Robust Subspace Clustering via Thresholding

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

July 2013; last revised August 2015

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 subspace clustering algorithm, which applies spectral clustering to an adjacency matrix obtained by thresholding the correlations between data points. In other words, the adjacency matrix is constructed from the nearest neighbors of each data point in spherical distance. A statistical performance analysis shows that the algorithm exhibits robustness to additive noise and succeeds even when the subspaces intersect. Specifically, our results reveal an explicit tradeoff between the affinity of the subspaces and the tolerable noise level. We furthermore prove that the algorithm succeeds even when the data points are incompletely observed with the number of missing entries allowed to be (up to a log-factor) linear in the ambient dimension. We also propose a simple scheme that provably detects outliers, and we 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 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 [3]. This and similar insights for other types of data have motivated research on finding low-dimensional structure in high-dimensional data. A prevalent low-dimensional structure is that of data points lying in a union of low-dimensional subspaces. The problem of finding the assignments of the data points to these (unknown) subspaces is referred to as subspace clustering [4] or hybrid linear modeling [5]. An example application of subspace clustering is the following. Given a set of images of faces under varying illumination conditions, cluster the images such that each of the resulting clusters corresponds to a single person [6]. Other application areas include unsupervised learning, image representation and segmentation [7], computer vision, specifically motion segmentation [8, 9], and disease detection [10]; we refer to [4] for a more complete list.

Often the data available is corrupted by noise and contains outliers. The general subspace clustering problem we consider takes this into account and can be formalized as follows. Suppose

Parts of this paper were presented at the 2013 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP) [1] and at the 2013 IEEE International Symposium on Information Theory (ISIT) [2].

we are given a set of N data points in $\mathbb{R}^m$, denoted by $\mathcal{X}$, and assume that

$$\mathcal{X} = \mathcal{X}_1 \cup ... \cup \mathcal{X}_L \cup \mathcal{O}.$$

Here, $\mathcal{O}$ denotes a set of outliers and the $n_\ell := |\mathcal{X}_\ell|$ points in $\mathcal{X}_\ell$ are given by

$$\mathbf{x}_j^{(\ell)} = \mathbf{y}_j^{(\ell)} + \mathbf{e}_j^{(\ell)}, \quad j = 1, ..., n_\ell \tag{1}$$

where $\mathbf{y}_j^{(\ell)} \in S_\ell$ with S_ℓ a d_ℓ-dimensional linear subspace of $\mathbb{R}^m$ and $\mathbf{e}_j^{(\ell)} \in \mathbb{R}^m$ is noise. The association of the points in $\mathcal{X}$ with the $\mathcal{X}_\ell$ and $\mathcal{O}$, the number of subspaces L, their dimensions d_ℓ, and their orientations are all unknown. We want to cluster the data points in $\mathcal{X}$, i.e., find their assignments to the sets $\mathcal{X}_1, ..., \mathcal{X}_L, \mathcal{O}$. Once these assignments have been identified, it is straightforward to extract approximations (recall that we have access to noisy data 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 [4] for an excellent survey. Spectral clustering methods (see [11] 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 $\mathbf{A} \in \mathbb{R}^{N \times N}$, where the (i,j)th entry of $\mathbf{A}$ measures the similarity between the data points $\mathbf{x}_i, \mathbf{x}_j \in \mathcal{X}$. A typical measure of similarity is, e.g., $e^{-\text{dist}(\mathbf{x}_i, \mathbf{x}_j)}$, where $\text{dist}(\cdot, \cdot)$ is some distance measure [4]. The association of the points in $\mathcal{X}$ to the subspaces S_ℓ is then estimated by applying spectral clustering to $\mathbf{A}$.

As noted in [12] there are only a few subspace clustering algorithms that are computationally tractable *and* known to succeed *provably* under non-restrictive conditions such as, e.g., intersecting subspaces. A notable exception is the sparse subspace clustering (SSC) algorithm proposed by Elhamifar and Vidal [13, 14], which applies spectral clustering to an adjacency matrix $\mathbf{A}$ obtained by sparsely representing each data point in terms of all the other data points through ℓ_1-minimization. SSC provably succeeds (in a sense to be made precise later) in the noiseless case under very general conditions, as shown by Soltanolkotabi and Candès in [15] via an elegant (geometric function) analysis. Most importantly, the results in [15] reveal that SSC succeeds even when the subspaces S_ℓ intersect (the linear subspaces S_ℓ and S_k are said to intersect if $S_\ell \cap S_k \neq \{\mathbf{0}\}$).

Analytical performance results for subspace clustering of noisy data are even more scarce. Vidal noted in [4] 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 recently in [12]. Specifically, the robust SSC (RSSC) algorithm in [12] replaces the ℓ_1-minimization step in SSC by an ℓ_1-penalized least squares, i.e., Lasso, step and provably succeeds under Gaussian noise and under very general conditions on the orientations of the subspaces S_ℓ. To construct the adjacency matrix $\mathbf{A}$, SSC for the noiseless and RSSC for the noisy case require the solution of N ℓ_1-minimization and N Lasso instances, respectively, each in N variables; this poses significant computational challenges for large data sets.

Contributions: We present a simple and computationally efficient subspace clustering algorithm, which applies spectral clustering to an adjacency matrix $\mathbf{A}$ obtained by thresholding correlations between the data points in $\mathcal{X}$. In other words, $\mathbf{A}$ is constructed from the nearest neighbors of each data point in spherical distance. The resulting algorithm is termed thresholding-based subspace clustering (TSC).

For our analytical results, we consider a semi-random data model with deterministic subspaces and the data points sampled at random from these subspaces. Specifically, we sample uniformly

at random from the intersection of the unit sphere and the corresponding subspace. The gist of the results we obtain is that TSC succeeds *provably*—even when the data is corrupted by additive Gaussian noise or incompletely observed—provided that the subspaces are sufficiently distinct and $\mathcal{X}$ contains sufficiently many points from each subspace. The measure of success we use in the noisy case and for incomplete observations is an intermediate performance measure as it does not address the clustering error, i.e., the fraction of misclassified points, directly. Rather, it guarantees that in the graph G with adjacency matrix $\mathbf{A}$, for each ℓ, the nodes corresponding to the points in $\mathcal{X}_\ell$ are connected to other nodes corresponding to points in $\mathcal{X}_\ell$ only. The same performance measure was used in [15, 12, 16, 17] for SSC, RSSC, LRR (low-rank representation), and SSC-orthogonal matching pursuit (SSC-OMP). In the noiseless case, we obtain significantly stronger results which come in terms of conditions guaranteeing that the clustering error is zero. This is accomplished by analyzing the connectivity properties of the random nearest neighbor graph induced by the statistical data model we employ. The corresponding results (Theorems 1 and 2) apply, however, to a smaller range of parameters d_ℓ, n_ℓ when compared to ensuring no false connections only (Corollary 1).

Our results for noisy data reflect the intuition that the more distinct the orientations of the subspaces, the more noise TSC tolerates. What is more, we find that TSC can succeed even under massive noise, provided that the subspaces are sufficiently low-dimensional. In practical applications the data points to be clustered are often incompletely observed, due to, e.g., scratches on images. Assuming that the orientation of the subspaces and the points on the subspaces are random, we prove that TSC can succeed even when the number of (arbitrary) missing entries in each data vector is (up to a log-factor) linear in the ambient dimension. Finally, we propose a simple scheme for outlier detection and we report corresponding analytical performance guarantees. Numerical results on synthetic data, on handwritten digits taken from the MNIST data base [18], and on images of faces taken from the extended Yale Face Database B [19, 20] complement our analytical results.

Relation to previous work: Lauer and Schnorr [21] 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 [21]. The local subspace affinity algorithm [22] and the spectral local best-fit flats (SLBF) algorithm [5] are based on spectral clustering applied to an adjacency matrix that is built from the nearest neighbors of each data point in Euclidean distance. Liu et al. [16] consider spectral clustering applied to an adjacency matrix obtained from a low-rank representation (LRR) of the data points through nuclear norm minimization. The performance analysis conducted in [16, specifically Theorem 3.1] shows that LRR succeeds provided that the subspaces S_ℓ are independent (the linear subspaces S_ℓ are said to be 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, the nuclear norm minimization required by LRR results in significant computational complexity. The analytical conditions guaranteeing success of SSC, and RSSC for the noisy case, reported in [15] and [12], respectively, are very similar to those found for TSC in this paper. TSC is, however, computationally much less demanding than SSC/RSSC. These complexity savings may come at the cost of clustering performance. Experiments on real and synthetic data, many of which are reported in Section 8, show that while there are situations where TSC outperforms SSC, SSC outperforming TSC is more common. Dyer et al. [17] propose to substitute the ℓ_1-minimization step in SSC by an orthogonal matching pursuit (OMP) step, and derive performance guarantees for the resulting SSC-OMP algorithm.

Lerman and Zhang [23] consider the problem of recovering multiple subspaces from data drawn

from a distribution on the union of these subspaces and pose recovery as a non-convex optimization problem. No computationally tractable algorithm to solve this recovery problem [24] seems to be available, though.

The problem of fitting a single low-dimensional subspace to a data set consisting of a modest number of noisy inliers and a large number of outliers was considered in [24], along with a convex programming algorithm with analytical performance guarantees. Chen and Lerman [25, 26] propose subspace clustering algorithms based on spectral clustering, termed Spectral Curvature Clustering (SCC) and Theoretical SCC (TSCC), along with a strategy for outlier detection, and provide corresponding probabilistic performance analyses.

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 the performance of TSC. Section 6 describes an outlier detection scheme and contains corresponding performance results. In Section 7, we compare our analytical performance results for TSC to analytical performance results for SSC/RSSC and further subspace clustering algorithms. Section 8 contains numerical results on synthetic and on real data, including a comparison of TSC to SSC/RSSC.

We discuss the various settings (noiseless, noisy, incomplete observations, and outliers) in an isolated fashion to keep the exposition accessible. All proofs are relegated to 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 qth entry. For the matrix $\mathbf{A}$, $\mathbf{A}_{ij}$ designates the entry in its ith row and jth 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, and $\|\mathbf{A}\|_F := (\sum_{i,j} |\mathbf{A}_{ij}|^2)^{1/2}$ its Frobenius norm. $\mathbf{I}_m$ denotes the $m \times m$ identity matrix. $\log(\cdot)$ is the natural logarithm, and $x \wedge y$ stands for the minimum of x and y. $\mathcal{L}(\cdot)$ denotes the Lebesgue measure. For the set $\mathcal{T}$, $|\mathcal{T}|$ designates its cardinality and $\overline{\mathcal{T}}$ is its complement. The set $\{1, ..., N\}$ is denoted by $[N]$. We write $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ for a Gaussian random vector with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. The unit sphere in $\mathbb{R}^m$ is $\mathbb{S}^{m-1} := \{\mathbf{x} \in \mathbb{R}^m \colon \|\mathbf{x}\|_2 = 1\}$. $1_{\{A\}}(\cdot)$ denotes the indicator function of the set A. For notational convenience, we use the shorthand $\max_{k\neq\ell}$ for $\max_{k\in[L]\colon k\neq\ell}$ and $\max_{k,\ell\colon k\neq\ell}$ for $\max_{k,\ell\in[L]\colon k\neq\ell}$. Similarly, $\max_{k\neq\ell,j}$ is shorthand for $\max_{k\in[L]\colon k\neq\ell, j\in[n_k]}$. We let $n_{\min} = \min_{\ell\in[L]} n_\ell$, $n_{\max} = \max_{\ell\in[L]} n_\ell$, and $d_{\max} = \max_{\ell\in[L]} d_\ell$. For random variables X, Y, we write $X \sim Y$ to indicate that X and Y have the same distribution. We say that a subgraph H of a graph G is connected if any two nodes in H can be joined by a path that has all intermediate nodes lying in H. The subgraph H of G is called a connected component of G if H is connected and if there are no connections between nodes in H and the remaining nodes in G [11]. The k-nearest neighbor graph of a set of points $\{\mathbf{a}_1, ..., \mathbf{a}_n\}$ with respect to the metric s is the undirected graph with vertex set $\{\mathbf{a}_1, ..., \mathbf{a}_n\}$ and edges between $\mathbf{a}_i$ and $\mathbf{a}_j$ if either $\mathbf{a}_i$ is among the k nearest neighbors of $\mathbf{a}_j$ or $\mathbf{a}_j$ is among the k nearest neighbors of $\mathbf{a}_i$, in both cases with respect to the metric s.

2 The TSC algorithm

The formulation of the thresholding-based subspace clustering (TSC) algorithm provided below assumes that outliers have already been removed from the data set $\mathcal{X}$, e.g., through the outlier detection scheme described in Section 6, and that the data points in $\mathcal{X}$ are normalized. The latter

assumption is relevant for Steps 1 and 2 below and is not restrictive as the data points can be normalized prior to clustering.

TSC algorithm. *Given a set of data points $\mathcal{X}$, an estimate of the number of subspaces $\hat{L}$ (estimation of L from $\mathcal{X}$ is discussed in Section 2.3), 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 $\mathcal{T}_j \subset [N] \backslash \{j\}$ (recall that $N = |\mathcal{X}|$) of cardinality q defined through*

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

Step 2: *Let $\mathbf{z}_j \in \mathbb{R}^N$ be the vector with ith entry $\exp(-2 \arccos(|\langle \mathbf{x}_j, \mathbf{x}_i \rangle|))$ 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 $\mathbf{Z} = [\mathbf{z}_1 \cdots \mathbf{z}_N]$.*

Step 4: *Apply normalized spectral clustering [11, 27] to $(\mathbf{A}, \hat{L})$.*

Since $\arccos(z)$ is decreasing in z for $z \in [0, 1]$, the set $\mathcal{T}_j$ is the set of q nearest neighbors of $\mathbf{x}_j$ with respect to the metric[1] $\tilde{s}(\mathbf{x}_i, \mathbf{x}_j) \coloneqq \arccos(|\langle \mathbf{x}_i, \mathbf{x}_j \rangle|)$. TSC is therefore built on the premise, explained in Sections 2.1 and 2.2, that the vectors close to $\mathbf{x}_j$ in terms of the distance $\tilde{s}$ also lie in the subspace $\mathbf{x}_j$ lies in. This can be formalized in terms of the q-nearest neighbor graph with respect to the distance $\tilde{s}$, i.e., the graph G with adjacency matrix $\mathbf{A}$, simply referred to as "the graph G" in the remainder of the paper. If each connected component in the graph G corresponds to exactly one of the sets $\mathcal{X}_\ell$, and if $\hat{L} = L$, then (normalized) spectral clustering yields correct segmentation of the data (i.e., it delivers the oracle segmentation $\mathcal{X} = \mathcal{X}_1 \cup ... \cup \mathcal{X}_L$ of $\mathcal{X}$) [11, Prop. 4; Sec. 7] and the clustering error will be zero. Even when the connected components of G do not correspond to the $\mathcal{X}_\ell$ exactly, but the weights in the adjacency matrix $\mathbf{A}$ corresponding to pairs of points that belong to different subspaces are small enough, TSC may still cluster the data correctly. The numerical results in Section 8 demonstrate that the spectral clustering step can cope with such imperfections.

In the noiseless case we will be able to establish conditions that ensure zero clustering error. In the noisy case we will work with an intermediate, albeit sensible, performance measure, also employed to assess the performance of the clustering algorithms considered in [17, 16, 15, 12]. This performance measure is formalized through the following property:

No false connections property. *G has no false connections if, for all $\ell \in [L]$, the nodes in G corresponding to $\mathcal{X}_\ell$ are connected to other nodes corresponding to $\mathcal{X}_\ell$ only.*

Ensuring the absence of false connections, does, however, not guarantee that the connected components in G correspond to the $\mathcal{X}_\ell$, as the points in a given set $\mathcal{X}_\ell$ may split up into two or more distinct clusters. TSC (with input parameter q) counters this problem by imposing that each node is connected to at least q other nodes and choosing q not too small relative to the n_ℓ. Taking q too large, however, increases the chances of points from different sets $\mathcal{X}_\ell$ being clustered together, thereby violating the no false connections property. Our analytical performance results for the noiseless case ensure correct segmentation of $\mathcal{X}$ by guaranteeing that G has no false connections and the subgraphs corresponding to the $\mathcal{X}_\ell$ are connected, provided that q is sufficiently large relative to the values $\log n_\ell$ and sufficiently small relative to the n_ℓ. The specific choice of q within this range will be seen to be irrelevant in terms of the *analytical* performance guarantees we obtain, but it does have an impact on the actual performance of TSC in practice.

[1] $\tilde{s}$ is not a distance metric in the strict sense as $\tilde{s}(\mathbf{x}, -\mathbf{x}) = 0$, but $-\mathbf{x} \neq \mathbf{x}$ for $\mathbf{x} \neq \mathbf{0}$. It satisfies, however, the defining properties of a pseudo-distance metric [28].

2.1 Measuring similarity via $\tilde{s}$

To see that $\tilde{s}(\mathbf{x}_i, \mathbf{x}_j) = \arccos(|\langle \mathbf{x}_i, \mathbf{x}_j \rangle|)$ leads to a sensible similarity measure for subspace clustering, consider the noiseless case, suppose that the subspaces S_ℓ are orthogonal to each other, and take $q \leq \min_\ell(|\mathcal{X}_\ell| - d_\ell)$. Then, G has no false connections thanks to $\langle \mathbf{x}_p, \mathbf{x}_j \rangle = 0$ for all $\mathbf{x}_p \in \mathcal{X}_\ell, \mathbf{x}_j \in \mathcal{X}_k$, $\ell \neq k$, while for each ℓ, there are at least $|\mathcal{X}_\ell| - d_\ell$ inner products $\langle \mathbf{x}_p, \mathbf{x}_j \rangle$ with $\mathbf{x}_p, \mathbf{x}_j \in \mathcal{X}_\ell$ that are non-zero, as no more than d_ℓ points in a d_ℓ-dimensional subspace can be orthogonal to each other. The analytical results in the following sections show that G can actually satisfy the no false connections property under much more general conditions, in particular even when the subspaces intersect. What lies beneath these results is the fact that for the statistical data model used throughout the paper, the magnitude of the inner product between the data points from the same subspace is typically larger than that between data points from different subspaces. This is also true in many practical problems, as e.g., the numerical results on clustering handwritten digits in Section 8.2 show. A different rationale for $\tilde{s}$ leading to a suitable similarity measure for subspace clustering is based on sparse signal representation theory, and is given next.

2.2 Least-squares TSC

A natural substitute for Step 2 in the TSC algorithm is to construct $\mathbf{z}_j$ from the best linear approximation of $\mathbf{x}_j$ in terms of the points indexed by $\mathcal{T}_j$. Specifically, let $\mathbf{X}_{\mathcal{T}_j}$ be the matrix whose columns are the vectors in $\mathcal{X}$ indexed by $\mathcal{T}_j$, and substitute Step 2 by:

Step 2-LS: Set the entries of $\mathbf{z}_j \in \mathbb{R}^N$ indexed by $\mathcal{T}_j$ to the absolute values of $\mathbf{X}_{\mathcal{T}_j}^\dagger \mathbf{x}_j$ and all other entries of $\mathbf{z}_j$ to zero.

The TSC algorithm with Step 2 replaced by Step 2-LS will henceforth be referred to as least squares (LS)-TSC.

The LS-variant of the TSC algorithm allows us to elicit a relationship between TSC, SSC, and SSC-OMP, with the common element being given by the insight that all three algorithms build their adjacency matrix based on sparse signal representation theory. To see this, we first note that each data point in a d_ℓ-dimensional subspace S_ℓ can be represented as a linear combination of at most d_ℓ other data points in S_ℓ. A possible approach to measuring similarity, put forward in [13, 14], finds a sparse representation of each data point $\mathbf{x}_j \in \mathcal{X}_\ell$ in terms of all other data points $\mathcal{X}_1 \cup \ldots \cup \mathcal{X}_L \backslash \{\mathbf{x}_j\}$, and uses the absolute values of the corresponding representation coefficients to quantify the similarity between $\mathbf{x}_j$ and all other data points. The hope is that the non-zeros in this representation correspond to points in $\mathcal{X}_\ell \setminus \{\mathbf{x}_j\}$, see Figure 1 for an illustration. SSC, SSC-OMP, and LS-TSC implement this idea by finding a sparse linear representation of $\mathbf{x}_j$ in terms of points in $\mathcal{X} \setminus \{\mathbf{x}_j\}$ via ℓ_1-minimization, OMP, and Steps 1 and 2-LS above, respectively. Note that Steps 1 and 2-LS above yield a sparse (if q is small) linear representation of $\mathbf{x}_j$ in terms of q points in $\mathcal{X} \setminus \{\mathbf{x}_j\}$.

