Necessary and Sufficient Conditions for Sketched Subspace Clustering

Daniel Pimentel-Alarcón[1], Laura Balzano[2], Robert Nowak[1]
[1]University of Wisconsin-Madison, [2]University of Michigan-Ann Arbor

***Abstract*— This paper is about an interesting phenomenon: two r-dimensional subspaces, even if they are orthogonal to one an other, can appear identical if they are only observed on a subset of coordinates. Understanding this phenomenon is of particular importance for many modern applications of subspace clustering where one would like to subsample in order to improve computational efficiency. Examples include real-time video surveillance and datasets so large that cannot even be stored in memory. In this paper we introduce a new metric between subspaces, which we call *partial coordinate discrepancy*. This metric captures a notion of similarity between subsampled subspaces that is not captured by other distance measures between subspaces. With this, we are able to show that subspace clustering is theoretically possible in lieu of coherence assumptions using only $r+1$ rows of the dataset at hand. This gives precise information-theoretic necessary and sufficient conditions for sketched subspace clustering. This can greatly improve computational efficiency without compromising performance. We complement our theoretical analysis with synthetic and real data experiments.**

I. Introduction

In subspace clustering (SC), one is given a data matrix $\mathbf{X}$ whose columns lie in the union of several (unknown) r-dimensional subspaces, and aims to infer these subspaces and cluster the columns in $\mathbf{X}$ accordingly [1]. The union of subspaces model is a powerful and flexible model that applies to a wide variety of practical applications, ranging from computer vision [2] to network inference [3], [4], compression [5], recommender systems and collaborative filtering [6], [7]. Hence there is growing attention to this problem. As a result, existing theory and methods can handle outliers [8]–[13], noisy measurements [14], privacy concerns [15], data constraints [16], and missing data [17]–[21], among other difficulties.

Yet, in many relevant applications, such as real-time video surveillance, or cases where $\mathbf{X}$ is too large to even store in memory, SC remains infeasible due to computational constraints. In applications like these, it is essential to handle big datasets in a computationally efficient manner, both in terms of storage and processing time.

Fortunately, studies regarding missing data show that under this model, very large datasets can be accurately represented using a very small number of its entries [17]–[21]. With this in mind, recent studies (e.g., [22]) explore the idea of projecting the data (e.g., subsampling or sketching) as alternatives to reduce computational costs (time and storage).

On this matter, it was recently shown that if the subspaces are sufficiently incoherent and separated, and the columns are well-spread over the subspaces, then the popular sparse subspace clustering (SSC) algorithm [23] will find the correct clustering using certain sketches of the data (e.g., gaussian projection, row subsampling, and the fast Johnson-Lindenstrauss transform) [24]. However, in general, these conditions are unverifiable.

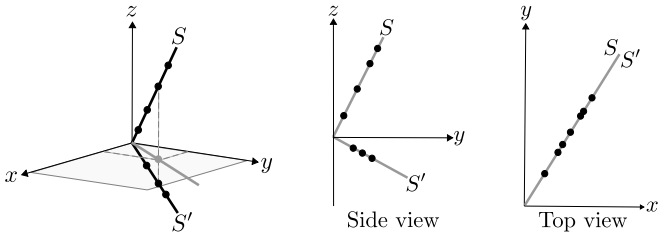

Fig. 1: **Left:** The columns in $\mathbf{X}$ (represented by points) lie in the union of two 1-dimensional subspaces in $\mathbb{R}^3$. We want to cluster these points using only a few coordinates (to improve computational costs). This can be done if we use coordinates (y,z), as in the **center**. The main difficulty is that the subspaces may be equal in certain coordinates. In this example, the subspaces are equal on the (x,y) coordinates. So if we use coordinates (x,y), as in the **right**, then all columns will appear to lie in the same subspace, and clustering would be impossible. We do not know beforehand the coordinates in which the subspaces are different. Searching for such coordinates could result in combinatorial complexity, defeating the purpose of subsampling.

In this paper we show that almost every $\mathbf{X}$ can be theoretically clustered using as few as $r+1$ rows (the minimum required) of a generic rotation of $\mathbf{X}$. The subtlety of this result is that the underlying subspaces may be equal in certain coordinates. This means that if we sample a column of $\mathbf{X}$ in a set of coordinates where the underlying subspaces are equal, one would be unable to determine (based on these observations) to which subspace it truly belongs. See Figure 1 to build some intuition.

To give a concrete example, consider images as in Figure 2. It has been shown that the face images of the same individual under varying illumination lie near a low-dimensional subspace [25]. Hence SC can be used to classify faces. However, some coordinates (e.g., the corner pixels) are equal across many individuals. If we only sampled those coordinates, we would be unable to cluster. Moreover, those coordinates would only obstruct clustering while consuming computational resources.

To the best of our knowledge, none of the existing distance measures between subspaces captures this notion of *partial coordinate similarity*. For instance, Example 1 in Section II shows that orthogonal subspaces (maximally apart with

Fig. 2: Images from the Extended Yale B dataset [26]. Each row has images of the same individual under varying illumination. The vectorized images of each individual lie near a 9-dimensional subspace [25], so the whole dataset lies near a union of subspaces. Some coordinates (e.g., the corner pixels) are equal across many individuals. If we only sampled those coordinates, we would be unable to subspace cluster.

respect to the principal angle distance, the affinity distance, and the subspace incoherence distance [10]) can be identical in certain coordinates. In this paper we study this phenomenon to derive precise information-theoretic necessary and sufficient conditions for sketched subspace clustering.

To this end we first introduce a new distance measure between subspaces that captures this relationship between subspaces, which we call *partial coordinate discrepancy*. This allows us to show that if we generically rotate $\mathbf{X}$, its columns will lie in subspaces that are different on all subsets of more than r coordinates with probability 1. In other words, generic rotations maximize partial coordinate discrepancy. This will imply that $\mathbf{X}$ can be clustered using only a sketch, that is, a few rows of a generic rotation of $\mathbf{X}$. We complement our theoretical analysis with experiments using synthetic and real data, showing the performance and advantages of sketching.

Organization of the paper

In Section II we formally state the problem, introduce our new distance measure between subspaces, and give our main results. In Section III we make several remarks about our distance measure. In Section IV we present experiments to support our results. We leave all proofs to Section V.

II. Model and main results

Let $\mathcal{U} := \{S^k\}_{k=1}^K$ be a set of r-dimensional subspaces of $\mathbb{R}^d$, and $\mathbf{X}$ be a $d \times n$ data matrix whose columns lie in the union of the subspaces in $\mathcal{U}$. Let $\mathbf{X}^k$ denote the matrix with all the columns of $\mathbf{X}$ corresponding to S^k. Assume:

A1 The columns of $\mathbf{X}^k$ are drawn independently according to an absolutely continuous distribution with respect to the Lebesgue measure on S^k.

A2 $\mathbf{X}^k$ has at least $r+1$ columns.

