Robust Subspace Clustering via Thresholding

Reinhard Heckel and Helmut Bölcskei
Dept. of IT & EE, ETH Zurich, Switzerland

Abstract

The problem of clustering noisy and incompletely observed high-dimensional data points into a union of low-dimensional subspaces and a set of outliers is considered. The number of subspaces, their dimensions, and their orientations are assumed unknown. We propose a simple low-complexity clustering algorithm based on thresholding correlations between the data points followed by spectral clustering. A statistical performance analysis shows that this algorithm succeeds even when the subspaces intersect and when the dimensions of the subspaces scale (up to a log-factor) linearly in the ambient dimension. Moreover, we prove that the algorithm succeeds even when the data points are incompletely observed with the number of missing observations scaling (up to a log-factor) linearly in the ambient dimension. The algorithm is shown to be robust in the presence of structured (i.e., sparse) and of unstructured additive noise. Our results reveal an explicit tradeoff between the affinity of the subspaces and the tolerable noise level. Finally, we propose a simple scheme that provably detects outliers, and present numerical results on real and synthetic data.

1 Introduction

One of the major challenges in modern data analysis is to extract relevant information from large high-dimensional data sets. The relevant features in high-dimensional data sets are often of limited complexity, or, more specifically, have low-dimensional structure. For example, images of faces are high-dimensional as the number of pixels is typically large, whereas the set of images of a given face under varying illumination conditions approximately lies in a 9-dimensional linear subspace [1]. This and similar findings for other types of data have motivated research on finding low-dimensional representations of high-dimensional data [33]. A prevalent low-dimensional structure is that of data points lying in a union of (unknown) low-dimensional subspaces. The problem of finding the assignments of the data points to these subspaces is known as subspace clustering [33] or hybrid linear modeling. An example application of subspace clustering is the following: Given a set of images of faces, cluster the images such that each of the resulting clusters corresponds to a single person [13]. Other application areas include unsupervised learning, image representation and segmentation [14], computer vision, specifically motion segmentation [34, 25], and disease detection [16]; we refer to [33] and the references therein for a more complete list.

Often the data available is corrupted by noise and contains outliers. The general subspace clustering problem we consider can therefore be formulated as follows. Suppose we are given a set of $N$ data points in $\mathbb{R}^m$, denoted by $\mathcal{X}$, and assume that

$$\mathcal{X} = \mathcal{X}_1 \cup \ldots \cup \mathcal{X}_L \cup \mathcal{O},$$
where \( O \) denotes a set of outliers and the \( n_l := |X_l| \) points in \( X_l, l \in [L] \), are given by

\[
x_{j}^{(l)} = y_{j}^{(l)} + e_{j}^{(l)},
\]

where \( e_{j}^{(l)} \in \mathbb{R}^m \) is noise and \( y_{j}^{(l)} \in S_l \) with \( S_l \) a \( d_l \)-dimensional subspace of \( \mathbb{R}^m \). The association of the points in \( \mathcal{X} \) with the \( X_l \) and \( O \), the number of subspaces \( L \), their dimensions \( d_l \), and their orientations are all unknown. We want to cluster the (noisy) points in \( \mathcal{X} \), i.e., find the assignments of the points in \( \mathcal{X} \) to the sets \( X_1, ..., X_L, O \). Once these associations have been identified, it is straightforward to extract approximations (recall that we have access to noisy observations only) of the subspaces \( S_l \) through principal component analysis (PCA).

Numerous approaches to subspace clustering have been proposed in the literature, including algebraic, statistical, and spectral clustering methods; we refer to [33] for an excellent overview. Spectral clustering methods (see [36] for an introduction) have found particularly widespread use thanks to their excellent performance properties and efficient implementations. At the heart of spectral clustering lies the construction of an adjacency matrix \( A \in \mathbb{R}^{N \times N} \), where the \((i,j)\)th entry of \( A \) measures the similarity between the data points \( x_i, x_j \in \mathcal{X} \). A typical measure of similarity is, e.g., \( e^{-\text{dist}(x_i, x_j)} \), where \( \text{dist}(\cdot, \cdot) \) is some distance measure [33]. The association of the points in \( \mathcal{X} \) to the subspaces \( S_l \) is obtained by finding the connected components in the graph \( G \) which adjacency matrix \( A \). This is accomplished via a singular value decomposition of the Laplacian of \( G \) followed by \( k \)-means clustering [36].

As noted in [28], there are very few algorithms that are computationally tractable and succeed provably under non-restrictive conditions (such as overlapping subspaces). A notable exception is the sparse subspace clustering (SSC) algorithm by Elhamifar and Vidal [6, 7], which applies spectral clustering to an adjacency matrix \( A \) whose clever construction is inspired by ideas from sparse signal recovery. SSC provably succeeds, in the noiseless case, under very general conditions, as shown in [27] via an elegant (geometric function) analysis. Most importantly, the statistical analysis in [27] reveals that SSC succeeds even when the subspaces \( S_l \) intersect (the linear subspaces \( S_l \) and \( S_k \) are said to intersect if \( S_l \cap S_k \neq \{0\} \)).

Analytical performance results for subspace clustering in the presence of noise are even more scarce. Vidal noted in [33] that “the development of theoretically sound algorithms [...] in the presence of noise and outliers is a very important open challenge”. A significant step towards addressing this challenge was reported in [28], posted while [11] was submitted. The robust SSC (RSSC) algorithm in [28] essentially replaces the \( \ell_1 \)-minimization steps in SSC by \( \ell_1 \)-penalized least squares, i.e., LASSO, steps and provably clusters data points corrupted by Gaussian noise under quite general conditions on the orientations of the subspaces \( S_l \) and on the number of points in the subspaces. To construct the adjacency matrix \( A \), SSC for the noiseless and RSSC for the noisy case require the solution of \( N \ \ell_1 \)-minimization and \( N \) LASSO instances, respectively, each in \( N \) variables; this poses significant computational challenges for large data sets.

**Contributions** We introduce a simple, computationally efficient, subspace clustering algorithm, which applies spectral clustering to an adjacency matrix \( A \) obtained by thresholding correlations between the data points in \( \mathcal{X} \). Specifically, \( A \) is constructed from the \( q \) nearest neighbors of each data point in \( \mathcal{X} \), with correlation serving as the underlying distance measure. The resulting algorithm is termed thresholding based subspace clustering (TSC).

The gist of the results we obtain is that TSC provably succeeds (even when the data is noisy or incompletely observed) if the subspaces are sufficiently distinct and if \( \mathcal{X} \) contains sufficiently many points from each subspace. Specifically, under additive Gaussian noise, TSC is shown to
succeed under very general conditions on the relative orientations of the subspaces and the number of points from each subspace contained in $\mathcal{X}$. In particular, the subspaces are allowed to intersect. Moreover, in the noiseless case, TSC is shown to succeed even when the dimensions of the subspaces scale (up to a log-factor) linearly in the ambient dimension. While SSC for the noiseless case and RSSC for the noisy case share these desirable properties, TSC is computationally much less demanding. Specifically, the construction of the adjacency matrix $A$ in the TSC algorithm requires the computation of $N^2$ inner products followed by thresholding only, as opposed to solving $N$ $\ell_1$-minimization and $N$ LASSO instances, each in $N$ variables, for SSC and RSSC, respectively.

Our results for the case of noisy data nicely reflect the intuition that the more distinct the orientations of the subspaces, the more noise TSC tolerates. What is more, TSC can succeed even under massive noise on the data, provided the subspaces are sufficiently low-dimensional. We also provide results which show that TSC provably succeeds under the influence of sparse noise, which models, e.g., corruptions or occlusions of images in face recognition [37], or mistracking of feature points in motion segmentation [25].

In practical applications the data points to be clustered are often incompletely observed, due to, e.g., scratches on images. The literature is essentially void of corresponding analytical performance results. We prove that TSC succeeds even when the number of missing entries in each data vector scales (up to a log-factor) linearly in the ambient dimension.

Finally, we propose a simple scheme for outlier detection which provably succeeds, even in the noisy case. Numerical results on synthetic data and on clustering handwritten digits (MNIST data set of handwritten digits [19]) and images of faces (extended Yale Face Database B [9, 21]) complement our theoretical results.

Relation to previous work

Lauer and Schnorr [18] also apply spectral clustering to an adjacency matrix constructed from correlations between data points, albeit, without thresholding. More importantly, no analytical performance results are available for the algorithm in [18]. Liu et al. [24] consider spectral clustering applied to an adjacency matrix $A$ obtained from a low-rank representation (LRR) of the data points through a nuclear norm minimization. A deterministic performance analysis reported in [24, Theorem 3.1] shows that LRR succeeds provided the subspaces are independent (the linear subspaces $S_i$ are called independent if the dimension of their (set) sum is equal to the sum of their dimensions), which implies that the subspaces must not intersect. Moreover, minimizing the nuclear norm results in significant computational challenges for large data sets.

Dyer et al. [4] propose to substitute the $\ell_1$-minimization step in SSC by an orthogonal matching pursuit (OMP) step, and derive deterministic recovery conditions for the resulting algorithm. In [5], the $\ell_1$-minimization step in SSC is replaced by an $\ell_2$-minimization step, which results in a least-squares representation of each data point in terms of all the other data points. While numerical results in [5] demonstrate the effectiveness of this approach, no analytical performance results are available. The local subspace affinity algorithm [38] and the spectral local best-fit flats (SLBF) algorithm [39] are based on spectral clustering and rely on an adjacency matrix $A$ built by finding the nearest neighbors in Euclidean distance (cf. [33]).

Lerman and Zhang [23] pose the problem of the recovery of multiple subspaces from data drawn from a distribution on a union of subspaces as a non-convex optimization problem. No computationally tractable algorithm to solve this optimization problem [22] seems to be available, though.

The problem of fitting a single low-dimensional subspace to a data set which consists of a modest number of noisy inliers and a large number of outliers was considered in [22]. A convex
programming algorithm proposed in [22] is shown to provably succeed.

Outline of the paper  The remainder of this paper is organized as follows. In Section 2, we introduce the TSC algorithm. Sections 3 and 4 contain analytical performance results for the noiseless and the noisy case, respectively. In Section 5, we analyze the impact of incompletely observed data points on TSC. Section 6 describes an outlier detection scheme along with corresponding analytical performance guarantees. In Section 7, we put our analytical results into perspective with respect to other analytical performance results for subspace clustering available in the literature. Section 8 contains numerical results on synthetic and on real data, including a performance comparison to SSC/RSSC. We decided to discuss the various settings (noiseless, noisy, incomplete observations, and outliers) in an isolated fashion to keep the exposition accessible. Proofs are given in the appendices.

Notation  We use lowercase boldface letters to denote (column) vectors, e.g., $\mathbf{x}$, and uppercase boldface letters to designate matrices, e.g., $\mathbf{A}$. The superscript $^T$ stands for transposition. For the vector $\mathbf{x}$, $x_q$ denotes its $q$th entry. For the matrix $\mathbf{A}$, $A_{ij}$ designates the entry in its $i$th row and $j$th column, $\mathbf{A}^\dagger := (\mathbf{A}^T \mathbf{A})^{-1} \mathbf{A}^T$ is its pseudo-inverse, $\|\mathbf{A}\|_{2\to 2} := \max_{\|\mathbf{v}\|_2 = 1} \|\mathbf{A}\mathbf{v}\|_2$ its spectral norm, $\|\mathbf{A}\|_F := (\sum_{i,j} |A_{ij}|^2)^{1/2}$ its Frobenius norm, and $\mathbf{I}_m$ denotes the $m \times m$ identity matrix. $\log(\cdot)$ is the natural logarithm, and $x \wedge y$ denotes the minimum of $x$ and $y$. For the set $\mathcal{T}$, $|\mathcal{T}|$ designates its cardinality and $\overline{\mathcal{T}}$ stands for its complement. The set $\{1, \ldots, N\}$ is denoted by $[N]$. We write $\mathcal{N}(\boldsymbol{\mu}, \Sigma)$ for a Gaussian random vector with mean $\boldsymbol{\mu}$ and covariance matrix $\Sigma$. The unit sphere in $\mathbb{R}^m$ is $S^{m-1} := \{ \mathbf{x} \in \mathbb{R}^m : \|\mathbf{x}\|_2 = 1 \}$. $1_{\{A\}}(\cdot)$ denotes the indicator function of the set $A$. For notational convenience, we use the shorthand $\max_{k \neq l}$ for $\max_{k \in [L]}: k \neq l$ and $\max_{k,l: k \neq l}$ for $\max_{k,l \in [L]: k \neq l}$.

2 The TSC algorithm

The formulation of the thresholding based subspace clustering (TSC) algorithm provided below assumes that outliers have already been removed from $\mathcal{X}$, e.g., through the outlier detection scheme described in Section 6, and that the data points in $\mathcal{X}$ are either normalized or of comparable norm. The latter assumption is relevant for Step 1 below and is not restrictive as the data points can be normalized prior to clustering.

**TSC algorithm.** Given a set of data points $\mathcal{X}$ and the parameter $q$ (the choice of $q$ is discussed below), perform the following steps:

**Step 1:** For every $\mathbf{x}_j \in \mathcal{X}$, identify the set $\mathcal{T}_j \subset [N] \setminus j$ of cardinality $q$ defined through

$$\langle \mathbf{x}_i, \mathbf{x}_j \rangle \geq \langle \mathbf{x}_p, \mathbf{x}_j \rangle \quad \text{for all } i \in \mathcal{T}_j \text{ and all } p \notin \mathcal{T}_j.$$

**Step 2:** Let $\mathbf{z}_j \in \mathbb{R}^N$ be the vector with $i$th entry $|\langle \mathbf{z}_j, \mathbf{x}_i \rangle|$ if $i \in \mathcal{T}_j$, and 0 if $i \notin \mathcal{T}_j$.

**Step 3:** Construct the adjacency matrix $\mathbf{A}$ according to $\mathbf{A} = \mathbf{Z} + \mathbf{Z}^T$, where $\mathbf{Z} = [\mathbf{z}_1, \ldots, \mathbf{z}_N]$.

**Step 4:** Estimate the number of subspaces using the eigengap heuristic [36] according to $\hat{L} = \arg \max_{i \in [N-1]} (\lambda_{i+1} - \lambda_i)$, where $\lambda_1 \leq \lambda_2 \leq \ldots \leq \lambda_N$ are the eigenvalues of the normalized Laplacian of the graph with adjacency matrix $\mathbf{A}$.

**Step 5:** Apply normalized spectral clustering [36] to $(\mathbf{A}, \hat{L})$.

The hope is that the indices in $\mathcal{T}_j$ correspond to vectors in the subspace $\mathbf{x}_j$ lies in. If for each $j \in [N]$ the indices in $\mathcal{T}_j$ correspond to vectors in the subspace $\mathbf{x}_j$ lies in, then each connected
component in the graph $G$ with adjacency matrix $A$ corresponds only to points that lie in the same subspace. The idea behind estimating $L$ according to Step 4 above is the following. By a well known result in spectral graph theory [29] the number of zero eigenvalues of the Laplacian of $G$ is equal to the number of connected components of $G$. In the noisy case, where the eigenvalues are not exactly zero, a robust estimator for $L$ is the eigengap heuristic [36] (which is also used for SSC in [27]). Finally, segmentation of the data into subspaces is obtained by applying spectral clustering to $(A, \hat{L})$.

The following definition is motivated by the desire that each connected component in the graph with adjacency matrix $A$ correspond to points within one subspace.

**Definition 1.** The subspace detection property holds for $\mathcal{X} = \mathcal{X}_1 \cup \ldots \cup \mathcal{X}_L$ and adjacency matrix $A$ if

1. $A_{ij} \neq 0$ only if $x_i$ and $x_j$ belong to the same set $\mathcal{X}_l$
2. for every $i \in [N]$, $A_{ij} \neq 0$ for at least $q$ points $x_j$ that belong to the same set $\mathcal{X}_l$ as $x_i$.

This subspace detection property is akin to the $\ell_1$ subspace detection property introduced in [27]. The corresponding notion in [28] is that of $A$ having “no false discoveries” and at least $q$ “true discoveries” in each row (column). The subspace detection property guarantees that each node in the graph $G$ (with adjacency matrix $A$) is connected to at least $q$ other nodes, all of which correspond to points within the same subspace. Owing to the spectral clustering step, the subspace detection property does not guarantee that TSC yields the correct associations of the data points to the subspaces. Conversely, even when the subspace detection property does not hold strictly, but the $A_{ij}$ for pairs $x_i, x_j$ belonging to different subspaces are “small enough”, TSC may still cluster the data correctly, as will be seen in the numerical results in Section 8.