The formal relationship between Steps 2 and 2-LS is brought out by noting that the non-zero entries of $\mathbf{z}_j$ in Step 2 are given by element-wise application of $\exp(-2\arccos(|\cdot|))$ to the vector $\mathbf{X}_{\mathcal{T}_j}^T \mathbf{x}_j$ whereas the nonzero entries of $\mathbf{z}_j$ in Step 2-LS are obtained by element-wise application of $|\cdot|$ to the entries of a weighted version of $\mathbf{X}_{\mathcal{T}_j}^T \mathbf{x}_j$, namely $\mathbf{X}_{\mathcal{T}_j}^\dagger \mathbf{x}_j = (\mathbf{X}_{\mathcal{T}_j}^T \mathbf{X}_{\mathcal{T}_j})^{-1} \mathbf{X}_{\mathcal{T}_j}^T \mathbf{x}_j$.

As our analytical performance results depend on connectivity properties of the graph G only, and not on the weights assigned to the edges of G (i.e., the values of the non-zero entries of $\mathbf{A}$), it follows immediately that the corresponding statements hold true verbatim for LS-TSC. Owing to the spectral clustering Step 4, the values of the non-zero entries of $\mathbf{A}$ do, however, make a difference in terms of practical performance. Corresponding numerical results will be provided in Section 8.

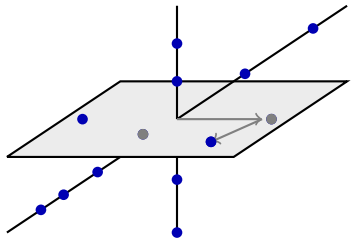

Figure 1: Each data point in a d_ℓ-dimensional subspace S_ℓ can be represented as a linear combination of at most d_ℓ other points from S_ℓ.

2.3 Estimation of the number of subspaces

The number of zero eigenvalues of the normalized Laplacian of the graph G is equal to the number of connected components of G [29]. It is therefore sensible to estimate the number of subspaces L as the multiplicity of the eigenvalue 0 of the normalized Laplacian of G. In practice, however, weights in the adjacency matrix $\mathbf{A}$ corresponding to pairs $\mathbf{x}_i, \mathbf{x}_j$ that belong to different subspaces might be non-zero, but possibly small, in which case the number of connected components in G may be smaller than L. This will result in eigenvalues that are not exactly equal to zero, but possibly small. A robust estimator for L taking this into account is the so-called eigengap heuristic [11]: $\hat{L} = \arg\max_{i \in [N-1]} (\lambda_{i+1} - \lambda_i)$, where $\lambda_1 \leq \lambda_2 \leq ... \leq \lambda_N$ are the eigenvalues of the normalized Laplacian of G.

We note that while satisfying the no false connections property does not say anything about the quality of the estimate $\hat{L}$, establishing that the connected components in G correspond to the $\mathcal{X}_\ell$, as done below in the noiseless case, automatically guarantees that $\hat{L} = L$.

3 Performance results for the noiseless case

We first consider noiseless data sets (i.e., $\mathbf{x}_j^{(\ell)} = \mathbf{y}_j^{(\ell)}$ in (1)) that have no outliers. 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. Specifically, we represent the data points in S_ℓ by $\mathbf{x}_j^{(\ell)} = \mathbf{U}^{(\ell)} \mathbf{a}_j^{(\ell)}$ where $\mathbf{U}^{(\ell)} \in \mathbb{R}^{m \times d_\ell}$ is an orthonormal basis for the d_ℓ-dimensional subspace S_ℓ and the $\mathbf{a}_j^{(\ell)} \in \mathbb{R}^{d_\ell}$ are i.i.d. uniformly distributed on $\mathbb{S}^{d_\ell - 1}$ (throughout the paper, whenever we say that the $\mathbf{a}_j^{(\ell)}$ or the $\mathbf{e}_j^{(\ell)}$ are i.i.d., we actually mean i.i.d. across j *and* ℓ). Therefore, the data points $\mathbf{x}_j^{(\ell)} = \mathbf{U}^{(\ell)} \mathbf{a}_j^{(\ell)}$ are distributed uniformly on $\{\mathbf{x} \in S_\ell \colon \|\mathbf{x}\|_2 = 1\}$, which ensures that the points are spread out on the subspaces, and avoids degenerate situations where data points lie in preferred directions. For example, suppose that the points on say, a two-dimensional subspace S_1, are skewed towards two (distinct) directions. Then, there are two sensible segmentations. One is to assign the points corresponding to each direction to separate clusters, the other to assign all points to one cluster.

Our results will be expressed in terms of two different notions of affinity between subspaces, namely

$$\mathrm{aff}_\infty(S_k, S_\ell) := \left\| \mathbf{U}^{(k)^T} \mathbf{U}^{(\ell)} \right\|_{2 \to 2}$$

and

$$\text{aff}(S_k, S_\ell) := \frac{1}{\sqrt{d_k \wedge d_\ell}} \left\| \mathbf{U}^{(k)^T} \mathbf{U}^{(\ell)} \right\|_F .$$

The relation between the affinity notions $\text{aff}_\infty(\cdot)$ and $\text{aff}(\cdot)$ is brought out by expressing them in terms of the principal angles between S_k and S_ℓ according to

$$\text{aff}_\infty(S_k, S_\ell) = \cos(\theta_1) \tag{2}$$

and

$$\text{aff}(S_k, S_\ell) = \frac{\sqrt{\cos^2(\theta_1) + ... + \cos^2(\theta_{d_k \wedge d_\ell})}}{\sqrt{d_k \wedge d_\ell}} \tag{3}$$

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

Definition 1. *The principal angles $\theta_1, ..., \theta_{d_k \wedge d_\ell}$ between the subspaces S_k and S_ℓ are defined recursively according to*

$$\cos(\theta_j) = \langle \mathbf{v}_j, \mathbf{u}_j \rangle , \text{ where } (\mathbf{v}_j, \mathbf{u}_j) = \arg\max \langle \mathbf{v}, \mathbf{u} \rangle$$

with the maximization carried out over all $\mathbf{v} \in S_k \colon \|\mathbf{v}\|_2 = 1$, $\mathbf{u} \in S_\ell \colon \|\mathbf{u}\|_2 = 1$, subject to $\langle \mathbf{v}, \mathbf{v}_i \rangle = 0$ and $\langle \mathbf{u}, \mathbf{u}_i \rangle = 0$ for all $i = 1, ..., j-1$ (for $j = 1$, this constraint is void).

Note that $0 \leq \text{aff}(S_k, S_\ell) \leq \text{aff}_\infty(S_k, S_\ell) \leq 1$. If S_k and S_ℓ intersect in p dimensions, i.e., if $S_k \cap S_\ell$ is p-dimensional, then $\cos(\theta_1) = ... = \cos(\theta_p) = 1$ [30]. Hence, if S_k and S_ℓ intersect in $p \geq 1$ dimensions, we have $\text{aff}_\infty(S_k, S_\ell) = 1$ and $\text{aff}(S_k, S_\ell) \geq \sqrt{p/(d_k \wedge d_\ell)}$. We finally note that the affinity notion [15, Definition 2.6] and [12, Definition 1.2], relevant to the analysis of SSC and RSSC, is equal to $\text{aff}(\cdot, \cdot)$.

We are now ready to state our first main result.

Theorem 1. *Suppose that $\mathcal{X}_\ell, \ell \in [L]$, is obtained by choosing n_ℓ points i.i.d. uniformly from $\{\mathbf{x} \in S_\ell \colon \|\mathbf{x}\|_2 = 1\}$, independently across ℓ, and let $\mathcal{X} = \mathcal{X}_1 \cup ... \cup \mathcal{X}_L$. Pick $\rho \in [0, 1)$ and suppose that $n_\ell \geq n_0$, for all $\ell \in [L]$, where n_0 is a constant that depends on $d_{\max}$ and ρ only. Pick $\gamma > 1$ and suppose that $q \in [c_2 \gamma \log n_{\max}, n_{\min}^\rho]$ with $c_2 = 6(12\pi)^{d_{\max}-1}$. If*

$$\max_{k,\ell \colon k \neq \ell} \text{aff}_\infty(S_k, S_\ell) < 1 \tag{4}$$

then TSC delivers the correct segmentation of $\mathcal{X}$ with probability at least $1 - \sum_{\ell=1}^L \left(n_\ell e^{-c_1(n_\ell - 1)} + 2{n_\ell}^{-\gamma+1} \right)$, where c_1 is a numerical constant.

Theorem 1 states that TSC delivers the correct segmentation of $\mathcal{X}$ with high probability if the subspaces do not intersect (recall that $\text{aff}_\infty(S_k, S_\ell) = 1$ if and only if S_k and S_ℓ intersect in at least one dimension) and if $\mathcal{X}$ contains sufficiently many points from each subspace ($n_\ell \geq n_0$, for all $\ell \in [L]$). Intuitively we expect that clustering becomes easier when the n_ℓ increase. To see that Theorem 1 confirms this intuition, set $n_\ell = n$, for all $\ell \in [L]$, and note that the probability of correct segmentation in Theorem 1 increases in n.

Theorem 1 furthermore shows that TSC delivers the correct segmentation of $\mathcal{X}$ asymptotically in the number of points in $\mathcal{X}$ from each subspace, n_ℓ, even when the n_ℓ scale differently (in a sense made precise below), and/or the number of subspaces, L, grows faster than one or more of the

n_ℓ. To see this, fix the d_ℓ, and let $n_\ell = n^{\kappa_\ell}$, $L = n^\kappa$ for numerical constants κ_ℓ and κ (possibly $\kappa > \kappa_\ell$, in which case L grows faster than n_ℓ), and let $n \to \infty$. Choose γ such that $(\gamma-1)\kappa_{\min} > \kappa$ where $\kappa_{\min} \coloneqq \min_\ell \kappa_\ell$. With $\kappa_{\max} = \max_\ell \kappa_\ell$, for $q \in [c_2\gamma\kappa_{\max}\log n,\, n^{\kappa_{\min}\rho}]$ with c_2 and γ from Theorem 1 (the interval is guaranteed to be nonempty for n sufficiently large as c_2 does not depend on n) it then follows that TSC yields correct segmentation with probability at least

$$1 - \sum_{\ell=1}^{L}\left(n^{\kappa_\ell}e^{-c_1(n^{\kappa_\ell}-1)} + 2n^{-(\gamma-1)\kappa_\ell}\right) \geq 1 - \left(n^{\kappa_{\min}+\kappa}e^{-c_1(n^{\kappa_{\min}}-1)} + 2n^{-(\gamma-1)\kappa_{\min}+\kappa}\right)$$

which tends to 1 as $n \to \infty$.

The proof of Theorem 1 is based on the realization that the graph G is a random graph owing to the random data model. Specifically, the proof is effected by showing that the connected components in G correspond to the $\mathcal{X}_\ell$ with probability satisfying the probability estimate in Theorem 1. As for the choice of q in Theorem 1, the upper bound on q is used to establish that G has no false connections, i.e., each $\mathbf{x}_j \in \mathcal{X}_\ell$ is connected to points in $\mathcal{X}_\ell$ only, for all ℓ. An upper bound on q is also necessary as obviously $q > n_{\min}$ results in G necessarily having false connections. The lower bound on q is needed to ensure that, in addition, the subgraphs $G(\mathcal{X}_\ell)$ corresponding to the $\mathcal{X}_\ell$ are connected, and hence the $G(\mathcal{X}_\ell)$ form connected components. In fact, the lower bound on q (as a function of $n_{\max}$) is order-wise necessary for the $G(\mathcal{X}_\ell)$ to be connected. Specifically, there exists a constant c that does not depend on n_ℓ, such that for $q = c\log n_\ell$, $G(\mathcal{X}_\ell)$ is not connected with probability 1 as $n_\ell \to \infty$ (not shown here). The exponential dependency of the constant $c_2 = 6(12\pi)^{d_{\max}-1}$ on $d_{\max}$ requires that the n_ℓ be exponential in the d_ℓ as this is necessary for the interval $[c_2\gamma\log n_{\max}, n_{\min}^\rho]$ of admissible values for q to be non-empty. While this restricts the range of parameters d_ℓ, n_ℓ, Theorem 1 applies to, the statement in Theorem 1 is strongest possible as it guarantees that the clustering error is zero as opposed to ensuring no false connections only. Zero clustering error ensures that *every* point in the data set is clustered correctly. We finally note that the exponential dependency of c_2 on $d_{\max}$ appears to be an artifact of our proof technique, as indicated by numerical results (not shown here). In fact, these numerical results suggest that c_2 may even be a decreasing function of $d_{\max}$.

Theorem 1 does not apply to subspaces that intersect as $\operatorname{aff}_\infty(S_k, S_\ell) = 1$ in this case. We can, however, find a statement analogous to Theorem 1, but in terms of $\operatorname{aff}(S_k, S_\ell)$, which applies to intersecting subspaces.

Theorem 2. *Suppose that $\mathcal{X}_\ell, \ell \in [L]$, is obtained by choosing n_ℓ points i.i.d. uniformly from $\{\mathbf{x} \in S_\ell \colon \|\mathbf{x}\|_2 = 1\}$, independently across ℓ, and let $\mathcal{X} = \mathcal{X}_1 \cup \ldots \cup \mathcal{X}_L$. Suppose furthermore that $q \in [c_1 \log n_{\max}, n_{\min}/6]$ with $c_1 = 18(12\pi)^{d_{\max}-1}$. If*

$$\max_{k,\ell\colon\, k\neq\ell} \operatorname{aff}(S_k, S_\ell) \leq \frac{1}{15\log N} \tag{5}$$

then TSC delivers the correct segmentation of $\mathcal{X}$ with probability at least $1 - 10/N - \sum_{\ell=1}^{L}(n_\ell e^{-c(n_\ell-1)} + 2n_\ell^{-2})$, where $c > 0$ is a numerical constant.

The interpretation of Theorem 2 is analogous to that of Theorem 1 with the important difference that the right hand side (RHS) of (5), as opposed to the RHS of (4), decreases, albeit very slowly, in the n_ℓ as $N = \sum_\ell n_\ell$. At first sight this is counterintuitive as we expect that clustering becomes easier when the number of points in each subspace increases. However, our statement guarantees

that *every* point in the data set is clustered correctly, even though the subspaces are allowed to intersect (cf. (5)). As the total number of points, N, increases, we would expect that the probability that at least *one* point is close to an intersection of two subspaces, and therefore misclustered, increases. Ensuring that the success probability increases in the n_ℓ, therefore leads to the affinity condition (5) becoming stricter as N increases.

Again, the exponential dependency of the constant $c_1 = 18(12\pi)^{d_{\max}-1}$ on $d_{\max}$ requires that the n_ℓ be exponential in the d_ℓ. If one is content with satisfying the (weaker) no false connections property only, this dependency on $d_{\max}$ vanishes by virtue of a lower bound on q not being needed. Specifically, this leads to the following result.

Corollary 1. *Suppose that $\mathcal{X}_\ell, \ell \in [L]$, is obtained by choosing n_ℓ points i.i.d. uniformly from $\{\mathbf{x} \in S_\ell \colon \|\mathbf{x}\|_2 = 1\}$, independently across ℓ, and let $\mathcal{X} = \mathcal{X}_1 \cup \ldots \cup \mathcal{X}_L$. Suppose furthermore that $q \leq n_{\min}/6$. If*

$$\max_{k,\ell \colon k \neq \ell} \operatorname{aff}(S_k, S_\ell) \leq \frac{1}{15 \log N}$$

then G has no false connections with probability at least $1 - \frac{10}{N} - \sum_{\ell \in [L]} n_\ell e^{-c(n_\ell - 1)}$, where $c > 0$ is a numerical constant.

Note that Corollary 1 does not require any relation between the n_ℓ and the d_ℓ, in particular the n_ℓ can be linear in the d_ℓ. At first sight this might seem surprising as nearest neighbor algorithms often suffer from the *curse of dimensionality* [31], manifested by the neighborhood of a point in a high-dimensional space no longer being local [31], e.g., the vast majority of points chosen i.i.d. on a high-dimensional unit sphere are essentially orthogonal to each other. Although TSC is a nearest neighbor algorithm, it relies only on the premise that the vectors close to a given data point $\mathbf{x}_j$ also lie in the subspace $\mathbf{x}_j$ lies in. This premise only requires the affinities between the subspaces S_ℓ to be sufficiently small, does not rely on a certain relation between the n_ℓ and the d_ℓ, and does not break down when the subspaces are high-dimensional. To see all this, we next provide a back-of-the-envelope argument establishing the no false connections property under a (dimension-independent) condition on the affinity of the subspaces. The proofs of the no false connections property in Theorems 1–4 and Corollary 1 are essentially formal versions of the argument below.

For ease of exposition, we set $d_\ell = d$, $n_\ell = |\mathcal{X}_\ell| = n$, for all ℓ, and we take the $\mathbf{a}_i^{(\ell)}$ in $\mathbf{x}_i^{(\ell)} = \mathbf{U}^{(\ell)} \mathbf{a}_i^{(\ell)}$ to be i.i.d. $\mathcal{N}(\mathbf{0}, (1/d)\mathbf{I}_d)$ (recall that $\mathbf{U}^{(\ell)}$ is an orthonormal basis for S_ℓ). As the corresponding direction vectors $\mathbf{x}_i^{(\ell)} / \left\|\mathbf{x}_i^{(\ell)}\right\|_2$ are distributed uniformly on $\{\mathbf{x} \in S_\ell \colon \|\mathbf{x}\|_2 = 1\}$ and $\left\|\mathbf{x}_i^{(\ell)}\right\|_2^2 = \left\|\mathbf{a}_i^{(\ell)}\right\|_2^2$ concentrates around its expectation $\mathbb{E}\left[\left\|\mathbf{a}_i^{(\ell)}\right\|_2^2\right] = 1$, this model is conceptually equivalent to the $\mathbf{a}_i^{(\ell)}$ being i.i.d. on the unit sphere, as assumed in the formal statements throughout the paper. The program of the back-of-the-envelope calculation below is as follows. We use the fact that the absolute value of the inner product between data points from within a given subspace concentrates around $c/\sqrt{d}$, whereas the absolute value of the inner product between data points from different subspaces, S_k and S_ℓ, concentrates around a value $\leq \operatorname{aff}_\infty(S_k, S_\ell) c/\sqrt{d}$. Thus, even when the subspace dimension d is large, the maximum inner product between data points from a given subspace will still be larger than the largest inner product between data points from different subspaces, provided that the affinity is sufficiently small. More formally, the no false connections property holds if for $\mathbf{x}_i \in \mathcal{X}_\ell$, the corresponding set $\mathcal{T}_i$ from Step 2 of the TSC algorithm corresponds to points in $\mathcal{X}_\ell$ only, for all $\mathbf{x}_i$, and for all ℓ. The set $\mathcal{T}_i$ contains indices corresponding to points in $\mathcal{X}_\ell$ only if the qth largest value in the set $\left\{\left|\left\langle \mathbf{x}_j^{(\ell)}, \mathbf{x}_i^{(\ell)} \right\rangle\right|, j \neq i\right\}$ exceeds the largest value in the set $\left\{\left|\left\langle \mathbf{x}_j^{(k)}, \mathbf{x}_i^{(\ell)} \right\rangle\right|, j, k \neq \ell\right\}$. Conditioned on $\mathbf{a}_i^{(\ell)}$, the random variable $\left\langle \mathbf{x}_j^{(\ell)}, \mathbf{x}_i^{(\ell)} \right\rangle = \left\langle \mathbf{a}_j^{(\ell)}, \mathbf{a}_i^{(\ell)} \right\rangle$ is

zero-mean Gaussian with variance $\left\|\mathbf{a}_i^{(\ell)}\right\|_2^2/d$. A standard result from order statistics [32] shows that, with high probability, the qth largest value in the set $\left\{\left|\left\langle \mathbf{x}_j^{(\ell)}, \mathbf{x}_i^{(\ell)}\right\rangle\right|, j \neq i\right\}$ is no larger than

$$c_1 \sqrt{\log(n/q)} \frac{\left\|\mathbf{a}_i^{(\ell)}\right\|_2}{\sqrt{d}}. \tag{6}$$

Next, consider data points $\mathbf{x}_j^{(k)}, \mathbf{x}_i^{(\ell)}$ from different subspaces (i.e., $k \neq \ell$), and note that

$$\begin{aligned}
\left|\left\langle \mathbf{x}_j^{(k)}, \mathbf{x}_i^{(\ell)}\right\rangle\right| &= \left|\left\langle \mathbf{U}^{(k)}\mathbf{a}_j^{(k)}, \mathbf{U}^{(\ell)}\mathbf{a}_i^{(\ell)}\right\rangle\right| \\
&\leq \left|\left\langle \mathbf{a}_j^{(k)}, \mathbf{a}_i^{(\ell)}\right\rangle\right| \left\|\mathbf{U}^{(k)^T}\mathbf{U}^{(\ell)}\right\|_2 \\
&= \left|\left\langle \mathbf{a}_j^{(k)}, \mathbf{a}_i^{(\ell)}\right\rangle\right| \mathrm{aff}_\infty(S_k, S_\ell).
\end{aligned}$$

As before, $\left\langle \mathbf{a}_j^{(k)}, \mathbf{a}_i^{(\ell)}\right\rangle$ is Gaussian with variance $\left\|\mathbf{a}_i^{(\ell)}\right\|_2^2/d$. Again, it follows that the largest value in the set $\left\{\left|\left\langle \mathbf{x}_j^{(k)}, \mathbf{x}_i^{(\ell)}\right\rangle\right|, j,k \neq \ell\right\}$ is smaller than

$$c_2 \sqrt{\log((L-1)n)} \frac{\left\|\mathbf{a}_i^{(\ell)}\right\|_2}{\sqrt{d}} \mathrm{aff}_\infty(S_k, S_\ell) \tag{7}$$

with high probability. We can hence expect TSC to succeed (with high probability) if (7) is smaller than (6) which leads to

$$\mathrm{aff}_\infty(S_k, S_\ell) \leq \frac{c_1\sqrt{\log(n/q)}}{c_2\sqrt{\log((L-1)n)}}.$$

The RHS of this condition does not depend on the dimension of the subspaces, d, which explains why TSC does not suffer from the curse of dimensionality. Note that while $\mathrm{aff}_\infty(S_k, S_\ell)$ does depend on d through the subspaces S_k and S_ℓ, it can easily be small for large d (e.g., for orthogonal subspaces S_k, S_ℓ of dimension d in $\mathbb{R}^m, m = 2d$, $\mathrm{aff}_\infty(S_k, S_\ell) = 0$, or consider Lemma 4 in Appendix D for a more interesting example, which shows that for L subspaces with random orientations $\mathrm{aff}_\infty(S_k, S_\ell)$, for all pairs $S_k, S_\ell, k \neq \ell$, is close to zero with high probability provided that $m \geq O(d + \log L)$).

