpha]^\mathsf{T}$ agrees with these projections, even though $S \notin \mathcal{U}^\star$.*

Fortunately, we can guarantee that only one of the subspaces in $\mathcal{U}^\star$ will agree with the given set of projections if both conditions hold, i.e., if (a) there is only one r-dimensional subspace that agrees with these projections, *and* (b) all the given projections correspond to the same subspace $S_k^\star$. To see this, observe that if (b) holds, then it is trivially true that $S_k^\star$ will agree with all the given projections. If in addition (a) holds, we automatically know that $S_k^\star$ is the only subspace that agrees with the projections. Notice that these conditions are also necessary in the sense that if either (a) or (b) fail, it cannot be guaranteed that only one of the subspaces in $\mathcal{U}^\star$ will agree with the given set of projections, as explained in Examples 5 and 6.

We will thus characterize the sets of projections that satisfy conditions (a) and (b). To this end, we will use the $d \times N_\tau$ binary matrix $\mathbf{\Omega}_\tau$ to encode the information about the given projections. Recall that $\boldsymbol{\omega}_i$ denotes the i^{th} column of $\mathbf{\Omega}_\tau$, and that $\mathbf{\Omega}_\tau$ is a matrix formed with a subset of the columns in $\breve{\mathbf{\Omega}}$. Let the nonzero entries of $\boldsymbol{\omega}_i$ indicate the canonical coordinates involved in the i^{th} projection. Let $\mathcal{K} := \{k_i\}_{i=1}^{N_\tau}$ be a multiset of indices in $\{1,\dots,K\}$ indicating that the i^{th} given projection (or equivalently, the column of $\breve{\mathbf{X}}_{\breve{\mathbf{\Omega}}}$ corresponding to $\boldsymbol{\omega}_i$) corresponds to the $k_i{}^{\text{th}}$ subspace in $\mathcal{U}^\star$. In Example 5, $\mathcal{K} = \{1,1,1,2\}$. Recall that $S_{\boldsymbol{\omega}_i}^\star$ denotes the restriction of $S_{k_i}^\star$ to the nonzero coordinates in $\boldsymbol{\omega}_i$. We will use $\mathcal{S}$ as shorthand for $\mathcal{S}(\mathcal{U}^\star, \mathcal{K}, \mathbf{\Omega}_\tau)$, which denotes the set of all r-dimensional subspaces S of $\mathbb{R}^d$ that satisfy $S_{\boldsymbol{\omega}_i} = S_{\boldsymbol{\omega}_i}^\star \; \forall\, i$. In words, $\mathcal{S}$ is the set of all r-dimensional subspaces matching projections of *some* of the subspaces in $\mathcal{U}^\star$ (indexed by $\mathcal{K}$) on $\mathbf{\Omega}_\tau$. Notice that $\mathcal{S}$ may be empty.

CONDITIONS TO GUARANTEE (a)

The conditions to guarantee that there is only one subspace consistent with a set of projections are given by Theorem 1 by (Pimentel-Alarcón et al., 2015b). We restate this result with some adaptations to our context as follows.

Lemma 4. *Let **A1** hold. W.p.* 1 *there is* at most one *subspace in* $\mathcal{S}(\mathcal{U}^\star, \mathcal{K}, \mathbf{\Omega}_\tau)$ *if and only if there is a matrix* $\mathbf{\Omega}_{\tau-i}$ *formed with* $d-r$ *columns of* $\mathbf{\Omega}_\tau$ *that satisfies* (ii)*.*

Lemma 4 states that $d-r$ projections *onto the right canonical coordinates* are sufficient to guarantee that there is only one subspace consistent with these projections. Intuitively, this means that if you have $d-r$ *good* projections, there is only one way that you can *stitch* them together into one subspace. In the union of subspaces setting, though, these *good* projections could correspond to different subspaces in $\mathcal{U}^\star$, so when we *stitch* them together, we could end up with a false subspace. This is what happened in Examples 3 and 5. That is why, in addition to guaranteeing that (a) there is only one subspace consistent with the given set of projections, we also need to guarantee that (b) all of these projections come from the same subspace.

CONDITIONS TO GUARANTEE (b)

The following lemma states that $d-r+1$ projections (onto the right canonical coordinates) guarantee that a set of given projections correspond to the same subspace in $\mathcal{U}^\star$, i.e., that $k_i = k_j$ for every $j \in \{1, \dots, d-r+1\}$.