It is easily seen that for the noiseless case the subspace detection property is satisfied for subspaces that are orthogonal to each other, as long as $q \leq \min_l(|\mathcal{X}_l| - d_l)$. Specifically, for subspaces that are orthogonal to each other, $\langle x_p, x_j \rangle = 0$ for all $x_p \in \mathcal{X}_l$, $x_j \in \mathcal{X}_k$, $l \neq k$, while for each $l$, there are at least $|\mathcal{X}_l| - d_l$ inner products $\langle x_p, x_j \rangle$ with $x_p, x_j \in \mathcal{X}_l$ that are not equal to zero (since no more than $d_l$ points in a $d_l$-dimensional subspace can be orthogonal to each other). The results in the following sections show that the subspace detection property can be satisfied under much more general conditions, in particular also when the subspaces intersect.

Recall that $q$ is an input parameter to the TSC algorithm. Choosing $q$ too small/large will lead to over/under-estimation of the number of subspaces $L$. Our analytical performance results ensure that the subspace detection property holds given that $q$ is sufficiently small relative to the $n_l$. Once this condition is satisfied, the specific choice of $q$ is irrelevant in terms of our performance guarantees, in particular it does not depend on the $d_l$ and the noise variance. Indeed, we will see, in the numerical results in Section 8, that TSC is not overly sensitive to the specific choice of $q$. Finally, note that TSC has only one input parameter. SSC shares this desirable feature and has only the LASSO parameter $\lambda$ as input parameter.

Finally, we note that a natural substitute for Step 2 in the TSC algorithm is to construct $z_j$ from the best linear approximation of $x_j$ through points indexed by $T_j$. Specifically, let $X_{T_j}$ be the matrix whose columns are the vectors in $\mathcal{X}$ indexed by $T_j$. Substitute Step 2 by:

**Step 2b:** Set the entries of $z_j \in \mathbb{R}^N$ indexed by $T_j$ to the absolute values of $X_{T_j}^T x_j$ and set all other entries to zero.

We call the TSC algorithm with Step 2 replaced by Step 2b TSCb. The formal relationship between Steps 2 and 2b is brought out by noting that $|z_{T_j}| = |X_{T_j}^T x_j| = |(X_{T_j} X_{T_j}^T)^{-1} X_{T_j}^T x_j|$, $i \in [q]$, and realizing that $z_{T_j}$ from the original Step 2 is the vector containing the absolute values of
Our analytical performance results stated below also apply to TSCb since the subspace detection property depends on the location of the non-zero entries of $A$ only, which in turn depends on the sets $T_j$ obtained in Step 1 only. The numerical properties of TSC and TSCb will be compared in Section 8. We note that TSCb is similar, in spirit, to SSC, as the Steps 1 and 2b above yield a (sparse) linear representation of $x_j$ in terms of $q$ points in $\mathcal{X} \setminus x_j$. SSC finds a sparse linear representation of $x_j$ in terms of points in $\mathcal{X} \setminus x_j$ via $l_1$-minimization.

We are now ready to state our main performance results.

### 3 Performance results for the noiseless case

In order to elicit the impact of the relative orientations of the subspaces $S_l$ on the performance of TSC, we take the $S_l$ to be deterministic and choose the points within the $S_l$ randomly. To this end, we represent the data points in $S_l$ by

$$y_j^{(l)} = U^{(l)} a_j^{(l)} \quad (2)$$

where $U^{(l)} \in \mathbb{R}^{m \times d_l}$ is a basis for the $d_l$-dimensional subspace $S_l$, and the $a_j^{(l)} \in \mathbb{R}^{d_l}$ are random. Unless explicitly stated otherwise, we take each matrix $U^{(l)}$ to be deterministic and orthonormal. We will consider the following two data models.

1. **The $a_j^{(l)}$ are i.i.d. uniformly distributed on $S_l^{d_l-1}$** (throughout the paper, whenever we say that the $a_j^{(l)}$ or $e_j^{(l)}$ are i.i.d., we actually mean i.i.d. across $j$ and $l$). Since each $U^{(l)}$ is orthonormal, the data points $y_j^{(l)} = U^{(l)} a_j^{(l)}$ are distributed uniformly on $\{y \in S_l : \|y\|_2 = 1\}$.

2. **The $a_j^{(l)}$ are i.i.d. $\mathcal{N}(0,(1/d_l)I_{d_l})$.** As the corresponding direction vectors $y_j^{(l)}/\|y_j^{(l)}\|_2$ are distributed uniformly on $\{y \in S_l : \|y\|_2 = 1\}$ and $\|y_j^{(l)}\|_2 = \|a_j^{(l)}\|_2$ concentrates around its expectation $\mathbb{E} \left[\|a_j^{(l)}\|_2^2\right] = 1$, this model is conceptually equivalent to the first model. The raison d’être for this model is that it is sometimes analytically more tractable.

We first consider clustering of noise-free data sets that have no outliers. Before stating concrete results, let us start with a back-of-the-envelope calculation to get an idea of the nature of the performance guarantees we can expect to find. For ease of exposition, we set $d_l = d$ and $n_l = |X_l| = n$, for all $l$. We take the $a_j^{(l)}$ to be i.i.d. $\mathcal{N}(0,(1/d)I_{d})$ distributed. By Definition 1, the subspace detection property can only be satisfied (for a given $q$) if $A_{lj} \neq 0$ only if $x_i$ and $x_j$ belong to the same set $X_l$. This is the case, if for each $x_i^{(l)} \in X_l$, and for each $X_l$, the set $T_l$ (associated with $x_i^{(l)}$) corresponds to points in $X_l$ only. Consider a specific $x_i^{(l)}$ with associated set $T_i$ and define $z_j^{(k)} := |\langle x_j^{(k)}, x_i^{(l)} \rangle|$. Note that, for simplicity of exposition, the notation $z_j^{(k)}$ does not reflect dependence on $x_i^{(l)}$. From Step 1 in the TSC algorithm it follows that $T_i$ corresponds to points in $X_l$ only if

$$z_i^{(l)} > \max_{k \neq i,j} z_j^{(k)} \quad (3)$$

where the order statistics $z_i^{(l)} \leq z_j^{(l)} \leq \ldots \leq z_{(n-l)}^{(l)}$ are defined by sorting the $\{z_j^{(l)}\}_{j \in [n]}$ \setminus $i$ in ascending order. We seek a sufficient condition on the relative orientations of the subspaces $S_l$ and the number of points from each subspace contained in $X$, for (3) to hold. Conditioned on $x_i^{(l)}$,
and hence on $a_i^{(l)}$, $\max_{k \neq l, j} z_j^{(k)}$ is the maximum of the absolute values of independent Gaussian random variables with variances $\|U^{(k)}^TU^{(l)}a_i^{(l)}\|_2^2$ and is thus with high probability smaller than (see Lemma 5 in Appendix G)

$$c' \max_{k \neq l} \left\| \frac{U^{(k)}^TU^{(l)}a_i^{(l)}}{2} \right\|_2 \sqrt{\log n + \log(L - 1)}$$

(4)

with an absolute constant $c'$. On the other hand, conditioned on $x_i^{(l)}$, the random variable $z_i^{(l)}$ is, with high probability, larger than (see Lemma 4 in Appendix G)

$$\tilde{c} \left\| \frac{U^{(l)}^TU^{(l)}a_i^{(l)}}{2} \right\|_2 \sqrt{\log(n - 1) - \log(q - 1)}$$

(5)

where $\tilde{c}$ is an absolute constant. We can hence expect TSC to succeed (with high probability) if (4) is smaller than (5) for all $a_i^{(l)}$ which leads to the condition

$$\max_{k, l : k \neq l} \left\| \frac{U^{(k)}^TU^{(l)}a_i^{(l)}}{2} \right\|_2 < \frac{\tilde{c}}{c} \sqrt{\log(n - 1) - \log(q - 1)} \sqrt{\log n + \log(L - 1)}$$

for all $i \in [n]$. (6)

Indeed, the rigorous performance guarantees for TSC stated below depend on the relative orientations of the subspaces $S_l$ through the quantity on the left hand side (LHS) of (6). Specifically, the more distinct the relative orientations of the subspaces, the easier it becomes to satisfy (6).

Inspired by the LHS of (6), we define two different notions of affinity between subspaces, on which the formal recovery guarantees stated below will be seen to depend, namely

$$\text{affp}(S_k, S_l) := \left\| \frac{U^{(k)}^TU^{(l)}}{2} \right\|_{2 \rightarrow 2}$$

and

$$\text{aff}(S_k, S_l) := \frac{1}{\sqrt{d_k \wedge d_l}} \left\| \frac{U^{(k)}^TU^{(l)}}{F} \right\|.$$ 

Note that the affinity notion [27, Definition 2.6] and [28, Definition 1.2], relevant to the analysis of SSC and RSSC, is equivalent to $\text{aff}(\cdot, \cdot)$. The relation between the affinity notions $\text{affp}(\cdot, \cdot)$ and $\text{aff}(\cdot, \cdot)$ is brought out by expressing them in terms of principal angles between $S_k$ and $S_l$ according to

$$\text{affp}(S_k, S_l) = \cos(\theta_1)$$

(7)

while

$$\text{aff}(S_k, S_l) = \frac{\sqrt{\cos^2(\theta_1) + ... + \cos^2(\theta_{d_k \wedge d_l})}}{\sqrt{d_k \wedge d_l}}$$

(8)

where $\theta_1, ..., \theta_{d_k \wedge d_l}$ with $0 \leq \theta_1 \leq ... \leq \theta_{d_k \wedge d_l} \leq \pi/2$ denotes the principal angles between $S_k$ and $S_l$, defined as follows.

**Definition 2.** The principal angles $\theta_1, ..., \theta_{d_k \wedge d_l}$ between the subspaces $S_k$ and $S_l$ are defined recursively according to

$$\cos(\theta_j) = \langle v_j, u_j \rangle,$$

where $(v_j, u_i) = \text{arg max} \langle v, u \rangle$

with the maximization carried out over all $v \in S_k : \|v\|_2 = 1$, $u \in S_l : \|u\|_2 = 1$, subject to $\langle v, v_i \rangle = 0$ and $\langle u, u_i \rangle = 0$ for all $i = 1, ..., j - 1$ (for $j = 1$, this constraint is void).
Note that \(0 \leq \text{aff}(S_k, S_l) \leq \text{affp}(S_k, S_l) \leq 1\). If \(S_k\) and \(S_l\) intersect in \(p\) dimensions, i.e., if \(S_k \cap S_l\) is \(p\)-dimensional, then \(\cos(\theta_1) = \ldots = \cos(\theta_p) = 1\). Hence, if \(S_k\) and \(S_l\) intersect in \(p \geq 1\) dimensions we have \(\text{affp}(S_k, S_l) = 1\) and \(\text{aff}(S_k, S_l) \geq \sqrt{p/(d_k \wedge d_l)}\).

We are now ready to state our first main result.

**Theorem 1.** Suppose that \(X, l \in [L]\), is obtained by choosing \(n_l = \rho q_l\), with \(\rho_l \geq 10/3\), points corresponding to \(S_l\) at random according to \(x_j^{(l)} = U^{(l)}a_j^{(l)}, j \in [n_l]\), where the \(a_j^{(l)}\) are i.i.d. \(N(0, (1/d_l)I_{d_l})\), and set \(X = X_1 \cup \ldots \cup X_L\). If

\[
\max_{k, l: k \neq l} \text{aff}(S_k, S_l) \leq c_1 \frac{\sqrt{\log \rho_{\min}}}{\sqrt{\log N}} \tag{9}
\]

with \(\rho_{\min} = \min_l \rho_l\), then the subspace detection property holds (for \(X\)) with probability at least

\[
1 - \sum_{l \in [L]} n_l \left( e^{-c_2(n_l^{-1})} + 2 \left( \sum_{k \neq l} n_k \right)^{-2} \right)
\]

where \(c_1\) and \(c_2\) are absolute constants satisfying \(0 < c_1, c_2 < 1\).

Theorem 1 states that TSC succeeds with high probability if \(\max_{k, l: k \neq l} \text{aff}(S_k, S_l)\) is sufficiently small, and if \(X\) contains sufficiently many points from each subspace (i.e., \(\rho_l = n_l/q \geq 10/3\), for all \(l\)). Intuitively, we expect that clustering becomes easier when the \(n_l\) increase. To see that Theorem 1 confirms this intuition, set \(n_l = n\), for all \(l\), and observe that the probability of success in Theorem 1 increases in \(n\); at the same time, for a fixed number of subspaces \(L\), the RHS of (9) increases in \(\rho_{\min}\) (recall that \(n_l = \rho_l q_l\)), and therefore (9) becomes easier to satisfy. Theorem 1 does not impose any restrictions on the dimensions of the subspaces; the only dependence on the subspaces \(S_l\) is via the affinity in the clustering condition (9). It is interesting to see that the clustering condition (9) implies (6), the condition derived in our back-of-the-envelope calculation. Note that Theorem 1 does not apply to subspaces that intersect as \(\text{affp}(S_k, S_l) = 1\) in that case and the RHS of (9) is strictly smaller than 1. We next present a result analogous to Theorem 1 that applies to intersecting subspaces.

**Theorem 2.** Suppose that \(X, l \in [L]\), is obtained by choosing \(n_l = \rho q_l\), with \(\rho_l \geq 6\), points corresponding to \(S_l\) at random according to \(x_j^{(l)} = U^{(l)}a_j^{(l)}, j \in [n_l]\), where the \(a_j^{(l)}\) are i.i.d. uniform on \(S^{d_l-1}\), and set \(X = X_1 \cup \ldots \cup X_L\). If

\[
\max_{k, l: k \neq l} \text{aff}(S_k, S_l) \leq \frac{1}{15 \log N} \tag{10}
\]

with \(m \geq 6 \log N\), then the subspace detection property holds (for \(X\)) with probability at least

\[
1 - 10/N - \sum_{l \in [L]} n_l e^{-c(n_l^{-1})}, \text{ where } c > 0 \text{ is an absolute constant.}
\]

The interpretation of Theorem 2 is analogous to that of Theorem 1 with the important difference that the RHS of (10), as opposed to the RHS of (9), decreases, albeit very slowly, in the \(n_l\) since \(N = \sum_l n_l\). For SSC a result in the spirit of Theorem 2 was reported in \([27, \text{Theorem 2.8}].\)

### 4 Impact of noise

In many practical applications the data points to be clustered are corrupted by measurement noise, typically modeled as additive Gaussian noise. Other noise sources require a structured noise model,
that describes, e.g., occlusions of images in face recognition [37], mistracking of feature points in motion segmentation [25], or corruptions in images. We will deal both with unstructured and structured noise and, as in the last section, take the subspaces to be deterministic and the data points in the subspaces to be random.

### 4.1 Unstructured (additive white Gaussian) noise

**Theorem 3.** Suppose that \( X_l, l \in [L], \) is obtained by choosing \( n_l = \rho l g \), with \( \rho_l \geq 6 \), points corresponding to \( S_l \) at random according to \( x_j^{(l)} = U^{(l)} a_j^{(l)} + e_j^{(l)}, j \in [n_l], \) where the \( a_j^{(l)} \) are i.i.d. uniform on \( S^{d_l - 1} \) and the \( e_j^{(l)} \) are i.i.d. \( N(0, \sigma^2/m)I_m \). Set \( X = X_1 \cup \ldots \cup X_L \) and let \( d_{\text{max}} = \max_l d_l \). If

\[
\max_{k,l: k \neq l} \text{aff}(S_k, S_l) + \frac{\sigma(1 + \sigma) \sqrt{d_{\text{max}}}}{\sqrt{\log N}} \leq \frac{1}{15 \log N} \tag{11}
\]

with \( m \geq 6 \log N \), then the subspace detection property holds (for \( X \)) with probability at least \( 1 - \frac{10}{N} - \sum_{l \in [L]} n_l e^{-c(n_l - 1)} \), where \( c > 0 \) is an absolute constant.

Theorem 3 states that TSC succeeds with high probability if the additive noise variance and the affinities between the subspaces are sufficiently small, and if \( X \) contains sufficiently many points from each subspace. Condition (11) nicely reflects the intuition that the more distinct the orientations of the subspaces, the more noise TSC tolerates. What is more, Condition (11) reveals that TSC can succeed even when the noise variance \( \sigma^2 \) is large, i.e., when \( \sigma^2 = E \left[ \| e_j^{(l)} \|_2^2 \right] > \| U^{(l)} a_j^{(l)} \|_2^2 = 1 \), provided that \( d_{\text{max}}/m \) is sufficiently small. This means that TSC can succeed even under massive noise, provided the dimensions of the subspaces are not too large.