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. It is therefore of interest to analyze the performance of TSC applied to noisy data.

Theorem 3. *Suppose that $\mathcal{X}_\ell, \ell \in [L]$, is obtained by choosing n_ℓ points corresponding to S_ℓ at random according to $\mathbf{x}_j^{(\ell)} = \mathbf{y}_j^{(\ell)} + \mathbf{e}_j^{(\ell)}, j \in [n_\ell]$, where the $\mathbf{y}_j^{(\ell)}$ are chosen i.i.d. uniformly from $\{\mathbf{y} \in S_\ell \colon \|\mathbf{y}\|_2 = 1\}$, independently across ℓ, and the $\mathbf{e}_j^{(\ell)}$ are i.i.d. $\mathcal{N}(\mathbf{0}, (\sigma^2/m)\mathbf{I}_m)$, independent of the $\mathbf{y}_j^{(\ell)}$. Let $\mathcal{X} = \mathcal{X}_1 \cup \ldots \cup \mathcal{X}_L$ and suppose that $q \leq n_{\min}/6$. If*

$$\max_{k,\ell\colon k\neq\ell} \mathrm{aff}(S_k, S_\ell) + \frac{\sigma(1+\sigma)}{\sqrt{\log N}} \frac{\sqrt{d_{\max}}}{\sqrt{m}} \leq \frac{1}{15 \log N} \tag{8}$$

with $m \geq 6 \log N$, then G has no false connections with probability at least $1 - \frac{10}{N} - \sum_{\ell\in[L]} n_\ell e^{-c(n_\ell - 1)}$, where $c > 0$ is a numerical constant.

First, note that, unlike in the noiseless case, the data points $\mathbf{x}_j^{(\ell)}$ in Theorem 3 do not have unit norm. However, since $\mathbf{e}_j^{(\ell)}$ concentrates around its mean, the norms $\|\mathbf{x}_j^{(\ell)}\|_2$ are close to each other with high probability. TSC also applies to points that are unnormalized, with the only difference that $\exp(-2\arccos(|\langle \mathbf{x}_j, \mathbf{x}_i\rangle|))$ in Step 2 has to be replaced by $\exp(-2\arccos(|\langle \mathbf{x}_j, \mathbf{x}_i\rangle| /(\|\mathbf{x}_j\|_2\|\mathbf{x}_i\|_2)))$. Second, note that Theorem 3, unlike the results in the noiseless case in Theorems 1 and 2 only ensures the absence of false connections in G and hence does not guarantee zero clustering error. Theorem 3 states that TSC succeeds (in the sense of G having no false connections) with high probability if $\mathcal{X}$ contains sufficiently many points from each subspace (see the probability estimate in Theorem 3) and if the additive noise variance and the affinities between the subspaces are sufficiently small.

Condition (8) nicely reflects the intuition that the more distinct the orientations of the subspaces the more noise TSC tolerates. What is more, Condition (8) reveals that TSC can succeed even under massive noise, i.e., even if $\sigma^2 = \mathbb{E}\left[\left\|\mathbf{e}_j^{(\ell)}\right\|_2^2\right] > \left\|\mathbf{y}_j^{(\ell)}\right\|_2^2 = 1$, provided that the dimensions of the subspaces are sufficiently small relative to the ambient dimension.

The intuition behind the factor $\sigma(1+\sigma)\sqrt{d_{\max}/m}$ in (8), made rigorous in the proof of Theorem 3, is as follows. Assume, for simplicity, that $d_\ell = d$, for all ℓ, and consider the most favorable situation of subspaces that are orthogonal to each other, i.e., $\operatorname{aff}(S_k, S_\ell) = 0$, for all pairs (k, ℓ) with $k \neq \ell$. Recall that 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 $\langle \mathbf{x}_j, \mathbf{x}_i\rangle = \langle \mathbf{y}_j, \mathbf{y}_i\rangle + \langle \mathbf{e}_j, \mathbf{e}_i\rangle + \langle \mathbf{y}_j, \mathbf{e}_i\rangle + \langle \mathbf{e}_j, \mathbf{y}_i\rangle$. Then, under the statistical data model of Theorem 3, we have $\left(\mathbb{E}\left[|\langle \mathbf{y}_j, \mathbf{y}_i\rangle|^2\right]\right)^{1/2} = \frac{1}{\sqrt{d}}$ if $\mathbf{y}_j, \mathbf{y}_i \in S_\ell$ and $\langle \mathbf{y}_j, \mathbf{y}_i\rangle = 0$ if $\mathbf{y}_j \in S_k$ and $\mathbf{y}_i \in S_\ell$, with $k \neq \ell$. If the terms $\langle \mathbf{e}_j, \mathbf{e}_i\rangle$, $\langle \mathbf{y}_j, \mathbf{e}_i\rangle$, and $\langle \mathbf{e}_j, \mathbf{y}_i\rangle$ are all small relative to $\frac{1}{\sqrt{d}}$, we have a margin on the order of $\frac{1}{\sqrt{d}}$ to distinguish pairs of points from within a given subspace from pairs of points from different subspaces. Indeed, $\langle \mathbf{y}_j, \mathbf{e}_i\rangle$ and $\langle \mathbf{e}_j, \mathbf{y}_i\rangle$ are small relative to $\frac{1}{\sqrt{d}}$ if $\frac{\sigma}{\sqrt{m}}$ is small relative to $\frac{1}{\sqrt{d}}$ (cf. (49)), while $\frac{\sigma^2}{\sqrt{m}}$ being small relative to $\frac{1}{\sqrt{d}}$ ensures that $\langle \mathbf{e}_j, \mathbf{e}_i\rangle$ is small relative to $\frac{1}{\sqrt{d}}$ (cf. (53)). These two conditions are obviously satisfied when $\sigma(1+\sigma)\sqrt{d/m}$ is small.

5 Incomplete data

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. It is therefore of significant interest to understand the impact of incomplete observations on the performance of TSC. Corresponding results for deterministic subspaces will necessarily depend on the specific orientations of the subspaces and will hence take on a form which makes it difficult to draw insightful conclusions. To make the problem analytically more tractable, we assume both the orientations of the subspaces as well as the data points in the subspaces to be random. Specifically, we will take the basis matrices $\mathbf{U}^{(\ell)}$ of the subspaces S_ℓ to be i.i.d. Gaussian random matrices, which ensures that each $\mathbf{U}^{(\ell)}$ is approximately orthonormal with high probability (rather than the $\mathbf{U}^{(\ell)}$ being strictly orthonormal as in the previous sections). For simplicity of exposition, throughout this section, we take the subspaces S_ℓ to have equal dimension d and let the number of points in each of the subspaces be n. We furthermore set the values of the unobserved entries in each data vector to zero and keep working in the original m-dimensional ambient space. As the TSC algorithm depends on inner products between the data points only this ensures that the missing observations will result in zero contributions.

Theorem 4. *Suppose that $\mathcal{X}_\ell$ is obtained by choosing n points corresponding to S_ℓ according to $\mathbf{x}_j^{(\ell)} = \mathbf{U}^{(\ell)}\mathbf{a}_j^{(\ell)}, j \in [n]$, where the $\mathbf{a}_j^{(\ell)}$ are i.i.d. uniform on $\mathbb{S}^{d-1}$, and set $\mathcal{X} = \mathcal{X}_1 \cup ... \cup \mathcal{X}_L$. Let the entries of the $\mathbf{U}^{(\ell)} \in \mathbb{R}^{m\times d}$ be i.i.d. $\mathcal{N}(0, 1/m)$. Pick $\rho \in [0, 1)$ and suppose that $n \geq n_0$, where n_0 is a constant that depends on d and ρ only. Suppose furthermore that $q \leq n^\rho$, and assume that in each $\mathbf{x}_j \in \mathcal{X}$ up to s arbitrary entries (possibly different for different $\mathbf{x}_j$) are unobserved, i.e., set to 0. If*

$$m \geq 3c_4 d + s\left(c_4 \log\left(\frac{me}{2s}\right) + c_3\right) + c_4 \log L \tag{9}$$

then G has no false connections with probability at least $1 - Lne^{-c_1(n-1)}$, where $c_1, c_2, c_3, c_4 > 0$ are numerical constants. If $s = 0$, (9) reduces to $m \geq 3c_4 d + c_4 \log L$.

Theorem 4 shows that the number of missing entries in the data vectors is allowed to be (up to a log-factor) linear in the ambient dimension. We can furthermore conclude that TSC succeeds (in the sense of G having no false connections) with high probability even when the dimensions of the subspaces are linear in the ambient dimension. This should, however, be taken with a grain of salt as the fully random subspace model ensures that the subspaces are approximately pairwise orthogonal with high probability, and hence the affinities between the subspaces are close to zero.

6 Outlier detection

We discuss the noiseless and the noisy case separately as the corresponding outlier models differ slightly. Moreover, the proof for the noiseless case is very simple and insightful and thus warrants 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 do not exhibit low-dimensional structure. Here, this is conceptualized by assuming random outliers distributed uniformly on $\mathbb{S}^{d-1}$, the unit sphere of $\mathbb{R}^m$. As before, the inliers are assumed to be distributed uniformly on $S_\ell \cap \mathbb{S}^{d_\ell - 1}$. The outlier detection criterion we employ is based on the following observation. The maximum inner product between an outlier and any other point (be it outlier or inlier) in $\mathcal{X}$ is, with high probability, smaller than $c\sqrt{\log N}/\sqrt{m}$, as made rigorous in the proof of Theorem 5 below. We therefore classify $\mathbf{x}_j$ as an outlier if

$$\max_{i\in[N]\setminus\{j\}} |\langle \mathbf{x}_i, \mathbf{x}_j\rangle| < c\sqrt{\log N}/\sqrt{m}. \tag{10}$$

The maximum inner product between any point $\mathbf{x}_j \in \mathcal{X}_\ell$ and the points in $\mathcal{X}_\ell \setminus \{\mathbf{x}_j\}$ is unlikely to be smaller than $1/\sqrt{d_{\max}}$, as formalized in the proof of Theorem 5. Hence, an inlier is unlikely to be misclassified as an outlier if $c\sqrt{\log N}/\sqrt{m} \leq 1/\sqrt{d_{\max}}$, i.e., if $d_{\max}/m$ is sufficiently small relative to $1/\sqrt{\log N}$. The following result formalizes this insight.

Theorem 5. *Suppose that the set of outliers, $\mathcal{O}$, is obtained by choosing N_0 outliers i.i.d. uniformly on $\mathbb{S}^{m-1}$, and that $\mathcal{X}_\ell, \ell \in [L]$, is obtained by choosing n_ℓ points i.i.d. uniformly from $\{\mathbf{x} \in S_\ell \colon \|\mathbf{x}\|_2 = 1\}$, independently across ℓ. Set $\mathcal{X} = \mathcal{X}_1 \cup ... \cup \mathcal{X}_L \cup \mathcal{O}$ and declare $\mathbf{x}_j \in \mathcal{X}$ to be an outlier if (10) holds with $c = \sqrt{6}$. Then, with $N = N_0 + \sum_\ell n_\ell$, all outliers are detected with probability at least $1 - 2N_0/N^2$. Furthermore, provided that*

$$\frac{d_{\max}}{m} \leq \frac{1}{6\log N} \tag{11}$$

no inlier in S_ℓ is misclassified as an outlier with probability at least

$$1 - n_\ell e^{-\frac{1}{2}\log\left(\frac{\pi}{2}\right)(n_\ell - 1)}. \tag{12}$$

Theorem 5 states that under Condition (11) and provided that the set $\mathcal{X}$ contains sufficiently many points from each subspace (cf. (12)), outlier detection succeeds with high probability, i.e., every outlier is detected and no inlier is misclassified as an outlier. Note that this result does not make any assumptions on the orientations of the subspaces S_ℓ.

Since (11) can be rewritten as $N_0 \leq e^{\frac{m}{6d_{\max}}} - \sum_\ell n_\ell$, it follows that outlier detection succeeds even if the number of outliers is exponential in $m/d_{\max}$.

Finally, note that the outlier detection rule (10) is very natural as it simply classifies those points as outliers whose (spherical) distance to *all* other points, and hence also to their individual nearest neighbors, is large. The scheme provably works as the nearest neighbor of each inlier is typically much closer than the nearest neighbor of each outlier. The idea of performing outlier detection based on nearest neighbor distance properties appeared previously e.g. in [33] (not in the context of subspace clustering though), where outliers are detected based on the connectivity properties of mutual[2] nearest neighbor graphs.

6.2 Noisy case

We next consider outlier detection under additive noise on the data points. To keep the analysis simple, we change the outlier model slightly. Specifically, we assume the outliers to be $\mathcal{N}(\mathbf{0}, (1/m)\mathbf{I}_m)$ distributed. Conceptually, this outlier model is equivalent to the one used in Section 6.1, as the directions of the outliers in the present model, i.e., $\mathbf{x}_i/\|\mathbf{x}_i\|_2$, are uniformly distributed on $\mathbb{S}^{m-1}$, and $\|\mathbf{x}_i\|_2$ concentrates around 1. We furthermore normalize the (noisy) data points such that the norm of the inliers also concentrates around 1. This guarantees that outlier detection is not trivially accomplished by exploiting differences in the norms between inliers and outliers.

Theorem 6. *Suppose that the set of outliers, $\mathcal{O}$, is obtained by choosing N_0 outliers i.i.d. $\mathcal{N}(\mathbf{0}, (1/m)\mathbf{I}_m)$, and that $\mathcal{X}_\ell, \ell \in [L]$, is obtained by choosing n_ℓ points corresponding to S_ℓ according to $\mathbf{x}_j^{(\ell)} = \frac{1}{\sqrt{1+\sigma^2}}\left(\mathbf{y}_j^{(\ell)} + \mathbf{e}_j^{(\ell)}\right), j \in [n_\ell]$, where the $\mathbf{y}_j^{(\ell)}$ are chosen i.i.d. uniformly from $\{\mathbf{y} \in S_\ell \colon \|\mathbf{y}\|_2 = 1\}$, independently across ℓ, and the $\mathbf{e}_j^{(\ell)}$ are i.i.d. $\mathcal{N}(\mathbf{0}, (\sigma^2/m)\mathbf{I}_m)$. Let $\mathcal{X} = \mathcal{X}_1 \cup \ldots \cup \mathcal{X}_L \cup \mathcal{O}$ and declare $\mathbf{x}_j \in \mathcal{X}$ to be an outlier if (10) holds with $c = 2.3\sqrt{6}$. Then, with $N = N_0 + \sum_l n_\ell$, assuming $m \geq 6 \log N$, all outliers are detected with probability at least $1 - 3\frac{N_0}{N^2}$. Furthermore, provided that*

$$\frac{d_{\max}}{m} \leq \frac{c_1}{(1+\sigma^2)^2 \log N} \tag{13}$$

where c_1 is a numerical constant, no inlier belonging to S_ℓ is misclassified as an outlier with probability at least

$$1 - n_\ell e^{-\frac{1}{2}\log\left(\frac{\pi}{2}\right)(n_\ell - 1)} - n_\ell^2 \frac{7}{N^3}. \tag{14}$$

Theorem 6 shows that outlier detection can succeed even under massive noise provided that $d_{\max}/m$ is sufficiently small.

[2] In a mutual k-nearest neighbor graph, the points $\mathbf{x}_i$ and $\mathbf{x}_j$ are connected if $\mathbf{x}_i$ is among the k-nearest neighbors of $\mathbf{x}_j$ *and* $\mathbf{x}_j$ is among the k-nearest neighbors of $\mathbf{x}_i$.

7 Comparison with SSC/RSSC and other algorithms

As mentioned in the introduction, there are only a few subspace clustering algorithms that are both computationally tractable and succeed *provably* under non-restrictive conditions. Notable exceptions are the SSC algorithm [13, 14], and for the noisy case the RSSC algorithm [12] (an algorithm analogous to the RSSC algorithm was also studied in [14]). Since our analytical performance results are in the spirit of those for SSC and RSSC in [15, 12]—in particular we use the same statistical data model—we next compare our findings to those in [15, 12]. Analytical performance guarantees for SSC in the fully deterministic case can be found in [14].

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 ℓ_1-minimization or Lasso, 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, essentially sharing the *analytical* performance guarantees of SSC and RSSC. The complexity savings may, however, come at the cost of actual performance. Specifically, while there are situations where TSC outperforms SSC, SSC outperforming TSC is more common, as will be seen in the numerical results in Section 8.

Concerning analytical performance guarantees, for SSC in the noiseless case, a result along the lines of Theorem 2 was reported in [15, Theorem 2.8], with the corresponding clustering condition in [15, Theorem 2.8] being identical (up to constants and log-factors) to our condition (5). However, the statement in [15, Theorem 2.8] is weaker than that in Theorem 2 as it does not pertain to the clustering error directly, but rather ensures no false connections only. To prove that the clustering error is zero, we additionally establish that the subgraphs corresponding to the $\mathcal{X}_\ell$ are connected. As already mentioned, this requires a lower bound on q, which entails that the n_ℓ be exponential in the d_ℓ. While this restricts the range of parameters d_ℓ, n_ℓ Theorems 1 and 2 apply to, the corresponding statements are strongest possible as they guarantee that the clustering error is zero as opposed to ensuring no false connections only. Again, as mentioned before, this exponential dependency appears to be an artifact of the proof technique we employ.

In the noisy case for RSSC a result analogous to our Theorem 3 was reported in [12, Theorem 3.1], with the corresponding clustering condition in [12, Theorem 3.1] being identical (again up to constants and log-factors) to our condition (8) with $\sigma(1+\sigma)$ in (8) replaced by σ. We note, however, that [12] requires σ to be bounded in the sense of $\sigma \leq c$, for some constant c, an assumption not needed in our case. If we take σ to satisfy $\sigma \leq c$, the factor $\sigma(1+\sigma)$ in Condition (8) above can be replaced by $\sigma(1+c)$ and we would get a clustering condition that is equivalent (again up to constants and log-factors) to that in [12]. A result concerning clustering of incompletely observed data paralleling our Theorem 4 does not seem to be available for SSC. The outlier detection scheme proposed in [15] in the context of SSC is based on the premise that outliers can not be represented sparsely in terms of the other data points. This scheme succeeds (i.e., every outlier is detected and no inlier is misclassified as an outlier) with probability at least $1-N_0 e^{-c\frac{n}{\log N}} - \sum_{\ell=1}^{L} n_\ell e^{-\sqrt{d_\ell}\sqrt{n_\ell - 1}}$ under the condition $N_0 \leq \min\{e^{c\sqrt{m}}/m, m \min_\ell (n_\ell/d_\ell)^{cm/d_\ell}\} - \sum_{\ell=1}^{L} n_\ell$, while our outlier detection scheme succeeds under Condition (9), i.e., $N_0 \leq e^{\frac{m}{6d_{\max}}} - \sum_{\ell=1}^{L} n_\ell$, with probability at least $1 - \frac{2N_0}{N^2} - \sum_{\ell=1}^{L} n_\ell e^{-\frac{1}{2}\log\left(\frac{\pi}{2}\right)(n_\ell - 1)}$. For both algorithms the number of outliers can be exponential in $m/d_{\max}$, the success probability increases in the n_ℓ, and the n_ℓ can be linear in the d_ℓ.