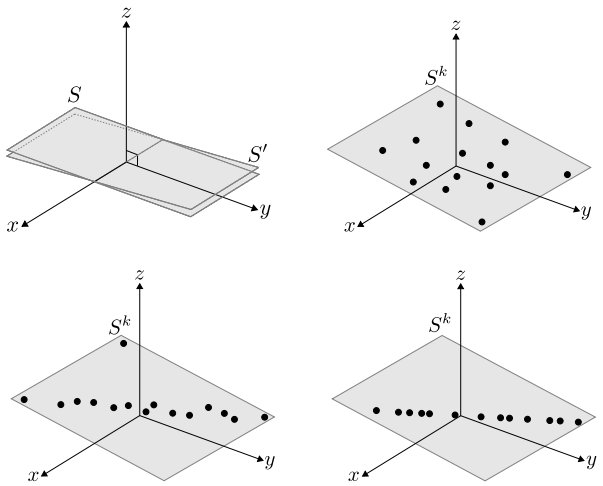

Fig. 3: Typical SC assumptions require (i) that the subspaces are sufficiently separated; this would discard subspaces that are too close, as in the **top-left**, (ii) that the subspaces are sufficiently incoherent; this would discard subspaces that are too aligned with the canonical axes, as in the **top-left**, and (iii) that the columns of $\mathbf{X}^k$ are well-spread over S^k, as in the **top-right**; this would discard cases where the distribution of columns over S^k is skewed, as in the **bottom** (left and right) [10]. In contrast, assumption **A1** allows any collection of subspaces, including nearby and coherent subspaces, as in the **top-left**. **A1** only requires that the columns of $\mathbf{X}^k$ are drawn generically, as in the **top-right** and **bottom-left**. **A1** excludes ill-conditioned samples with Lebesgue measure zero, as in the **bottom-right**, where all columns lie in a line (when S^k is a plane).

A1 essentially requires that the columns in $\mathbf{X}^k$ are drawn generically from S^k. This allows nearby and coherent subspaces, and skewed distributions of the columns. In contrast, typical SC assumptions require that the subspaces are sufficiently separated, that S^k is incoherent (not too aligned with the canonical axes), and that the columns are well-spread over S^k. See Figure 3 to build some intuition. **A2** is a fundamental requirement for subspace clustering, as K sets of r columns can be clustered into K arbitrary r-dimensional subspaces.

Recall that we want to cluster $\mathbf{X}$ using only a few of its rows. The restriction of an r-dimensional subspace in general position to $\ell \leq r$ coordinates is simply $\mathbb{R}^\ell$. So if $\mathbf{X}$ is sampled on r or fewer rows, *any* subspace in general position would agree with *all* the subsampled columns, making clustering impossible. It follows that $\mathbf{X}$ must be sampled on at least $\ell = r+1$ rows in order to be clustered. In other words, $\ell = r+1$ rows are necessary for sketched subspace clustering. We will now show that $\mathbf{X}$ can be clustered using only this bare minimum of rows, i.e., that $\ell = r+1$ is also theoretically sufficient. To this end, we first introduce our new notion of distance between subspaces, which we call *partial coordinate discrepancy*.

Let $[d]^\ell$ denote the collection of all subsets of $\{1,\dots,d\}$ with exactly ℓ distinct elements. Let $\mathrm{Gr}(r,\mathbb{R}^d)$ denote the Grassmann manifold of r-dimensional subspaces in $\mathbb{R}^d$, and let $\mathbb{1}_{\{\cdot\}}$ denote the indicator function. For any subspace,

matrix or vector that is compatible with a set $\boldsymbol{\omega} \in [d]^\ell$, we will use the subscript $\boldsymbol{\omega}$ to denote its restriction to the coordinates/rows in $\boldsymbol{\omega}$. For example, $\mathbf{X}_{\boldsymbol{\omega}} \in \mathbb{R}^{\ell \times n}$ denotes the restriction of $\mathbf{X}$ to the rows in $\boldsymbol{\omega}$, and $S^k_{\boldsymbol{\omega}} \subset \mathbb{R}^\ell$ denotes the restriction of S^k to the coordinates in $\boldsymbol{\omega}$.

Definition 1. *Given $S, S' \in \mathrm{Gr}(r, \mathbb{R}^d)$, define the* partial coordinate discrepancy *between S and S' as:*

$$\delta(S, S') \;:=\; \frac{1}{\binom{d}{r+1}} \sum_{\boldsymbol{\omega} \in [d]^{r+1}} \mathbb{1}_{\{S_{\boldsymbol{\omega}} \neq S'_{\boldsymbol{\omega}}\}}.$$

Example 1. *Consider the following 1-dimensional subspaces:*

$$S = \mathrm{span}\begin{bmatrix}1\\1\\1\\1\end{bmatrix} \qquad S' = \mathrm{span}\begin{bmatrix}1\\1\\-1\\-1\end{bmatrix}.$$

Then $\delta(S, S') = \frac{4}{6}$, because if $\boldsymbol{\omega} = \{1, 2\}$ or $\boldsymbol{\omega} = \{3, 4\}$, then $S_{\boldsymbol{\omega}} = S'_{\boldsymbol{\omega}} = \mathrm{span}[1\ 1]^\mathsf{T}$, but for any of the other 4 choices of $\boldsymbol{\omega}$, $S_{\boldsymbol{\omega}} \neq S'_{\boldsymbol{\omega}}$. In other words, S and S' would appear to be the same if they were only observed on the first two or the last two coordinates/rows. Notice that S and S' are orthogonal (maximally apart with respect to the principal angle distance, the affinity distance, and the subspace incoherence distance [10]), yet they are identical when restricted to certain coordinates.

Remark 1. *Notice that δ takes values in $[0, 1]$. One can interpret δ as the probability that two subspaces are different on $r+1$ coordinates chosen at random. For instance, if two subspaces are drawn independently according to the uniform measure over $\mathrm{Gr}(r, \mathbb{R}^d)$, then with probability 1 they will have $\delta = 1$.*

Example 1 shows that even orthogonal subspaces can appear identical if they are only sampled on a subset of coordinates. Existing measures of distance between subspaces fail to capture this notion of partial coordinate similarity. In contrast, δ is a distance measure (metric) that quantifies the partial coordinate similarity of two subspaces when restricted to subsets of coordinates. We formalize this in the next lemma. The proof is given in Section V.

Lemma 1. *Partial coordinate discrepancy is a metric over $\mathrm{Gr}(r, \mathbb{R}^d)$.*

Lemma 1 implies that two different subspaces must be different on at least one set $\boldsymbol{\omega}$ with $r+1$ coordinates. If subspaces $S, S' \in \mathcal{U}$ are different on $\boldsymbol{\omega}$, then columns corresponding to S and S' can be subspace clustered using only $\mathbf{X}_{\boldsymbol{\omega}}$ by iteratively trying combinations of $r+1$ columns in $\mathbf{X}_{\boldsymbol{\omega}}$. This is because under **A1**, a set of $r+1$ columns in $\mathbf{X}_{\boldsymbol{\omega}}$ will be linearly dependent if and only if they correspond to the same subspace in $\mathcal{U}$. This implies that we can cluster $\mathbf{X}$ using only $r+1$ rows. The challenge is to determine which rows to use. If the subspaces in $\mathcal{U}$ have $\delta = 1$ (i.e., they are different on all subsets of $r+1$ coordinates), then we can cluster $\mathbf{X}$ using any set of $r+1$ rows. But if δ is small, we would need to use *the right* rows, which could be hard to find. This matches the intuition that subspaces that are very similar are harder to cluster.

Fortunately, we will show that generic rotations yield maximal partial coordinate discrepancy. In other words, we will see that if we generically rotate the subspaces in $\mathcal{U}$, then the rotated subspaces will be different on all subsets of $r+1$ coordinates. This will imply that we can cluster $\mathbf{X}$ using *any* $r+1$ rows of a generic rotation of $\mathbf{X}$. To formalize these ideas, let $\boldsymbol{\Gamma} : \mathbb{R}^d \to \mathbb{R}^d$ denote a rotation operator. Assume

A3 The rotation angles of $\boldsymbol{\Gamma}$ are drawn independently according to an absolutely continuous distribution with respect to the Lebesgue measure on $(0, 2\pi)$.