The intuition behind the factor \( \sigma(1 + \sigma) \sqrt{d_{\text{max}}/m} \) in (11) is as follows. Assume, for simplicity, that \( d_l = d \), for all \( l \), and consider the most favorable situation of subspaces that are orthogonal to each other, i.e., \( \text{aff}(S_k, S_l) = 0 \), for all \( k \neq l \). TSC relies on the inner products between points within a given subspace to typically be larger than the inner products between points in distinct subspaces. First, note that \( \langle x_j, x_i \rangle = \langle y_j, y_i \rangle + \langle e_j, e_i \rangle + \langle y_j, e_i \rangle + \langle e_j, y_i \rangle \). Then, under the statistical data model of Theorem 3, we have \( E \left[ \| (y_j, y_i) \|^2 \right]^{1/2} \) is 

\[
\begin{align*}
\frac{1}{\sqrt{d}} & \quad \text{if } y_j, y_i \in S_l \text{ and } \langle y_j, y_i \rangle = 0 \text{ if } y_j \in S_k \text{ and } y_i \in S_l, \text{ with } k \neq l. \\
\end{align*}
\]

When the terms \( \langle e_j, e_i \rangle, \langle y_j, e_i \rangle, \) and \( \langle e_j, y_i \rangle \) are small relative to \( \frac{1}{\sqrt{d}} \), we have a margin on the order of \( \frac{1}{\sqrt{d}} \) to separate points within a given cluster from points in other clusters. Indeed, if \( \frac{\sigma^2}{\sqrt{m}} \) is small relative to \( \frac{1}{\sqrt{d}}, \langle y_j, e_i \rangle \) and \( \langle e_j, y_i \rangle \) are (sufficiently) small, while \( \frac{\sigma^2}{\sqrt{m}} \) being small relative to \( \frac{1}{\sqrt{d}} \) ensures that \( \langle e_j, e_i \rangle \) is (sufficiently) small. These two conditions are satisfied when \( \sigma(1 + \sigma) \sqrt{d/m} \) is (sufficiently) small.

### 4.2 Structured noise

Examples of noise sources which require a structured noise model are as follows. In face recognition problems errors due to occlusion and corruption are common [37]. In motion segmentation trajectories may be corrupted due to feature points that are mistracked [25]. The resulting errors in these two examples can be modeled as corruptions that are sparse in the identity basis [37, 25]. In the ensuing analysis we will be a bit more general by assuming that the data points are corrupted by errors that are sparse in a general orthonormal basis \( C \in \mathbb{R}^{m \times m} \) with Gaussian coefficient vectors. We consider two settings. First, we take both the orientations of the subspaces as well as the sparse support of the error signal to be random. Specifically, we will assume that the \( U^{(l)} \) are Gaussian.
random matrices, which ensures that the $U^{(l)}$ are approximately orthonormal with high probability. Second, we consider the case of deterministic subspaces and sparse error signals with deterministic support. We start with our result on the former setting.

**Theorem 4.** Suppose that $X_l, l \in [L]$, is obtained by choosing $n_l = \rho_1 q$, with $\rho_1 \geq 6$, points corresponding to $S_l$ at random according to $X^{(l)} = U^{(l)} a^{(l)} + C^{(l)} c^{(l)}$, $j \in [n_l]$, where the $a^{(l)} \in \mathbb{R}^{d}$ are i.i.d. uniform on $S^{d-1}$ and the $c^{(l)} \in \mathbb{R}^s$ are i.i.d. $\mathcal{N}(0, (\sigma^2/s) I_s)$. Let the entries of each matrix $U^{(l)} \in \mathbb{R}^{m \times d}$ be i.i.d. $\mathcal{N}(0, 1/m)$ and suppose that the $C^{(l)} \in \mathbb{R}^{m \times s}$ are chosen uniformly at random from all submatrices of $C \in \mathbb{R}^{m \times m}$ with $s$ columns, where $C$ is orthonormal. Set $X = X_1 \cup \ldots \cup X_L$. If

$$
\frac{c_3}{\sqrt{m}} \sqrt{3d + \log L} + 4 \frac{\sqrt{d}}{\sqrt{m}} \sigma (2 + \sigma) \leq \frac{1}{12 \log N}
$$

(12)

then the subspace detection property is satisfied (for $X$) with probability at least

$$
1 - \frac{12}{N} - N e^{-c(n-1)} - N^2 e^{-\frac{d}{2}} - N^2 e^{-\frac{18s}{m-s+1} \frac{s^2}{m^2}} - 4 e^{-c_1 m}
$$

(13)

where $c, c_1, c_3 > 0$ are absolute constants.

The first term in the clustering condition (12) accounts for the affinity between the random subspaces, the second term accounts for noise and is structurally equivalent to the corresponding term in the clustering condition for the case of additive white Gaussian noise (11). Since $s$ must be large for the probability estimate (13) to be close to one, our result suggests that sparse errors can be detrimental. On the other hand, since the term in the clustering condition (12) accounting for the noise, is structurally equivalent to (11), we expect structured noise according to the statistical model in Theorem 4 to have an effect similar to that of white Gaussian noise, as long as $s$ is sufficiently large. This result should, however, be taken with a grain of salt, as choosing the $U^{(l)}$ Gaussian ensures that for each pair $(U^{(l)}, C^{(l)}_j)$, the matrices $U^{(l)}$ and $C^{(l)}_j$ are close to orthogonal to each other, and, in addition, that aff($S_k, S_l$), for $k \neq l$, is close to zero, with high probability. This is a favorable situation for subspace clustering, which one is unlikely to encounter in practice. We therefore next consider the case of deterministic subspaces and deterministic $C^{(l)}_j$.

**Theorem 5.** Suppose that $X_l, l \in [L]$, is obtained by choosing $n_l = \rho_1 q$, with $\rho_1 \geq 6$, points corresponding to $S_l$ at random according to $X^{(l)} = U^{(l)} a^{(l)} + C^{(l)} c^{(l)}$, $j \in [n_l]$, where the $a^{(l)} \in \mathbb{R}^{d_l}$ are i.i.d. uniform on $S^{d_l-1}$, the $c^{(l)} \in \mathbb{R}^s$ are i.i.d. $\mathcal{N}(0, (\sigma^2/s) I_s)$, and the $C^{(l)} \in \mathbb{R}^{m \times s}$ are orthonormal. Set $X = X_1 \cup \ldots \cup X_L$ and let $d_{\max} = \max d_l$. If

$$
\max_{k \neq l} \text{aff}(S_k, S_l) + 2\sigma \frac{\sqrt{d_{\max}}}{\sqrt{s}} (\sigma + 2) \leq \frac{1}{12 \log N}
$$

(14)

then the subspace detection property is satisfied (for $X$) with probability at least

$$
1 - \frac{12}{N} - \frac{\rho_1}{\chi} - \sum_{l \in [L]} n_l e^{-c(n_l-1)}, \text{ where } c > 0 \text{ is an absolute constant.}
$$

Observe that the clustering condition (14) is structurally equivalent to (11) with $d/m$ substituted by $d/s$. Again, this result suggests that sparse errors can be detrimental. Moreover, it also suggests that more unstructured noise is better tolerated by TSC. This will be confirmed in the numerical results in Section 8.
5 Incomplete data and impact of subspace dimension

In practical applications the data points to be clustered are often incompletely observed, think of, e.g., images that exhibit scratches or have missing parts. We next investigate the impact of incompletely observed data points on the performance of TSC. We chose to work in the original \( m \)-dimensional signal space and set the missing entries in each data vector to zero. As the TSC algorithm depends on inner products between the data points only this ensures that the missing observations result in contributions equal to zero.

Understanding the impact of incompletely observed data points on clustering performance is obviously of significant importance. The literature seems, however, essentially void of corresponding analytical results. We start by noting that in the deterministic subspace setting any such result will necessarily depend on the specific orientations of the subspaces. To simplify things, we therefore assume both the orientations of the subspaces as well as the points in the subspaces to be random. Specifically, we will again assume that the \( U^{(l)} \) are Gaussian random matrices, which ensures that the \( U^{(l)} \) are approximately orthonormal with high probability. For simplicity of exposition, we furthermore take all subspaces to have equal dimension \( d \) and let the number of points in each of the subspaces be \( n \).

Theorem 6. Suppose that \( X_l, l \in [L] \), is obtained by choosing \( n = \rho q \), with \( \rho \geq 10/3 \), points corresponding to \( S_l \) at random according to \( x^{(l)}_j = U^{(l)} a^{(l)}_j, j \in [n] \), where the \( a^{(l)}_j \) are i.i.d. \( N(0, (1/d) I_d) \), and set \( X = X_1 \cup ... \cup X_L \). Assume that in each \( x_j \in X \) up to \( s \) entries (possibly different for different \( x_j \)) are set to 0. Let the entries of each \( U^{(l)} = R^{m \times d} \) be i.i.d. \( N(0, 1/m) \). If

\[
m \geq c_4 \frac{\log N}{\log \rho} \left( 3d + \log L + s \log \left( \frac{me}{2s} \right) \right) + sc_3
\]

then the subspace detection property holds (for \( X \)) with probability at least \( 1 - L n e^{-c_2(n-1)} - \frac{2L}{(L-1)^2 n} - 4e^{-c_1 m} \), where \( c_1, c_2, c_3, c_4 > 0 \) are absolute constants. If \( s = 0 \), (15) reduces to \( m \geq c_4 \frac{\log N}{\log \rho} (3d + \log L) \).

Strikingly, Theorem 6 shows that the number of missing entries in the data vectors is allowed to scale (up to a log-factor) linearly in the ambient dimension.

Impact of subspace dimension We can furthermore conclude, from Theorem 6, that TSC succeeds with high probability even when the dimensions of the subspaces scale (up to a log-factor) linearly in the ambient dimension. Drawing such a conclusion from Theorems 1 or 2, which are based on deterministic subspaces, seems difficult as the relation between \( m, d, \) and \( L \) is implicit in the affinity measures. These findings should, however, be taken with a grain of salt as the fully random subspace model ensures that the subspaces are approximately orthogonal to each other with high probability. In the case of \( s = 0 \), a result for SSC, analogous to Theorem 6, was reported in [27, Thm. 1.2]. We note, however, that [27, Theorem 1.2] assumes the \( U^{(l)} \) to be chosen uniformly from the set of all orthonormal matrices in \( R^{m \times d} \), whereas in Theorem 6 the \( U^{(l)} \) are Gaussian. Conceptually, these two models for the \( U^{(l)} \) are equivalent.

6 Outlier detection

We discuss the noisy and noise-free cases separately as the corresponding statistical models for the outliers differ slightly; moreover, the proof for the noiseless case is very simple and insightful and thus warrants an individual presentation.
6.1 Noise-free case

Outliers are data points that do not lie in one of the low-dimensional subspaces $S_l$ and have no low-dimensional linear structure. Here, this is conceptualized by assuming random outliers distributed uniformly on the unit sphere of $\mathbb{R}^m$. The outlier detection criterion we employ is based on the following observation. The maximum inner product between an outlier and any other point (outlier or inlier) in $X$ is, with high probability, smaller than $c\sqrt{\log N}/\sqrt{m}$. We therefore classify $x_j$ as an outlier if

$$\max_{i \in [N] \setminus j} |\langle x_i, x_j \rangle| < c\sqrt{\log N}/\sqrt{m}. \quad (16)$$

The maximum inner product between any point $x_j \in X_l$ and the points in $X_l \setminus x_j$ is unlikely to be smaller than $1/\sqrt{d_{\text{max}}}$. Hence, an inlier is unlikely to be misclassified as an outlier if $c\sqrt{\log N}/\sqrt{m} \leq 1/\sqrt{d_{\text{max}}}$, i.e., if $d_{\text{max}}/m$ is sufficiently small.

**Theorem 7.** Suppose that the set of outliers, $O$, is obtained by choosing $N_0$ outliers i.i.d. uniformly on $S^{m-1}$, and that $X_l, l \in [L]$, is obtained by choosing $n_l$ points corresponding to $S_l$ at random according to $x_j^{(l)} = U_j^{(l)}a_j^{(l)}, j \in [n_l]$, where the $a_j^{(l)}$ are i.i.d. uniform on $S^{d_l-1}$ and each $U_j^{(l)}$ is deterministic and orthonormal. Set $X = X_1 \cup \ldots \cup X_L \cup O$ and declare $x_j \in X$ to be an outlier if (16) holds with $c = \sqrt{6}$. Then, with $N = \sum_l n_l + N_0$, every outlier is detected with probability at least $1 - 2N_0/N^2$. Furthermore, with $d_{\text{max}} = \max_l d_l$, provided that

$$\frac{d_{\text{max}}}{m} \leq \frac{1}{6 \log N} \quad (17)$$

with probability at least $1 - n_l e^{-\frac{1}{2} \log \left( \frac{N}{m} \right)}$ no inlier in $S_l$ is misclassified as an outlier.

Theorem 7 states that, provided Condition (17) holds, and the set $X$ contains sufficiently many points from each subspace, outlier detection succeeds with high probability, i.e., every outlier is detected and no inlier is misclassified as an outlier. The dependence on $n_l$ is intuitive. If $n_l = 1$, there is no way of distinguishing the single inlier in $S_l$ from the outliers.

Since (17) can be rewritten as $N_0 \leq e^{6d_{\text{max}}} - \sum_l n_l$, it follows that outlier detection succeeds even if the number of outliers scales exponentially in $m/d_{\text{max}}$, i.e., if $d_{\text{max}}$ is kept constant, exponentially in the ambient dimension! Note that this result does not make any assumptions on the orientations of the subspaces $S_l$. The outlier detection scheme proposed in [27] in the context of SSC allows to identify outliers under a very similar condition. However, the scheme in [27] requires the solution of $N \ell_1$-minimization problems, each in $N$ variables, while the algorithm proposed here needs to compute $N^2$ inner products followed by thresholding only.

6.2 Noisy case

We next consider outlier detection under additive noise. To simplify the analysis, we change our outlier model slightly. Specifically, we assume the outliers to be distributed as $\mathcal{N}(0, (1/m)I_m)$. Conceptually, this outlier model is equivalent to that used in Section 6.1, as the directions of the outliers in the present model, i.e., $x_i/\|x_i\|_2$, are uniformly distributed on $S^{m-1}$, and $\|x_i\|_2$ concentrates around $1$. We furthermore employ a normalization which ensures that the inliers also have about unit norm. This guarantees that outlier detection is not trivially accomplished by exploiting differences in the norms between inliers and outliers.
Theorem 8. Suppose that \( \mathcal{O} \) is obtained by choosing \( N_0 \) outliers i.i.d. \( \mathcal{N}(\mathbf{0}, (1/m)I_m) \), and that \( \mathcal{X}_l, l \in [L] \), is obtained by choosing \( n_l \) points corresponding to \( S_l \) at random according to \( x^{(l)}_j = \frac{1}{\sqrt{1 + \sigma^2}}(U^{(l)}a^{(l)}_j + e^{(l)}_j), j \in [m] \), where the \( a^{(l)}_j \) are i.i.d. uniform on \( S^{d_l-1} \) and the \( e^{(l)}_j \) are i.i.d. \( \mathcal{N}(\mathbf{0}, (\sigma^2/m)I_m) \) and independent of the \( a^{(l)}_j \). Set \( \mathcal{X} = \mathcal{X}_1 \cup \ldots \cup \mathcal{X}_L \cup \mathcal{O} \) and declare \( x_j \in \mathcal{X} \) to be an outlier if (16) holds with \( c = 2.3\sqrt{6} \). Then, with \( N = \sum_l n_l + N_0 \), assuming \( m \geq 6 \log N \), every outlier is detected with probability at least \( 1 - 3N_0/N^2 \). Furthermore, with \( d_{\max} = \max_l d_l \), provided that

\[
d_{\max} \leq \frac{c_1}{(1 + \sigma^2)^2 \log N}
\]

(18)

where \( c_1 \) is an absolute constant, with probability at least \( 1 - n_l e^{-\frac{1}{2} \log \left( \frac{n_l}{2} \right)} - n_l \frac{7}{N^2} \) no inlier belonging to \( S_l \) is misclassified as an outlier.

Theorem 8 states that, provided Condition (18) holds, and \( \mathcal{X} \) contains sufficiently many points from each subspace, outlier detection succeeds with high probability. Our result furthermore shows that outlier detection can succeed even when \( \sigma^2 \) is large, provided that \( d_{\max}/m \) is sufficiently small. An outlier detection scheme for RSSC does not seem to be available.