In terms of input parameters, RSSC in [12] chooses the Lasso regularization parameter λ in a data-driven fashion, which makes the algorithm essentially parameterless. TSC in contrast has q as an input parameter. A variation of TSC with a data-driven choice of q was proposed in [34] and shown to lead to performance guarantees that are essentially equivalent to those reported in this

paper.

A comparison of the analytical performance results for RSSC (in particular [12, Theorem 3.1]) to those for a number of representative subspace clustering algorithms such as generalized PCA (GPCA) [35], K-flats [36], and LRR [16], can be found in [12, Section 5]. This comparison also features computational complexity 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 [12, Section 5] essentially carry over to TSC.

8 Numerical results

We use the following performance metrics.

- The clustering error (CE) measures the fraction of misclassified points and is defined as follows. Denote the estimate of the number of subspaces by $\hat{L}$. Let $\mathbf{c} \in [L]^N$ and $\hat{\mathbf{c}} \in [\hat{L}]^N$ be the original and estimated assignments of the points in $\mathcal{X}$ to the individual subspaces. The CE is then defined as
$$\mathrm{CE}(\hat{\mathbf{c}}, \mathbf{c}) = \min_{\pi} \left(1 - \frac{1}{N} \sum_{i=1}^{N} 1_{\{\pi(\hat{c}_i)=c_i\}} \right)$$
where the minimum is taken over all assignments $\pi \colon [L] \to [\hat{L}]$ (for $\hat{L} = L$, π is simply a permutation). Note that π appears naturally in the definition of the CE as the specific cluster indices are irrelevant to the CE. The problem of finding the optimal assignment π can be cast as finding the maximal matching of a weighted bipartite graph, which can be solved efficiently via the Hungarian algorithm [37].

- 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 so as to be able to discriminate between under- and overestimation. In principle, EL averaged over problem instances, may therefore 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) (for a given adjacency matrix $\mathbf{A}$) is defined as
$$\mathrm{FDE}(\mathbf{A}) = 1 - \frac{1}{N} \sum_{i=1}^{N} \frac{\|\mathbf{b}_{\mathbf{x}_i}\|_2}{\|\mathbf{b}_i\|_2}$$
where $\mathbf{b}_i$ is the ith column of the $N \times N$ adjacency matrix $\mathbf{A}$ and $\mathbf{b}_{\mathbf{x}_i}$ is the vector containing the entries of $\mathbf{b}_i$ corresponding to the set $\mathcal{X}_\ell$ the data point $\mathbf{x}_i$ lives in. The FDE measures to which extent points from different subspaces are connected in the graph G (with adjacency matrix $\mathbf{A}$), and equals zero if G has no false connections.

Throughout this section, we set $q = \max(3, \lceil n/20 \rceil)$ if the correct L is provided to TSC, and $q = 2\max(3, \lceil n/20 \rceil)$ if L is estimated according to the eigengap heuristic. Matlab code to reproduce the results in this section is available at http://www.nari.ee.ethz.ch/commth/research/.

8.1 Synthetic data

Throughout Section 8.1, unless explicitly stated otherwise, we take $n_\ell = n$ and $d_\ell = d$, for all ℓ, and generate the d-dimensional subspaces S_ℓ by drawing i.i.d. orthonormal basis matrices $\mathbf{U}^{(\ell)} \in \mathbb{R}^{m \times d}$ uniformly at random from the set of all orthonormal matrices in $\mathbb{R}^{m \times d}$.

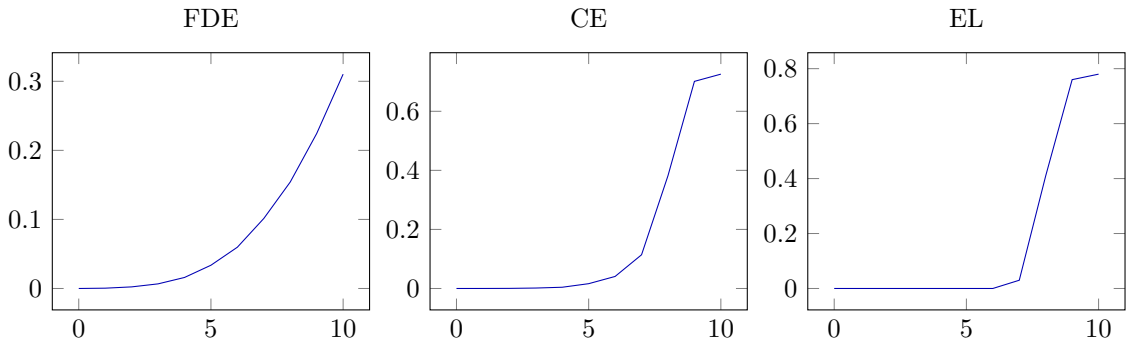

Figure 2: Clustering error metrics as a function of the dimension of the intersection, t, for clustering points taken from two 10-dimensional subspaces of $\mathbb{R}^{200}$.

8.1.1 Intersection of subspaces

We next demonstrate that, as predicted by Theorem 2, TSC can succeed even when the subspaces S_ℓ intersect. In order to facilitate comparison to SSC, we perform the same experiment as in [15, Sec. 5.1.2]. Specifically, we set $m = 200$, $d = 10$, and generate two subspaces, S_1 and S_2, at random through their defining bases $\mathbf{U}^{(1)}$ and $\mathbf{U}^{(2)}$ obtained as follows. We choose, uniformly at random, from the set of all sets of $2d-t$ orthonormal vectors in $\mathbb{R}^m$, a set of $2d-t$ orthonormal vectors, and identify the columns of $\mathbf{U}^{(1)}$ and $\mathbf{U}^{(2)}$ with the first and last d of these vectors, respectively. This ensures that the intersection of S_1 and S_2 is at least of dimension t. Next, we generate $n = 20d$ data points in each of the two subspaces according to $\mathbf{x}_i^{(\ell)} = \mathbf{U}^{(\ell)}\mathbf{a}_i^{(\ell)}$, with the $\mathbf{a}_i^{(\ell)}$ drawn i.i.d. uniformly on $\mathbb{S}^{d-1}$. For each $t = 0, ..., d$ the CE, EL, and FDE are obtained by averaging over 100 problem instances. From the results, shown in Figure 2, we can conclude that, as long as the dimension of the intersection of the subspaces is not too large, TSC does, indeed, yield a CE close to zero. The same experiment was performed for SSC in [15, Sec. 5.1.2] and delivered slightly better results.

8.1.2 Influence of d, n, and incomplete data

The goal of the next experiment is to elicit the impact 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 G has false connections. We generate $L = 10$ subspaces of $\mathbb{R}^{50}$, and vary their dimension d and the number n of points taken from each subspace. The individual data points are chosen according to the statistical model Theorem 2 is based on. For each pair (d, n), the FDE, CE, and EL are obtained by averaging over 20 problem instances. The results, depicted in Figure 3, show, as indicated in Section 2, that TSC can, indeed, succeed even when G has false connections (i.e., when the FDE is non-zero).

Next, we generate $L = 6$ subspaces of $\mathbb{R}^{50}$ by choosing their defining bases $\mathbf{U}^{(\ell)}$ as follows. We first draw $\mathbf{U} \in \mathbb{R}^{m\times d/3}$ (we restrict d to integer multiples of 3) uniformly from the set of all orthonormal matrices in $\mathbb{R}^{m\times d/3}$. Then, we choose $\tilde{\mathbf{U}}^{(\ell)} \in \mathbb{R}^{m\times 2d/3}, \ell \in [L]$, independently across ℓ and independently of $\mathbf{U}$, uniformly at random from the set of all orthonormal matrices in $\mathbb{R}^{m\times 2d/3}$ that are orthogonal to $\mathbf{U}$, and set $\mathbf{U}^{(\ell)} = [\tilde{\mathbf{U}}^{(\ell)}\ \mathbf{U}] \in \mathbb{R}^{m\times d}$. This ensures that the subspaces S_ℓ with basis matrices $\mathbf{U}^{(\ell)}$ intersect in at least $d/3$ dimensions and hence $\mathrm{aff}(S_k, S_\ell) \geq 1/\sqrt{3}$ for all $k, \ell \in [L], k \neq \ell$. The data points are chosen according to the statistical model Theorem 2 is based on. For each data point $\mathbf{x}_i$, we set the entries of $\mathbf{x}_i$ with indices in $\mathcal{D}_i$ to zero, where the sets $\mathcal{D}_i$ are chosen independently and uniformly at random from the set $\{\mathcal{D} \subseteq [m] \colon |\mathcal{D}| = s\}$. The results,

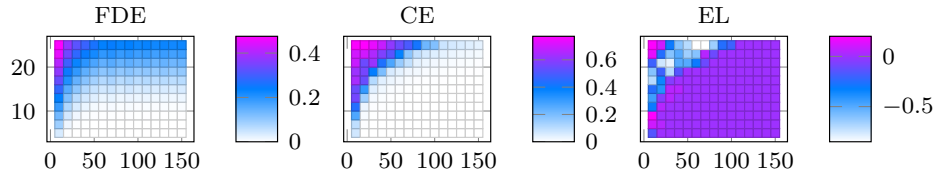

Figure 3: Clustering error metrics as a function of the dimension, d, of the subspaces on the vertical and the number of points taken from each subspace, n, on the horizontal axis, for $L = 10$ subspaces of $\mathbb{R}^{50}$.

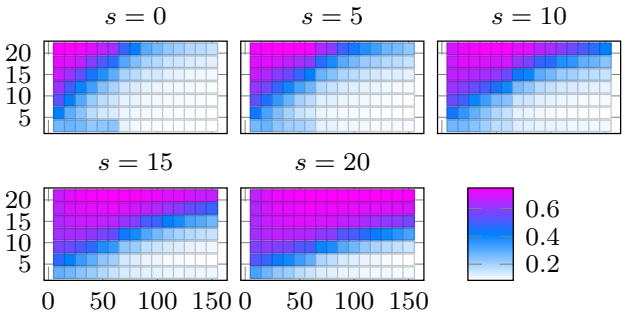

Figure 4: Clustering error as a function of the dimension, d, of the subspaces on the vertical and the number of points taken from each subspace, n, on the horizontal axis for s missing entries in the data vectors. The results are for $L = 6$ subspaces S_ℓ of $\mathbb{R}^{50}$, with $\mathrm{aff}(S_k, S_\ell) \geq 1/\sqrt{3}$ for all $k, \ell \in [L], k \neq \ell$.

summarized in Figure 4, show that TSC can succeed even when a large fraction of the entries in each data vector is missing.

8.1.3 Additive noise

We generate $L = 10$ subspaces of $\mathbb{R}^{50}$ and vary their dimension d and the number of points n taken from each subspace. The data points are subjected to additive noise before clustering. Specifically, we use the statistical data model Theorem 3 is based on. The results, depicted in Figure 5, show that TSC can succeed even when the noise variance is large.

In Section 4, 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 $\mathbb{R}^{400}$, each of dimension $d = 5$ (hence $d/m = 1/80$), and we 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, and the noise variance σ^2. The corresponding results, depicted in Figure 6, confirm the analytical predictions of Theorem 3.

8.1.4 Detection of outliers

In order to facilitate comparison with the outlier detection scheme proposed for SSC in [15], we perform our experiment with exactly the same parameters as used in [15, 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 5 is based on. The number of outliers is hence equal to the total number of inliers. We measure

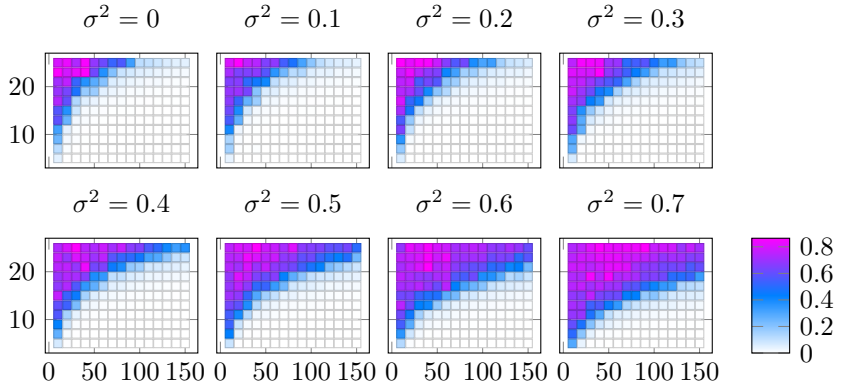

Figure 5: Clustering error for data points taken from $L = 10$ subspaces of $\mathbb{R}^{50}$ corrupted by additive Gaussian noise, as a function of the dimension, d, of the subspaces on the vertical and the number of points taken from each subspace, n, on the horizontal axis for different noise variances σ^2.

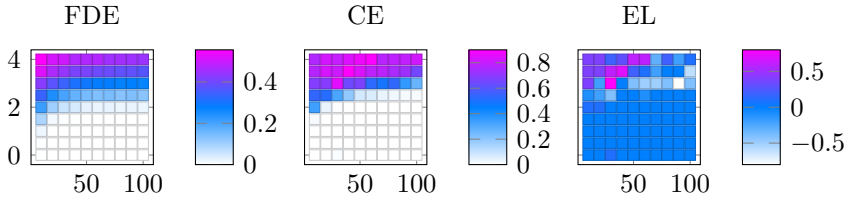

Figure 6: Clustering error metrics for points taken from $L = 5$ subspaces of $\mathbb{R}^{400}$, each of which is 5-dimensional, corrupted by additive Gaussian noise, as a function of the noise variance, σ^2, on the vertical and the number of points taken from each subspace, n, on the horizontal axis.

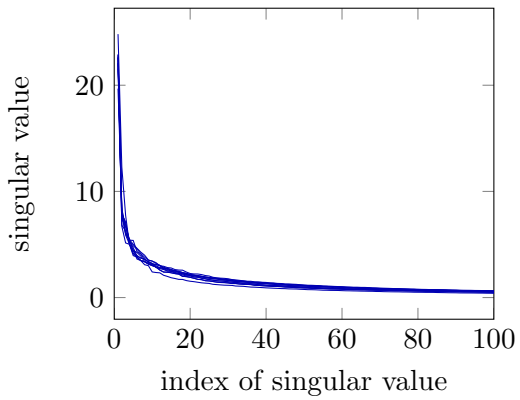

Figure 7: Singular values of the matrices with columns corresponding to the vectorized images of a given digit from the MNIST data base.

performance in terms of the misclassification error, 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 $\mathcal{X}$. We find a misclassification error of $\{0.017, 1.510^{-4}, 2.510^{-5}\}$ for $m = \{50, 100, 200\}$, respectively. The performance reported for SSC in [15] is similar.

8.2 Clustering handwritten digits

We next apply TSC to the problem of clustering handwritten digits. Specifically, we work with the MNIST test data set [18] that contains 10,000 centered 28×28 pixel images of handwritten digits. The assumption underlying the idea of posing this problem as a subspace clustering problem is that the vectorized images of the different handwritten versions of a single digit lie approximately in a low-dimensional subspace [38]. To validate this assumption, we compute the singular values of the matrices $\mathbf{X}_\ell$ with columns corresponding to the vectorized images of the ℓth digit, $\ell = 0, 1, ..., 9$, and sort them in descending order. The results, plotted in Figure 7, show that the singular values of the matrices $\mathbf{X}_\ell$, indeed, decay to zero rapidly ($m = 784$). As mentioned in Section 2, TSC is built on the premise that the vectors close to $\mathbf{x}_j$ in terms of the distance $\tilde{s}(\mathbf{x}_i,\mathbf{x}_j) = \arccos(|\langle \mathbf{x}_i, \mathbf{x}_j \rangle|)$ also lie in the subspace $\mathbf{x}_j$ lies in. As our analytical results in Sections 3 and 4 show, this premise is met (with high probability) for the statistical data model used throughout the paper. To see whether the premise is also met in practice, we compute $\exp(-\tilde{s}(\mathbf{x}_i, \mathbf{x}_j))$ for all pairs $\mathbf{x}_j, \mathbf{x}_i$ of vectorized images of the digits $\{1, 3, 7\}$ from the MNIST data set. In other words, we compute the adjacency matrix for $q = N$. The results, depicted in Figure 8, show that, indeed, $\exp(-\tilde{s}(\mathbf{x}_i, \mathbf{x}_j))$ for $\mathbf{x}_i, \mathbf{x}_j$ coming from the same digit is typically larger than for $\mathbf{x}_i, \mathbf{x}_j$ coming from different digits.

We compare the performance of TSC, LS-TSC, and SSC/RSSC. For SSC, we use the implementation from [14]. The empirical mean and variance of the CE are computed by averaging over 100 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. The results, summarized in Figure 9, show that SSC performs better than both TSC and LS-TSC when the data set contains few ($n \lesssim 80$) images of each digit, TSC and LS-TSC outperform SSC when the data set contains many ($n \gtrsim 80$) images of each digit. Note that the FDE for TSC is significantly smaller than that for SSC, even in the regime $n \lesssim 80$ where SSC performs better. This is a result of q being small, which yields a sparse adjacency matrix for TSC, and therefore increases the chance of the nonzero entries, indeed, corresponding to points within the same subspace.

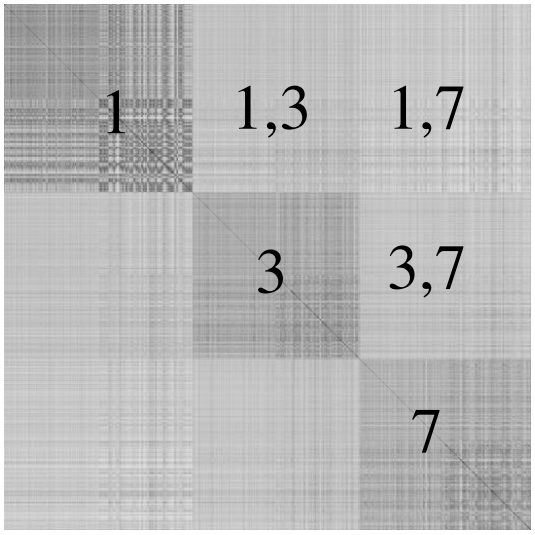

Figure 8: Matrix with entries $\mathbf{A}_{ij} = \exp(-\arccos(|\langle \mathbf{x}_j, \mathbf{x}_i \rangle|))$ for all pairs $\mathbf{x}_j, \mathbf{x}_i$ of vectorized images of the digits $\{1, 3, 7\}$ from the MNIST data base.

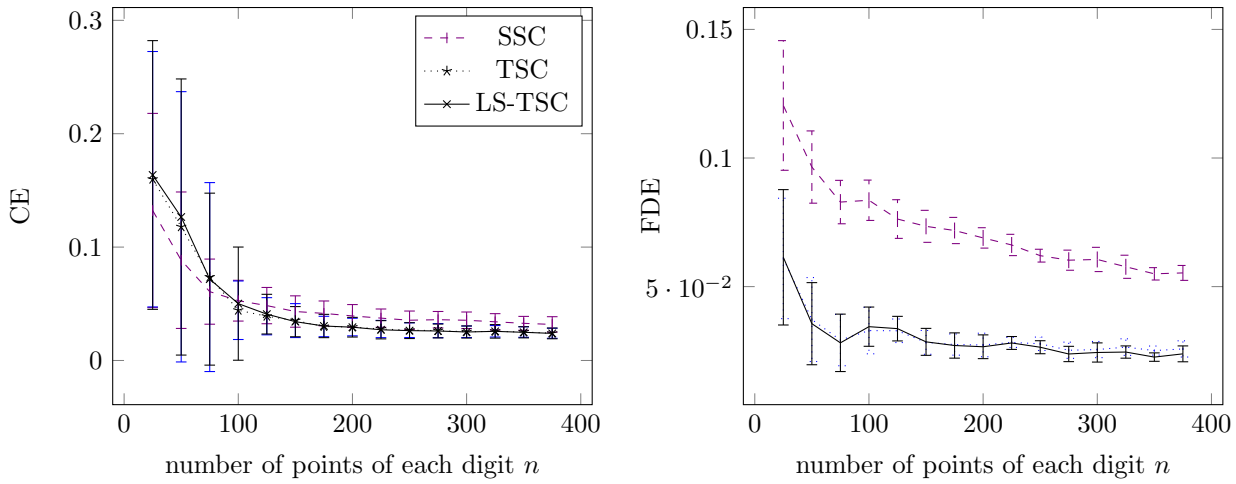

Figure 9: Empirical mean and standard deviation of the CE and FDE for handwritten digits from the MNIST data base.