Essentially, **A3** requires that $\boldsymbol{\Gamma}$ is a generic rotation. Equivalently, $\boldsymbol{\Gamma}$ can be considered as a generic $d \times d$ orthonormal matrix. Rotating $\mathbf{X}$ equates to left multiplying it by $\boldsymbol{\Gamma}$. Similarly, the rotation of a subspace S by $\boldsymbol{\Gamma}$ (which we will denote by $\boldsymbol{\Gamma}S$) is given by $\mathrm{span}\{\boldsymbol{\Gamma}\mathbf{U}\}$, where $\mathbf{U}$ is a basis of S. The next lemma states that rotating subspaces by a generic rotation yields subspaces with maximal partial coordinate discrepancy. The proof is given in Section V.

Lemma 2. *Let $\boldsymbol{\Gamma}$ denote a rotation operator drawn according to **A3**. Let S, S' be different subspaces in $\mathrm{Gr}(r, \mathbb{R}^d)$. Then $\delta(\boldsymbol{\Gamma}S, \boldsymbol{\Gamma}S') = 1$ with probability 1.*

Lemma 2 states that regardless of $\delta(S, S')$, we can rotate S and S' to obtain new subspaces with maximal partial coordinate discrepancy (i.e., subspaces that are different on all subsets of $r+1$ coordinates). See Figure 4 for some insight. Intuitively, a generic rotation distributes the local differences of S and S' across all coordinates. So as long as $S \neq S'$, then $(\boldsymbol{\Gamma}S)_{\boldsymbol{\omega}}$ will differ (at least by a little bit) from $(\boldsymbol{\Gamma}S')_{\boldsymbol{\omega}}$ for every $\boldsymbol{\omega} \in [d]^\ell$, with $\ell > r$. This implies that $\boldsymbol{\Gamma}\mathbf{X}$ can be perfectly clustered using *any* subset of $\ell > r$ rows of $\boldsymbol{\Gamma}\mathbf{X}$ (and clustering $\boldsymbol{\Gamma}\mathbf{X}$ is as good as clustering $\mathbf{X}$). This is summarized in our main result, stated in the next theorem. The proof is given in Section V.

Theorem 1. *Let **A1**-**A3** hold, and let $\boldsymbol{\omega} \in [d]^\ell$, with $\ell > r$. Let $\mathbf{X}'$ be a subset of the columns in $\mathbf{X}$. Transform and row-subsample $\mathbf{X}'$ to obtain $(\boldsymbol{\Gamma}\mathbf{X}')_{\boldsymbol{\omega}}$. Then with probability 1, the columns in $\mathbf{X}'$ lie in an r-dimensional subspace of $\mathbb{R}^d$ if and only if the columns in $(\boldsymbol{\Gamma}\mathbf{X}')_{\boldsymbol{\omega}}$ lie in an r-dimensional subspace of $\mathbb{R}^\ell$.*

Theorem 1 states that theoretically, $\mathbf{X}$ can be clustered using *any* $r+1$ rows of a generic rotation $\mathbf{X}' := \boldsymbol{\Gamma}\mathbf{X}$. Under **A1**-**A3**, perfectly clustering $\mathbf{X}'_{\boldsymbol{\omega}}$ is theoretically possible with probability 1 by iteratively trying combinations of $r+1$

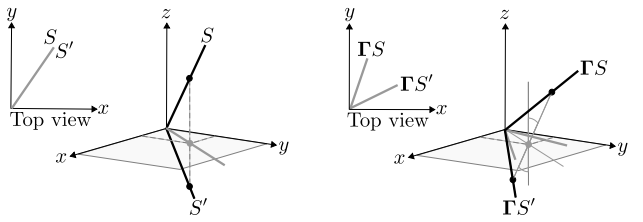

Fig. 4: **Left:** Two different subspaces (even orthogonal) can appear identical if they are only observed on a subset of coordinates. In this figure, S and S' are identical if they are only observed on the (x, y) coordinates (top view). **Right:** Lemma 2 shows that if we rotate S and S' generically, the rotated subspaces $\mathbf{\Gamma}S$ and $\mathbf{\Gamma}S'$ will be different on all subsets of more than r coordinates. In this figure, the rotated subspaces $\mathbf{\Gamma}S$ and $\mathbf{\Gamma}S'$ are different in all sets of $r + 1 = 2$ coordinates, including the (x, y) plane.

columns in $\mathbf{X}'_{\boldsymbol{\omega}}$ and verifying whether they are rank-r. This is because under **A1** and **A3**, a set of $r + 1$ columns in $\mathbf{X}'_{\boldsymbol{\omega}}$ will be linearly dependent if and only if they correspond to the same subspace. Nonetheless, this combinatorial SC algorithm can be computationally prohibitive, especially for large n.

In practice, we can use an algorithm such as sparse subspace clustering (SSC) [23]. This algorithm enjoys state-of-the-art performance, works well in practice, and has theoretical guarantees. The main idea behind SSC is that a column $\mathbf{x}$ in $\mathbf{X}$ lying in subspace S can be written as a linear combination of a few other columns in S (in fact, r or fewer). In contrast, it would require more columns from other subspaces to express $\mathbf{x}$ as their linear combination (as many as d). So SSC aims to find a sparse vector $\mathbf{c} \in \mathbb{R}^{n-1}$, such that $\mathbf{x} = (\mathbf{X}\backslash\mathbf{x})\mathbf{c}$. Here $\mathbf{X}\backslash\mathbf{x}$ denotes the $d \times (n-1)$ matrix formed with all the columns in $\mathbf{X}$ except $\mathbf{x}$. The nonzero entries in $\mathbf{c}$ index columns from the same subspace as $\mathbf{x}$. SSC aims to find such vector $\mathbf{c}$ by solving

$$\underset{\mathbf{c}\in\mathbb{R}^{n-1}}{\arg\min} \ \|\mathbf{c}\|_1 \quad \text{s.t.} \quad \mathbf{x} = (\mathbf{X}\backslash\mathbf{x})\mathbf{c}, \tag{1}$$

where $\|\cdot\|_1$ denotes the 1-norm, given by the sum of absolute values. SSC then uses spectral clustering on these coefficients to recover the clusters.

Unfortunately, the solution to (1) is not exact. In fact, a typical solution to (1) will have most entries close to zero, and only a few (yet more than r) relevant entries. If we only use $\ell = r + 1$ rows, the location of the relevant entries in $\mathbf{c}$ will be somewhat meaningless in the sense that they could correspond to columns from different subspaces, as it takes at most $r + 1$ linearly independent columns to represent a column in $\mathbb{R}^{r+1}$.

As the number of rows ℓ grows, the relevant entries in $\mathbf{c}$ are more likely to correspond to columns from the same subspace as $\mathbf{x}$. On the other hand, as ℓ grows, so does the computational complexity of (1). Without subsampling the rows, the computational complexity of SSC is $\mathcal{O}(dn^3)$. In contrast, using $\ell > r$ rows, the computational complexity of SSC will only be $\mathcal{O}(\ell n^3)$. Depending on d, n and r, this can bring substantial computational improvements. We thus want ℓ to be large enough such that the relevant entries in $\mathbf{c}$ reveal clusters of $\mathbf{X}$, but not so large that (1) is too computationally expensive.

In fact, we know from Wang et al. [24] that SSC will find the correct clustering using only $\ell = \mathcal{O}(r \log(rK^2) + \log n)$ rows if the following conditions hold (see Figure 3 to build some intuition):

(i) The angles between subspaces are sufficiently large.
(ii) The subspaces are sufficiently incoherent with the canonical basis, or the data is transformed by a gaussian projection or by the fast Johnson-Lindenstrauss transform [27].
(iii) The columns of $\mathbf{X}^k$ are well-spread over S^k.

On the other hand, Theorem 1 states that theoretically it is possible to cluster $\mathbf{X}$ using only $\ell = r+1$ rows, in lieu of these conditions. This reveals a gap between theory and practice that we further study in our experiments.