7 Comparison with SSC/RSSC and other algorithms

As mentioned in the introduction, there are very few clustering algorithms that are both computationally tractable and succeed provably under non-restrictive conditions. Notable exceptions are the SSC algorithm [27], and for the noisy case the RSSC algorithm [28]. Since our results are in the spirit of those for SSC and RSSC—in particular we use the same statistical data model—we next compare our findings to those in [27, 28].

While SSC and RSSC employ a “global” criterion for building the adjacency matrix \( \mathbf{A} \) by sparsely representing each data point in terms of all the other data points through \( \ell_1 \)-minimization, TSC is based on a “local” criterion, namely the comparison of inner products of pairs of data points. This makes TSC computationally much less demanding than SSC and RSSC, while, perhaps surprisingly, sharing the performance guarantees and clustering conditions of SSC and RSSC. Specifically, for SSC in the noiseless case and RSSC in the noisy case, results analogous to Theorems 2 and 3 were reported in [27, Theorem 2.8] and [28, Theorem 3.1], respectively. The clustering condition for SSC in [27, Theorem 2.8] is identical (up to constants and log-factors) to (10), and the clustering condition for RSSC in [28] is identical (up to constants and log-factors) to (11) with \( \sigma(1 + \sigma) \) in (11) replaced by \( \sigma \). We note, however, that [28] requires \( \sigma \) to be bounded in the sense of \( \sigma \leq c \), for some constant \( c \), an assumption not needed here. If we took \( \sigma \) to be bounded, i.e., \( \sigma \leq c \), the factor \( \sigma(1 + \sigma) \) in Condition (11) above would be replaced by \( \sigma(1 + c) \) and we would get a condition that is equivalent (up to constants and log-factors) to that in [28]. In [28] \( \sigma \) is assumed to be bounded in the sense of \( \sigma \leq c \) to ensure sufficiently many “true discoveries” in each row/column in the affinity matrix; \((i,j)\) is said to be a true discovery if \( A_{ij} \neq 0 \) for \( x_i \) and \( x_j \) belonging to the same subspace. Condition (11) in Theorem 3 guarantees that the subspace detection property is satisfied, and thus ensures (see ii. in Definition 1) \( q \) true discoveries in each row/column.

Owing to the simplicity of TSC, the proofs in this paper are conceptually and technically less involved than the proofs of the corresponding results for SSC and RSSC in [27, 28]. Finally, we provide performance guarantees for clustering of incompletely observed data points and structured noise. Corresponding results for SSC and RSSC do not seem to be available.
A comparison of the analytical performance results for RSSC (in particular [28, Theorem 3.1]) to those for a number of representative subspace clustering approaches such as generalized PCA (GPCA) [35], the K-flats algorithm [32], and LRR [24], can be found in [28, Section 5]. In addition, this comparison also features computational tractability and robustness aspects. As the main analytical performance results for TSC are structurally equivalent to those for SSC and RSSC the conclusions drawn in the comparison in [28, Section 5] carry over to TSC.

8 Numerical results
We use the following performance metrics.

- The clustering error (CE) measures the number of misclassified points relative to the total number of points and is defined as follows. Denote the estimate of the number of subspaces by $\hat{L}$ and note that possibly $\hat{L} \neq L$. Let $c \in [L]^N$ and $\hat{c} \in [\hat{L}]^N$ be the original and estimated assignments of the points in $X$ to the individual clusters. The clustering error is given by

$$CE(\hat{c}, c) = \min_{\pi} \left( 1 - \frac{1}{N} \sum_{i=1}^{N} 1_{\{\pi(c_i) = \hat{c}_i\}} \right)$$

where the minimum is with respect to all assignments $\pi : [L] \to [\hat{L}]$ (for $\hat{L} = L$, $\pi$ is simply a permutation). Note that $\pi$ naturally appears in this definition as the specific cluster indices are irrelevant to the clustering error. The problem of finding the optimal assignment $\pi$ can be cast as finding the maximal matching of a weighted bipartite graph, which can be solved efficiently via the Hungarian algorithm [30].

- The error in estimating the number of subspaces $L$ is denoted as $EL$ and takes the value 0 if the estimate $\hat{L}$ is correct, 1 if $L < \hat{L}$, and $-1$ if $L > \hat{L}$. We employ a signed error measure to be able to discriminate between under- and overestimation. In principle, $EL$ averaged over several problem instances, may equal zero, while $\hat{L} \neq L$ for each individual problem instance. However, as it turns out (in the numerical results below), for a given choice of problem parameters, we get that either $L < \hat{L}$ or $L > \hat{L}$ almost consistently.

- The feature detection error (FDE) is defined as

$$FDE(A) = \frac{1}{N} \sum_{i=1}^{N} \left( 1 - \|b_{x_i}\|_2/\|b_i\|_2 \right)$$

where $b_i$ is the $i$th column of the adjacency matrix $A$ and $b_{x_i}$ is the vector containing the entries of $b_i$ corresponding to the subspace $x_i$ lives in. The FDE measures to which extent points from different subspaces are connected in the graph with adjacency matrix $A$, and equals zero if the subspace detection property holds.

8.1 Synthetic data
Throughout this section, unless explicitly stated otherwise, we set, for simplicity, $n_l = n$ and $d_l = d$, for all $l$, and generate the $d$-dimensional subspaces $S_l$ by drawing independently (across $l$) a corresponding orthonormal basis $U^{(l)} \in \mathbb{R}^{m \times d}$ uniformly at random from the set of all orthonormal matrices in $\mathbb{R}^{m \times d}$. 
8.1.1 Impact of the choice of $q$

The goal of the first experiment is to demonstrate, as stated in Section 2, that TSC is not overly sensitive to the exact choice of $q$. We generate $L = 10$ subspaces in $\mathbb{R}^{50}$, each of dimension $d = 5$, and vary the number of points in each subspace $n$, as well as the parameter $q$. For a given choice of subspaces, the data points are chosen according to the statistical model Theorem 2 is based on. For each pair $(n,q)$ we average over 100 problem instances. The results, summarized in Figure 1, show that if $q$ is too small (i.e., $q \leq 3$ in this experiment), TSC fails, even though the FDE is close to zero. This is a consequence of too small $q$ leading to overestimation of $L$, which, in turn, results in points belonging to a single subspace being assigned to distinct clusters. For $q > n$, TSC can fail because points from different subspaces may be assigned to a single cluster. When $n$ is sufficiently large (for this experiment $n \geq 3d$) TSC is seen to succeed for $q \in \{4, \ldots, 30\}$. This shows that TSC is not overly sensitive to the exact choice of $q$.

8.1.2 Intersection of subspaces

We next demonstrate that TSC can succeed even when the subspaces $S_l$ intersect, as proven in Theorem 2. To facilitate comparison to SSC, we perform the same experiment as in [27, Sec. 5.1.2]. Specifically, we set $m = 200$, $d = 10$, $q = d$, and generate two subspaces, $S_1$ and $S_2$, at random through their defining bases $U^{(1)}$ and $U^{(2)}$ as follows. We choose, uniformly at random, from the set of all $2d - t$ orthonormal vectors in $\mathbb{R}^m$, a set of $2d - t$ orthonormal vectors, and identify the columns of $U^{(1)}$ and $U^{(2)}$ with the last and first $d$ of those vectors, respectively. This ensures that the intersection of $S_1$ and $S_2$ is of dimension $t$. Next, we generate $n = 20d$ data points in each of the two subspaces according to $x_i^{(t)} = U^{(t)}a_i^{(t)}$, with the $a_i^{(t)}$ drawn i.i.d. uniformly on $S^{d-1}$. For each $t = 0, \ldots, d$ the CE, EL, and FDE are obtained by averaging over 100 problem instances. The results, shown in Figure 2, allow us to conclude that, as long as the dimension of the intersection of the subspaces is not too large, TSC does, indeed, succeed. An analogous experiment was performed for SSC in [27, Sec. 5.1.2] and delivered comparable results.

8.1.3 Influence of $d$, $n$, and incomplete data

The goal of the next experiment is to elicit the dependence of $d$, $n$, and the number of missing entries in the data points on clustering performance, and to furthermore demonstrate that TSC can succeed even when the subspace detection property is not satisfied strictly. We generate $L = 10$ subspaces of $\mathbb{R}^{50}$, set $q = d$, and vary the dimension $d$ of the subspaces and the number of points in each subspace $n = d\gamma$ by varying the parameter $\gamma$. The data points are chosen according to the
statistical model Theorem 2 is based on. For each pair \((d, \gamma)\), the FDE, CE, and EL are obtained by averaging over 10 problem instances. The results, depicted in Figure 3, show, as indicated in Section 2, that TSC can, indeed, succeed even when the subspace detection property does not hold strictly.

Finally, we perform the same experiment, but set the entries of \(x_i\) with indices in \(D_i\) to zero, where the sets \(D_i\) are chosen uniformly at random from the set \(\{D \subseteq [m] : |D| = s\}\) and independently for each \(x_i\). The results, summarized in Figure 4, show that TSC succeeds even when a large fraction of the entries in each data vector is missing.

### 8.1.4 Unstructured noise

We next perform the same experiment as in Section 8.1.3, but subject the data points to additive noise before clustering. Specifically, the data points are chosen according to the statistical model Theorem 3 is based on. The results, depicted in Figure 5, show that TSC can succeed even when the noise variance is large, provided that \(n\) is sufficiently large.

In Section 4.1, we found that TSC can succeed even under massive noise (i.e., if \(\sigma^2 > 1\),
provided that $d/m$ is sufficiently small. To demonstrate this effect numerically, we generate $L = 5$ subspaces in $\mathbb{R}^{100}$, each of dimension $d = 2$ (hence $d/m = 1/50$), and choose the data points again according to the statistical model Theorem 3 is based on. We vary the number of points in each subspace $n = d\gamma$ by varying $\gamma$, and we also vary the noise variance. Moreover, we set $q = 0.65n$. The results, depicted in Figure 6, show that, indeed, even when $\sigma^2 > 1$, TSC can succeed.

### 8.1.5 Sparse noise

Finally, we investigate the impact of sparse corruptions on clustering performance. We generate $L = 15$ subspaces, and vary the dimension of the subspaces, $d$, and the number of points in each subspace $n = d\gamma$ as a fraction of $d$, again by varying $\gamma$. The points in each subspace are chosen according to the statistical model Theorem 4 is based on, with $\sigma^2 = 0.4$ and $C = I_m$. The results, depicted in Figure 7, corroborate the results in Section 4.2, namely that the effect of sparse corruptions is drastic if $s$ is small, and similar to that for additive unstructured noise if $s$ is large.

### 8.1.6 Detection of outliers

In order to facilitate a comparison with the outlier detection scheme proposed for SSC in [27], we perform our experiment with exactly the parameters used in [27, Sec. 5.2]. Specifically, we set $d = 5$, vary $m \in \{50, 100, 200\}$, and generate $L = 2m/d$ subspaces at random. We choose $n$ inliers per
subspace and a total of $N_0 = Ln$ outliers according to the statistical model Theorem 7 is based on. Note that this results in the number of outliers being equal to the number of inliers. We measure performance in terms of the misclassification probability, defined as the number of misclassified points (i.e., outliers misclassified as inliers and inliers misclassified as outliers) divided by the total number of points in $X$. We find a misclassification error probability of $\{0.017, 1.510^{-4}, 2.510^{-5}\}$ for $m = \{50, 100, 200\}$, respectively. Similar performance was reported for the scheme proposed in [27] for SSC.

8.2 Clustering handwritten digits

We next apply TSC to the problem of clustering handwritten digits. Specifically, we work with the MNIST data set of handwritten digits [19], and we use the training set that contains 60,000 centered $20 \times 20$ pixel images of handwritten digits. The assumption underlying the idea of posing the problem of clustering handwritten digits as a subspace clustering problem is that the images of a single digit lie in a low-dimensional subspace of unknown dimension and orientation.

We compare the performance of TSC and TSCb to the performance of SSC/RSSC. The empirical mean and variance of the clustering error are computed by averaging over 50 of the following problem instances. We choose the digits $\{2, 4, 8\}$ and for each digit we choose $n$ images uniformly at random from the set of all images of that digit. In order to make the LASSO step in SSC computationally more tractable, we first apply PCA to the corresponding images to reduce the dimension of the images from $M = 400$ to $m = 50$. We choose $q = 10$ for TSC and set the LASSO penalty to $\lambda = 10$ for SSC. The choice $\lambda = 10$ is motivated by the fact that it yields the best performance among $\lambda \in \{0.5, 1, 2, 5, 10, 20, 30\}$ for $n = 200$. The results, summarized in Figure 8, show that TSC performs best, TSCb second best, and SSC performs slightly worse than TSCb. Note that the same experiment, for $n = 200$, was performed in [39, Sec. 3.4] for several further subspace clustering algorithms. Comparing the results in [39, Tab. 8] to those obtained for TSC shows that TSC is competitive with respect to state-of-the-art subspace clustering algorithms such as GPCA [35], $K$-flats [32], and SLBF [39].
8.3 Clustering faces

We next apply TSC to the problem of clustering images of faces under varying illumination conditions. The motivation for applying TSC to this problem comes from the insight that images of a given face under varying illumination conditions can be approximated well by a 9-dimensional linear subspace [1]. Each 9-dimensional subspace \( S_l \) would then contain the images corresponding to a given person. We work with the extended Yale Face Database B [9, 21], and compare our results to those reported in [39, Tab. 5] for the same experiment. The extended Yale Face Database B contains 168 \( \times \) 192 pixel (cropped) images of 38 persons, each taken under 64 different illumination conditions. We resized the images to \( 84 \times 96 \) pixels using the imresize function in Matlab. In each problem instance, we chose \( L \) persons uniformly at random from the set of 38 persons. As pointed out in [39, Section 3.3] the subspaces corresponding to different persons are extremely close to each other, which makes clustering hard. Discrimination between the clusters (and hence persons) can be improved through preprocessing of the data as described in [39, Section 3.3]. Specifically, we remove the first three principal components of \( X \), where \( X \) is the matrix whose columns are the data points in the data set \( \mathcal{X} \). For each \( L \), we apply TSC with \( q = 10 \), and the CE was obtained by averaging over 500 problem instances. The results are summarized in Figure 9. Comparing Figure 9 to [39, Tab. 5] (which contains performance results for a number of subspace clustering algorithms, such as SSC [6, 7], GPCA [35], K-flats [32], and SLBF [39]) shows that the performance of TSC is comparable to that of state of the art subspace clustering algorithms. We hasten to add, however, that in this experiment SSC performs better than TSC.

A Proof of Theorem 1

Theorem 1 is a special case of the following, more general, result, which is a key ingredient of the proof of Theorem 6 as well.

**Theorem 9.** Suppose that \( \mathcal{X}_l, l \in [L] \), is obtained by choosing \( n_l = \rho_l q \), with \( \rho_l \geq 10/3 \), points corresponding to \( S_l \) at random according to \( x_j^{(l)} = U_j^{(l)} a_j^{(l)}, j \in [n_l] \), where the \( a_j^{(l)} \) are i.i.d. \( \mathcal{N}(\mathbf{0}, (1/d_l) \mathbf{I}_{d_l}) \),
and set \( \mathcal{X} = \mathcal{X}_1 \cup \ldots \cup \mathcal{X}_L \). Assume that in each \( x_j \in \mathcal{X} \) up to \( s \) entries (possibly different for different \( x_j \)) are missing, i.e., set to 0. Assume that the \( U^{(l)} \) satisfy

\[
\min_{l, D: |D| \leq 2s, \|a\|_2 = 1} \left\| U^{(l)T}_D U^{(l)} a \right\|_2 \geq \eta 
\]

and

\[
\max_{k, l: k \neq l, D: |D| \leq 2s, \|a\|_2 = 1} \left\| U^{(k)T}_D U^{(l)} a \right\|_2 \leq \delta 
\]

where \( U^{(l)}_D \in \mathbb{R}^{m \times d_l} \) is the matrix obtained from \( U^{(l)} \in \mathbb{R}^{m \times d_l} \) by setting the rows with indices in \( D \subseteq [m] \) to zero. If

\[
\frac{\delta}{\eta} \leq c_1 \frac{\sqrt{\log \rho_{\text{min}}}}{\sqrt{\log N}}
\]

with \( \rho_{\text{min}} = \min_l \rho_l \), then the subspace detection property holds (for \( \mathcal{X} \)) with probability at least

\[
1 - \sum_{l \in [L]} n_l \left( e^{-c_2(n_l - 1)} + 2 \left( \sum_{k \neq l} n_k \right)^{-2} \right)
\]

where \( c_1 \) and \( c_2 \) are absolute constants satisfying \( 0 < c_1 < 1 \) and \( c_2 > 0 \).