Lemma 5. *Let **A1** hold. Suppose* $\mathcal{S}(\mathcal{U}^\star, \mathcal{K}, \mathbf{\Omega}_\tau)$ *is nonempty, and* $\mathbf{\Omega}_\tau$ *has* $d-r+1$ *columns. W.p.* 1*, if* $\mathbf{\Omega}_\tau$ *satisfies* (i)*, then* $k_i = k_j$ *for every* i, j*.*

Notice that the conditions of Lemma 5 imply those of Lemma 4. This means that once we know that $d-r+1$ projections correspond to the same subspace in $\mathcal{U}^\star$, we automatically know that there is only one subspace consistent with these projections, and it can only be one of the subspaces in $\mathcal{U}^\star$. In order to prove Lemma 5, let $\mathbf{a}_{\boldsymbol{\omega}_i} \in \mathbb{R}^{r+1}$ denote a nonzero vector orthogonal to $S^\star_{\boldsymbol{\omega}_i}$, and recall that $\mathbf{a}_i$ is the vector in $\mathbb{R}^d$ with the entries of $\mathbf{a}_{\boldsymbol{\omega}_i}$ in the nonzero locations of $\boldsymbol{\omega}_i$ and zeros elsewhere. Let $\mathbf{A}$ denote be the $d \times (d-r+1)$ matrix with $\{\mathbf{a}_i\}_{i=1}^{d-r+1}$ as its columns. We will use $\mathbf{A}'$ to denote a matrix formed with a subset of the columns in $\mathbf{A}$, and $\boldsymbol{i}$ to denote the indices of such columns, i.e. $\boldsymbol{i} := \{i : \mathbf{a}_i \in \mathbf{A}'\}$.

We say that $\mathbf{A}'$ is *minimally linearly dependent* if the columns in $\mathbf{A}'$ are linearly dependent, but every proper subset of the columns in $\mathbf{A}'$ is linearly independent. Recall that $n(\mathbf{A}')$ and $m(\mathbf{A}')$ denote the number of columns and the number of nonzero rows in $\mathbf{A}'$. We first determine when will *some* projections correspond to the same subspace of $\mathcal{U}^\star$, i.e., when will $k_i = k_j$ for *some* pairs (i, j).

Lemma 6. *Let **A1** hold. W.p.* 1*, if* $\mathbf{A}'$ *is minimally linearly dependent, then* $k_i = k_j$ *for every* $i, j \in \boldsymbol{i}$*.*

Proof. Let $\mathbf{A}' = [\mathbf{A}''\ \mathbf{a}_i]$ be minimally linearly dependent. Then $\mathbf{A}''\boldsymbol{\beta} = \mathbf{a}_i$ for some $\boldsymbol{\beta} \in \mathbb{R}^{n(\mathbf{A}'')}$, where every entry of $\boldsymbol{\beta}$ is nonzero. On the other hand, $\mathbf{a}_i$ is a nonzero function of $S^\star_{k_i}$. Similarly, every column $\mathbf{a}_j$ of $\mathbf{A}''$ is a nonzero function of $S^\star_{k_j}$. Under **A1**, w.p.1 the subspaces in $\mathcal{U}^\star$ keep no relation between each other, so $\mathbf{A}''\boldsymbol{\beta} = \mathbf{a}_i$ can only hold if $S^\star_{k_i} = S^\star_{k_j}\ \forall\ i, j \in \boldsymbol{i}$, i.e., if $k_i = k_j\ \forall\ i, j \in \boldsymbol{i}$. □

Now we can determine when will *all* the projections correspond to the same subspace of $\mathcal{U}^\star$.

Lemma 7. *Let **A1** hold. W.p.* 1*, if* $\mathbf{A}$ *is minimally linearly dependent, then* $k_i = k_j$ *for every* i, j*.*

The next lemma uses Lemma 2 in (Pimentel-Alarcón et al., 2015b) (which we state here as Lemma 9) to characterize when will $\mathbf{A}$ be minimally linearly dependent.

Lemma 8. *Let **A1** hold. W.p.* 1*,* $\mathbf{A}$ *is minimally linearly dependent if* $\mathcal{S}(\mathcal{U}^\star, \mathcal{K}, \mathbf{\Omega}_\tau) \neq \emptyset$ *and every matrix* $\mathbf{A}'$ *formed with a* proper *subset of the columns in* $\mathbf{A}$ *satisfies* $m(\mathbf{A}') \geq n(\mathbf{A}') + r$*.*

Lemma 9. *Let **A1** hold. W.p.* 1*, the columns in* $\mathbf{A}'$ *are linearly independent if* $m(\mathbf{A}'') \geq n(\mathbf{A}'') + r$ *for every matrix* $\mathbf{A}''$ *formed with a subset of the columns in* $\mathbf{A}'$*.*

Proof. (Lemma 8) Suppose every matrix $\mathbf{A}'$ formed with a proper subset of the columns in $\mathbf{A}$ satisfies $m(\mathbf{A}') \geq n(\mathbf{A}') + r$. By Lemma 9, every proper subset of the columns in $\mathbf{A}$ is linearly independent. To see that the columns in $\mathbf{A}$ are linearly dependent, recall that $\ker \mathbf{A}^\mathsf{T}$ contains every element of $\mathcal{S}$ (see Section 3 in (Pimentel-Alarcón et al., 2015b)). Therefore, $\mathbf{A}$ contains at most $d-r$ linearly independent columns (otherwise $\dim \ker \mathbf{A}^\mathsf{T} < r$, and $\mathcal{S}$ would be empty). But since $\mathbf{A}$ has $d-r+1$ columns, we conclude that they are linearly dependent. □

We now give the proofs of Lemma 5 and Theorem 3.