We have shown that theoretically, conditions (i)-(iii) are sufficient but not necessary. It remains an open question whether there exists a polynomial time algorithm that can provably cluster without these requirements.

III. About δ and other distances

In this section we make several remarks about partial coordinate discrepancy and its relation to other distances between subspaces. First recall the definition of principal angle distance between two subspaces [28].

Definition 2 (Principal angle distance)**.** *Let S, S' be subspaces in $\mathrm{Gr}(r, \mathbb{R}^d)$ with orthonormal bases $\mathbf{U}, \mathbf{U}'$. The* principal angle distance *between S and S' is defined as*

$$\theta(S, S') \ := \ \|\mathbf{U}_\perp^\mathsf{T}\mathbf{U}'\|_2,$$

where $\mathbf{U}_\perp$ is an orthonormal basis of $S^\perp$.

It is intuitive that when data are generated from subspaces that are close to one another, it is difficult to cluster these data correctly. Typically, other results use the principal angle distance to measure how *close* subspaces are. For example, in the previous section we discussed that if conditions (i)-(iii) hold, then $\mathcal{O}(r \log(rK^2) + \log n)$ rows are sufficient for clustering [14]. Condition (i) essentially requires that θ is sufficiently large.

The partial coordinate discrepancy δ is an other useful metric. Here we used it to show that theoretically, $r+1$ rows are necessary and sufficient for clustering in lieu of these assumptions. We now wish to compare δ and θ. We will see that subspaces close in one metric can in general be far in the other. We believe this is an important observation for bridging the gap between the sufficient oversampling of the rows required when using θ and the necessary and sufficient condition of Theorem 1.

In our study, we will analyze δ using bases of subspaces, so let us first show that δ shares the important property of being basis independent. To see this, let $\mathbf{U}, \mathbf{U}' \in \mathbb{R}^{d\times r}$ denote bases of S, S'. Notice that $S_{\boldsymbol{\omega}} = S'_{\boldsymbol{\omega}}$ if and only if there exists a matrix $\mathbf{B} \in \mathbb{R}^{r\times r}$ such that $\mathbf{U}'_{\boldsymbol{\omega}} = \mathbf{U}_{\boldsymbol{\omega}}\mathbf{B}$. Now suppose that instead of $\mathbf{U}$, we choose an other basis $\mathbf{V}$ of S. Since $\mathbf{U}$

and $\mathbf{V}$ are both bases of S, there must exists a full-rank matrix $\mathbf{\Theta} \in \mathbb{R}^{r\times r}$ such that $\mathbf{U} = \mathbf{V}\mathbf{\Theta}$. As before, $S_{\boldsymbol{\omega}} = S'_{\boldsymbol{\omega}}$ if and only if there exists a matrix $\mathbf{B}' \in \mathbb{R}^{r\times r}$ such that $\mathbf{U}'_{\boldsymbol{\omega}} = \mathbf{V}_{\boldsymbol{\omega}}\mathbf{B}'$. Now observe that if $\exists\ \mathbf{B}$ such that $\mathbf{U}'_{\boldsymbol{\omega}} = \mathbf{U}_{\boldsymbol{\omega}}\mathbf{B}$, then $\exists\ \mathbf{B}'$ (namely $\mathbf{B}' = \mathbf{\Theta}\mathbf{B}$) such that $\mathbf{U}'_{\boldsymbol{\omega}} = \mathbf{V}\mathbf{B}'$. Similarly, if $\exists\ \mathbf{B}'$ such that $\mathbf{U}'_{\boldsymbol{\omega}} = \mathbf{V}_{\boldsymbol{\omega}}\mathbf{B}'$, then $\exists\ \mathbf{B}$ (namely $\mathbf{B} = \mathbf{\Theta}^{-1}\mathbf{B}'$) such that $\mathbf{U}'_{\boldsymbol{\omega}} = \mathbf{U}_{\boldsymbol{\omega}}\mathbf{B}$.

With this, we can now study the relationship between partial coordinate discrepancy and principal angle distance. The next example shows that two subspaces may be close with respect to θ, but far with respect to δ.

Example 2 (Small θ may coincide with large δ)**.** *Consider a subspace S spanned by $\mathbf{U} \in \mathbb{R}^{d\times r}$. Let $\epsilon > 0$ be given, and let $\mathbf{U}' = \mathbf{U} + \epsilon$. It is easy to see that $\theta(S, S') \to 0$ as $\epsilon \to 0$. In contrast, $\delta(\mathbf{U}, \mathbf{U}') = 1$ for every ϵ.*

Conversely, the next example shows that two subspaces may be close with respect to δ, but far with respect to θ.

Example 3 (Small δ may coincide with large θ)**.** *Consider two subspaces $S, S' \in \mathrm{Gr}(r, \mathbb{R}^d)$ spanned by*

$$\mathbf{U} = \begin{bmatrix}\mathbf{I}\\ \mathbf{I}\\ \mathbf{0}\end{bmatrix} \quad \text{and} \quad \mathbf{U}' = \begin{bmatrix}\mathbf{I}\\ -\mathbf{I}\\ \mathbf{0}\end{bmatrix}.$$

where $\mathbf{I}$ denotes the identity matrix. For $r \ll d$, $\delta(S, S')$ will be close to zero, because the two subspaces differ on only $r+1$ subsets of the first $2r$ coordinates. However, the subspaces are orthogonal and so the principal angle distance is maximal; $\theta(S, S') = 1$.

Examples 2 and 3 show that in general, subspaces close in one metric can be far in the other. However, for subspaces that are incoherent with the canonical axes, there is an interesting relation between δ and θ. Recall that coherence is a parameter indicating how aligned a subspace is with the canonical axes [29]. More precisely,

Definition 3 (Coherence)**.** *Let $S \in \mathrm{Gr}(r, \mathbb{R}^d)$. Let $\mathbf{P}_S$ denote the projection operator onto S, and $\mathbf{e}_i$ the i^{th} canonical vector in $\mathbb{R}^d$. The standard coherence parameter $\mu \in [1, \frac{d}{r}]$ of S is defined as*

$$\mu := \frac{d}{r}\max_{1\le i\le d}\|\mathbf{P}_S\mathbf{e}_i\|_2^2.$$

Intuitively, an incoherent subspace (small μ) will be well-spread over all the canonical directions. Equivalently, the magnitude of the rows of its bases will not vary too much. In this case, if δ is small, we can also expect θ to be small. The following example demonstrates one such scenario.

Example 4 (An example where small δ, μ imply small θ)**.** *Suppose that S and S' are spanned by orthogonal bases $\mathbf{U}, \mathbf{U}'$ respectively. Suppose they have η coordinates on which they span the same subspace; for η close to d, this will result in a small δ. Suppose the coherence for each subspace is bounded by μ_0,* i.e.,

$$\frac{d}{r}\max_{1\le i\le d}\|\mathbf{P}_S\mathbf{e}_i\|_2^2 = \frac{d}{r}\max_{1\le i\le d}\|\mathbf{U}_i\|_2^2 \le \mu_0$$

where $\mathbf{U}_i$ is the i^{th} row of $\mathbf{U}$. Further suppose that if we subsample the basis only on the η coordinates the two subspaces have in common, we can lower bound their inner product:

$$\left\|\sum_{i=1}^{\eta}\mathbf{U}_i^T\mathbf{U}'_i\right\|_2 \ge c_0\frac{\eta}{d}.$$

This is essentially another incoherence condition that will hold with $c_0 \approx 1$ when the subspaces are highly incoherent with the canonical basis. Then

$$\theta(S, S') \le 1 - \left(c_0\frac{\eta}{d} - (d-\eta)\frac{\mu_0 r}{d}\right)^2$$

when $c_0\frac{\eta}{d} - (d-\eta)\frac{\mu_0 r}{d} > 0$.