**Proof of Theorem 1.** Theorem 1 follows as a special case of Theorem 9 by setting \( s = 0 \) and choosing the \( U^{(l)} \) to be orthonormal. For \( s = 0 \), \( U^{(l)}_D = U^{(l)} \) and since the \( U^{(l)} \) are orthonormal, \( U^{(l)T}_D U^{(l)} = U^{(l)T} U^{(l)} = I_{d_l} \). Thus (19) is satisfied with \( \eta = 1 \), and the LHS of (20) reduces to \( \max_{k, l: k \neq l} \text{affp}(S_k, S_l) \), in turn (20) and (21) reduce to (9), which then establishes the result.

**Proof of Theorem 9.** As stated in Section 3, the subspace detection property is satisfied if for each \( x_i^{(l)} \in \mathcal{X}_i \), and for each \( \mathcal{X}_i \), we have

\[
z_{(n_l - q)}^{(l)} > \max_{k \neq l, j} z_j^{(k)}
\]

where \( z_j^{(k)} := |\langle x_j^{(k)}, x_i^{(l)} \rangle| \) and the order statistics \( z_1^{(l)} \leq z_2^{(l)} \leq \ldots \leq z_{(n_l - 1)}^{(l)} \) are defined by sorting the \( \{z_j^{(l)}\}_{j \in [n_l] \setminus i} \) in ascending order. For simplicity of exposition the notation \( z_j^{(k)} \) does not reflect the dependence on \( x_i^{(l)} \). Our proof proceeds in two steps. First, we upper-bound the probability of (23) being violated for a given \( x_i^{(l)} \). A union bound over all \( N \) vectors \( x_i^{(l)}, i \in [n_l], l \in [L] \), then yields the final result. Note that

\[
z_j^{(k)} = |\langle x_j^{(k)}, x_i^{(l)} \rangle| = |\langle U^{(k)}_D a_j^{(k)}, U^{(l)}_E a_i^{(l)} \rangle| = |\langle a_j^{(k)}, U^{(k)T}_D U^{(l)}_E a_i^{(l)} \rangle|.
\]

Here the sets \( D, E \subseteq [m] \) contain the indices of the entries of \( x_j^{(k)} \) and \( x_i^{(l)} \), respectively, that are missing (i.e., set to zero). Thus, conditioned on \( a_i^{(l)} \), \( z_j^{(k)} \) is distributed as

\[
\left\| U^{(k)T}_D U^{(l)}_E a_i^{(l)} \right\|_2 \approx \frac{z_j^{(k)}}{\sqrt{|D|}}.
\]
where the $\tilde{z}^{(k)}_j$ are the absolute values of independent (over $j$ and $k$) $\mathcal{N}(0,1)$ distributed random variables. Condition (19) guarantees that for every $\mathcal{D}, \mathcal{E}$ with $|\mathcal{D}|, |\mathcal{E}| \leq s$, for every $l$, and for every $a_i^{(l)}$, we have $\|U_D^{(l)} U_E^{(l)} a_i^{(l)}\|_2 \leq \eta \|a_i^{(l)}\|_2$. We can, therefore, bound the probability that (23), conditioned on $a_i^{(l)}$, does not hold, according to

$$P\left[\tilde{z}^{(l)}_{(n_l-q)} \leq \max_{k \neq l, j} \tilde{z}^{(k)}_j\right] \leq P\left[\tilde{z}^{(l)}_{(n_l-q)} \eta \|a_i^{(l)}\|_2 \leq \max_{k \neq l, j, \mathcal{D}, \mathcal{E} : |\mathcal{D}|, |\mathcal{E}| \leq s} \|U_D^{(k)} U_E^{(l)} a_i^{(l)}\|_2 \tilde{z}^{(k)}_j\right] \leq P\left[\tilde{z}^{(l)}_{(n_l-q)} \leq \frac{\delta}{\eta} \max_{k \neq l, j} \tilde{z}^{(k)}_j \right] \leq P\left[\tilde{z}^{(l)}_{(n_l-q)} \leq \frac{\delta}{\eta} \max_{k \neq l, j} \tilde{z}^{(k)}_j \right] \leq \frac{\delta}{\eta} \sqrt{\theta \log \rho_l}$$

where we used (20) to conclude that $\|U_D^{(k)} U_E^{(l)} a_i^{(l)}\|_2 \leq \delta \|a_i^{(l)}\|_2$ for all $\mathcal{D}, \mathcal{E}$ with $|\mathcal{D}|, |\mathcal{E}| \leq s$. Next, we note that for random variables $X$ and $Y$ and constants $\phi$ and $\varphi$ satisfying $\phi \leq \varphi$, we have that

$$P[X \leq Y] \leq P[\{X \leq \phi\} \cup \{\varphi \leq Y\}] \leq P[X \leq \phi] + P[\varphi \leq Y].$$

Applying (25) to (24), with $\phi = \sqrt{\theta \log \rho_l}$, where $\theta > 0$ is an absolute constant and $\varphi = \frac{\delta}{\eta} \sqrt{6 \log \left(\sum_{k \neq l} n_k\right)}$, yields

$$P\left[\tilde{z}^{(l)}_{(n_l-q)} \leq \frac{\delta}{\eta} \max_{k \neq l, j} \tilde{z}^{(k)}_j \right] \leq P\left[\tilde{z}^{(l)}_{(n_l-q)} \leq \sqrt{\theta \log \rho_l}\right] + P\left[\sqrt{6 \log \left(\sum_{k \neq l} n_k\right)} \leq \max_{k \neq l, j} \tilde{z}^{(k)}_j \right].$$

Note that this requires $\phi \geq \varphi$, i.e.,

$$\sqrt{\theta \log \rho_l} \geq \frac{\delta}{\eta} \sqrt{6 \log \left(\sum_{k \neq l} n_k\right)}$$

which is implied by (21) with $c_1 = \sqrt{\theta/6}$ upon noting that $\sum_{k \neq l} n_k \leq N$. Using $\rho_l = n_q^2 = n_q - 1 = \frac{1}{1 - \gamma}$ with $\gamma = \frac{n_q - 1}{n_q}$ and applying Lemma 4 in Appendix G yields

$$P\left[\tilde{z}^{(l)}_{(n_l-q)} \leq \sqrt{\theta \log \rho_l}\right] \leq P\left[\tilde{z}^{(l)}_{(n_l-q)} \leq \sqrt{\theta \log \left(\frac{1}{1 - \gamma}\right)}\right] \leq e^{-c_2(n_q-1)}.$$

Here, $c_2$ is an absolute constant which satisfies $c_2 > 0$ if $\rho_l > 10/3$, a condition which ensures $\gamma > \frac{7}{10}$, needed for Lemma 4. Application of (27) and Lemma 5 in Appendix G to (26) yields

$$P\left[\tilde{z}^{(l)}_{(n_l-q)} \leq \max_{k \neq l, j} \tilde{z}^{(k)}_j \right] \leq e^{-c_2(n_q-1)} + \frac{2}{\left(\sum_{k \neq l} n_k\right)^2}.$$ 

Taking the union bound over all $x_i^{(l)}$, $i \in [n_l], l \in [L]$, yields the desired lower bound on the probability of the subspace detection property to hold. 

\[\square\]
B Proof of Theorems 2 and 3

Theorem 2 follows from Theorem 3 by setting $\sigma = 0$. We therefore need to prove Theorem 3 only. We again start by noting that, as stated in Section 3, the subspace detection property is satisfied if for each $x_i^{(l)} \in X_i$, and for each $X_i$, we have

$$z^{(l)}_{(n_i - q)} > \max_{k \neq l,j} z^{(k)}_j$$  \hspace{1cm} (29)

where $z^{(k)}_j = |\langle x_j^{(k)}, x_i^{(l)} \rangle|$. Next, we upper-bound the probability of (29) being violated. A union bound over all $N$ vectors $x_i^{(l)}, i \in [n_i], l \in [L]$, will, as before, yield the final result. We start by setting

$$\tilde{z}^{(k)}_j := \left| \langle a_j^{(k)}, U^{(k)T} U^{(l)} a_i^{(l)} \rangle \right|$$  \hspace{1cm} (30)

and noting that

$$z^{(k)}_j = \left| \langle x_j^{(k)}, x_i^{(l)} \rangle \right| = \left| \langle a_j^{(k)}, U^{(k)T} U^{(l)} a_i^{(l)} \rangle + e_j^{(k)} \right|$$  \hspace{1cm} (31)

with

$$e_j^{(k)} := \langle e_j^{(k)}, e_i^{(l)} \rangle + \langle e_j^{(k)}, U^{(l)} a_i^{(l)} \rangle + \langle U^{(k)} a_j^{(k)}, e_i^{(l)} \rangle.$$  \hspace{1cm} (32)

Recalling that $z^{(l)}_1 \leq z^{(l)}_2 \leq \ldots \leq z^{(l)}_{(n_i - 1)}$ are the order statistics of $\{z_j^{(l)}\}_{j \in [n_i]}$, it follows that

$$\tilde{z}^{(l)}_{(n_i - q)} - \max_{j \neq i} |e_j^{(l)}| \leq z^{(l)}_{(n_i - q)}$$

and hence the probability of (29) being violated can be upper-bounded as

$$P \left[ z^{(l)}_{(n_i - q)} \leq \max_{k \neq l,j} z^{(k)}_j \right] \leq P \left[ \tilde{z}^{(l)}_{(n_i - q)} - \max_{j \neq i} |e_j^{(l)}| \leq \max_{k \neq l,j} \tilde{z}^{(k)}_j + \max_{k \neq l,j} |e_j^{(k)}| \right]$$

$$\leq P \left[ \tilde{z}^{(l)}_{(n_i - q)} \leq \frac{\nu}{\sqrt{d_i}} \right]$$

$$+ P \left[ \alpha + 2\epsilon \leq \max_{j \neq i} |e_j^{(l)}| + \max_{k \neq l,j} \tilde{z}^{(k)}_j + \max_{k \neq l,j} |e_j^{(k)}| \right]$$  \hspace{1cm} (33)

$$\leq P \left[ \tilde{z}^{(l)}_{(n_i - q)} \leq \frac{\nu}{\sqrt{d_i}} \right] + P \left[ \max_{k \neq l,j} \tilde{z}^{(k)}_j \geq \alpha \right]$$

$$+ P \left[ \max_{j \neq i} |e_j^{(l)}| \geq \epsilon \right] + P \left[ \max_{k \neq l,j} |e_j^{(k)}| \geq \epsilon \right]$$  \hspace{1cm} (34)

where $\alpha, \epsilon, \nu$ are chosen later and for (33) we used (25) with $\phi = \frac{\nu}{\sqrt{d_i}}$ and $\varphi = \alpha + 2\epsilon$, which leads to the assumption $\alpha + 2\epsilon \leq \frac{\nu}{\sqrt{d_i}}$, resolved below. We next upper-bound the individual terms in (34) to get the following results proved at the end of this appendix:

Step 1 Setting $\epsilon = \frac{2\sigma(1+\sigma)}{\sqrt{m}} \beta$, we have for all $\beta \geq \frac{1}{\sqrt{2\pi}}$ satisfying $\beta/\sqrt{m} \leq 1$ that

$$P \left[ |e_j^{(k)}| \geq \epsilon \right] \leq 7e^{-\frac{\beta^2}{7}}.$$  \hspace{1cm} (35)
Step 2  Setting
\[ \alpha = \frac{\beta(1 + \beta)}{\sqrt{d_l}} \max_{k \neq l} \frac{1}{\sqrt{d_k}} \|U^{(k)}T U^{(l)}\|_F \]  \hspace{1cm} (36)
we have for all \( \beta \geq 0 \) that
\[ P \left( \max_{k \neq l,j} z_j^{(k)} \geq \alpha \right) \leq \sum_{k \in [L] \setminus l} (1 + 2n_k) e^{-\frac{\beta^2}{2}} \leq 3N e^{-\frac{\beta^2}{2}}. \]  \hspace{1cm} (37)

Step 3  For \( \nu = 2/3 \) and \( n_l/q = \rho_l \geq 6 \), there is a constant \( c = c(\rho_l, \nu) > 1/20 \) such that
\[ P \left( \tilde{z}_{(n_l-q)} \leq \frac{\nu}{\sqrt{d_l}} \right) \leq e^{-c(n_l-1)}. \]  \hspace{1cm} (38)

Before presenting the detailed arguments leading to (35), (37), and (38), we show how the proof is completed. Setting \( \beta = \sqrt{6 \log N} \) and using (35), (37) (note that \( \beta/\sqrt{m} \leq 1 \) is satisfied since, by assumption, \( m \geq 6 \log N \)), and (38) in (34) yields
\[ P \left( \tilde{z}_{(n_l-q)} \leq \max_{k \neq l,j} z_j^{(k)} \right) \leq e^{-c(n_l-1)} + 3N e^{-\frac{\beta^2}{2}} + 7Ne^{-\frac{\beta^2}{2}} = \frac{10}{N^2} + e^{-c(n_l-1)}. \]  \hspace{1cm} (39)

Taking the union bound over all vectors \( x_i^{(l)}, i \in [n_l], l \in [L] \), yields the desired lower bound on the probability of the subspace detection property to hold.

Recall that for (33) we imposed the condition \( \alpha + 2\epsilon \leq \nu \sqrt{d_l} \). With our choices for \( \epsilon, \alpha, \) and \( \nu \) in Step 1, 2, and 3, respectively, this condition becomes
\[ \beta(1 + \beta) \max_{k \neq l} \frac{1}{\sqrt{d_k}} \|U^{(k)}T U^{(l)}\|_F + 4\sigma(1 + \sigma) \sqrt{d_l} \sqrt{m} \beta \leq \frac{2}{3}. \]  \hspace{1cm} (40)

Next, note that
\[ (1 + \beta) \leq 3.2 \sqrt{\log N} \leq 4\sqrt{\log N} \]  \hspace{1cm} (41)
as a consequence of \( N \geq 6 \), by assumption. Therefore, (40) is implied by
\[ \max_{k \neq l} \frac{1}{\sqrt{d_k}} \|U^{(k)}T U^{(l)}\|_F + \frac{\sigma(1 + \sigma)}{\sqrt{\log N}} \frac{\sqrt{d_l}}{\sqrt{m}} \beta \leq \frac{2}{3} \cdot 4\sqrt{6 \log N} \]
which, in turn, is implied by (11). This concludes the proof.

It remains to prove the bounds (35), (37), and (38).

Step 1, proof of (35)  Conditional on \( a_j^{(k)} \), with \( \|U^{(k)}a_j^{(k)}\|_2 = 1 \), we have \( \langle U^{(k)}a_j^{(k)}, e_i^{(l)} \rangle \sim \mathcal{N}(0, \sigma^2/m) \). Using Lemma 3 in Appendix G, for \( \beta \geq \frac{1}{\sqrt{2\pi}} \), we hence get
\[ P \left( \left| \langle U^{(k)}a_j^{(k)}, e_i^{(l)} \rangle \right| > \frac{\sigma}{\sqrt{m}} \beta \right) \leq 2e^{-\frac{\beta^2}{2}}. \]  \hspace{1cm} (42)
Next, we upper-bound \( \Pr \left( |\langle e_j^{(k)}, e_i^{(l)} \rangle| \geq \phi \right) \). Conditional on \( e_i^{(l)} \), we have \( \langle e_j^{(k)}, e_i^{(l)} \rangle \sim \mathcal{N}(0, \frac{\sigma}{m} \| e_i^{(l)} \|_2^2) \). Lemma 3 yields, for \( \beta \geq \frac{1}{\sqrt{2\pi}} \), that
\[
\Pr \left[ \left| \langle e_j^{(k)}, e_i^{(l)} \rangle \right| \geq \beta \frac{\sigma}{\sqrt{m}} \| e_i^{(l)} \|_2 \right] \leq 2e^{-\frac{\beta^2}{2}}. \tag{43}
\]
Since \( \beta/\sqrt{m} \leq 1 \), by assumption, we get
\[
\Pr \left[ \| e_i^{(l)} \|_2 \geq 2\sigma \right] \leq \Pr \left[ \| e_i^{(l)} \|_2 \geq \left( 1 + \frac{\beta}{\sqrt{m}} \right) \sigma \right] \leq e^{-\frac{\beta^2}{2}} \tag{44}
\]
where the second inequality follows from (103). Next, note that for random variables \( X, Y \) (possibly dependent) and a constant \( \phi \), we have
\[
\Pr \left[ X \geq \phi \right] = \Pr \left[ \{ X \geq \phi \} \cup \{ X \geq \phi \} \cup \{ Y \geq X \geq \phi \} \right] \leq \Pr \left[ \{ X \geq \phi \} \cup \{ Y \geq \phi \} \right] \leq \Pr \left[ X \geq Y \right] + \Pr \left[ Y \geq \phi \right]. \tag{45}
\]
Combining (43) and (44) via (45) yields
\[
\Pr \left[ \left| \langle e_j^{(k)}, e_i^{(l)} \rangle \right| \geq \frac{2\sigma^2}{\sqrt{m}} \right] \leq \Pr \left[ \left| \langle e_j^{(k)}, e_i^{(l)} \rangle \right| \geq \beta \frac{\sigma}{\sqrt{m}} \| e_i^{(l)} \|_2 \right] + \Pr \left[ \| e_i^{(l)} \|_2 \geq 2\sigma \right] \leq 3e^{-\frac{\beta^2}{2}}. \tag{46}
\]
Finally, combining (42) and (46) via a union bound yields (35).