L	2	3	5	8	10
CE, TSC, orig. dat.	12.42%	19.85%	29.17%	36.84%	39.84%
FDE, TSC, orig. dat.	0.0248	0.0419	0.0648	0.0863	0.0971
CE, TSC, whitening	8.06%	9%	10.14%	12.58%	17.86%
FDE, TSC, whitening	0.0154	0.0245	0.0384	0.0525	0.0591
CE, LSA	32.8%	52.29%	58.02%	59.19%	60.42%
CE, SCC	16.62%	38.16%	58.90%	66.11%	73.02%
CE, LRR	9.52%	19.52%	34.16%	41.19%	38.85%
CE, LatLRR	2.54%	4.21%	6.9%	14.34%	22.92%
CE, LRSC	5.32%	8.47%	12.24%	23.72%	30.36%
CE, SSC	1.86%	3.1%	4.31%	5.85%	10.94%

Table 1: CE and FDE for clustering faces for TSC. The CEs for LSA, SCC, LRR, LatLRR, LRSC, and SSC are taken from [14, Table 5].

8.3 Clustering faces

We finally apply TSC to the problem of clustering images of faces taken under varying illumination conditions. The motivation for applying TSC to this problem stems from the insight that the vectorized images of a given face taken under varying illumination conditions lie approximately in a 9-dimensional linear subspace [3]. 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 [19, 20], which contains 192×168 pixel images of 38 persons, each taken under 64 different illumination conditions. To be able to compare our results to those reported in [14] for SSC, SCC, Local Subspace Affinity (LSA) [22], Low-Rank Subspace Clustering (LRSC) [39], and LatLRR [40], we apply TSC to exactly the same data sets as used in [14, Sec. 7.2]. The averages of the CE and the FDE we obtain for $L = 2, 3, 5, 8, 10$ subjects are reported in Table 1, along with the values from [14, Table 5]. Comparing these results to [14, Table 5] shows that TSC performs better than LSA and SCC, but worse than LRR, LatLRR, LRSC, and SSC, with the latter exhibiting the best performance in the group LSC, SCC, LRR, LatLRR, LRSC, TSC, SSC.

As pointed out in [5, Section 3.3] the subspaces corresponding to different persons are extremely close to each other, which renders the corresponding clustering problem hard. Discrimination between the clusters (and hence persons) can be improved through preprocessing of the data set as described in [5, Section 3.3]. Specifically, the preprocessed data set $\tilde{\mathcal{X}}$ is obtained by removing the first two principal components of $\mathbf{X}$, where $\mathbf{X}$ is the matrix whose columns are the data points in $\mathcal{X}$, and taking the points in $\tilde{\mathcal{X}}$ as the columns of the resulting matrix. We applied TSC with preprocessing to the same data sets as used in [14]. Comparing the corresponding results, summarized in the first four rows in Table 1, to the results reported in [14, Table 5], and reproduced in our Table 1, for completeness, we can see that TSC with preprocessing performs better than LSA, SCC, and LRR applied to the raw data, but worse than LatLRR for $L = 2, 3, 5$, LRSC for $L = 2, 3$, and SSC for all L considered, in all cases applied to the raw data. We note that TSC with preprocessing remains computationally less demanding than the other algorithms without preprocessing.

Acknowledgments

We would like to thank Eirikur Agustsson for helpful discussions, in particular a result he obtained inspired our proof of Lemma 2. Moreover, we would like to thank Mahdi Soltanolkotabi for helpful and inspiring discussions.

A Proof of Theorem 1

The $\mathcal{X}_\ell$ in Theorem 1 are obtained by choosing n_ℓ points uniformly from $\{\mathbf{x} \in S_\ell \colon \|\mathbf{x}\|_2 = 1\}$. As mentioned previously, this is equivalent to choosing the points according to $\mathbf{x}_j^{(\ell)} = \mathbf{U}^{(\ell)}\mathbf{a}_j^{(\ell)}, j \in [n_\ell]$, where the $\mathbf{a}_j^{(\ell)}$ are i.i.d. uniform on $\mathbb{S}^{d_\ell - 1}$, and $\mathbf{U}^{(\ell)} \in \mathbb{R}^{m \times d_\ell}$ is an orthonormal basis for the subspace S_ℓ.

The proof is effected by showing that the connected components in the (random) graph G with adjacency matrix $\mathbf{A}$ (constructed by the TSC algorithm) correspond to the $\mathcal{X}_\ell$ with high probability. As mentioned previously, normalized spectral clustering will identify these components perfectly [11, Prop. 4] and hence yield correct segmentation of $\mathcal{X}$.

We prove that the connected components in G correspond to the $\mathcal{X}_\ell$ by showing that G has no false connections and the subgraphs $G(\mathcal{X}_\ell)$ corresponding to the $\mathcal{X}_\ell$ are connected, for all ℓ. To this end, we define the events NFC $\coloneqq \{G$ has no false connections$\}$ and C $\coloneqq \{G(\mathcal{X}_\ell)$ is connected, for all $\ell\}$ and upper-bound the probability $\mathrm{P}\left[\overline{\mathrm{C} \text{ and } \mathrm{NFC}}\right]$. This will be accomplished by exploiting the fact that conditioned on NFC, owing to $\langle \mathbf{x}_i, \mathbf{x}_j \rangle = \langle \mathbf{a}_i, \mathbf{a}_j \rangle$, for $\mathbf{x}_i, \mathbf{x}_j \in \mathcal{X}_\ell$ (by orthonormality of the $\mathbf{U}^{(\ell)}$), $G(\mathcal{X}_\ell)$ is the q-nearest neighbor graph of $\mathcal{X}_\ell$ with respect to the distance $\arccos(|\langle \mathbf{a}_i, \mathbf{a}_j \rangle|)$. An analysis of the connectivity properties of $G(\mathcal{X}_\ell)$ will then yield an upper bound on $\mathrm{P}\left[\overline{\mathrm{C}}|\mathrm{NFC}\right]$ which together with an upper bound on $\mathrm{P}\left[\overline{\mathrm{NFC}}\right]$ delivers the final result according to

$$\begin{aligned} \mathrm{P}\left[\overline{\mathrm{C} \text{ and } \mathrm{NFC}}\right] &= \mathrm{P}\left[\overline{\mathrm{C}} \text{ or } \overline{\mathrm{NFC}}\right] \\ &= \mathrm{P}\left[\overline{\mathrm{NFC}}\right] + \mathrm{P}\left[\overline{\mathrm{C}} \text{ and } \mathrm{NFC}\right] \\ &\leq \mathrm{P}\left[\overline{\mathrm{NFC}}\right] + \mathrm{P}\left[\overline{\mathrm{C}}|\mathrm{NFC}\right]. \end{aligned} \tag{15}$$

We proceed by establishing the upper bounds on the terms in the RHS of (15).

We will use Lemma 1 below, proven in Appendix A.1, to upper-bound $\mathrm{P}\left[\overline{\mathrm{NFC}}\right]$. The lemma is also a key ingredient of the proof of Theorem 4 pertaining to incomplete data, and is hence stated in a form general enough to cover that case as well.

Lemma 1. *Suppose that $\mathcal{X}_\ell$ is obtained by choosing n_ℓ points in S_ℓ according to $\mathbf{x}_j^{(\ell)} = \mathbf{U}^{(\ell)}\mathbf{a}_j^{(\ell)}, j \in [n_\ell]$, where the $\mathbf{a}_j^{(\ell)}$ are i.i.d. uniform on $\mathbb{S}^{d_\ell - 1}$, $\mathbf{U}^{(\ell)} \in \mathbb{R}^{m \times d_\ell}$ (not necessarily orthonormal), and let $\mathcal{X} = \mathcal{X}_1 \cup \ldots \cup \mathcal{X}_L$. Assume that in each $\mathbf{x}_j \in \mathcal{X}$ up to s arbitrary entries (possibly different for different $\mathbf{x}_j$) are unobserved, i.e., set to 0. Pick $\rho \in [0, 1)$ and suppose that $n_\ell \geq n_0$, for all $\ell \in [L]$, where n_0 is a constant that depends on $d_{\max}$ and ρ only. Suppose that $q \leq n_{\min}^\rho$ and*

$$\frac{\max_{k,\ell \colon k \neq \ell, \mathcal{D} \colon |\mathcal{D}| \leq 2s} \left\| \mathbf{U}_{\mathcal{D}}^{(k)^T} \mathbf{U}^{(\ell)} \right\|_{2 \to 2}}{\min_{\ell, \mathcal{D} \colon |\mathcal{D}| \leq 2s, \|\mathbf{a}\|_2 = 1} \left\| \mathbf{U}_{\mathcal{D}}^{(\ell)^T} \mathbf{U}^{(\ell)} \mathbf{a} \right\|_2} < 1 \tag{16}$$

where $\mathbf{U}_{\mathcal{D}}^{(\ell)} \in \mathbb{R}^{m \times d_\ell}$ is the matrix obtained from $\mathbf{U}^{(\ell)}$ by setting the rows with indices in $\mathcal{D}$ to zero. Then, G has no false connections with probability at least $1 - \sum_{\ell=1}^{L} n_\ell e^{-c_1(n_\ell - 1)}$, where $c_1 > 0$ is a numerical constant.

It follows from Lemma 1 with $s = 0$ that

$$\mathrm{P}\left[\overline{\mathrm{NFC}}\right] \leq \sum_{\ell=1}^{L} n_\ell e^{-c_1(n_\ell - 1)}. \tag{17}$$

To see this note that for $s = 0$ (16) reduces to (4). Specifically, for $s = 0$, $\mathbf{U}_{\mathcal{D}}^{(\ell)} = \mathbf{U}^{(\ell)}$ and since the $\mathbf{U}^{(\ell)}$ are orthonormal we have ${\mathbf{U}_{\mathcal{D}}^{(\ell)}}^T \mathbf{U}^{(\ell)} = {\mathbf{U}^{(\ell)}}^T \mathbf{U}^{(\ell)} = \mathbf{I}_{d_\ell}$. The denominator in (16) therefore equals 1, and the numerator reduces to $\max_{k,\ell:\, k\neq\ell} \left\| {\mathbf{U}^{(k)}}^T \mathbf{U}^{(\ell)} \right\|_{2\to 2} = \max_{k,\ell:\, k\neq\ell} \mathrm{aff}_\infty(S_k, S_\ell)$ which establishes the equivalence of (16) and (4).

It remains to upper-bound $\mathrm{P}\left[\overline{\mathrm{C}}|\mathrm{NFC}\right]$. By a union bound argument, we get

$$\mathrm{P}\left[\overline{\mathrm{C}}|\mathrm{NFC}\right] \leq \sum_{\ell=1}^{L} \mathrm{P}[G(\mathcal{X}_\ell) \text{ is not connected } |\mathrm{NFC}]\,. \tag{18}$$

As mentioned above, conditioned on NFC, $G(\mathcal{X}_\ell)$ is the q-nearest neighbor graph of $\mathcal{X}_\ell$ with pseudo-distance metric $\arccos(|\langle \mathbf{a}_i, \mathbf{a}_j \rangle|)$ (recall that, conditioned on NFC, we have $\langle \mathbf{x}_i, \mathbf{x}_j \rangle = \langle \mathbf{a}_i, \mathbf{a}_j \rangle$ for $\mathbf{x}_i, \mathbf{x}_j \in \mathcal{X}_\ell$). It is this insight that allows us to find upper bounds on the terms $\mathrm{P}[G(\mathcal{X}_\ell)$ is not connected $|\mathrm{NFC}]$ as formalized in Lemma 2 below, which is proven in Appendix A.2.

Lemma 2. *Let $\mathbf{a}_1, ..., \mathbf{a}_n \in \mathbb{R}^d$ be drawn i.i.d. uniformly on $\mathbb{S}^{d-1}$, $d > 1$, and let $\tilde{G}$ be the corresponding $\tilde{k}$-nearest neighbor graph with respect to the pseudo-distance metric $\tilde{s}(\mathbf{a}_i, \mathbf{a}_j) = \arccos(|\langle \mathbf{a}_i, \mathbf{a}_j \rangle|)$. Then, with $\tilde{k} \geq \gamma\, 6(12\pi)^{d-1} \log n$, for every $\gamma > 1$, we have*

$$\mathrm{P}\left[\tilde{G} \text{ is connected}\right] \geq 1 - \frac{2}{n^{\gamma-1}\gamma \log n}.$$

Using Lemma 2 we can then conclude that $\mathrm{P}[G(\mathcal{X}_\ell)$ is not connected $|\mathrm{NFC}] \leq 2{n_\ell}^{-\gamma+1}$ (using $\gamma \log n_\ell \geq \log n_\ell \geq \log n_0 \geq 1$) provided that $q \geq \gamma\, 6(12\pi)^{d_\ell - 1} \log n_\ell$, which is satisfied by the assumption $q \in [6(12\pi)^{d_{\max}-1} \gamma \log n_{\max}, n_{\min}^{\rho}]$. Inserting into (18) yields

$$\mathrm{P}\left[\overline{\mathrm{C}}|\mathrm{NFC}\right] \leq \sum_{\ell=1}^{L} 2{n_\ell}^{-\gamma+1}. \tag{19}$$

Finally, combining the upper bounds (17) and (19) in (15), we get

$$\mathrm{P}\left[\overline{\mathrm{C} \text{ and } \mathrm{NFC}}\right] \leq \sum_{\ell=1}^{L} \left(n_\ell e^{-c_1(n_\ell - 1)} + 2{n_\ell}^{-\gamma+1} \right)$$

as desired.

A.1 Proof of Lemma 1

We need to show that G has no false connections, i.e., for each $\mathbf{x}_i^{(\ell)} \in \mathcal{X}_\ell$, the associated set $\mathcal{T}_i$ corresponds to points in $\mathcal{X}_\ell$ only, for all ℓ. This is accomplished by proving that for $\mathbf{x}_i^{(\ell)} \in \mathcal{X}_\ell$, we have

$$z_{(n_\ell - q)}^{(\ell)} > \max_{k\neq\ell, j} z_j^{(k)}. \tag{20}$$

Here, $z_j^{(k)} \coloneqq \left|\left\langle \mathbf{x}_j^{(k)}, \mathbf{x}_i^{(\ell)} \right\rangle\right|$ and $z_{(1)}^{(\ell)} \leq z_{(2)}^{(\ell)} \leq \ldots \leq z_{(n_\ell - 1)}^{(\ell)}$ are the order statistics of $\{z_j^{(\ell)}\}_{j \in [n_\ell] \setminus \{i\}}$. Note that, for simplicity of exposition, the notation $z_j^{(k)}$ does not reflect dependence on $\mathbf{x}_i^{(\ell)}$. Next, we upper-bound the probability of (20) being violated. A union bound over all N vectors $\mathbf{x}_i^{(\ell)}, \ell \in [L], i \in [n_\ell]$, then yields the final result. We first note that for $k \neq \ell$, by the Cauchy-Schwarz inequality,

$$\begin{aligned}
z_j^{(k)} &= \left|\left\langle \mathbf{x}_j^{(k)}, \mathbf{x}_i^{(\ell)} \right\rangle\right| = \left|\left\langle \mathbf{U}_{\mathcal{D}}^{(k)} \mathbf{a}_j^{(k)}, \mathbf{U}_{\mathcal{E}}^{(\ell)} \mathbf{a}_i^{(\ell)} \right\rangle\right| \\
&= \left|\left\langle \mathbf{a}_j^{(k)}, {\mathbf{U}_{\mathcal{D}}^{(k)}}^T \mathbf{U}_{\mathcal{E}}^{(\ell)} \mathbf{a}_i^{(\ell)} \right\rangle\right| \leq \left\|\mathbf{a}_j^{(k)}\right\|_2 \left\|{\mathbf{U}_{\mathcal{D}}^{(k)}}^T \mathbf{U}_{\mathcal{E}}^{(\ell)} \mathbf{a}_i^{(\ell)}\right\|_2 \\
&\leq \left\|{\mathbf{U}_{\mathcal{D}}^{(k)}}^T \mathbf{U}_{\mathcal{E}}^{(\ell)}\right\|_{2\to 2} \left\|\mathbf{a}_j^{(k)}\right\|_2 \left\|\mathbf{a}_i^{(\ell)}\right\|_2 \\
&\leq \max_{k,\ell\colon\, k\neq\ell, \mathcal{D}\colon\, |\mathcal{D}|\leq 2s} \left\|{\mathbf{U}_{\mathcal{D}}^{(k)}}^T \mathbf{U}^{(\ell)}\right\|_{2\to 2}
\end{aligned}$$

where the sets $\mathcal{D}, \mathcal{E} \subset [m]$ contain the indices of the unobserved entries (set to zero) in $\mathbf{x}_j^{(k)}$ and $\mathbf{x}_i^{(\ell)}$, respectively, and $\mathbf{U}_{\mathcal{D}}^{(\ell)}, \mathbf{U}_{\mathcal{E}}^{(\ell)} \in \mathbb{R}^{m \times d_\ell}$ are the matrices obtained from $\mathbf{U}^{(\ell)} \in \mathbb{R}^{m \times d_\ell}$ by setting the rows with indices in $\mathcal{D}$ and $\mathcal{E}$, respectively, to zero. Since the distribution of $\mathbf{a}_j^{(\ell)}$ is rotationally invariant, we get, for a fixed $\mathbf{a}_i^{(\ell)}$ with unit norm, that

$$\begin{aligned}
z_j^{(\ell)} &= \left|\left\langle \mathbf{a}_j^{(\ell)}, {\mathbf{U}_{\mathcal{D}}^{(\ell)}}^T \mathbf{U}_{\mathcal{E}}^{(\ell)} \mathbf{a}_i^{(\ell)} \right\rangle\right| \\
&= \left|\left\langle \mathbf{a}_j^{(\ell)}, \frac{{\mathbf{U}_{\mathcal{D}}^{(\ell)}}^T \mathbf{U}_{\mathcal{E}}^{(\ell)} \mathbf{a}_i^{(\ell)}}{\left\|{\mathbf{U}_{\mathcal{D}}^{(\ell)}}^T \mathbf{U}_{\mathcal{E}}^{(\ell)} \mathbf{a}_i^{(\ell)}\right\|_2} \right\rangle\right| \left\|{\mathbf{U}_{\mathcal{D}}^{(\ell)}}^T \mathbf{U}_{\mathcal{E}}^{(\ell)} \mathbf{a}_i^{(\ell)}\right\|_2 \\
&\sim \left|\left\langle \mathbf{a}_j^{(\ell)}, \mathbf{a}_i^{(\ell)} \right\rangle\right| \left\|{\mathbf{U}_{\mathcal{D}}^{(\ell)}}^T \mathbf{U}_{\mathcal{E}}^{(\ell)} \mathbf{a}_i^{(\ell)}\right\|_2 \\
&\geq \min_{\ell, \mathcal{D}\colon\, |\mathcal{D}|\leq 2s, \|\mathbf{a}\|_2 = 1} \left\|{\mathbf{U}_{\mathcal{D}}^{(\ell)}}^T \mathbf{U}^{(\ell)} \mathbf{a}\right\|_2 \underbrace{\left|\left\langle \mathbf{a}_j^{(\ell)}, \mathbf{a}_i^{(\ell)} \right\rangle\right|}_{\tilde{z}_j^{(\ell)} \coloneqq}.
\end{aligned}$$

This allows us to conclude that, for all $z \in \mathbb{R}$,

$$\mathrm{P}\left[z_j^{(\ell)} \leq z\right] \leq \mathrm{P}\left[\min_{\ell, \mathcal{D}\colon\, |\mathcal{D}|\leq 2s, \|\mathbf{a}\|_2 = 1} \left\|{\mathbf{U}_{\mathcal{D}}^{(\ell)}}^T \mathbf{U}^{(\ell)} \mathbf{a}\right\|_2 \tilde{z}_j^{(\ell)} \leq z\right]$$

and hence the probability of (20) being violated can be upper-bounded according to

$$\mathrm{P}\left[z_{(n_\ell - q)}^{(\ell)} \leq \max_{k \neq \ell, j} z_j^{(k)}\right] \leq \mathrm{P}\left[\tilde{z}_{(n_\ell - q)}^{(\ell)} \leq 1 - \eta\right] \tag{21}$$

which, owing to (16), holds for an $\eta > 0$. Next, observe that

$$\begin{aligned}
\mathrm{P}\Big[\tilde{z}^{(\ell)}_{(n_\ell - q)} \leq 1-\eta\Big] &= \mathrm{P}\big[\text{there exists a set } I \subset [n_\ell] \setminus \{i\} \\
&\qquad \text{with } |I| = n_\ell - q \text{ such that } \tilde{z}^{(\ell)}_j \leq 1-\eta \text{ for all } j \in I\big] \\
&\leq \binom{n_\ell - 1}{n_\ell - q} \max_{I\colon |I| = n_\ell - q} \mathrm{P}\Big[\tilde{z}^{(\ell)}_j \leq 1-\eta, \text{ for all } j \in I\Big] \qquad (22) \\
&\leq \left(e\frac{n_\ell - 1}{q-1}\right)^{q-1} p^{n_\ell - q} \qquad (23)
\end{aligned}$$

with $p = \mathrm{P}\Big[\tilde{z}^{(\ell)}_j \leq 1-\eta\Big]$ (recall that the $\tilde{z}^{(\ell)}_j$ are i.i.d.), where we used a union bound to get (22) and $\binom{n}{n-k} = \binom{n}{k} \leq \left(\frac{en}{k}\right)^k$ [41] for (23). Since (23) is increasing in q, and $q \leq n^\rho_{\min} \leq n^\rho_\ell$, by assumption, setting $\varrho = \frac{n_\ell - 1}{n^\rho_\ell - 1}$, we obtain

$$\begin{aligned}
\mathrm{P}\Big[\tilde{z}^{(\ell)}_{(n_\ell - q)} \leq 1-\eta\Big] &\leq (e\varrho)^{\frac{n_\ell - 1}{\varrho}} p^{(n_\ell - 1)\left(1 - \frac{1}{\varrho}\right)} \\
&= \left((e\varrho)^{\frac{1}{\varrho}} p^{1-\frac{1}{\varrho}}\right)^{n_\ell - 1} \\
&\leq e^{-(n_\ell - 1)c_1}
\end{aligned}$$

where the last inequality holds for a constant $c_1 > 0$, provided that $(e\varrho)^{\frac{1}{\varrho}} p^{1-\frac{1}{\varrho}} < 1$, i.e., if $(e\varrho)^{-\frac{1}{\varrho - 1}} > p = \mathrm{P}\Big[\tilde{z}^{(\ell)}_j \leq 1-\eta\Big]$. This inequality can be satisfied for every given $p < 1$ by taking ϱ sufficiently large. Since the pdf of $\tilde{z}^{(\ell)}_j = \left|\left\langle \mathbf{a}^{(\ell)}_j, \mathbf{a}^{(\ell)}_i \right\rangle\right|$ is given by $f(z) = \frac{2}{\sqrt{\pi}} \frac{\Gamma(d_\ell/2)}{\Gamma((d_\ell - 1)/2)} (1 - z^2)^{\frac{d_\ell - 3}{2}} 1_{\{|z| \leq 1\}}$ and $\eta > 0$, we, indeed, have $p < 1$. As $\varrho = \frac{n_\ell - 1}{n^\rho_\ell - 1}$ is increasing in n_ℓ and $n_\ell \geq n_0$, by assumption, ϱ can, indeed, be made sufficiently large provided that n_0 is large enough.