From this example our intuition is confirmed: if η is very close to d, $c_0 \approx 1$, and μ_0 is constant, the term in the parentheses is near 1 and the angle is small. To see how we get the bound on $\theta(S, S')$, first note that $\theta(S, S') = 1 - \|\mathbf{U}^T\mathbf{U}'\|_2^2$, and we can bound the second term from below.

$$\begin{aligned}\|\mathbf{U}^T\mathbf{U}'\|_2 &= \left\|\sum_{i=1}^{d}\mathbf{U}_i^T\mathbf{U}'_i\right\|_2 = \left\|\sum_{i=1}^{\eta}\mathbf{U}_i^T\mathbf{U}'_i + \sum_{i=\eta+1}^{d}\mathbf{U}_i^T\mathbf{U}'_i\right\|_2\\ &\ge c_0\frac{\eta}{d} - \left\|\sum_{i=\eta+1}^{d}\mathbf{U}_i^T\mathbf{U}'_i\right\|_2 \qquad (2)\\ &\ge c_0\frac{\eta}{d} - \sum_{i=\eta+1}^{d}\|\mathbf{U}_i\|_2\left\|\mathbf{U}'_i\right\|_2 \ge c_0\frac{\eta}{d} - (d-\eta)\frac{\mu_0 r}{d}\end{aligned}$$

where we used the triangle inequality, matrix norm inequality, and step (2) follows by assumption.

This illustrates a case where, if the subspaces in $\mathcal{U}$ have low coherence and their partial coordinate discrepancy is small, the angle between them will also be small.

Existing analyses show that practical SC algorithms tend to fail if θ is small [23]. It follows that for incoherent subspaces, if δ is small, SC can be very hard in practice. This is illustrated in Figure 5, which shows that the clustering performance of practical algorithms declines as δ decreases.

IV. Experiments

Theorem 1 shows that one can cluster $\mathbf{X}$ using only $r+1$ rows of $\mathbf{\Gamma}\mathbf{X}$. As discussed in Section II, practical algorithms like SSC may require more than these bare minimum number of rows. In this section we present experiments to study the gap between what is theoretically possible and what is practically possible with state-of-the-art algorithms.

In Section III we also explained that for incoherent subspaces, the partial coordinate discrepancy δ and the principal angle distance θ have a tight relation: if δ is small, then θ is small too. Existing analyses show that practical SC algorithms tend to fail if θ is small [23]. It follows that for incoherent subspaces, if δ is small, SC can be very hard in practice. The experiments of this section support these results.

In our experiments, we will compare the following approaches to subspace clustering:

(a) Cluster $\mathbf{X}$ directly (full-data).
(b) Cluster $\ell > r$ rows of $\mathbf{\Gamma}\mathbf{X}$.

To compare things vis-à-vis, we will study the cases above using the sparse subspace clustering (SSC) algorithm [23]. We chose SSC because it enjoys state-of-the-art performance, works well in practice, and has theoretical guarantees. In all our experiments we use the SSC code provided by their authors [23].

A. Simulations

We will first use simulations to study the cases above as a function of the ambient dimension d, the partial coordinate discrepancy δ of the subspaces in $\mathcal{U}$, and the number of rows used ℓ. To obtain subspaces with a specific δ, we first generated a $d \times r$ matrix $\mathbf{V}$ with entries drawn i.i.d. from the standard Gaussian distribution. Subspaces generated this way have low coherence. Then, for $k = 1, \dots, K$, we selected the k^{th} set of δ' rows in $\mathbf{V}$ (i.e., rows $(k-1)\delta'+1, \dots, k\delta'$) and replaced them with other entries, also drawn i.i.d. from the standard Gaussian distribution. This yields K bases, which will span the subspaces in $\mathcal{U}$. This way, the bases of any S and S' in $\mathcal{U}$ will differ on exactly $2\delta'$ rows. It follows that $\delta(S, S')$ is equal to the probability of selecting any of these $2\delta'$ rows in $r+1$ draws (without replacement). That is,

$$\delta(S, S') \;=\; 1 - \frac{\binom{d-2\delta'}{r+1}}{\binom{d}{r+1}} \qquad \text{for every } S, S' \in \mathcal{U}. \tag{3}$$

Unfortunately, (3) gives little intuition of how small or large δ is. We will thus upper bound δ by a small number that is easily interpretable. To do this, we will use the next simple bound, which gives a clear idea of how small δ is in our experiments. A derivation is given in Section V.

$$\delta(S, S') \leq \frac{(r+1)(2\delta' - r)}{d - r} = \mathcal{O}\left(\frac{r\delta'}{d}\right). \tag{4}$$

In each trial of our experiments, we generated a set $\mathcal{U}$ of $K = 5$ subspaces, each of dimension $r = 5$, using the procedure described above. Next we generated a matrix $\mathbf{X}$ with $n_k = 100$ columns from each subspace. The coefficients of each column in $\mathbf{X}$ are drawn i.i.d. from the standard Gaussian distribution. Matrices generated this way satisfy **A1** and **A2**. To measure accuracy, we find the best matching between the identified clusters and the original sets.

In our first simulation we study the dependency on δ' (which gives a proxy of δ through (4)) and ℓ, with $d = 10^5$ fixed. The results are summarized in Figure 5 (top-left). This figure shows the gap between theory and practice. Theorem 1 shows that theoretically, all these trials can be perfectly clustered. This figure shows, as predicted in Section III, that for incoherent subspaces, clustering becomes harder in practice as δ' (and hence δ) shrinks. Observe that as δ' grows, fewer rows suffice for accurate clustering. For example, in this experiment, SSC consistently succeeds with $\ell = \delta'$.

Next we study the cases above as a function of d and δ', with $\ell = \delta'$. The results are summarized in Figure 5 (top-right). This also shows a gap between theory and practice. Figure 5 shows, as predicted in Section III, that for incoherent subspaces, if δ' (and hence δ) is too small, the angle between the subspaces in $\mathcal{U}$ will be small, whence clustering can be hard in practice. In this experiment, we also record the minimum δ' and ℓ required for a clustering accuracy of at least 95%. Figure 5 (bottom-left) shows that for incoherent subspaces, as d grows, one only requires a vanishing partial coordinate discrepancy δ and a vanishing fraction of rows ℓ/d to succeed.

In our last simulation we study the computation time required required to cluster $\mathbf{X}$ directly (full-data), and to cluster $\ell = 20$ rows of $\mathbf{\Gamma X}$ as a function of d. In this experiment, we fix $\ell = \delta' = 20$, known from our previous experiment to produce 100% accuracy for a wide range of d. Unsurprisingly, Figure 5 (bottom-right) shows that if