**Step 2, proof of (37)** We first upper-bound the probability that \( \max_j \tilde{z}_j^{(k)} \) (with \( \tilde{z}_j^{(k)} \) as defined in (30)) is larger than a constant, which then yields, via a union bound, an upper bound on the probability that \( \max_{k \neq l, j} \tilde{z}_j^{(k)} \) is larger than a constant. For convenience, we define \( B = U^{(k)^T} U^{(l)} \).

With this notation, \( \tilde{z}_j^{(k)} = |\langle a_j^{(k)}, B a_i^{(l)} \rangle| \). We have [27, Proof of Lem. 7.5]
\[
\Pr \left[ \left\| B a_i^{(l)} \right\|_2 \geq \frac{\| B \|_F}{\sqrt{d_l}} + \kappa \right] \leq e^{-\frac{d_l}{4\| B \|_F^2} \left( \| B \|_F \right)^2}. \tag{47}
\]
Setting \( \kappa = \beta \| B \|_F / \sqrt{d_l} \) in (47) yields
\[
\Pr \left[ \left\| B a_i^{(l)} \right\|_2 \geq \frac{1 + \beta}{\sqrt{d_l}} \| B \|_F \right] \leq e^{-\frac{d_l}{2\| B \|_F^2} \left( \| B \|_F \right)^2} \leq e^{-\frac{\beta^2}{2}}. \tag{48}
\]
By Theorem 12 in Appendix G we have
\[
\Pr \left[ \left| \langle a_j^{(k)}, B a_i^{(l)} \rangle \right| > \frac{\beta}{\sqrt{d_k}} \left\| B a_i^{(l)} \right\|_2 \right] \leq 2e^{-\frac{\beta^2}{2}}. \tag{49}
\]
Using (45) with $X = \max_j \tilde{z}_j^{(k)}$, $\phi = \frac{\beta}{\sqrt{d_k}} \frac{1 + \beta}{\sqrt{d_l}} \|B\|_F$, and $Y = \frac{\beta}{\sqrt{d_k}} \|B_j^{(l)}\|_2$, we get

$$
P \left[ \max_j \tilde{z}_j^{(k)} \geq \frac{\beta}{\sqrt{d_k}} \frac{1 + \beta}{\sqrt{d_l}} \|B\|_F \right] \leq P \left[ \max_j \left| \left\langle a_j^{(k)}, B_j^{(l)} \right\rangle \right| > \frac{\beta}{\sqrt{d_k}} \|B_j^{(l)}\|_2 \right] + P \left[ \|B_j^{(l)}\|_2 \geq \frac{1 + \beta}{\sqrt{d_l}} \|B\|_F \right]
$$

$$
\leq P \left[ \|B_j^{(l)}\|_2 \geq \frac{1 + \beta}{\sqrt{d_l}} \|B\|_F \right] + \sum_{j \in [n_k]} P \left[ \left| \left\langle a_j^{(k)}, B_j^{(l)} \right\rangle \right| > \frac{\beta}{\sqrt{d_k}} \|B_j^{(l)}\|_2 \right]
$$

(50)

where a union bound is used to obtain (50), and (51) follows from (48) and (49). Taking the union bound over $k \in [L] \backslash l$ concludes the proof of (37).

**Step 3, proof of** (38) **We first note that the pdf of** $\langle a_j^{(l)}, a_i^{(l)} \rangle$ **is given by** $f(z) = \frac{1}{\sqrt{\pi} \Gamma((d_l - 1)/2)} (1 - z^2)^{(d_l - 3)/2} 1_{\{|z| \leq 1\}}$. Hence, we get (recall that $\tilde{z}_j^{(l)} = |\langle a_j^{(l)}, a_i^{(l)} \rangle|$)

$$
P \left[ \tilde{z}_j^{(l)} \leq \frac{\nu}{\sqrt{d_l}} \right] = \frac{2}{\sqrt{\pi}} \Gamma((d_l - 1)/2) \int_0^{\sqrt{\nu}} (1 - z^2)^{(d_l - 3)/2} 1_{\{|z| \leq 1\}} dz
$$

$$
\leq \frac{2}{\sqrt{\pi}} \Gamma((d_l - 1)/2) \frac{\nu}{\sqrt{d_l}} \leq \sqrt{\frac{2}{\pi}} \nu
$$

(52)

where the last inequality follows from [8, Eq. 8.1]. Next, observe that,

$$
P \left[ \tilde{z}_{(n_l-q)}^{(l)} \leq \frac{\nu}{\sqrt{d_l}} \right] = P \left[ \text{there exists a set } I \subset [n_l - 1] \text{ with } |I| = n_l - q \right.
$$

such that $\tilde{z}_i^{(l)} \leq \frac{\nu}{\sqrt{d_l}}$ for all $i \in I$

$$
\leq \binom{n_l - 1}{n_l - q} \max_{|I| = n_l - q} P \left[ \tilde{z}_i^{(l)} \leq \frac{\nu}{\sqrt{d_l}}, \text{ for all } i \in I \right]
$$

(53)

$$
\leq \left( \frac{n_l - 1}{q - 1} \right)^{q - 1} \left( p^{n_l - q} \right)^{n_l - q} \left( e \phi \right)^{\frac{n_l - 1}{q - 1}} (n_l - 1) \left( 1 - \frac{1}{q - 1} \right)
$$

(54)

$$
\leq e^{- (n_l - 1) \left( \log \left( \frac{1}{\phi} \right) \left( 1 - \frac{1}{q - 1} \right) - \frac{1}{q - 1} \log(e \phi) \right)} = e^{- c(\phi, \nu)(n_l - 1)}
$$

(55)

where we used a union bound to get (53), $(\binom{n}{k}) \leq (\frac{n}{k})^k$ [3] for (54), and (52) yields (55); we also set $\phi = \frac{n_l - 1}{q - 1}$ for convenience. Here $c(\phi, \nu)$ satisfies $c(\phi, \nu) > 1/20$ for $\nu = 2/3$ and $\phi \geq 6$. Note that $\phi = \frac{n_l - 1}{q - 1} \geq \frac{n_l}{q} = \rho_l \geq 6$, where the last inequality holds by assumption.
C Proofs of Theorems 4 and 5

Theorems 4 and 5 rely on the following result.

**Theorem 10.** Suppose that $X_i, l \in [L]$, is obtained by choosing $n_l = \rho_l q$, with $\rho_l \geq 6$, points corresponding to $S_i$ at random according to $X_j = U(l) a_j + C_j^{\ell} c_j$, $j \in [n_l]$, where the $a_j^{\ell} \in \mathbb{R}^{d_l}$ are i.i.d. uniform on $S_l^{d_l-1}$, the $c_j^{\ell} \in \mathbb{R}^s$ are i.i.d. $\mathcal{N}(0, (\sigma^2/s) I_s)$, and $C_j^{\ell} \in \mathbb{R}^{mxs}$. Set $X = X_1 \cup \ldots \cup X_L$ and let $d_{\max} = \max_l d_l$. If

$$\max_{k,l: k \neq l} \text{aff}(S_k, S_l) + \xi \leq \frac{1}{12 \log N}$$

with

$$\xi = 2\sigma \sqrt{d_{\max}} \sqrt{s} \left( \sigma \max_{k,j: (k,j) \neq (l,i)} \frac{1}{\sqrt{s}} \left\| C_j^{(k)} C_i^{(l)} \right\|_F + 2 \max_{k,l,i, j} \frac{1}{\sqrt{d_k}} \left\| U^{(k)} C_i^{(l)} \right\|_F \right)$$

then the subspace detection property is satisfied (for $X$) with probability at least $1 - \frac{12}{N} - \sum_{l \in [L]} n_l e^{-c(n_l-1)}$, where $c > 0$ is an absolute constant.

**Proof of Theorem 5.** Theorem 5 follows from Theorem 10 by noting that $\left\| C_j^{(k)} C_i^{(l)} \right\|_F \leq \sqrt{s}$ and $\left\| U^{(k)} C_i^{(l)} \right\|_F \leq \sqrt{d_k}$ (recall that $C_j^{(l)}$ is a submatrix of the orthonormal matrix $C$ and thus has orthonormal columns). Thus, $\xi$ is upper-bounded by $2\sigma \sqrt{d_{\max}} (\sigma + 2)$, and hence (57) is implied by (14).

**Proof of Theorem 4.** Theorem 4 follows from Theorem 10 by bounding the probability that the clustering condition (57) is violated.

First, set $\delta = \frac{c_1}{\sqrt{m}} \sqrt{3d} + \log L$ with $c_3 = \sqrt{c_2}$, where $c_2$ is the constant in Lemma 2 in Appendix D. Using (12), we can bound the probability that (57) is violated according to

$$P\left[ \max_{k,l: k \neq l} \text{aff}(S_k, S_l) + \epsilon \geq \delta + 4 \frac{\sqrt{d}}{\sqrt{m}} \sigma (2 + \sigma) \right] \leq P\left[ \max_{k,l: k \neq l} \text{aff}(S_k, S_l) \geq \delta \right]$$

$$+ \sum_{(k,j) \neq (l,i)} P\left[ \frac{1}{\sqrt{s}} \left\| C_j^{(k)} C_i^{(l)} \right\|_F \geq 2 \sqrt{\frac{s}{m}} \right]$$

$$+ \sum_{k,l,i} P\left[ \frac{1}{\sqrt{s}} \left\| U^{(k)} C_i^{(l)} \right\|_F \geq 2 \sqrt{\frac{d}{m}} \right].$$

First, we upper-bound the probability in (60). Observe that $\left\| U^{(k)} C_i^{(l)} \right\|_F$ is the $\ell_2$-norm of an $\mathcal{N}(0, (1/m) I_{ds})$ random vector (recall that $C_i^{(l)}$ is a submatrix of the orthonormal matrix $C$ and thus has orthonormal columns), and therefore we have, according to (103) in Appendix G, that

$$P\left[ \frac{1}{\sqrt{s}} \left\| U^{(k)} C_i^{(l)} \right\|_F \geq \frac{\sqrt{ds} + \beta}{\sqrt{ms}} \right] \leq e^{-\beta^2 \frac{d}{2}}$$

which, with $\beta = \sqrt{ds}$, reduces to

$$P\left[ \frac{1}{\sqrt{s}} \left\| U^{(k)} C_i^{(l)} \right\|_F \geq 2 \sqrt{\frac{d}{m}} \right] \leq e^{-\frac{d}{2}}.$$
Next, we upper-bound the probability in (59). Since \( C_j^{(k)} \) and \( C_i^{(l)} \) are submatrices of the orthonormal matrix \( C \), \( \| C_j^{(k)T} C_i^{(l)} \|_F^2 \) is simply given by the number of columns in \( C_j^{(k)} \) and \( C_i^{(l)} \) that are equal. Since the \( C_j^{(k)} \) are chosen uniformly at random from all submatrices of \( C \) with \( s \) columns, this number is simply the number of successes in \( s \) draws without replacement from a population of size \( m \) containing \( s \) successes, and thus follows a hypergeometric distribution [26]. According to [26, Eq. 1.3], we have for \( \kappa > 0 \) that

\[
P \left[ \| C_j^{(k)T} C_i^{(l)} \|_F^2 \geq s \left( \frac{m}{s} + \kappa \right) \right] \leq e^{-2s \frac{m-s+1}{s} \kappa^2}
\]

and we get, with \( \kappa = 3s/m \), that

\[
P \left[ \frac{1}{\sqrt{s}} \| C_j^{(k)T} C_i^{(l)} \|_F \geq 2 \sqrt{s/m} \right] \leq e^{-18s \frac{m-s+1}{3s} \frac{s^2}{m^2}}.
\]

Finally, we have, with \( \text{aff}(S_k, S_l) \leq \text{affp}(S_k, S_l) \), that

\[
P \left[ \max_{k,l: k \neq l} \text{aff}(S_k, S_l) \geq \delta \right] \leq P \left[ \max_{k,l: k \neq l} \text{affp}(S_k, S_l) \geq \delta \right] \leq 4e^{-c_1 m}
\]

where the last inequality follows by application of Lemma 2 in Appendix D. Note that this requires \( m \geq \frac{1}{\pi^2} (3d + \log L) \) which is ensured by the definition of \( \delta \) above.

Using (61), (62), and (63) in (60) yields the following upper bound for the probability of (57) being violated:

\[
N^2 e^{-\frac{dn}{2}} + N^2 e^{-18s \frac{m-s+1}{3s} \frac{s^2}{m^2}} + 4e^{-c_1 m}
\]

which concludes the proof. \( \square \)

**Proof of Theorem 10.** The proof is identical to the proof of Theorem 3 in Appendix B, with Step 1 in Appendix B substituted by the following statement proven below: Setting

\[
\epsilon = \beta (1 + \beta) \frac{\sigma}{\sqrt{s}} \left( \sigma \max_{k, j: (k,j) \neq (i,j)} \frac{1}{\sqrt{s}} \left\| C_j^{(k)T} C_i^{(l)} \right\|_F + 2 \max_{k,l,i} \frac{1}{\sqrt{d_k}} \left\| U_j^{(k)T} C_i^{(l)} \right\|_F \right)
\]

we have for \( \beta \geq 0 \), that

\[
P \left[ |e_j^{(k)}| \geq \epsilon \right] \leq 9e^{-\frac{\beta^2}{2}}
\]

where

\[
e_j^{(k)} = \left\langle C_j^{(k)} e_j^{(k)}, C_i^{(l)} c_i^{(l)} \right\rangle + \left\langle C_j^{(k)} e_j^{(k)}, U^{(l)} a_i^{(l)} \right\rangle + \left\langle U^{(k)} a_j^{(k)}, C_i^{(l)} c_i^{(l)} \right\rangle.
\]

Note that (64) is equivalent to (35), with, however, a different choice of \( \epsilon \). Using (64) instead of (35) in the proof of Theorem 3, (37), and (38) in (34), and choosing \( \alpha \) and \( \beta \) as in the proof of Theorem 3 yields

\[
P \left[ z_j^{(l)} \leq \max_{k \neq l} z_j^{(k)} \right] \leq \frac{12}{N^2} + e^{-c(n+1)}.
\]
Recall that for (39) we imposed the condition $\alpha + 2\epsilon \leq \nu / \sqrt{d_l}$ with $\nu = 2/3$, which here becomes

$$\max_{k,l: k \neq l} \text{aff}(S_k, S_l) + 2\sigma \frac{\sqrt{d_l}}{\sqrt{s}} \left( \sigma \max_{k,j: (k,j) \neq (l,i)} \frac{1}{\sqrt{s}} \| C_j^{(k)} C_i^{(l)} \|_F \right) + 2\max_{k,l,i} \frac{1}{\sqrt{d_k}} \| U^{(k)} C_i^{(l)} \|_F \leq \frac{2}{3} (1 + \beta).$$ (66)

The proof is concluded by using $(1 + \beta) \leq 3.2\sqrt{\log N}$ (cf. (41)) to show that

$$\frac{2}{3(1 + \beta)} \geq \frac{1}{12 \log N},$$ (67)

which, in turn, yields that (66) is implied by (57).