A.2 Proof of Lemma 2

Our proof is inspired by ideas from [42, 33, 43] dealing with the connectivity of nearest neighbor graphs for points chosen randomly in the plane. Here, we study the connectivity of nearest neighbor graphs $\tilde{G}$ for points chosen randomly on the unit sphere $\mathbb{S}^{d-1}$. The main idea of our proof is as follows. We first partition the unit sphere into M regions $R_1, ..., R_M$ of equal area and small diameter. Then we show that, for every given point $\mathbf{a}_i$, all points in the regions neighboring the region that contains $\mathbf{a}_i$ are among the $\tilde{k}$ nearest neighbors of $\mathbf{a}_i$. Next, we show that all regions R_m contain at least one point, which combined with the fact that $R_1, ..., R_M$ is the partitioning of a contiguous area, implies that $\tilde{G}$ is connected, as desired.

We start by introducing the spherical distance metric s for points $\mathbf{x}, \mathbf{y} \in \mathbb{S}^{d-1}$ as

$$s(\mathbf{x}, \mathbf{y}) \coloneqq \arccos(\langle \mathbf{x}, \mathbf{y} \rangle)$$

and defining the spherical cap around $\mathbf{p} \in \mathbb{S}^{d-1}$ of spherical radius $\theta \in [0, \pi/2]$ as

$$C(\mathbf{p}, \theta) \coloneqq \{\mathbf{x} \in \mathbb{S}^{d-1} \colon s(\mathbf{x}, \mathbf{p}) \leq \theta\}.$$

The distance metrics s and $\tilde{s}$ are related according to

$$\begin{aligned}
\tilde{s}(\mathbf{x}, \mathbf{y}) &= \arccos(|\langle \mathbf{x}, \mathbf{y} \rangle|) \\
&= \min(\arccos(\langle \mathbf{x}, \mathbf{y} \rangle), \arccos(-\langle \mathbf{x}, \mathbf{y} \rangle)) \\
&= \min(s(\mathbf{x}, \mathbf{y}), s(-\mathbf{x}, \mathbf{y})). \qquad (24)
\end{aligned}$$

In the following, whenever we refer to the points in a region Q, we actually mean the points in $\{\mathbf{a}_1, ..., \mathbf{a}_n\}$ that lie in Q, i.e., $\{\mathbf{a}_1, ..., \mathbf{a}_n\} \cap Q$. We denote by $\#(Q)$ the number of points in Q, and by $N(C(\mathbf{a}_i, \theta))$ the number of points in $C(\mathbf{a}_i, \theta)$, excluding $\mathbf{a}_i$, i.e., $N(C(\mathbf{a}_i, \theta)) \coloneqq |C(\mathbf{a}_i, \theta) \cap \{\mathbf{a}_1, ...\mathbf{a}_{i-1}, \mathbf{a}_{i+1}, ..., \mathbf{a}_n\}|$. Note that the points contained in $C(\mathbf{a}_i, \theta) \setminus \{\mathbf{a}_i\}$ are the $N(C(\mathbf{a}_i, \theta))$ nearest neighbors of $\mathbf{a}_i$ with respect to (w.r.t.) the distance s. Later in the proof we will need the following relation between the nearest neighbors of a point $\mathbf{a}_i$ w.r.t. the distance s and the nearest neighbors of $\mathbf{a}_i$ w.r.t. the distance $\tilde{s}$: The points contained in $(C(\mathbf{a}_i, \theta) \setminus \{\mathbf{a}_i\}) \cup (C(-\mathbf{a}_i, \theta) \setminus \{-\mathbf{a}_i\})$ are the $N(C(\mathbf{a}_i, \theta)) + N(C(-\mathbf{a}_i, \theta))$ nearest neighbors of $\mathbf{a}_i$ w.r.t. the distance $\tilde{s}$. To see this, first note that by (24) every point $\mathbf{a}_j$ in $C(\mathbf{a}_i, \theta) \cup C(-\mathbf{a}_i, \theta)$ satisfies $\tilde{s}(\mathbf{a}_i, \mathbf{a}_j) \leq \theta$. Since $\theta \leq \pi/2$ the caps $C(\mathbf{a}_i, \theta)$ and $C(-\mathbf{a}_i, \theta)$ are non-overlapping so that the total number of points in $(C(\mathbf{a}_i, \theta) \setminus \{\mathbf{a}_i\}) \cup (C(-\mathbf{a}_i, \theta) \setminus \{-\mathbf{a}_i\})$ is given by $N(C(\mathbf{a}_i, \theta)) + N(C(-\mathbf{a}_i, \theta))$.

We proceed to partitioning the unit sphere $\mathbb{S}^{d-1}$ into M non-overlapping regions of equal area and small diameter. Such a partitioning was described in [44, 45] and has found applications, e.g., in theoretical computer science [46].

Lemma 3 (extracted from the proof of Lemma 6.2 in [45])**.** *For each $d > 1$, there exists a partitioning $\mathrm{FS}(d, M) = \{R_1, ..., R_M\}$ of the unit sphere $\mathbb{S}^{d-1}$ into M non-overlapping regions $R_1, ..., R_M$ of equal area, with the spherical diameter of each R_m satisfying $\sup\{s(\mathbf{x}, \mathbf{y}) \colon \mathbf{x}, \mathbf{y} \in R_m\} \leq \theta^\star$. Here,*

$$\theta^\star \coloneqq 8\Theta(\mathcal{L}(\mathbb{S}^{d-1})/M) \tag{25}$$

where $\Theta(\cdot)$ is the inverse function of $\mathcal{L}(C(\mathbf{p}, \theta))$ w.r.t. θ (recall that $\mathcal{L}(\cdot)$ denotes the Lebesgue measure, and note that $\mathcal{L}(C(\mathbf{p}, \theta))$ is independent of $\mathbf{p} \in \mathbb{S}^{d-1}$).

Let $\mathrm{FS}(d, M) = \{R_1, ..., R_M\}$ be a partition of the unit sphere according to Lemma 3. Connectivity of $\tilde{G}$ will now be established by showing that each point $\mathbf{a}_i \in R_m$ is connected to all points that lie in neighboring regions of R_m, and in addition, all regions contain at least one point. To this end, define the events $A \coloneqq \{\#(R_m) > 0, \text{ for all } m \in [M]\}$ and $B_m \coloneqq \{\#(C(\mathbf{c}_m, 3\theta^\star)) \leq k\}$ where $C(\mathbf{c}_m, 3\theta^\star)$ is the spherical cap around an arbitrary, but fixed point $\mathbf{c}_m \in R_m$, with $\theta^\star$ given by (25), and $k \coloneqq \tilde{k}/2$. We assume for expositional simplicity that $\tilde{k}$ is even (the proof applies with minor changes to general k by setting $k \coloneqq \lfloor \tilde{k}/2 \rfloor$). The proof is then effected by showing that i) on $A \cap \left(\cap_{m=1}^M B_m\right)$, $\tilde{G}$ is connected and ii) upper-bounding the probability that $A \cap \left(\cap_{m=1}^M B_m\right)$ does not hold.

By Lemma 3, the spherical cap $C(\mathbf{a}_i, 2\theta^\star)$ around a given $\mathbf{a}_i \in R_m$ contains all neighboring regions of R_m, and, since $C(\mathbf{a}_i, 2\theta^\star) \subset C(\mathbf{c}_m, 3\theta^\star)$ (see Figure 10 for an illustration), on B_m we have $N(C(\mathbf{a}_i, 2\theta^\star)) \leq k$, for all $\mathbf{a}_i \in R_m$. All points in $C(\mathbf{a}_i, 2\theta^\star) \setminus \{\mathbf{a}_i\}$ are hence among the k nearest neighbors of $\mathbf{a}_i$ w.r.t. the distance s. W.l.o.g., suppose that $-\mathbf{a}_i \in R_{m'}$. By (24), on $B_m \cap B_{m'}$ all points in $(C(\mathbf{a}_i, \theta) \setminus \{\mathbf{a}_i\}) \cup (C(-\mathbf{a}_i, \theta) \setminus \{-\mathbf{a}_i\})$ are therefore among the $\tilde{k} = 2k$ nearest neighbors of $\mathbf{a}_i$ w.r.t. the distance $\tilde{s}$ (see the paragraph below (24)). On A, each R_m contains at least one point; thus on $A \cap B_m \cap B_{m'}$, each neighboring region of R_m and $R_{m'}$ contains at least one of the $\tilde{k}$ nearest neighbors of $\mathbf{a}_i$ w.r.t. the distance $\tilde{s}$. Therefore, on $A \cap \left(\cap_{m,m'=1}^M B_m \cap B_{m'}\right) = A \cap \left(\cap_{m=1}^M B_m\right)$, each point $\mathbf{a}_i \in R_m$ is connected with all points in the neighboring regions of R_m and each region contains at least one point. As this holds for all points $\mathbf{a}_1, ..., \mathbf{a}_n$, $\tilde{G}$ is connected.

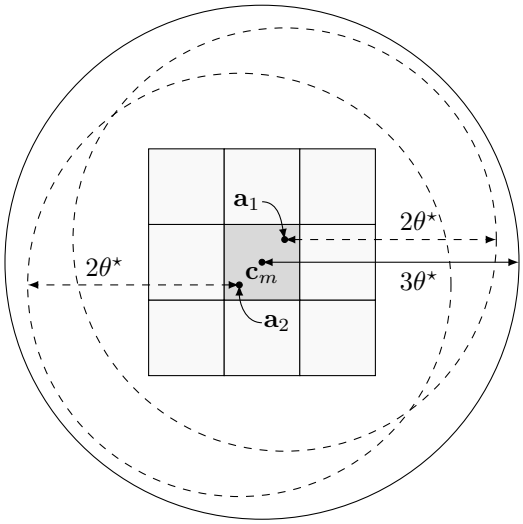

Figure 10: R_m (gray region) along with the spherical caps $C(\mathbf{a}_1, 2\theta^\star)$, $C(\mathbf{a}_2, \theta^\star)$, and $C(\mathbf{c}_m, 3\theta^\star)$.

It remains to upper-bound the probability of $\overline{A \cap \left(\cap_{m=1}^{M} B_m\right)}$. We first note that

$$\begin{aligned}
\mathrm{P}\left[\overline{A \cap \left(\cap_{m=1}^{M} B_m\right)}\right] &= \mathrm{P}\left[\bar{A} \cup \left(\bigcup_{m=1}^{M} \bar{B}_m\right)\right] \\
&\leq \mathrm{P}\left[\bar{A}\right] + \sum_{m=1}^{M} \mathrm{P}\left[\bar{B}_m\right]
\end{aligned}$$

and start by upper-bounding $\mathrm{P}\left[\bar{A}\right]$. Set

$$M = \frac{n}{\gamma \log n} \tag{26}$$

where $\gamma > 1$ is the constant in the statement of Lemma 2. Observe that

$$\begin{aligned}
\mathrm{P}\left[\bar{A}\right] &= \mathrm{P}\left[\cup_{m=1}^{M} \{\#(R_m) = 0\}\right] \leq \sum_{m=1}^{M} \mathrm{P}[\#(R_m) = 0] \\
&= \sum_{m=1}^{M} \left(1 - \frac{1}{M}\right)^n \qquad (27) \\
&\leq M e^{-n/M} = \frac{n}{\gamma \log n} e^{-\gamma \log n} = \frac{n^{1-\gamma}}{\gamma \log n} \qquad (28)
\end{aligned}$$

where in (27) we used the fact that the n points are chosen i.i.d. and the probability of a given point ending up in R_m is $1/M$.

We next upper-bound $\mathrm{P}\left[\bar{B}_m\right]$. To this end set $k = 3np$. We establish later that this choice of k satisfies $\tilde{k} = 2k \leq \gamma c_2 \log n$, for a constant c_2 depending on d only. Since the k'-nearest neighbor graph of $\{\mathbf{a}_1, ..., \mathbf{a}_n\}$ with $k' \geq \tilde{k}$ is connected if the $\tilde{k}$-nearest neighbor graph $\tilde{G}$ is connected, this will yield the desired result.

First note that $\#(C(\mathbf{c}_m, 3\theta^\star))$ is binomially distributed with parameters (n, p), where $p \coloneqq \mathcal{L}(C(\mathbf{c}_m, 3\theta^\star))/\mathcal{L}(\mathbb{S}^{d-1})$. By a tail bound on the binomial distribution [47, Thm. 1] we obtain, with $t = 2np$, that

$$\begin{aligned}\mathrm{P}\left[\bar{B}_m\right] &= \mathrm{P}[\#(C(\mathbf{c}_m, 3\theta^\star)) > np + t] \\ &\leq e^{-\frac{t^2}{2(np+t/3)}} = e^{-\frac{6}{5}np} \leq e^{-np}. \end{aligned} \tag{29}$$

Since $R_m \subset C(\mathbf{c}_m, 3\theta^\star)$, we have

$$p = \frac{\mathcal{L}(C(\mathbf{c}_m, 3\theta^\star))}{\mathcal{L}(\mathbb{S}^{d-1})} \geq \frac{\mathcal{L}(R_m)}{\mathcal{L}(\mathbb{S}^{d-1})} = \frac{1}{M} = \frac{\gamma \log n}{n}.$$

By a union bound we thus get

$$\mathrm{P}\left[\bigcup_{m=1}^{M} \bar{B}_m\right] \leq Me^{-np} \leq Me^{-n/M} = \frac{n^{1-\gamma}}{\gamma \log n}. \tag{30}$$

Combining (28) and (30) yields $\mathrm{P}\left[\bar{A} \cup \left(\bigcup_{m=1}^{M} \bar{B}_m\right)\right] \leq \frac{2}{n^{\gamma-1}\gamma \log n}$.

It remains to show that there exists a constant c_2 (depending on d only) such that $\tilde{k} = 2k \leq \gamma\, c_2 \log n$. This is accomplished by upper-bounding $\mathcal{L}(C(\mathbf{c}_m, 3\theta^\star))$ and using this upper bound to establish that $k = 3np = 3n\mathcal{L}(C(\mathbf{c}_m, 3\theta^\star))/\mathcal{L}(\mathbb{S}^{d-1}) \leq \gamma \frac{c_2}{2} \log n$. To this end, we first upper-bound $\theta^\star$ in (25) and then use this bound to upper-bound $\mathcal{L}(C(\mathbf{c}_m, 3\theta^\star))$. By [45, Eq. 5.9], we have

$$\begin{aligned}\theta^\star = 8\Theta(\mathcal{L}(\mathbb{S}^{d-1})/M) &\leq 8 \arcsin\left(\left(\frac{\mathcal{L}(\mathbb{S}^{d-1})}{\mathcal{L}(\mathbb{S}^{d-2})}\frac{d-1}{M}\right)^{\frac{1}{d-1}}\right) \\ &\leq 4\pi\left(\frac{\mathcal{L}(\mathbb{S}^{d-1})}{\mathcal{L}(\mathbb{S}^{d-2})}\frac{d-1}{M}\right)^{\frac{1}{d-1}} \end{aligned} \tag{31}$$

where we used $\arcsin(x) \leq \frac{\pi}{2}x$, for $0 \leq x \leq 1$. We next establish that the argument of arcsin in (31) is, indeed, smaller than 1. Using $\mathcal{L}(\mathbb{S}^{d-1}) = \frac{2\pi^{d/2}}{\Gamma(d/2)}$ (e.g., [45, p. 1]) and $\frac{\Gamma(\frac{d-1}{2})}{\Gamma(\frac{d}{2})} \leq \sqrt{2}\frac{\sqrt{d}}{d-1}$ (e.g., [48, Eq. 8.1]), we obtain

$$\frac{\mathcal{L}(\mathbb{S}^{d-1})}{\mathcal{L}(\mathbb{S}^{d-2})}\frac{d-1}{M} = \sqrt{\pi}\frac{\Gamma(\frac{d-1}{2})}{\Gamma(\frac{d}{2})}\frac{d-1}{M} \leq \sqrt{2\pi}\frac{\sqrt{d}}{M} = \sqrt{2\pi d}\frac{\gamma \log n}{n} \leq 6(12\pi)^{d-1}\frac{\gamma \log n}{n} \leq 1$$

where we used $\sqrt{2\pi d} \leq 6(12\pi)^{d-1}$ for $d \geq 1$, and the last inequality holds by the assumption $n \geq \tilde{k} \geq \gamma\, 6(12\pi)^{d-1} \log n$.

Application of [45, Eq. 5.2] and subsequently of (31) yields

$$\mathcal{L}(C(\mathbf{c}_m, 3\theta^\star)) \leq \frac{\mathcal{L}(\mathbb{S}^{d-2})}{d-1}(3\theta^\star)^{d-1} \leq \frac{\mathcal{L}(\mathbb{S}^{d-2})}{d-1}(12\pi)^{d-1}\frac{\mathcal{L}(\mathbb{S}^{d-1})}{\mathcal{L}(\mathbb{S}^{d-2})}\frac{d-1}{M} = (12\pi)^{d-1}\frac{\mathcal{L}(\mathbb{S}^{d-1})}{M}.$$

We thus have

$$k = 3np = 3n\,\frac{\mathcal{L}(C(\mathbf{c}_m, 3\theta^*))}{\mathcal{L}(\mathbb{S}^{d-1})} \leq 3 \cdot (12\pi)^{d-1}\,\gamma\, \log n$$

and hence $k \leq \gamma \frac{c_2}{2} \log n$ with $c_2 = 6(12\pi)^{d-1}$, as desired.

B Proof of Theorem 2 and Corollary 1

Analogously to Theorem 1, the proof of Theorem 2 is established by upper-bounding the probability $\mathrm{P}\left[\overline{\mathrm{C \text{ and } NFC}}\right]$ according to

$$\mathrm{P}\left[\overline{\mathrm{C \text{ and } NFC}}\right] \leq \mathrm{P}\left[\overline{\mathrm{NFC}}\right] + \mathrm{P}\left[\overline{\mathrm{C}}|\mathrm{NFC}\right] \tag{32}$$

where $\mathrm{NFC} = \{G \text{ has no false connections}\}$ and $\mathrm{C} = \{G(\mathcal{X}_\ell) \text{ is connected, for all } \ell\}$, as in the proof of Theorem 1. We start by upper-bounding $\mathrm{P}\left[\overline{\mathrm{NFC}}\right]$. Since $q \leq n_{\min}/6$ and (8) for $\sigma = 0$ reduces to (5) it follows from Theorem 3 (the assumption $m \geq 6 \log N$, i.e., $\sqrt{6 \log N}/\sqrt{m} = \beta/\sqrt{m} \leq 1$ relevant for Step 1 in the proof of Theorem 3 is not needed owing to $\sigma = 0$) that

$$\mathrm{P}\left[\overline{\mathrm{NFC}}\right] \leq \frac{10}{N} + \sum_{\ell \in [L]} n_\ell e^{-c(n_\ell - 1)} \tag{33}$$

where $c > 0$ is a numerical constant.