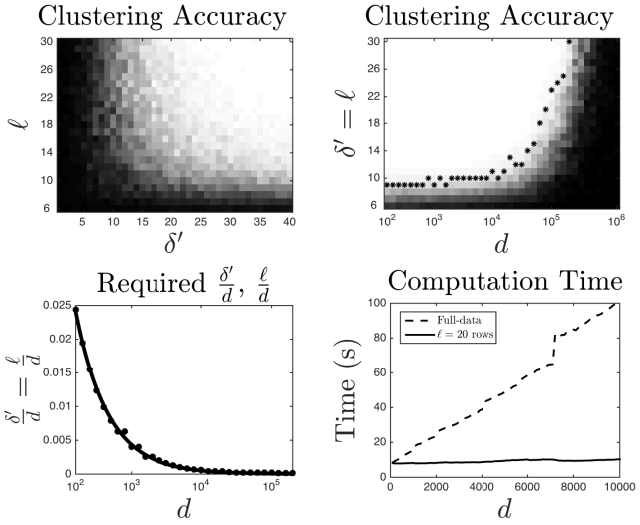

Fig. 5: Proportion of correctly classified points by SSC, using only $\ell > r$ rows of $\mathbf{\Gamma X}$, with $K = 5$ subspaces, each of dimension $r = 5$, and $n_k = 100$ columns per subspace. The color of each pixel indicates the average over 100 trials (the lighter the better). White represents 100% accuracy, and black represents 20%, which amounts to random guessing. Theorem 1 states that theoretically, all these trials can be perfectly clustered. This shows a gap between theory and practice. **Top-Left:** Transition diagram as a function of δ' (which gives a proxy of the partial coordinate discrepancy δ through (4)), and the number of used rows ℓ, with fixed ambient dimension $d = 10^5$. As discussed in Section III, for incoherent subspaces, clustering becomes harder in practice as δ' shrinks. Observe that as δ' grows, fewer rows suffice for accurate clustering. **Top-Right:** Transition diagram as a function of d and δ', using only $\ell = \delta'$ rows. All pixels above the black point in each column have at least 95% accuracy. These points represent the minimum δ' and ℓ required for a clustering accuracy of at least 95%. As discussed in Section III, for incoherent subspaces, if δ' (and hence δ) is too small, the angle between the subspaces in $\mathcal{U}$ will be too small, whence clustering can be hard in practice. **Bottom-Left:** Partial coordinate discrepancy δ (upper bounded by $\leq \mathcal{O}(r\delta'/d)$) and fraction of rows ℓ/d required by 3SC for a clustering accuracy of at least 95%. The curve is the best exponential fit to these points. This curve represents the discriminant between 95% accuracy (above curve) and less than 95% accuracy (below curve). This shows that for incoherent subspaces, as d grows, one only requires a vanishing partial coordinate discrepancy δ and a vanishing fraction of rows ℓ/d to succeed. **Bottom-Right:** Time required to cluster $\mathbf{X}$ directly (full-data), and to cluster $\ell = 20$ rows of $\mathbf{\Gamma X}$ as a function of the ambient dimension d (average over 100 trials). In all of these trials, both options achieve 100% accuracy.

we only use a constant number of rows, the computation time is virtually unaffected by the ambient dimension, unlike standard (full-data) algorithms. This can thus bring computational complexity orders of magnitude lower (depending on d and n) than standard (full-data) techniques.

B. Real Data

We now evaluate the performance of sketching on a real life problem where the phenomenon of partial coordinate similarity arises naturally: classifying faces. To this end we use the Extended Yale B dataset [26], which consists of face images of 38 individuals with a fixed pose under varying illumination (see Figure 6). As discussed in [23], shadows and specularities in these images can be modeled as sparse errors. So as a preprocessing step, we first apply the augmented Lagrange multiplier method [30] for robust principal component analysis on the images of each individual (using code provided by the authors). This will remove the sparse errors, so that the vectorized images of each individual lie near a 9-dimensional subspace [25]. Hence, the matrix $\mathbf{X}$ containing all the vectorized images images lies near a union of 38, 9-dimensional subspaces.

Observe that these images are very similar on several regions. For example, the lower corners are mostly dark. Distinct subspaces can thus appear to be the same if they are only observed on the coordinates corresponding to these pixels. If we only use a few rows of $\mathbf{X}$ (without rotating), there is a positive probability of selecting these coordinates. In this case, we would be unable to determine the right clustering. Fortunately, Lemma 2 shows that the columns of a generic rotation of $\mathbf{X}$ will lie near a union of subspaces that will be different on all subsets of $\ell > r$ coordinates (maximal partial coordinate discrepancy). This implies, as shown in Theorem 1, that the clusters of the original $\mathbf{X}$ will be the same as the clusters of any $\ell > r$ rows of the rotated $\mathbf{X}$. This means that we can cluster $\mathbf{X}$ using any $\ell > r$ coordinates of a rotation of $\mathbf{X}$. This is verified by the following experiment.

In this experiment we study classification accuracy as a function of the number of individuals, or equivalently the number of subspaces K, and as a function of the number of rows ℓ used for clustering. We do this replicating the experiment in [23]: we first divide all individuals into four groups, corresponding to individuals $\{1,\dots,10\}$, $\{11,\dots,20\}$, $\{21,\dots,30\}$ and $\{31,\dots,38\}$. Next we cluster all possible choices of $K \in \{2,3,5,8,10\}$ individuals for the first three groups, and $K \in \{2,3,5,8\}$ individuals for the last group. We repeat this experiment for different choices of ℓ, and record the classification accuracy. The results are summarized in Figure 6. They show that one can achieves the same performance as standard (full-data) methods, using only a small fraction of the data. This results in computational advantages (time and memory).

V. Proofs

In this section we give the proofs of all our statements.

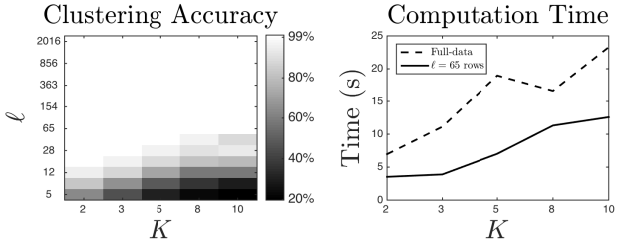

Fig. 6: **Left:** Proportion of correctly classified images from the Extended Yale B dataset [26] (see Figure 2), as a function of the number of individuals, or equivalently the number of subspaces K, and as a function of the number of rows ℓ used for clustering. In particular, $\ell = d = 2016$ corresponds to standard (full-data) SSC. **Right:** Computation time as a function of the number of individuals K, with $\ell = 65$ fixed (known from the center figure to achieve the same accuracy as standard SSC). Recall that the computational complexities of SSC and sketching are $\mathcal{O}(dn^3)$ and $\mathcal{O}(\ell n^3)$, respectively. Here $d = 2016$ and $n = 38K$. This shows that sketching achieves the same accuracy as standard SSC in only a fraction of the time. This gap becomes more evident as d and n grow, as shown in Figure 5.

Proof of Lemma 1

We need to show that δ satisfies the three properties of a metric. Let $S, S', S'' \in \mathrm{Gr}(r, \mathbb{R}^d)$.

(i) It is easy to see that if $S = S'$, then $\delta(S,S') = 0$. To obtain the converse, suppose $\delta(S,S') = 0$. Let $\boldsymbol{\upsilon} = \{1,\dots,r\}$, and let $\boldsymbol{\omega}_i = \boldsymbol{\upsilon} \cup i$, with $i = r+1,\dots,d$. Take bases $\mathbf{U}, \mathbf{U}'$ of S, S', such that $\mathbf{U}_{\boldsymbol{\omega}_1} = \mathbf{U}'_{\boldsymbol{\omega}_1}$. We can do this because $\delta(S,S') = 0$, which implies $S_{\boldsymbol{\omega}} = S'_{\boldsymbol{\omega}}$ for every $\boldsymbol{\omega} \in [d]^{r+1}$, including $\boldsymbol{\omega}_1$. Next observe that for $i = r+2,\dots,d$, since $S_{\boldsymbol{\omega}_i} = S'_{\boldsymbol{\omega}_i}$ and $\mathbf{U}_{\boldsymbol{\upsilon}} = \mathbf{U}'_{\boldsymbol{\upsilon}}$, it must be that $\mathbf{U} = \mathbf{U}'$ on the i^{th} row (otherwise $S_{\boldsymbol{\omega}_i} \neq S'_{\boldsymbol{\omega}_i}$). We thus conclude that $\mathbf{U} = \mathbf{U}'$, which implies $S = S'$.