It remains to prove (64). For expositional convenience, we set $B = C_i^{(l)} C_j^{(k)}$. We start by upper-bounding $P \left[ \left\langle B c_j^{(k)}, c_i^{(l)} \right\rangle \geq \phi \right]$. Conditional on $c_j^{(k)}$ we have $\left\langle B c_j^{(k)}, c_i^{(l)} \right\rangle \sim N(0, \beta \| B c_j^{(k)} \|_2^2)$. Using Lemma 3, we get

$$P \left[ \left\langle B c_j^{(k)}, c_i^{(l)} \right\rangle \geq \beta \frac{\sigma}{\sqrt{s}} \| B c_j^{(k)} \|_2 \right] \leq 2e^{-\beta^2 \frac{s}{s}}.$$ (68)

Invoking (45) with $X = \left\langle B c_j^{(k)}, c_i^{(l)} \right\rangle$, $\phi = \| B \|_F \sigma^2 \frac{\sqrt{1 + \beta}}{\sqrt{s}}$, and $Y = \beta \frac{\sigma}{\sqrt{s}} \| B c_j^{(k)} \|_2$, we obtain

$$P \left[ \left\langle B c_j^{(k)}, c_i^{(l)} \right\rangle \geq \| B \|_F \sigma^2 \frac{\sqrt{1 + \beta}}{\sqrt{s}} \right] \leq P \left[ \left\langle B c_j^{(k)}, c_i^{(l)} \right\rangle \geq \beta \frac{\sigma}{\sqrt{s}} \| B c_j^{(k)} \|_2 \right] + P \left[ \| B c_j^{(k)} \|_2 \geq \frac{\sqrt{1 + \beta}}{\sqrt{s}} \| B \|_F \right] \leq 3e^{-\beta^2 \frac{s}{s}}.$$ (69)

where we used (68) and (102) to get (69).

Next, for expositional convenience, set $D = U^{(l)} C_j^{(k)}$. Conditional on $c_j^{(k)}$, it follows from Theorem 12 that

$$P \left[ \left\langle D c_j^{(k)}, a_i^{(l)} \right\rangle \geq \beta \frac{\sigma}{\sqrt{d_l}} \| D c_j^{(k)} \|_2 \right] \leq 2e^{-\beta^2 \frac{s}{s}}.$$ (70)

By (102) we have

$$P \left[ \| D c_j^{(k)} \|_2 \geq \| D \|_F (1 + \beta) \frac{\sigma}{\sqrt{d_l}} \right] \leq e^{-\frac{\beta^2}{2}}.$$ (71)

Again invoking (45) with $X = \left\langle D c_j^{(k)}, a_i^{(l)} \right\rangle$, $\phi = \| D \|_F (1 + \beta) \frac{\sigma}{\sqrt{d_l}} \frac{\beta}{\sqrt{s}}$, and $Y = \frac{\beta}{\sqrt{d_l}} \| D c_j^{(k)} \|_2$, we get

$$P \left[ \left\langle D c_j^{(k)}, a_i^{(l)} \right\rangle \geq \| D \|_F (1 + \beta) \frac{\sigma}{\sqrt{d_l}} \frac{\beta}{\sqrt{s}} \right] \leq P \left[ \left\langle D c_j^{(k)}, a_i^{(l)} \right\rangle \geq \beta \frac{\sigma}{\sqrt{d_l}} \| D c_j^{(k)} \|_2 \right] + P \left[ \| D c_j^{(k)} \|_2 \geq \| D \|_F (1 + \beta) \frac{\sigma}{\sqrt{s}} \right] \leq 3e^{-\beta^2 \frac{s}{s}}.$$ (72)

where we used (70) and (71) to get (72). Finally, (64) is obtained by combining (69) and (72) via a union bound. 

\[\square\]
D Proof of Theorem 6

Theorem 6 follows from Theorem 9 and the following result from [8], which builds on the concentration inequality the Johnson-Lindenstrauss Lemma is based on, and a covering argument.

Lemma 1 ([8, Eq. 9.12 with \( \rho = \frac{2}{c_1^2} \)]. Let \( U \) be a \( p \times s \) random matrix which satisfies for some \( \tilde{c} > 0 \), for every \( x \in \mathbb{R}^s \), and every \( t \in (0, 1) \),

\[
P\left[ \left\| Ux \right\|_2^2 - \left\| x \right\|_2^2 \geq t^2 \left\| x \right\|_2^2 \right] \leq 2e^{-c_1^2t^2p}.
\] (73)

Then, we have

\[
P\left[ \left\| U^T U - I_s \right\|_{2 \rightarrow 2} \geq \delta \right] \leq 2e^{-0.6\delta^2p+3s}. \tag{74}
\]

We note that (73) is satisfied for random matrices with i.i.d. \( \mathcal{N}(0, 1/p) \) entries as well as for random matrices with i.i.d. subgaussian entries of variance 1/p [8, Lem. 9.8] (a random variable \( x \) is subgaussian [8, Sec. 7.4] if its tail probability satisfies \( P[|x| > t] \leq c_1 e^{-c_2t^2} \) for constants \( c_1, c_2 > 0 \)).

Lemma 2. Let the entries of the \( U^{(l)} \in \mathbb{R}^{m \times d} \), \( l \in [L] \), be i.i.d. \( \mathcal{N}(0, 1/m) \), and let \( U^{(l)}_D \in \mathbb{R}^{m \times d} \) be the matrix obtained from \( U^{(l)} \in \mathbb{R}^{m \times d} \) by setting the rows with indices in \( D \subseteq [m] \) to zero. Then, we have for \( \delta \in (0, 1) \) that

\[
\min_{l, D: \ |D| \leq 2s, \ |a|_2 = 1} \left\| U^{(l)}_D^T U^{(l)} a \right\|_2 \geq (1 - \delta) \frac{m - 2s}{m} \tag{75}
\]

and

\[
\max_{k \neq l, D: \ |D| \leq 2s, \ |a|_2 = 1} \left\| U^{(k)}_D^T U^{(l)} a \right\|_2 \leq \delta \tag{76}
\]

with probability at least \( 1 - 4e^{-c_1m} \), provided that

\[
m \geq \frac{c_2}{\delta^2} \left( 3d + \log L + s \log \left( \frac{me}{2s} \right) \right) + c_3 s \tag{77}
\]

where \( c_1, c_2, c_3 > 0 \) are absolute constants.

Before proving Lemma 2, we show how the proof of Theorem 6 can be completed using Lemma 2. Set \( \delta = \frac{c_2}{\rho^2} \frac{m - 2s}{c_3} \sqrt{\frac{\log p}{\log N}} \), where \( c_1 \) is the constant in (21) and \( c_3 \) is the constant in (15). With this choice of \( \delta \), (15) (with \( c_4 = c_2 (2c_3/(c_1(c_3 - 2)))^2 \)) implies (77) and hence, for \( \eta = (1 - \delta) \frac{m - 2s}{m} \), according to Lemma 2, (19) and (20) are satisfied with probability \( \geq 1 - 4e^{-\tilde{c}m} \), where \( \tilde{c} \) is an absolute constant. Next, observe that using \( c_1 \leq 1 \) (by assumption) and \( \rho \leq N \), we have \( \delta = \frac{c_2}{\rho^2} \frac{m - 2s}{c_3} \sqrt{\frac{\log p}{\log N}} \leq \frac{1}{2} \). Thus, \( \frac{\delta}{1 - \delta} \leq 2\delta \), and hence

\[
\frac{\delta}{\eta} = \frac{\delta}{1 - \delta} \frac{m}{m - 2s} \leq 2\delta \frac{m}{m - 2s} \leq 2\delta \frac{c_3}{c_3 - 2} = c_1 \sqrt{\frac{\log p}{\log N}}
\]

where we used that \( \frac{m}{m - 2s} \leq \frac{c_3}{c_3 - 2} \) as a consequence of \( m \geq c_3 s \) (from (15)). We therefore established that (21) holds with probability \( \geq 1 - 4e^{-\tilde{c}m} \), and application of Theorem 9 concludes the proof.
Proof of Lemma 2. We start with (76). First, note that since the rows of $U_D^{(k)}$ indexed by $D$ have all entries equal to zero by definition, we have $U_D^{(k)} U^{(l)} = V_i^T V_j$, where $V_i \in \mathbb{R}^{p \times d}$ and $V_j \in \mathbb{R}^{p \times d}$, with $p = m - |D|$, denote the restrictions of $U^{(k)}$ and $U^{(l)}$, respectively, to the rows indexed by $[m] \setminus D$. Set $V_i = \sqrt{m/p} V_i$, let $U = [V_i \ V_j] \in \mathbb{R}^{p \times 2d}$, and note that the entries of $U$ are i.i.d. $\mathcal{N}(0, 1/p)$. Using $m \geq p$, we have

$$
\|V_i^T V_j\|_{2 \to 2} \leq \frac{m}{p} \|V_i^T V_j\|_{2 \to 2} = \|\tilde{V}_i \tilde{V}_j\|_{2 \to 2} \leq \|U^T U - I_{2d}\|_{2 \to 2}
$$

where the last inequality follows from the fact that $\tilde{V}_i \tilde{V}_j$ is a submatrix of $U^T U - I_{2d}$ [31, Ex. 3.4]. Therefore, we get

$$
P[\|V_i^T V_j\|_{2 \to 2} \geq \delta] \leq P[\|U^T U - I_{2d}\|_{2 \to 2} \geq \delta] \leq 2e^{-c_0 \delta^2 p + 3d} \quad (78)
$$

$$
\leq 2e^{-c_0 \delta^2 (m-2s) + 6d} \quad (79)
$$

where $c_0$ is an absolute constant, (78) follows from Lemma 1, and (79) is a consequence of $p \geq m-2s$.

Taking the union bound over all pairs $(i, j)$, i.e., over all pairs $(k, l)$ with $k, l \in [L]$ and for each of those pairs $(k, l)$ over all $D \subseteq [m]$ with $|D| = 2s$, i.e., over $(m/2s)^2 \leq (mc^2/(2s))$ sets, we obtain

$$
P[\max_{i \neq j} \|V_i^T V_j\|_{2 \to 2} \geq \delta] \leq L^2 \left(\frac{mc}{2s}\right)^{2s} 2e^{-c_0 \delta^2 (m-2s) + 6d} \quad (80)
$$

$$
= 2e^{-c_0 \delta^2 (m-2s) + 6d + 2 \log L + 2s \log \left(\frac{mc}{2s}\right)}
$$

$$
= 2e^{-c_0 \delta^2 m - 2s - \frac{2}{c_0} \log (L + s \log \left(\frac{mc}{2s}\right))} \quad (81)
$$

$$
\leq 2e^{-c_1 m} \quad (82)
$$

where the last inequality holds for $c_1 > 0$ provided that (77) (with $c_2$ and $c_3$ sufficiently large) holds, as (77) implies that the term in the square brackets in (81) equals $mc_4$ for a constant $c_4 > 0$. We therefore established that (76) holds with probability $\geq 1 - 2e^{-c_1 m}$, provided that (77) is satisfied.

We next prove (75). With the same notation as above, applying Lemma 1 to $\tilde{V}_i$ (recall that the entries of $V_i$ are i.i.d. $\mathcal{N}(0, 1/p)$), we get

$$
P[\|\tilde{V}_i^T \tilde{V}_i - I_d\|_{2 \to 2} \geq \delta] \leq 2e^{-c_0 \delta^2 p + 3d}.
$$

Next, taking the union bound over all $L$ subspaces and over all $D \subseteq [m]$ with $|D| \leq 2s$, yields

$$
P[\max_i \|\tilde{V}_i^T \tilde{V}_i - I_d\|_{2 \to 2} \geq \delta] \leq L \left(\frac{mc}{2s}\right)^{2s} 2e^{-c_0 \delta^2 p + 3d} \quad (83)
$$

$$
\leq 2e^{-c_1 m} \quad (84)
$$

where we used that the RHS of (83) is smaller than the RHS of (80) (recall that $p \geq m - 2s$) and therefore (84) follows from the steps leading from (80) to (82). Note that for every $B \in \mathbb{R}^{m \times d}$ and every $a \in \mathbb{R}^d$, we have

$$
\|a\|_2^2 - \|Ba\|_2^2 = \langle (I_d - B^T B) a, a \rangle \leq \|B^T B - I_d\|_{2 \to 2} \|a\|_2 \quad (85)
$$
where we used that \( \mathbf{I}_d - \mathbf{B}^T \mathbf{B} \) is self adjoint [15]. Since there is an \( \mathbf{a} \in \mathbb{R}^d \) that satisfies (85) with equality, we have
\[
1 - \|\mathbf{B}^T \mathbf{B} - \mathbf{I}_d\|_{2\rightarrow 2} = \min_{\|\mathbf{a}\|_2 = 1} \|\mathbf{B} \mathbf{a}\|_2^2 = \min_{\|\mathbf{a}\|_2 = 1} \|\mathbf{B}^T \mathbf{B} \mathbf{a}\|_2.
\] (86)

Applying (86), we get
\[
\min_{i, \|\mathbf{a}\|_2 = 1} \|\mathbf{V}_i^T \mathbf{V}_i \mathbf{a}\|_2 = \frac{p}{m} \min_{i, \|\mathbf{a}\|_2 = 1} \|\mathbf{\tilde{V}}_i^T \mathbf{\tilde{V}}_i \mathbf{a}\|_2
= \frac{p}{m} \left(1 - \max_i \|\mathbf{\tilde{V}}_i^T \mathbf{\tilde{V}}_i - \mathbf{I}_d\|_{2\rightarrow 2}\right).
\] (87)

From (87), and \( p \geq m - 2s \), it follows that
\[
P\left[ \min_{i, \|\mathbf{a}\|_2 = 1} \|\mathbf{V}_i^T \mathbf{V}_i \mathbf{a}\|_2 \leq \frac{m - 2s}{m} (1 - \delta) \right] \leq P\left[ \min_{i, \|\mathbf{a}\|_2 = 1} \|\mathbf{V}_i^T \mathbf{V}_i \mathbf{a}\|_2 \leq \frac{p}{m} (1 - \delta) \right]
= P\left[ \max_i \|\mathbf{\tilde{V}}_i^T \mathbf{\tilde{V}}_i - \mathbf{I}_d\|_{2\rightarrow 2} \geq \delta \right] \leq 2e^{-c_1 m}
\]
where the last inequality follows by application of (84). We therefore established (75), which concludes the proof.

### E Proof of Theorem 7

The proof consists of two parts, corresponding to the two statements in Theorem 7. First, we bound the probability that the outlier detection scheme does not detect all outliers, and then we bound the probability that one or more of the inliers are misclassified.

We start by bounding the probability that the outlier detection scheme fails to detect a given outlier. A union bound over all \( N_0 \) outliers will then yield a bound on the probability that the outlier detection scheme does not detect all outliers. Let \( \mathbf{x}_j \) be an outlier. The probability that (16) with \( c = \sqrt{6} \) is violated for \( \mathbf{x}_j \) can be upper-bounded as
\[
P\left[ \max_{i \in [N] \setminus j} |\langle \mathbf{x}_i, \mathbf{x}_j \rangle| > \frac{\sqrt{6 \log N}}{\sqrt{m}} \right] \leq \sum_{i \in [N] \setminus j} P\left[ |\langle \mathbf{x}_i, \mathbf{x}_j \rangle| > \frac{\sqrt{6 \log N}}{\sqrt{m}} \|\mathbf{x}_j\|_2 \right]
\leq 2Ne^{-3\log N} = \frac{2}{N^2}
\] (88)
where we used the union bound and \( \|\mathbf{x}_j\|_2 = 1 \) in the first inequality and Theorem 12 in the second. Taking the union bound over all \( N_0 \) outliers, we have thus established that the probability of our scheme failing to detect one or more outliers is at most \( 2N_0/N^2 \).

Next, we bound the probability that the outlier detection scheme misclassifies a given inlier as an outlier. A union bound over all \( n_l \) inliers in \( S_l \) will then yield a bound on the probability that the outlier detection scheme misclassifies one or more of the inliers in \( S_l \) as an outlier. Consider the inlier \( \mathbf{x}_j \in S_l \). With
\[
\max_{i \in [N] \setminus j} |\langle \mathbf{x}_i, \mathbf{x}_j \rangle| \geq \max_{i \in [n_l] \setminus j} |\langle \mathbf{x}_i^{(l)}, \mathbf{x}_j^{(l)} \rangle| = \max_{i \in [n_l] \setminus j} |\langle \mathbf{a}_i^{(l)}, \mathbf{a}_j^{(l)} \rangle|\]
and using (17), i.e., $\sqrt{6} \log N / \sqrt{m} \leq 1 / \sqrt{d_{\text{max}}} \leq 1 / \sqrt{d_l}$, the probability that (16) holds with $c = \sqrt{6}$ for $x_j$ can be bounded as

$$\Pr \left[ \max_{i \in [N]} |\langle x_i, x_j \rangle| \leq \frac{\sqrt{6} \log N}{\sqrt{m}} \right] \leq \prod_{i \in [N]} \Pr \left[ |\langle a_i^{(l)}, a_j^{(l)} \rangle| \leq \frac{1}{\sqrt{d_l}} \right].$$

Then, it follows from (52) with $\nu = 1$. Taking the union bound over all inliers proves that the probability of the outlier detection scheme misclassifying one or more of the inliers in $S_l$ as an outlier is at most $n_6e^{-\frac{1}{2} \log(\frac{d}{\nu})(n_i-1)}$, which concludes the proof.