We next upper-bound $\mathrm{P}\left[\overline{\mathrm{C}}|\mathrm{NFC}\right]$. In Appendix A we established that (cf. (19) with $\gamma = 3$)

$$\mathrm{P}\left[\overline{\mathrm{C}}|\mathrm{NFC}\right] \leq \sum_{\ell \in [L]} 2{n_\ell}^{-2} \tag{34}$$

provided that $q \geq 3 \cdot 6(12\pi)^{d_\ell - 1} \log n_\ell$, for all ℓ, which is satisfied by the assumption $q \in [c_1 \log n_{\max}, n_{\min}/6]$ with $c_1 = 18(12\pi)^{d_{\max}-1}$. Using (33) and (34) in (32) finally yields

$$\mathrm{P}\left[\overline{\mathrm{C \text{ and } NFC}}\right] \leq 10/N + \sum_{\ell=1}^{L} \left(n_\ell e^{-c(n_\ell - 1)} + 2{n_\ell}^{-2} \right)$$

as desired.

Corollary 1 follows directly from (33).

C Proof of Theorem 3

As in the proof of Lemma 1, we show that G has no false connections by establishing that for each $\mathbf{x}_i^{(\ell)} \in \mathcal{X}_\ell$ the associated set $\mathcal{T}_i$ corresponds to points in $\mathcal{X}_\ell$ only. Again, this is accomplished by showing that

$$z^{(\ell)}_{(n_\ell - q)} > \max_{k \neq \ell, j} z_j^{(k)} \tag{35}$$

where $z_j^{(k)} = \left|\left\langle \mathbf{x}_j^{(k)}, \mathbf{x}_i^{(\ell)} \right\rangle\right|$. Next, we upper-bound the probability of (35) being violated. A union bound over all N vectors $\mathbf{x}_i^{(\ell)}, i \in [n_\ell], \ell \in [L]$, will, as before, yield the final result. We start by setting

$$\tilde{z}_j^{(k)} := \left|\left\langle \mathbf{a}_j^{(k)}, {\mathbf{U}^{(k)}}^T \mathbf{U}^{(\ell)} \mathbf{a}_i^{(\ell)} \right\rangle\right| \tag{36}$$

and noting that

$$z_j^{(k)} = \left|\left\langle \mathbf{a}_j^{(k)}, {\mathbf{U}^{(k)}}^T \mathbf{U}^{(\ell)} \mathbf{a}_i^{(\ell)} \right\rangle + e_j^{(k)}\right| \tag{37}$$

with

$$e_j^{(k)} := \left\langle \mathbf{e}_j^{(k)}, \mathbf{e}_i^{(\ell)} \right\rangle + \left\langle \mathbf{e}_j^{(k)}, \mathbf{U}^{(\ell)} \mathbf{a}_i^{(\ell)} \right\rangle + \left\langle \mathbf{U}^{(k)} \mathbf{a}_j^{(k)}, \mathbf{e}_i^{(\ell)} \right\rangle . \tag{38}$$

Now recall that $z_{(1)}^{(\ell)} \leq z_{(2)}^{(\ell)} \leq ... \leq z_{(n_\ell - 1)}^{(\ell)}$ are the order statistics of $\{z_j^{(\ell)}\}_{j \in [n_\ell] \setminus \{i\}}$. It follows that

$$\tilde{z}_{(n_\ell - q)}^{(\ell)} - \max_{j \neq i} |e_j^{(\ell)}| \leq z_{(n_\ell - q)}^{(\ell)}$$

and hence the probability of (35) being violated can be upper-bounded according to

$$\begin{aligned}
\mathrm{P}\left[z_{(n_\ell - q)}^{(\ell)} \leq \max_{k \neq \ell, j} z_j^{(k)} \right] &\leq \mathrm{P}\left[\tilde{z}_{(n_\ell - q)}^{(\ell)} - \max_{j \neq i} \left| e_j^{(\ell)} \right| \leq \max_{k \neq \ell, j} \tilde{z}_j^{(k)} + \max_{k \neq \ell, j} \left| e_j^{(k)} \right| \right] \\
&\leq \mathrm{P}\left[\tilde{z}_{(n_\ell - q)}^{(\ell)} \leq \frac{\nu}{\sqrt{d_\ell}} \right] \\
&\quad + \mathrm{P}\left[\alpha + 2\epsilon \leq \max_{j \neq i} \left| e_j^{(\ell)} \right| + \max_{k \neq \ell, j} \tilde{z}_j^{(k)} + \max_{k \neq \ell, j} \left| e_j^{(k)} \right| \right] \qquad (39) \\
&\leq \mathrm{P}\left[\tilde{z}_{(n_\ell - q)}^{(\ell)} \leq \frac{\nu}{\sqrt{d_\ell}} \right] + \mathrm{P}\left[\max_{k \neq \ell, j} \tilde{z}_j^{(k)} \geq \alpha \right] \\
&\quad + \underbrace{\mathrm{P}\left[\max_{j \neq i} \left| e_j^{(\ell)} \right| \geq \epsilon \right] + \mathrm{P}\left[\max_{k \neq \ell, j} \left| e_j^{(k)} \right| \geq \epsilon \right]}_{\leq \sum_{(j,k) \neq (i,\ell)} \mathrm{P}\left[\left| e_j^{(k)} \right| \geq \epsilon \right]} \qquad (40)
\end{aligned}$$

where α, ϵ, and ν are chosen later. In (39) and (40) we used that for random variables X and Y, possibly dependent, and constants ϕ and φ satisfying $\phi \geq \varphi$, we have

$$\begin{aligned}
\mathrm{P}[X \leq Y] &\leq \mathrm{P}[\{X \leq \phi\} \cup \{\varphi \leq Y\}] \\
&\leq \mathrm{P}[X \leq \phi] + \mathrm{P}[\varphi \leq Y] .
\end{aligned} \tag{41}$$

Specifically, in (39) we used (41) with $\phi = \frac{\nu}{\sqrt{d_\ell}}$ and $\varphi = \alpha + 2\epsilon$, which leads to the assumption $\alpha + 2\epsilon \leq \frac{\nu}{\sqrt{d_\ell}}$, resolved below. We next upper-bound the individual terms in (40) to get the following results proven at the end of this appendix:

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

$$\mathrm{P}\left[\left| e_j^{(k)} \right| \geq \epsilon \right] \leq 7 e^{-\frac{\beta^2}{2}} . \tag{42}$$

Step 2: Setting

$$\alpha = \frac{\beta(1+\beta)}{\sqrt{d_\ell}} \max_{k \neq \ell} \frac{1}{\sqrt{d_k}} \left\| \mathbf{U}^{(k)T} \mathbf{U}^{(\ell)} \right\|_F \tag{43}$$

we have for all $\beta \geq 0$ that

$$\mathrm{P}\left[\max_{k \neq \ell, j} \tilde{z}_j^{(k)} \geq \alpha \right] \leq \sum_{k \in [L] \setminus \{\ell\}} (1 + 2n_k) e^{-\frac{\beta^2}{2}} \leq 3N e^{-\frac{\beta^2}{2}} . \tag{44}$$

Step 3: For $\nu = 2/3$ and $n_\ell \geq 6q$, there is a constant $c = c(\nu) > 1/20$ such that

$$\mathrm{P}\left[\tilde{z}^{(\ell)}_{(n_\ell - q)} \leq \frac{\nu}{\sqrt{d_\ell}}\right] \leq e^{-c(n_\ell - 1)}. \tag{45}$$

Before presenting the detailed arguments leading to (42), (44), and (45), we show how the proof is completed. Setting $\beta = \sqrt{6 \log N}$ and using (42), (44) (note that $\beta \leq \sqrt{m}$ is satisfied since, by assumption, $m \geq 6 \log N$), and (45) in (40) yields

$$\begin{aligned}\mathrm{P}\left[z^{(\ell)}_{(n_\ell - q)} \leq \max_{k \neq \ell, j} z^{(k)}_j\right] &\leq e^{-c(n_\ell - 1)} + 3Ne^{-\frac{\beta^2}{2}} + 7Ne^{-\frac{\beta^2}{2}} \\ &= \frac{10}{N^2} + e^{-c(n_\ell - 1)}. \end{aligned} \tag{46}$$

Taking the union bound over all vectors $\mathbf{x}^{(\ell)}_i, i \in [n_\ell], \ell \in [L]$, yields the desired lower bound on G having no false connections.

Recall that for (39) we imposed the condition $\alpha + 2\epsilon \leq \frac{\nu}{\sqrt{d_\ell}}$. With our choices for ϵ, α, and ν in Steps 1, 2, and 3, respectively, this condition becomes

$$\beta(1+\beta) \max_{k \neq \ell} \frac{1}{\sqrt{d_k}} \left\| \mathbf{U}^{(k)^T} \mathbf{U}^{(\ell)} \right\|_F + 4\sigma(1+\sigma) \frac{\sqrt{d_\ell}}{\sqrt{m}} \beta \leq \frac{2}{3}. \tag{47}$$

Next, note that $(1+\beta) \leq 4\sqrt{\log N}$ as a consequence of $N \geq 6$ ($N = \sum_{\ell=1}^{L} n_\ell$, and $n_\ell \geq 6q \geq 6$, for all ℓ), by assumption. Therefore, (47) is implied by

$$\max_{k \neq \ell} \frac{1}{\sqrt{d_k}} \left\| \mathbf{U}^{(k)^T} \mathbf{U}^{(\ell)} \right\|_F + \frac{\sigma(1+\sigma)}{\sqrt{\log N}} \frac{\sqrt{d_\ell}}{\sqrt{m}} \leq \frac{2}{3 \cdot 4\sqrt{6 \log N}}$$

which, in turn, is implied by (8). This concludes the proof.

It remains to prove the bounds (42), (44), and (45).

Step 1, proof of (42): By an argument of the form (41), we get

$$\begin{aligned}\mathrm{P}\left[\left|e^{(k)}_j\right| \geq \epsilon\right] \leq \mathrm{P}&\left[\left|\left\langle \mathbf{e}^{(k)}_j, \mathbf{e}^{(\ell)}_i \right\rangle\right| \geq \frac{2\sigma^2}{\sqrt{m}} \beta\right] \\ &+ \mathrm{P}\left[\left|\left\langle \mathbf{e}^{(k)}_j, \mathbf{U}^{(\ell)} \mathbf{a}^{(\ell)}_i \right\rangle\right| \geq \frac{\sigma}{\sqrt{m}}\right] \\ &+ \mathrm{P}\left[\left|\left\langle \mathbf{U}^{(k)} \mathbf{a}^{(k)}_j, \mathbf{e}^{(\ell)}_i \right\rangle\right| \geq \frac{\sigma}{\sqrt{m}}\right]. \end{aligned} \tag{48}$$

We next upper-bound the probabilities in (48). Conditional on $\mathbf{a}^{(k)}_j$, with $\left\| \mathbf{U}^{(k)} \mathbf{a}^{(k)}_j \right\|_2 = 1$, we have $\left\langle \mathbf{U}^{(k)} \mathbf{a}^{(k)}_j, \mathbf{e}^{(\ell)}_i \right\rangle \sim \mathcal{N}(0, \sigma^2/m)$. Using Lemma 6 in Appendix G, for $\beta \geq \frac{1}{\sqrt{2\pi}}$, we hence get

$$\mathrm{P}\left[\left|\left\langle \mathbf{U}^{(k)} \mathbf{a}^{(k)}_j, \mathbf{e}^{(\ell)}_i \right\rangle\right| \geq \frac{\sigma}{\sqrt{m}} \beta\right] \leq 2e^{-\frac{\beta^2}{2}}. \tag{49}$$

Next, we upper-bound the first term on the RHS of (48). Conditional on $\mathbf{e}^{(\ell)}_i$, we have $\left\langle \mathbf{e}^{(k)}_j, \mathbf{e}^{(\ell)}_i \right\rangle \sim \mathcal{N}\left(0, \frac{\sigma^2}{m} \left\| \mathbf{e}^{(\ell)}_i \right\|_2^2\right)$. Lemma 6 yields, for $\beta \geq \frac{1}{\sqrt{2\pi}}$, that

$$\mathrm{P}\left[\left|\left\langle \mathbf{e}^{(k)}_j, \mathbf{e}^{(\ell)}_i \right\rangle\right| \geq \beta \frac{\sigma}{\sqrt{m}} \left\| \mathbf{e}^{(\ell)}_i \right\|_2\right] \leq 2e^{-\frac{\beta^2}{2}}. \tag{50}$$

Since $\beta = \sqrt{6 \log N} \leq \sqrt{m}$, by assumption, we get

$$\mathrm{P}\left[\left\|\mathbf{e}_i^{(\ell)}\right\|_2 \geq 2\sigma\right] \leq \mathrm{P}\left[\left\|\mathbf{e}_i^{(\ell)}\right\|_2 \geq \left(1 + \frac{\beta}{\sqrt{m}}\right)\sigma\right] \leq e^{-\frac{\beta^2}{2}} \tag{51}$$

where the second inequality follows from (90). Next, note that for random variables X, Y, possibly dependent, and a constant ϕ, we have

$$\begin{aligned}\mathrm{P}[X \geq \phi] &= \mathrm{P}[\{X \geq Y \geq \phi\} \cup \{X \geq \phi \geq Y\} \cup \{Y \geq X \geq \phi\}] \\ &\leq \mathrm{P}[\{X \geq Y\} \cup \{Y \geq \phi\}] \\ &\leq \mathrm{P}[X \geq Y] + \mathrm{P}[Y \geq \phi]\,.\end{aligned} \tag{52}$$

Combining (50) and (51) via (52) yields

$$\mathrm{P}\Big[\underbrace{\left|\left\langle \mathbf{e}_j^{(k)}, \mathbf{e}_i^{(\ell)}\right\rangle\right|}_{X} \geq \underbrace{\frac{2\sigma^2}{\sqrt{m}}\beta}_{\phi}\Big] \leq \mathrm{P}\Big[\left|\left\langle \mathbf{e}_j^{(k)}, \mathbf{e}_i^{(\ell)}\right\rangle\right| \geq \underbrace{\beta\frac{\sigma}{\sqrt{m}}\left\|\mathbf{e}_i^{(\ell)}\right\|_2}_{Y}\Big] + \mathrm{P}\left[\left\|\mathbf{e}_i^{(\ell)}\right\|_2 \geq 2\sigma\right] \leq 3e^{-\frac{\beta^2}{2}}. \tag{53}$$

Finally, using (49) and (53) in (48) gives the desired result (42).

Step 2, proof of (44): We first upper-bound the probability of $\max_j \tilde{z}_j^{(k)}$, for a given k, to exceed a constant, which then yields, via a union bound over k, an upper bound on the probability of $\max_{k \neq \ell, j} \tilde{z}_j^{(k)}$ exceeding a constant. For convenience, we set $\mathbf{B} \coloneqq \mathbf{U}^{(k)^T}\mathbf{U}^{(\ell)}$ so that $\tilde{z}_j^{(k)} = \left|\left\langle \mathbf{a}_j^{(k)}, \mathbf{B}\mathbf{a}_i^{(\ell)}\right\rangle\right|$. We start by noting [15, Proof of Lem. 7.5] that

$$\mathrm{P}\left[\left\|\mathbf{B}\mathbf{a}_i^{(\ell)}\right\|_2 \geq \frac{\|\mathbf{B}\|_F}{\sqrt{d_\ell}} + \kappa\right] \leq e^{-d_\ell \frac{\kappa^2}{2\|\mathbf{B}\|_{2\to 2}^2}}. \tag{54}$$

Setting $\kappa = \beta\|\mathbf{B}\|_F/\sqrt{d_\ell}$ in (54) yields

$$\mathrm{P}\left[\left\|\mathbf{B}\mathbf{a}_i^{(\ell)}\right\|_2 \geq \frac{1+\beta}{\sqrt{d_\ell}}\|\mathbf{B}\|_F\right] \leq e^{-\frac{\beta^2}{2}\frac{\|\mathbf{B}\|_F^2}{\|\mathbf{B}\|_{2\to 2}^2}} \leq e^{-\frac{\beta^2}{2}}. \tag{55}$$

By Proposition 1 in Appendix G, we have

$$\mathrm{P}\left[\left|\left\langle \mathbf{a}_j^{(k)}, \mathbf{B}\mathbf{a}_i^{(\ell)}\right\rangle\right| > \frac{\beta}{\sqrt{d_k}}\left\|\mathbf{B}\mathbf{a}_i^{(\ell)}\right\|_2\right] \leq 2e^{-\frac{\beta^2}{2}}. \tag{56}$$

Now, using (52) with $X = \max_j \tilde{z}_j^{(k)}$, $\phi = \frac{\beta}{\sqrt{d_k}}\frac{1+\beta}{\sqrt{d_\ell}}\|\mathbf{B}\|_F$, and $Y = \frac{\beta}{\sqrt{d_k}}\left\|\mathbf{B}\mathbf{a}_i^{(\ell)}\right\|_2$, we get

$$\begin{aligned}\mathrm{P}\left[\max_j \tilde{z}_j^{(k)} \geq \frac{\beta}{\sqrt{d_k}}\frac{1+\beta}{\sqrt{d_\ell}}\|\mathbf{B}\|_F\right] &\leq \mathrm{P}\left[\max_j \left|\left\langle \mathbf{a}_j^{(k)}, \mathbf{B}\mathbf{a}_i^{(\ell)}\right\rangle\right| \geq \frac{\beta}{\sqrt{d_k}}\left\|\mathbf{B}\mathbf{a}_i^{(\ell)}\right\|_2\right] \\ &\quad + \mathrm{P}\left[\left\|\mathbf{B}\mathbf{a}_i^{(\ell)}\right\|_2 \geq \frac{1+\beta}{\sqrt{d_\ell}}\|\mathbf{B}\|_F\right] \\ &\leq \sum_{j\in[n_k]} \mathrm{P}\left[\left|\left\langle \mathbf{a}_j^{(k)}, \mathbf{B}\mathbf{a}_i^{(\ell)}\right\rangle\right| \geq \frac{\beta}{\sqrt{d_k}}\left\|\mathbf{B}\mathbf{a}_i^{(\ell)}\right\|_2\right] \\ &\quad + \mathrm{P}\left[\left\|\mathbf{B}\mathbf{a}_i^{(\ell)}\right\|_2 \geq \frac{1+\beta}{\sqrt{d_\ell}}\|\mathbf{B}\|_F\right] \qquad (57) \\ &\leq (1+2n_k)e^{-\frac{\beta^2}{2}} \qquad (58)\end{aligned}$$

where a union bound is used to obtain (57), and (58) follows from (55) and (56). Taking the union bound over $k \in [L]\backslash\{\ell\}$ concludes the proof of (44).

Step 3, proof of (45): We first note that the pdf of $\tilde{z}_j^{(\ell)} = \langle \mathbf{a}_j^{(\ell)}, \mathbf{a}_i^{(\ell)} \rangle$ is given by $f(z) = \frac{1}{\sqrt{\pi}} \frac{\Gamma(d_\ell/2)}{\Gamma((d_\ell-1)/2)} (1-z^2)^{\frac{d_\ell-3}{2}} 1_{\{|z|\leq 1\}}$. Hence, we get

$$\begin{aligned} \mathrm{P}\left[\tilde{z}_j^{(\ell)} \leq \frac{\nu}{\sqrt{d_\ell}}\right] &\leq \frac{2}{\sqrt{\pi}} \frac{\Gamma(d_\ell/2)}{\Gamma((d_\ell-1)/2)} \int_0^{\frac{\nu}{\sqrt{d_\ell}}} (1-z^2)^{\frac{d_\ell-3}{2}} 1_{\{z\leq 1\}} dz \\ &\leq \frac{2}{\sqrt{\pi}} \frac{\Gamma(d_\ell/2)}{\Gamma((d_\ell-1)/2)} \frac{\nu}{\sqrt{d_\ell}} \leq \underbrace{\sqrt{\frac{2}{\pi}}\,\nu}_{p_\nu :=} \end{aligned} \tag{59}$$

where the last inequality follows from [48, Eq. 8.1]. Next, observe that,

$$\begin{aligned} \mathrm{P}\left[\tilde{z}_{(n_\ell-q)}^{(\ell)} \leq \frac{\nu}{\sqrt{d_\ell}}\right] &= \mathrm{P}\big[\text{there exists a set } I \subset [n_\ell]\backslash\{i\} \text{ with} \\ &\qquad |I| = n_\ell - q \text{ such that } \tilde{z}_j^{(\ell)} \leq \frac{\nu}{\sqrt{d_\ell}} \text{ for all } j \in I\big] \\ &\leq \binom{n_\ell - 1}{n_\ell - q} \max_{I\colon |I|=n_\ell-q} \mathrm{P}\left[\tilde{z}_j^{(\ell)} \leq \frac{\nu}{\sqrt{d_\ell}}, \text{ for all } j \in I\right] &(60) \\ &\leq \left(e\frac{n_\ell - 1}{q-1}\right)^{q-1} \left(\mathrm{P}\left[\tilde{z}_j^{(\ell)} \leq \frac{\nu}{\sqrt{d_\ell}}\right]\right)^{n_\ell - q} &(61) \\ &\leq \left(e\frac{n_\ell - 1}{q-1}\right)^{q-1} p_\nu^{n_\ell - q} = (e\varrho)^{\frac{n_\ell-1}{\varrho}} p_\nu^{(n_\ell-1)\left(1-\frac{1}{\varrho}\right)} &(62) \\ &= \exp\left(-(n_\ell - 1)\underbrace{\left(\log\left(\frac{1}{p_\nu}\right)\left(1-\frac{1}{\varrho}\right) - \frac{1}{\varrho}\log(e\varrho)\right)}_{c(\varrho,\nu) :=}\right) &(63) \end{aligned}$$

where we used a union bound to get (60), $\binom{n}{n-k} = \binom{n}{k} \leq \left(\frac{en}{k}\right)^k$ [41] and the fact that the $\tilde{z}_j^{(\ell)}$ are i.i.d. for (61), and (59) yields (62); we also set $\varrho := \frac{n_\ell - 1}{q-1}$ for notational convenience. Here, $c(\varrho, \nu)$ satisfies $c(\varrho, \nu) > 1/20$ for $\nu = 2/3$ and $\varrho \geq 6$, as desired. Note that $\varrho = \frac{n_\ell-1}{q-1} \geq \frac{n_\ell}{q} \geq 6$, where both inequalities follow from $n_\ell \geq 6q$, for all ℓ, which holds by assumption.