(ii) That $\delta(S,S') = \delta(S,S')$ follows immediately from the definition.

(iii) To see that δ satisfies the triangle inequality, write:

$$\begin{aligned}\delta(S,S') + \delta(S',S'') &= \frac{1}{\binom{d}{r+1}} \sum_{\boldsymbol{\omega}\in[d]^{r+1}} \left(\mathbb{1}_{\{S_{\boldsymbol{\omega}} \neq S'_{\boldsymbol{\omega}}\}} + \mathbb{1}_{\{S'_{\boldsymbol{\omega}} \neq S''_{\boldsymbol{\omega}}\}}\right) \\ &\geq \frac{1}{\binom{d}{r+1}} \sum_{\boldsymbol{\omega}\in[d]^{r+1}} \mathbb{1}_{\{S_{\boldsymbol{\omega}} \neq S'_{\boldsymbol{\omega}} \cup S'_{\boldsymbol{\omega}} \neq S''_{\boldsymbol{\omega}}\}} \\ &\geq \frac{1}{\binom{d}{r+1}} \sum_{\boldsymbol{\omega}\in[d]^{r+1}} \mathbb{1}_{\{S_{\boldsymbol{\omega}} \neq S''_{\boldsymbol{\omega}}\}} = \delta(S,S''),\end{aligned}$$

where the last inequality follows because $\{S = S' \cap S' = S''\}$ implies $\{S = S''\}$, whence $\mathbb{1}_{\{S_{\boldsymbol{\omega}} \neq S'_{\boldsymbol{\omega}} \cup S'_{\boldsymbol{\omega}} \neq S''_{\boldsymbol{\omega}}\}} = \mathbb{1}_{\{S_{\boldsymbol{\omega}} \neq S''_{\boldsymbol{\omega}}\}} = 0$, and in any other case, $\mathbb{1}_{\{S_{\boldsymbol{\omega}} \neq S'_{\boldsymbol{\omega}} \cup S'_{\boldsymbol{\omega}} \neq S''_{\boldsymbol{\omega}}\}} = 1 \geq \mathbb{1}_{\{S_{\boldsymbol{\omega}} \neq S''_{\boldsymbol{\omega}}\}}$. □

Proof of Lemma 2

We need to show that if $S \neq S'$, then $(\boldsymbol{\Gamma} S)_{\boldsymbol{\omega}} \neq (\boldsymbol{\Gamma} S')_{\boldsymbol{\omega}}$ for every $\boldsymbol{\omega} \in [d]^{r+1}$. Let $\mathbf{U}$ and $\mathbf{U}'$ denote bases of S and S'. Observe that $(\boldsymbol{\Gamma} S)_{\boldsymbol{\omega}} = (\boldsymbol{\Gamma} S')_{\boldsymbol{\omega}}$ if and only if there exists a matrix $\mathbf{B} \in \mathbb{R}^{r\times r}$ such that $(\boldsymbol{\Gamma}\mathbf{U}')_{\boldsymbol{\omega}} = (\boldsymbol{\Gamma}\mathbf{U})_{\boldsymbol{\omega}}\mathbf{B}$, or equivalently, if and only if $\boldsymbol{\Gamma}_{\boldsymbol{\omega}}\mathbf{U}' = \boldsymbol{\Gamma}_{\boldsymbol{\omega}}\mathbf{U}\mathbf{B}$, which we can rewrite as

$$\boldsymbol{\Gamma}_{\boldsymbol{\omega}}(\mathbf{U}' - \mathbf{U}\mathbf{B}) = \mathbf{0}. \tag{5}$$

Let $\boldsymbol{\upsilon}$ denote the subset with the first r elements in $\boldsymbol{\omega}$, and i denote the last element in $\boldsymbol{\omega}$. Then we can rewrite (5) as

$$\begin{bmatrix} \boldsymbol{\Gamma}_{\boldsymbol{\upsilon}} \\ \boldsymbol{\Gamma}_i \end{bmatrix} (\mathbf{U}' - \mathbf{U}\mathbf{B}) \;=\; \mathbf{0}. \tag{6}$$

Since $\boldsymbol{\Gamma}$ is drawn according to **A3**, the rows in $\boldsymbol{\Gamma}_{\boldsymbol{\upsilon}}$ are linearly independent with probability 1. Since $\mathbf{U}$ is a basis of an r-dimensional subspace, its r columns are also linearly independent. It follows that $\boldsymbol{\Gamma}_{\boldsymbol{\upsilon}}\mathbf{U}$ is a full-rank $r\times r$ matrix. So we can use the top block in (6) to obtain $\mathbf{B} = (\boldsymbol{\Gamma}_{\boldsymbol{\upsilon}}\mathbf{U})^{-1}\boldsymbol{\Gamma}_{\boldsymbol{\upsilon}}\mathbf{U}'$. We can plug this in the bottom part of (6) to obtain

$$\boldsymbol{\Gamma}_i(\mathbf{U}' - \mathbf{U}(\boldsymbol{\Gamma}_{\boldsymbol{\upsilon}}\mathbf{U})^{-1}\boldsymbol{\Gamma}_{\boldsymbol{\upsilon}}\mathbf{U}') \;=\; \mathbf{0}. \tag{7}$$

Recall that $(\boldsymbol{\Gamma}_{\boldsymbol{\upsilon}}\mathbf{U})^{-1} = (\boldsymbol{\Gamma}_{\boldsymbol{\upsilon}}\mathbf{U})^{\ddagger}/|\boldsymbol{\Gamma}_{\boldsymbol{\upsilon}}\mathbf{U}|$, where $(\boldsymbol{\Gamma}_{\boldsymbol{\upsilon}}\mathbf{U})^{\ddagger}$ and $|\boldsymbol{\Gamma}_{\boldsymbol{\upsilon}}\mathbf{U}|$ denote the adjugate and the determinant of $\boldsymbol{\Gamma}_{\boldsymbol{\upsilon}}\mathbf{U}$. Therefore, we may rewrite (7) as the following system of r polynomial equations:

$$\boldsymbol{\Gamma}_i(|\boldsymbol{\Gamma}_{\boldsymbol{\upsilon}}\mathbf{U}|\mathbf{U}' - \mathbf{U}(\boldsymbol{\Gamma}_{\boldsymbol{\upsilon}}\mathbf{U})^{\ddagger}\boldsymbol{\Gamma}_{\boldsymbol{\upsilon}}\mathbf{U}') \;=\; \mathbf{0}. \tag{8}$$

Observe that the left-hand side of (8) is just an other way to write $\boldsymbol{\Gamma}_i(\mathbf{U}' - \mathbf{U}\mathbf{B})$, where $\mathbf{B}$ is in terms of $\mathbf{U}, \mathbf{U}'$ and $\boldsymbol{\Gamma}_{\boldsymbol{\upsilon}}$. Since $S \neq S'$, there exists no $\mathbf{B} \in \mathbb{R}^{r\times r}$ such that $\mathbf{U}' = \mathbf{U}\mathbf{B}$. Equivalently, $(\mathbf{U}' - \mathbf{U}\mathbf{B}) \neq \mathbf{0}$. Since $\boldsymbol{\Gamma}$ is drawn according to **A3**, we conclude that the left hand side of (8) is a nonzero set of polynomials, and so (8) holds with probability zero.