## F Proof of Theorem 8

The basic structure of the proof is the same as that of the proof of Theorem 7. The individual steps are more technical, though, due to the additive noise term.

We start by bounding the probability that the outlier detection scheme fails to detect a given outlier. A union bound over all $N_0$ outliers will then yield a bound on the probability that the outlier detection scheme does not detect all outliers. Let $x_j$ be an outlier and set $\beta = \sqrt{6} \log N$. The probability that (16) with $c = 2.3\sqrt{6}$ is violated for $x_j$ can be upper-bounded as

$$\Pr \left[ \max_{i \in [N]} |\langle x_i, x_j \rangle| > \frac{2.3\beta}{\sqrt{m}} \right] \leq \sum_{i \in [N]} \Pr \left[ |\langle x_i, x_j \rangle| \geq \frac{2.3\beta}{\sqrt{m}} \right] \leq \sum_{i \in [N]} \left( \Pr \left[ |\langle x_i, x_j \rangle| \geq \frac{\beta}{\sqrt{m}} \|x_i\|_2 \right] + \Pr[\|x_i\|_2 \geq 2.3] \right)$$

where we used (45) with $X = |\langle x_i, x_j \rangle|$, $\phi = \frac{2.3\beta}{\sqrt{m}}$, and $Y = \frac{\beta}{\sqrt{m}} \|x_i\|_2$ to get (90). We next bound the first term in the sum in (90). Since $x_j \sim \mathcal{N}(0, (1/m)I_m)$, we have that, conditioned on $x_i$, $\langle x_j, x_i \rangle \sim \mathcal{N}(0, \|x_i\|_2^2/m)$. Hence, with $\beta = \sqrt{6} \log N$, it follows from (100) that

$$\Pr \left[ |\langle x_j, x_i \rangle| \geq \frac{\beta}{\sqrt{m}} \|x_i\|_2 \right] \leq 2e^{-\frac{\beta^2}{2}} = \frac{2}{N^3}, \quad (91)$$

We next bound the second term in the sum in (90) and treat the cases where $x_i$ is an inlier and where it is an outlier separately.

First, suppose that $x_i$ is an inlier. Since $\frac{1+2\sigma}{\sqrt{1+\sigma^2}} \leq 2.3$ for $\sigma \geq 0$, we have

$$\Pr[\|x_i\|_2 \geq 2.3] \leq \Pr \left[ \frac{1}{\sqrt{1+\sigma^2}} \left\| U^{(l)} a_i^{(l)} + e_i^{(l)} \right\|_2 \geq \frac{1+2\sigma}{\sqrt{1+\sigma^2}} \right] \leq \Pr \left[ \left\| U^{(l)} a_i^{(l)} \right\|_2 + \left\| e_i^{(l)} \right\|_2 \geq 1+2\sigma \right] \leq \Pr \left[ \left\| e_i^{(l)} \right\|_2 \geq 2\sigma \right] \leq \frac{1}{N^3},$$

(93)
where (92) follows from the triangle inequality, and to obtain (93) we applied (44) with $\beta = \sqrt{6 \log N}$ and used that $\frac{\beta}{\sqrt{m}} \leq 1$, by assumption.

Next, suppose that $x_i$ is an outlier. Applying (44) with $\sigma = 1$ (again using that $\frac{\beta}{\sqrt{m}} \leq 1$, by assumption), we have

$$P[\|x_i\|_2 \geq 2.3] \leq \frac{1}{N^3}. \quad (94)$$

Finally, combining (91), (93) (for $x_i$ an inlier), and (94) (for $x_i$ an outlier) in (90) yields

$$P\left[ \max_{i \in [N] \setminus j} |\langle x_i, x_j \rangle| > \frac{2.3\sqrt{6 \log N}}{\sqrt{m}} \right] \leq \sum_{i \in [N] \setminus j} \left( \frac{2}{N^3} + \frac{1}{N^3} \right) \leq N \cdot \frac{3}{N^3}. \quad (95)$$

Taking the union bound over all $N_0$ outliers yields the desired result, namely that the probability of our scheme not detecting all $N_0$ outliers is at most $N_0 \cdot \frac{3}{N^3}$.

Next, we bound the probability that the outlier detection scheme misclassifies a given inlier as an outlier. Consider the inlier $x_j$ belonging to subspace $S_l$. Then

$$\max_{i \in [N] \setminus j} |\langle x_i, x_j \rangle| = \max_{i \in [n] \setminus j} |\langle x_i^{(l)}, x_j^{(l)} \rangle| = \frac{1}{1 + \sigma^2} z_i^{(l)}$$

with $z_i^{(l)} = |\langle a_j^{(k)} U^{(k)} x_i^{(l)}, a_j^{(k)} U^{(l)} \rangle + e_j^{(k)}|$. Thus, for $\epsilon \geq 0$, under the assumption

$$\frac{1}{1 + \sigma^2} \left( \frac{1}{\sqrt{d_l}} - \epsilon \right) \geq \frac{2.3\sqrt{6 \log N}}{\sqrt{m}}$$

resolved below, we have

$$P\left[ \max_{i \in [N] \setminus j} |\langle x_j, x_i \rangle| \leq \frac{2.3\sqrt{6 \log N}}{\sqrt{m}} \right] \leq P\left[ \max_{i \in [n] \setminus j} \frac{1}{1 + \sigma^2} z_i^{(l)} \leq \frac{1}{1 + \sigma^2} \left( \frac{1}{\sqrt{d_l}} - \epsilon \right) \right]$$

$$= P\left[ \max_{i \in [n] \setminus j} z_i^{(l)} \leq \frac{1}{\sqrt{d_l}} - \epsilon \right]$$

$$\leq P\left[ \max_{i \in [n] \setminus j} z_i^{(l)} - \max_{i \in [n] \setminus j} |e_i^{(l)}| \leq \frac{1}{\sqrt{d_l}} - \epsilon \right]$$

$$\leq P\left[ \max_{i \in [n] \setminus j} z_i^{(l)} \leq \frac{1}{\sqrt{d_l}} \right] + P\left[ \epsilon \leq \max_{i \in [n] \setminus j} |e_i^{(l)}| \right] \quad (96)$$

where (96) follows with $e_i^{(l)}$ and $\tilde{z}_i^{(l)}$ as defined in (32) and (30), respectively, (97) follows from (25) with $X = \max_{i \in [n] \setminus j} z_i^{(l)} - \frac{1}{\sqrt{d_l}}$, $Y = \max_{i \in [n] \setminus j} |e_i^{(l)}| - \epsilon$, and $\phi = \varphi = 0$. Next, note that (52) with $\nu = 1$ yields

$$P\left[ \max_{i \in [n] \setminus j} \tilde{z}_i^{(l)} \leq \frac{1}{\sqrt{d_l}} \right] = \prod_{i \in [n] \setminus j} P\left[ \tilde{z}_i^{(l)} \leq \frac{1}{\sqrt{d_l}} \right]$$

$$\leq \prod_{i \in [n] \setminus j} \sqrt{\frac{2}{\pi}} = e^{-\frac{1}{3} \log(\frac{2}{\pi})} \approx 0.1.$$ 

(97)
Application of (98) and (35) with $\epsilon = \frac{2\sigma(1+\sigma)}{\sqrt{m}} \sqrt{6 \log N}$ (using that $\beta/\sqrt{m} \leq 1$, as verified below) to (97) yields

$$P \left[ \max_{i \in [N] \setminus J} |\langle x_j, x_i \rangle| \leq \frac{2.3 \sqrt{6 \log N}}{\sqrt{m}} \right] \leq e^{-\frac{1}{2} \log(N)(n_l-1) + n_l \frac{7}{N^3}}. \quad (99)$$

To see that $\frac{\beta}{\sqrt{m}} \leq 1$, note that for $c_1$ sufficiently small (i.e., $c_1 \leq 1/6$), we have from (18),

$$\frac{1}{m} \leq d_{\text{max}} \leq \frac{c_1}{(1+\sigma^2)^2 \log N} \leq \frac{c_1}{\log N}.$$

Taking a union bound over all inliers belonging to $S_l$ shows that the probability of the outlier detection scheme misclassifying one or more of the inliers belonging to $S_l$ as an outlier is at most

$$n_l \left( e^{-\frac{1}{2} \log(N)(n_l-1) + n_l \frac{7}{N^3}} \right).$$

Finally, we show that (95), for all $l \in [L]$, is implied by (18). Rewriting (95) yields

$$\frac{1}{1 + \sigma^2} \frac{1}{\sqrt{6 \log N}} \geq \sqrt{d_l} \left( 2.3 + \frac{2\sigma(1+\sigma)}{1+\sigma^2} \right).$$

Since $\frac{\sigma(1+\sigma)}{1+\sigma^2} \leq 1.3$ for $\sigma \geq 0$, (95) is implied by

$$\frac{1}{1 + \sigma^2} \frac{1}{\sqrt{6 \log N}} \geq \sqrt{d_{\text{max}}} \frac{4.9}{\sqrt{m}}$$

which is equal to (18) with $c_1 = \frac{1}{(4.9)^2 \cdot 6}$.

G Supplementary results

For convenience, in the following, we summarize tail bounds from the literature that are frequently used throughout this paper. We start with a well-known tail bound on Gaussian random variables.

**Lemma 3** ([17, Prop. 19.4.2]). Let $x \sim \mathcal{N}(0, 1)$. For $\beta \geq \frac{1}{\sqrt{2\pi}}$, we have

$$P[x \geq \beta] \leq e^{-\frac{\beta^2}{2}}. \quad (100)$$

**Theorem 11** ([20, Eq. 1.6]). Let $f$ be Lipschitz on $\mathbb{R}^m$ with Lipschitz constant $L \in \mathbb{R}$, i.e., $|f(b_1) - f(b_2)| \leq L\|b_1 - b_2\|_2$ for all $b_1, b_2 \in \mathbb{R}^m$, and let $x \sim \mathcal{N}(0, I_m)$. Then, for $\beta > 0$, we have

$$P[f(x) \geq \mathbb{E}[f(x)] + \beta] \leq e^{-\frac{\beta^2}{2L^2}}.$$
Applying the concentration inequality in Theorem 11 to \( f(x) = \|Bx\|_2 \) with \( B \in \mathbb{R}^{s \times m} \) which has \( L = \|B\|_{2 \rightarrow 2} \), and using that by Jensen’s inequality \((\mathbb{E}[\|Bx\|_2])^2 \leq \mathbb{E}[\|Bx\|_2]^2 = \|B\|_F^2 \), we get
\[
P[\|Bx\|_2 \geq \|B\|_F + \beta] \leq e^{-\frac{\beta^2}{2\|B\|_{2 \rightarrow 2}^2}}. \tag{101}
\]
With \( \|B\|_{2 \rightarrow 2} \leq \|B\|_F \) (101) becomes
\[
P[\|Bx\|_2 \geq \|B\|_F(1 + \beta)] \leq e^{-\frac{\beta^2}{2}}. \tag{102}
\]
We will frequently need (101) particularized to \( B = I_m \), which reads
\[
P[\|x\|_2 \geq \sqrt{m} + \beta] \leq e^{-\frac{\beta^2}{2}}. \tag{103}
\]

**Theorem 12** ([2, Theorem 1]). Let \( a \) be distributed uniformly on \( S^{m-1} \) and fix \( b \in \mathbb{R}^m \). Then, for \( \beta \geq 0 \), we have
\[
P[|\langle a, b \rangle| > \frac{\beta}{\sqrt{m}} \|b\|_2] \leq 2e^{-\frac{\beta^2}{2}}.
\]

**Lemma 4.** Consider the order statistics \( z(1) \leq z(2) \leq \ldots \leq z(n) \) for the set \( \{z_i\}_{i \in [n]} \), where \( z_i = |x_i| \) and the \( x_i \) are i.i.d. \( \mathcal{N}(0, 1) \), and set \( \gamma = r/n \). For \( \gamma > 0.7 \) there are constants \( c(\gamma) > 0 \) and \( \alpha(\gamma) > 0 \) such that
\[
P[z_{(r)} \leq \sqrt{\alpha(\gamma) \log(1/(1 - \gamma))}] \leq e^{-c(\gamma)n}. \tag{104}
\]

**Proof.**
\[
P[z_{(r)} \leq \beta] = P[\text{There exists a set } I \text{ with } |I| = r \text{ such that } z_i \leq \beta \text{ for all } i \in I]
\]
\[
\leq \binom{n}{r} \max_{|I| = r} P[z_i \leq \beta, \text{ for all } i \in I]
\]
\[
= \binom{n}{r} (P[z_i \leq \beta])^r
\]
\[
\leq \left( \frac{en}{n - r} \right)^{n-r} (P[z_i \leq \beta])^r
\]
\[
= \left( \frac{e}{1 - \gamma} \right)^{n(1 - \gamma)} (P[z_i \leq \beta])^{n\gamma} \tag{105}
\]
where we used \( \binom{n}{r} = \binom{n}{n-r} \) and \( \binom{n}{k} \leq \left( \frac{en}{k} \right)^k \) [3] in (105). We first consider the case where \( \gamma \) is large, i.e., close to one, set \( \beta = \sqrt{\alpha \log(1/(1 - \gamma))} \), and assume that \( \alpha \) is chosen such that \( \beta \geq 2 \). With \( 2e^{-\frac{\beta^2}{2}} \leq \left( 1 - \frac{1}{\beta^2} \right)^{\frac{1}{\beta}} \) for \( \beta \geq 2 \), and a standard lower bound on the \( Q \)-function [17, Prop. 19.4.2], we get
\[
\sqrt{\frac{2}{\pi}} e^{-\beta^2} = 2e^{-\frac{\beta^2}{2}} \frac{1}{\sqrt{2\pi}} e^{-\frac{\beta^2}{2}} \leq \left( 1 - \frac{1}{\beta^2} \right) \frac{1}{\beta\sqrt{2\pi}} e^{-\frac{\beta^2}{2}}
\]
\[
\leq Q(\beta) = \frac{1 - P[z_i \leq \beta]}{2}.
\]
Applying this to (106) yields

\[
P[z(r) \leq \beta] \leq \left( \frac{e}{1 - \gamma} \right)^{(1 - \gamma)} \left( 1 - 2\sqrt{\frac{2}{\pi} e^{-\beta^2}} \right)^{\gamma n} \\
= \left( \frac{e}{1 - \gamma} \right)^{(1 - \gamma)} \left( 1 - 2\sqrt{\frac{2}{\pi} (1 - \gamma)^\alpha} \right)^{\gamma n} \\
\leq e^{-c(\gamma, \alpha)n}
\]

(107)

(108)

where we used \( \beta = \sqrt{\alpha \log(1/(1 - \gamma))} \). For the last inequality, we assumed that the term in the outer brackets in (107) is smaller than 1; this is the case if \( \gamma \) is sufficiently close to 1 so that \( \alpha \) can be chosen sufficiently small. Else, we simply set \( \beta = 1/2 \), thus get \( \alpha = 1/(4 \log(1/(1 - \gamma))) \), and note that \( P[z_i \leq 1/2] < 0.383 \) (by numerical evaluation). Thus, with (106), we have

\[
P[z(r) \leq 1] = \left( \frac{e}{1 - \gamma} \right)^{(1 - \gamma)} \left( 0.383^\gamma \right)^n \\
\leq e^{-c(\gamma, \alpha)n}
\]

(109)

and we see (by numerical evaluation) that \( c(\gamma, \alpha) > 0 \) as long as \( \gamma \geq 0.7 \).

**Lemma 5.** Let \( z(n) = \max_{i \in [n]} z_i \), where \( z_i = |x_i| \) and the \( x_i \) are i.i.d. \( \mathcal{N}(0,1) \). Then, for \( \xi \geq 0 \), we have

\[
P[z(n) \geq \sqrt{2(1 + \xi) \log n}] \leq \frac{2}{n^2}.
\]

(110)

**Proof.** Observe that with \( \beta = \sqrt{2(1 + \xi) \log n} \), we have

\[
P[z(n) \geq \beta] = P[\{z_1 \geq \beta\} \cup \ldots \cup \{z_n \geq \beta\}] \\
\leq nP[z_i \geq \beta] \\
\leq 2ne^{-\beta^2/2} = 2e^{-\beta^2/2 + \log n} = \frac{2}{n^2}
\]

(111)

(112)

(113)

where (112) follows from the union bound, and (113) from Lemma 3.

**References**