Proof. (Lemma 5) Suppose $\mathbf{\Omega}_\tau$ satisfies (i). Under **A1**, an entry of $\mathbf{A}$ is nonzero if and only if the same entry of $\mathbf{\Omega}_\tau$ is nonzero, which implies $\mathbf{A}$ satisfies the conditions of Lemma 8. It follows that $\mathbf{A}$ is minimally linearly dependent, so by Lemma 7, $k_i = k_j\ \forall\ i, j$. □

Proof. (Theorem 3) Let $\mathbf{x}_{\boldsymbol{\omega}_i}$ denote the column of $\breve{\mathbf{X}}_{\breve{\mathbf{\Omega}}}$ corresponding to the i^{th} column of $\mathbf{\Omega}_\tau$, and suppose S fits $\mathbf{X}'_{\mathbf{\Omega}'}$. By definition, $\mathbf{x}_{\boldsymbol{\omega}_i} \in S_{\boldsymbol{\omega}_i}$. On the other hand, $\mathbf{x}_{\boldsymbol{\omega}_i} \in S^\star_{\boldsymbol{\omega}_i}$ by assumption, which implies $\mathbf{x}_{\boldsymbol{\omega}_i} \in S_{\boldsymbol{\omega}_i} \cap S^\star_{\boldsymbol{\omega}_i}$. Therefore, $S_{\boldsymbol{\omega}_i} = S^\star_{\boldsymbol{\omega}_i}$ w.p. 1 (because if $S_{\boldsymbol{\omega}_i} \neq S^\star_{\boldsymbol{\omega}_i}$, then $\mathbf{x}_{\boldsymbol{\omega}_i} \notin S_{\boldsymbol{\omega}_i} \cap S^\star_{\boldsymbol{\omega}_i}$ w.p. 1). Since this is true for every i, we conclude that $S \in \mathcal{S}$. Now assume $\mathbf{\Omega}_\tau$ satisfies (i). Then $\mathbf{\Omega}_\tau$ satisfies the conditions of Lemmas 4 and 5. By Lemma 5, $k_i = k_j$ for every i, j, which trivially implies $S^\star_{k_i} \in \mathcal{S}$. By Lemma 4, there is only one subspace in $\mathcal{S}$. This implies $S = S^\star_{k_i}$. □

Acknowledgements

This work was partially supported by AFOSR grant FA9550-13-1-0138.

References

Balzano, L., Szlam, A., Recht, B., and Nowak, R. K-subspaces with missing data. In *IEEE Statistical Signal Processing Workshop*, 2012.

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

Elhamifar, E. and Vidal, R. Sparse subspace clustering. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2009.

Elhamifar, E. and Vidal, R. Sparse subspace clustering: Algorithm, theory, and applications. In *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2013.

Eriksson, B., Balzano, L., and Nowak, R. High-rank matrix completion and subspace clustering with missing data. In *Conference on Artificial Intelligence and Statistics*, 2012.

Hu, H., Feng, J., and Zhou, J. Exploiting unsupervised and supervised constraints for subspace clustering. In *IEEE Pattern Analysis and Machine Intelligence*, 2015.

Kanatani, K. Motion segmentation by subspace separation and model selection. In *IEEE International Conference in Computer Vision*, 2001.

Liu, G., Lin, Z., and Yu, Y. Robust subspace segmentation by low-rank representation. In *International Conference on Machine Learning*, 2010.

Liu, G., Lin, Z., Yan, S., Sun, J., Yu, Y., and Ma, Y. Robust recovery of subspace structures by low-rank representation. In *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2013.

Mateos, G. and Rajawat, K. Dynamic network cartography: Advances in network health monitoring. In *IEEE Signal Processing Magazine*, 2013.

Peng, X., Yi, Z., and Tang, H. Robust subspace clustering via thresholding ridge regression. In *AAAI Conference on Artificial Intelligence*, 2015.

Pimentel-Alarcón, D., Balzano, L., and Nowak, R. On the sample complexity of subspace clustering with missing data. In *IEEE Statistical Signal Processing Workshop*, 2014.

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

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

Qu, C. and Xu, H. Subspace clustering with irrelevant features via robust dantzig selector. In *Advances in Neural Information Processing Systems*, 2015.

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

Soltanolkotabi, M. Algorithms and theory for clustering and nonconvex quadratic programming. In *PhD. Dissertation*, 2014.

Soltanolkotabi, M., Elhamifar, E., and Candès, E. Robust subspace clustering. In *Annals of Statistics*, 2014.

Vidal, R. Subspace clustering. In *IEEE Signal Processing Magazine*, 2011.

Wang, Y. and Xu, H. Noisy sparse subspace clustering. In *International Conference on Machine Learning*, 2013.

Wang, Y., Wang, Y., and Singh, A. Differentially private subspace clustering. In *Advances in Neural Information Processing Systems*, 2015.

Yang, C., Robinson, D., and Vidal, R. Sparse subspace clustering with missing entries. In *International Conference on Machine Learning*, 2015.

Zhang, A., Fawaz, N., Ioannidis, S., and Montanari, A. Guess who rated this movie: Identifying users through subspace clustering. In *Conference on Uncertainty in Artificial Intelligence*, 2012.