Since $(\boldsymbol{\Gamma}S)_{\boldsymbol{\omega}} = (\boldsymbol{\Gamma}S')_{\boldsymbol{\omega}}$ if and only if (8) holds, we conclude that with probability 1, $(\boldsymbol{\Gamma}S)_{\boldsymbol{\omega}} \neq (\boldsymbol{\Gamma}S')_{\boldsymbol{\omega}}$. Since $\boldsymbol{\omega}$ was arbitrary, we conclude that this is true for every $\boldsymbol{\omega} \in [d]^{r+1}$, as desired. □

Proof of Theorem 1

Recall that $\mathbf{X}^k$ denotes the matrix formed with all the columns in $\mathbf{X}$ corresponding to the k^{th} subspace in $\mathcal{U}$. Under **A1**-**A2**, with probability 1 the partition $\{\mathbf{X}^k\}_{k=1}^K$ is the only way to cluster the columns in $\mathbf{X}$ into K r-dimensional subspaces. This is because under **A1**, the columns in $\mathbf{X}$ will lie on intersections of the subspaces in $\mathcal{U}$ with probability zero. So any combination of more than r columns from different subspaces in $\mathcal{U}$ will lie in a subspace of dimension greater than r with probability 1.

Recall that $[d]^{\ell}$ denotes the set of all subsets of $\{1,\dots,d\}$ with exactly ℓ distinct elements, and that $\boldsymbol{\Gamma}$ denotes a generic rotation drawn according to **A3**. Let $\boldsymbol{\omega} \in [d]^{\ell}$, and define $(\boldsymbol{\Gamma}\mathcal{U})_{\boldsymbol{\omega}}$ as the set of rotated subspaces in $\mathcal{U}$, restricted to the coordinates in $\boldsymbol{\omega}$, i.e., $(\boldsymbol{\Gamma}\mathcal{U})_{\boldsymbol{\omega}} := \{(\boldsymbol{\Gamma}S^k)_{\boldsymbol{\omega}}\}_{k=1}^K$. Lemma 2 implies that all the subspaces in $(\boldsymbol{\Gamma}\mathcal{U})_{\boldsymbol{\omega}}$ are different. It is easy to see that the columns in $(\boldsymbol{\Gamma}\mathbf{X}^k)_{\boldsymbol{\omega}}$ lie in $(\boldsymbol{\Gamma}S^k)_{\boldsymbol{\omega}}$. By **A1** and **A3**, the columns in $(\boldsymbol{\Gamma}\mathbf{X})_{\boldsymbol{\omega}}$ will lie on intersections of the subspaces in $(\boldsymbol{\Gamma}\mathcal{U})_{\boldsymbol{\omega}}$ with probability zero. So any combination of more than r columns from different subspaces in $(\boldsymbol{\Gamma}\mathcal{U})_{\boldsymbol{\omega}}$ will lie in a subspace of dimension greater than r with probability 1. □

Derivation of (4)

We want to show that

$$\delta(S,S') \leq \frac{(r+1)(2\delta' - r)}{d-r}.$$

Recall that $\delta(S,S')$ is the probability that S and S' are different on a set of $r+1$ coordinates selected uniformly at random (without replacement). In the setup of Section IV, the bases $\mathbf{U}, \mathbf{U}'$ of S, S' are different on exactly $2\delta'$ rows. Then

$$\begin{aligned}
\delta(S,S') \;&=\; \mathsf{P}(1 \text{ out of } 2\delta' \text{ rows in } r+1 \text{ draws}) \\
&=\; \mathsf{P}\left(\bigcup_{\tau=1}^{r+1}\{1 \text{ out of } 2\delta' \text{ rows in } \tau^{\text{th}} \text{ draw}\}\right) \\
&\overset{(a)}{\leq} \sum_{\tau=1}^{r+1} \mathsf{P}(1 \text{ out of } 2\delta' \text{ rows in } \tau^{\text{th}} \text{ draw}) \\
&\overset{(b)}{=} \sum_{\tau=1}^{r+1}\sum_{\rho=0}^{\tau-1} \mathsf{P}(1 \text{ out of } 2\delta' \text{ rows in } \tau^{\text{th}} \text{ draw} \mid \\
&\qquad \rho \text{ out of } 2\delta' \text{ rows in first } \tau-1 \text{ draws})\cdot \\
&\qquad \mathsf{P}(\rho \text{ out of } 2\delta' \text{ rows in first } \tau-1 \text{ draws}) \\
&\overset{(c)}{=} \sum_{\tau=1}^{r+1}\sum_{\rho=0}^{\tau-1} \frac{2\delta'-r}{d-r}\cdot \\
&\qquad \mathsf{P}(\rho \text{ out of } 2\delta' \text{ rows in first } \tau-1 \text{ draws}),
\end{aligned}$$

where (a) follows by the union bound, (b) follows by the law of total probability, and (c) follows because the probability of selecting one of the $2\delta'$ distinct rows in the τ^{th} draw (without replacement) is smallest if $\tau = r+1$ and $\rho = r$, which corresponds to the case where the ratio (after r draws) of distinct rows $(2\delta'-r)$ versus equal rows $(d-r)$ is smallest. Continuing with the last equation, we have:

$$\begin{aligned}
\delta(S,S') \;&\leq\; \frac{2\delta'-r}{d-r}\sum_{\tau=1}^{r+1} \\
&\qquad \underbrace{\sum_{\rho=0}^{\tau-1} \mathsf{P}(\rho \text{ out of } 2\delta' \text{ rows in first } \tau-1 \text{ draws})}_{=1} \\
&\leq\; \frac{(r+1)(2\delta'-r)}{d-r},
\end{aligned}$$

as desired. □

VI. Acknowledgements

Work by L. Balzano was supported by ARO Grant W911NF1410634.

References

[1] R. Vidal, *Subspace clustering*, IEEE Signal Processing Magazine, 2011.

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

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

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

[5] W. Hong and J. Wright and K. Huang and Y. Ma, *Multi-scale hybrid linear models for lossy image representation*, IEEE Transactions on Image Processing, 2006.

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

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

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

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

[10] M. Soltanolkotabi and E. Candès, *A geometric analysis of subspace clustering with outliers*, Annals of Statistics, 2012.

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

[12] C. Qu and H. Xu, *Subspace clustering with irrelevant features via robust Dantzig selector*, Advances in Neural Information Processing Systems, 2015.

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

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

[15] Y. Wang, Y.-X. Wang and A. Singh, *Differentially private subspace clustering*, Advances in Neural Information Processing Systems, 2015.

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

[17] L. Balzano, B. Recht and R. Nowak, *High-dimensional matched subspace detection when data are missing*, IEEE International Symposium on Information Theory, 2010.

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

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

[20] D. Pimentel-Alarcón and R. Nowak, *The information-theoretic requirements of subspace clustering with missing data*, International Conference on Machine Learning, 2016.

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

[22] J. He, L. Balzano and A. Szlam, *Incremental gradient on the Grassmannian for online foreground and background separation in subsampled video*, Conference on Computer Vision and Pattern Recognition, 2012.

[23] E. Elhamifar and R. Vidal, *Sparse subspace clustering: algorithm, theory, and applications*, IEEE Transactions on Pattern Analysis and Machine Intelligence, 2013.

[24] Y. Wang, Y.-X. Wang and A. Singh, *A deterministic analysis of noisy sparse subspace clustering for dimensionality-reduced data*, International Conference on Machine Learning, 2015.

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

[26] K. Lee, J. Ho and D. Kriegman, *Acquiring linear subspaces for face recognition under variable lighting*, IEEE Transactions on Pattern Analysis and Machine Intelligence, 2005.

[27] N. Ailon and B. Chazelle, *The fast Johnson-Lindenstrauss transform and approximate nearest neighbors*, SIAM Journal on Computing, 2009.

[28] G. Golub and C. Loan, *Matrix Computations*, The Johns Hopkins University Press, 3rd edition, 1996.

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

[30] Z. Lin, R. Liu and Z. Su, *Linearized alternating direction method with adaptive penalty for low rank representation*, Advances in Neural Information Processing Systems, 2011.