D Proof of Theorem 4

Theorem 4 follows from Lemma 1 by establishing that the clustering condition (16) is satisfied for i.i.d. Gaussian random matrices $\mathbf{U}^{(\ell)} \in \mathbb{R}^{m\times d}$ with high probability. This is accomplished via the following lemma, which shows that certain submatrices of Gaussian matrices $\mathbf{U}^{(\ell)} \in \mathbb{R}^{m\times d}$ are approximately pairwise orthogonal, as long as m is sufficiently large relative to d.

Lemma 4. *Let the entries of the $\mathbf{U}^{(\ell)} \in \mathbb{R}^{m\times d}$, $\ell \in [L]$, be i.i.d. $\mathcal{N}(0, 1/m)$, and let $\mathbf{U}_{\mathcal{D}}^{(\ell)} \in \mathbb{R}^{m\times d}$ be the matrix obtained from $\mathbf{U}^{(\ell)} \in \mathbb{R}^{m\times d}$ by setting the rows with indices in $\mathcal{D} \subseteq [m]$ to zero. Then, we have for $\delta \in (0,1)$ with probability at least $1 - 4e^{-c'm}$, where c' is a numerical constant, that*

$$\min_{\ell,\,\mathcal{D}\colon |\mathcal{D}|\leq 2s,\, \|\mathbf{a}\|_2=1} \left\| \mathbf{U}_{\mathcal{D}}^{(\ell)^T} \mathbf{U}^{(\ell)} \mathbf{a} \right\|_2 \geq (1-\delta)\frac{m-2s}{m} \tag{64}$$

and

$$\max_{k,\ell:\ k\neq\ell,\,\mathcal{D}:\ |\mathcal{D}|\leq 2s} \left\| {\mathbf{U}_{\mathcal{D}}^{(k)}}^T \mathbf{U}^{(\ell)} \right\|_{2\to 2} \leq \delta \tag{65}$$

provided that

$$m \geq \frac{c_2}{\delta^2}\left(3d + \log L + s\log\left(\frac{me}{2s}\right)\right) + c_3 s \tag{66}$$

where $c_2, c_3 > 0$ are numerical constants.

Before proving Lemma 4, we show how the proof of Theorem 4 can be completed. Set $\delta = \frac{c_1}{2}\frac{c_3-2}{c_3}$, where c_3 is the constant in (9) and c_1 is a constant satisfying $c_1 < 1$. With this choice of δ, (9) (with $c_4 = c_2/\delta^2 = c_2\left(2c_3/(c_1(c_3-2))\right)^2$) implies (66) and hence, by Lemma 4, with probability $\geq 1 - 4e^{-c'm}$, we have

$$\frac{\max_{k,\ell:\ k\neq\ell,\mathcal{D}:\ |\mathcal{D}|\leq 2s} \left\| {\mathbf{U}_{\mathcal{D}}^{(k)}}^T \mathbf{U}^{(\ell)} \right\|_{2\to 2}}{\min_{l,\mathcal{D}:\ |\mathcal{D}|\leq 2s, \|\mathbf{a}\|_2=1} \left\| {\mathbf{U}_{\mathcal{D}}^{(\ell)}}^T \mathbf{U}^{(\ell)}\mathbf{a} \right\|_2} \leq \frac{\delta}{(1-\delta)\frac{m-2s}{m}} \leq 2\delta\frac{m}{m-2s} \leq 2\delta\frac{c_3}{c_3-2} = c_1 < 1 \tag{67}$$

where we used $\delta \leq \frac{1}{2}$ ($c_1 < 1$, by assumption, implies $\delta = \frac{c_1}{2}\frac{c_3-2}{c_3} \leq \frac{1}{2}$), and $\frac{m}{m-2s} \leq \frac{c_3}{c_3-2}$ as a consequence of $m \geq c_3 s$ (from (9)) and $c_3 > 2$ (c_3 can be chosen freely as long as $c_3 > 0$). We therefore established that (16) holds with probability $\geq 1 - 4e^{-c'm}$, and application of Lemma 1 concludes the proof.

Proof of Lemma 4. The proof relies on the following result from [48], which builds on a covering argument and the concentration inequality the Johnson-Lindenstrauss Lemma [49] is based on.

Lemma 5 ([48, Eq. 9.12 with $\rho = \frac{2}{e^3-1}$])**.** *Let $\mathbf{U}$ be a $p \times s$ random matrix satisfying, for some $\tilde{c} > 0$, for every $\mathbf{x} \in \mathbb{R}^s$, and for every $t \in (0,1)$,*

$$\mathrm{P}\left[\left|\|\mathbf{U}\mathbf{x}\|_2^2 - \|\mathbf{x}\|_2^2\right| \geq t\|\mathbf{x}\|_2^2\right] \leq 2e^{-\tilde{c}t^2 p}. \tag{68}$$

Then, we have

$$\mathrm{P}\left[\left\|\mathbf{U}^T\mathbf{U} - \mathbf{I}_s\right\|_{2\to 2} \geq \delta\right] \leq 2e^{-0.6\tilde{c}\delta^2 p + 3s}. \tag{69}$$

We note that (68) is satisfied, inter alia, for random matrices with i.i.d. $\mathcal{N}(0, 1/p)$ entries.

We show below that (65) and (64) hold individually with probability $\geq 1 - 2e^{-c'm}$. By a union bound, (65) and (64) thus hold simultaneously with probability $\geq 1 - 4e^{-c'm}$, as desired. We start with (65). First, note that since the rows of $\mathbf{U}_{\mathcal{D}}^{(k)}$ indexed by $\mathcal{D}$ have all entries equal to zero by definition, we have ${\mathbf{U}_{\mathcal{D}}^{(k)}}^T \mathbf{U}^{(\ell)} = \mathbf{V}_i^T\mathbf{V}_j$, where $\mathbf{V}_i \in \mathbb{R}^{p\times d}$ and $\mathbf{V}_j \in \mathbb{R}^{p\times d}$, with $p = m - |\mathcal{D}|$, denote the restrictions of $\mathbf{U}^{(k)}$ and $\mathbf{U}^{(\ell)}$, respectively, to the rows indexed by $[m]\backslash\mathcal{D}$. Set $\tilde{\mathbf{V}}_i = \sqrt{m/p}\mathbf{V}_i$, let $\mathbf{U} = [\tilde{\mathbf{V}}_i\ \tilde{\mathbf{V}}_j] \in \mathbb{R}^{p\times 2d}$, and note that the entries of $\mathbf{U}$ are i.i.d. $\mathcal{N}(0, 1/p)$. Using $m \geq p$, we have

$$\begin{aligned}\left\|\mathbf{V}_i^T\mathbf{V}_j\right\|_{2\to 2} &\leq \frac{m}{p}\left\|\mathbf{V}_i^T\mathbf{V}_j\right\|_{2\to 2} \\ &= \left\|\tilde{\mathbf{V}}_i^T\tilde{\mathbf{V}}_j\right\|_{2\to 2} \leq \left\|\mathbf{U}^T\mathbf{U} - \mathbf{I}_{2d}\right\|_{2\to 2}\end{aligned}$$

where the last inequality follows from the fact that $\tilde{\mathbf{V}}_i^T\tilde{\mathbf{V}}_j$ is a principal submatrix of $\mathbf{U}^T\mathbf{U}-\mathbf{I}_{2d}$ [50, Cor. 8.1.20]. Therefore, we get

$$\begin{aligned}\mathrm{P}\left[\left\|\mathbf{V}_i^T\mathbf{V}_j\right\|_{2\to 2}\geq\delta\right] &\leq \mathrm{P}\left[\left\|\mathbf{U}^T\mathbf{U}-\mathbf{I}_{2d}\right\|_{2\to 2}\geq\delta\right]\\ &\leq 2e^{-c_0\delta^2 p+6d} \qquad (70)\\ &\leq 2e^{-c_0\delta^2(m-2s)+6d} \qquad (71)\end{aligned}$$

where $c_0=0.6\tilde{c}$ ($\tilde{c}$ is the constant in Lemma 5), (70) follows from Lemma 5, and (71) 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 $k,\ell\in[L]$ and for each of those pairs (k,ℓ) over all $\mathcal{D}\subseteq[m]$ with $|\mathcal{D}|=2s$, i.e., over $\binom{m}{2s}\leq\left(\frac{me}{2s}\right)^{2s}$ sets, we obtain

$$\begin{aligned}\mathrm{P}\left[\max_{i\neq j}\left\|\mathbf{V}_i{}^T\mathbf{V}_j\right\|_{2\to 2}\geq\delta\right] &\leq L^2\left(\frac{me}{2s}\right)^{2s}2e^{-c_0\delta^2(m-2s)+6d} \qquad (72)\\ &= 2e^{-c_0\delta^2(m-2s)+6d+2\log L+2s\log\left(\frac{me}{2s}\right)}\\ &= 2e^{-c_0\delta^2\left[m-2s-\frac{2}{c_0\delta^2}\left(3d+\log L+s\log\left(\frac{me}{2s}\right)\right)\right]}\\ &\leq 2e^{-c_0\delta^2\left[m-2s-\frac{2}{c_0c_2}(m-c_3 s)\right]} \qquad (73)\\ &= 2e^{-c_0\delta^2\left[\left(1-\frac{2}{c_0c_2}\right)m+2s\left(\frac{c_3}{c_0c_2}-1\right)\right]}\\ &\leq 2e^{-c'm} \qquad (74)\end{aligned}$$

where we used (66) for (73), and (74) holds with $c'=c_0\delta^2\left(1-\frac{2}{c_0c_2}\right)$ provided that $c_3\geq c_0c_2$, which, in turn, is guaranteed by choosing c_3 sufficiently large (recall that c_3 can be chosen freely). Note that $c'=c_0\delta^2\left(1-\frac{2}{c_0c_2}\right)>0$ provided that $c_0c_2>2$, which holds if c_2 is chosen sufficiently large. This concludes the proof of (65) holding with probability $\geq 1-2e^{-c'm}$.

It remains to show that (64) holds with probability $\geq 1-2e^{-c'm}$. Applying Lemma 5 to $\tilde{\mathbf{V}}_i$ (recall that the entries of $\mathbf{V}_i$ are i.i.d. $\mathcal{N}(0,1/p)$), we get

$$\mathrm{P}\left[\left\|\tilde{\mathbf{V}}_i^T\tilde{\mathbf{V}}_i-\mathbf{I}_d\right\|_{2\to 2}\geq\delta\right]\leq 2e^{-c_0\delta^2 p+3d}.$$

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

$$\begin{aligned}\mathrm{P}\left[\max_i\left\|\tilde{\mathbf{V}}_i^T\tilde{\mathbf{V}}_i-\mathbf{I}_d\right\|_{2\to 2}\geq\delta\right] &\leq L\left(\frac{me}{2s}\right)^{2s}2e^{-c_0\delta^2 p+3d} \qquad (75)\\ &\leq 2e^{-c'm} \qquad (76)\end{aligned}$$

where we used the fact that the RHS of (75) is smaller than the RHS of (72) (recall that $p\geq m-2s$) and therefore (76) follows from (74). Next, note that for every $\mathbf{a}\in\mathbb{R}^d$, we have

$$\begin{aligned}\|\mathbf{a}\|_2^2-\left\|\tilde{\mathbf{V}}_i\mathbf{a}\right\|_2^2 &= \left\langle(\mathbf{I}_d-\tilde{\mathbf{V}}_i^T\tilde{\mathbf{V}}_i)\mathbf{a},\mathbf{a}\right\rangle\\ &\leq \left\|(\mathbf{I}_d-\tilde{\mathbf{V}}_i^T\tilde{\mathbf{V}}_i)\mathbf{a}\right\|_2\|\mathbf{a}\|_2\leq\left\|\tilde{\mathbf{V}}_i^T\tilde{\mathbf{V}}_i-\mathbf{I}_d\right\|_{2\to 2}\|\mathbf{a}\|_2^2.\end{aligned}$$

It follows that

$$1-\left\|\tilde{\mathbf{V}}_i^T\tilde{\mathbf{V}}_i-\mathbf{I}_d\right\|_{2\to 2}\leq\min_{\|\mathbf{a}\|_2=1}\left\|\tilde{\mathbf{V}}_i\mathbf{a}\right\|_2^2=\min_{\|\mathbf{a}\|_2=1}\left\|\tilde{\mathbf{V}}_i^T\tilde{\mathbf{V}}_i\mathbf{a}\right\|_2$$

and therefore (recall that $\tilde{\mathbf{V}}_i = \sqrt{m/p}\mathbf{V}_i$)

$$\begin{aligned}\min_{i,\,\|\mathbf{a}\|_2=1} \left\|\mathbf{V}_i^T\mathbf{V}_i\mathbf{a}\right\|_2 &= \frac{p}{m}\min_{i,\|\mathbf{a}\|_2=1}\left\|\tilde{\mathbf{V}}_i^T\tilde{\mathbf{V}}_i\mathbf{a}\right\|_2 \\ &\geq \frac{p}{m}\left(1-\max_i\left\|\tilde{\mathbf{V}}_i^T\tilde{\mathbf{V}}_i-\mathbf{I}_d\right\|_{2\to 2}\right). \end{aligned} \tag{77}$$

From (77), and $p \geq m-2s$, we get

$$\begin{aligned}\mathrm{P}\left[\min_{i,\|\mathbf{a}\|_2=1}\left\|\mathbf{V}_i^T\mathbf{V}_i\mathbf{a}\right\|_2 \leq (1-\delta)\frac{m-2s}{m}\right] &\leq \mathrm{P}\left[\min_{i,\|\mathbf{a}\|_2=1}\left\|\mathbf{V}_i^T\mathbf{V}_i\mathbf{a}\right\|_2 \leq (1-\delta)\frac{p}{m}\right] \\ &\leq \mathrm{P}\left[\max_i\left\|\tilde{\mathbf{V}}_i^T\tilde{\mathbf{V}}_i-\mathbf{I}_d\right\|_{2\to 2}\geq\delta\right] \leq 2e^{-c'm}\end{aligned}$$

where the last inequality follows by application of (76). □

E Proof of Theorem 5

The proof consists of two parts, corresponding to the two statements in Theorem 5. First, we bound the probability of the outlier detection scheme failing to detect one or more outliers, and then we bound the probability of one or more of the inliers being misclassified as an outlier.

We start by bounding the probability of the outlier detection scheme failing to detect a given outlier. A union bound over all N_0 outliers will then yield a bound on the probability of the outlier detection scheme failing to detect one or more outliers. Let $\mathbf{x}_j$ be an outlier. The probability of (10) with $c = \sqrt{6}$ being violated for $\mathbf{x}_j$, and therefore $\mathbf{x}_j$ being misclassified as an inlier, can be upper-bounded as

$$\begin{aligned}\mathrm{P}\left[\max_{i\in[N]\setminus\{j\}}|\langle\mathbf{x}_i,\mathbf{x}_j\rangle| > \frac{\sqrt{6\log N}}{\sqrt{m}}\right] &\leq \sum_{i\in[N]\setminus\{j\}}\mathrm{P}\left[|\langle\mathbf{x}_i,\mathbf{x}_j\rangle| > \frac{\sqrt{6\log N}}{\sqrt{m}}\|\mathbf{x}_i\|_2\right] \\ &\leq 2Ne^{-3\log N} = \frac{2}{N^2}\end{aligned} \tag{78}$$

where we used a union bound and $\|\mathbf{x}_i\|_2 = 1$ in the first inequality and Proposition 1 in Appendix G 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 of the outlier detection scheme misclassifying a given inlier $\mathbf{x}_j \in \mathcal{X}_\ell$ as an outlier. A union bound over all n_ℓ inliers in $\mathcal{X}_\ell$ will then complete the proof. For an inlier $\mathbf{x}_j \in \mathcal{X}_\ell$, we have

$$\max_{i\in[N]\setminus\{j\}}|\langle\mathbf{x}_i,\mathbf{x}_j\rangle| \geq \max_{i\in[n_\ell]\setminus\{j\}}\left|\left\langle\mathbf{x}_i^{(\ell)},\mathbf{x}_j^{(\ell)}\right\rangle\right| = \max_{i\in[n_\ell]\setminus\{j\}}\left|\left\langle\mathbf{a}_i^{(\ell)},\mathbf{a}_j^{(\ell)}\right\rangle\right|.$$

Using (11), i.e., $\sqrt{6\log N}/\sqrt{m} \leq 1/\sqrt{d_{\max}} \leq 1/\sqrt{d_\ell}$, the probability of (10) holding can then be

upper-bounded as

$$\begin{aligned}
\mathrm{P}\left[\max_{i\in[N]\setminus\{j\}} |\langle \mathbf{x}_i, \mathbf{x}_j\rangle| \le \frac{\sqrt{6\log N}}{\sqrt{m}}\right] &\le \mathrm{P}\left[\max_{i\in[n_\ell]\setminus\{j\}} \left|\left\langle \mathbf{a}_i^{(\ell)}, \mathbf{a}_j^{(\ell)}\right\rangle\right| \le \frac{1}{\sqrt{d_\ell}}\right] \\
&= \prod_{i\in[n_\ell]\setminus\{j\}} \mathrm{P}\left[\left|\left\langle \mathbf{a}_i^{(\ell)}, \mathbf{a}_j^{(\ell)}\right\rangle\right| \le \frac{1}{\sqrt{d_\ell}}\right] \\
&\le \prod_{i\in[n_\ell]\setminus\{j\}} \sqrt{\frac{2}{\pi}} = e^{-\frac{1}{2}\log\left(\frac{\pi}{2}\right)(n_\ell-1)} \qquad (79)
\end{aligned}$$

where (79) follows from (59) with $\nu = 1$. Taking the union bound over all inliers in $\mathcal{X}_\ell$ yields the desired upper bound $n_\ell e^{-\frac{1}{2}\log\left(\frac{\pi}{2}\right)(n_\ell-1)}$ on the outlier detection scheme misclassifying one or more of the inliers in $\mathcal{X}_\ell$ as an outlier.

F Proof of Theorem 6

The basic structure of the proof is the same as that of the proof of Theorem 5. The individual steps are, however, a bit more technical, owing to the additive noise term.

We start by bounding the probability of the outlier detection scheme failing to detect a given outlier. A union bound over all N_0 outliers will, as before, yield the desired result. Let $\mathbf{x}_j$ be an outlier and set $\beta = \sqrt{6\log N}$. The probability that (10) with $c = 2.3\sqrt{6}$ is violated for $\mathbf{x}_j$ can be upper-bounded as

$$\begin{aligned}
\mathrm{P}\left[\max_{i\in[N]\setminus\{j\}} |\langle \mathbf{x}_i, \mathbf{x}_j\rangle| > \frac{2.3\beta}{\sqrt{m}}\right] &\le \sum_{i\in[N]\setminus\{j\}} \mathrm{P}\left[|\langle \mathbf{x}_i, \mathbf{x}_j\rangle| > \frac{2.3\beta}{\sqrt{m}}\right] \\
&\le \sum_{i\in[N]\setminus\{j\}} \left(\mathrm{P}\left[|\langle \mathbf{x}_i, \mathbf{x}_j\rangle| \ge \frac{\beta}{\sqrt{m}}\|\mathbf{x}_i\|_2\right] + \mathrm{P}[\|\mathbf{x}_i\|_2 \ge 2.3]\right) \qquad (80)
\end{aligned}$$

where we applied (52) with $X = |\langle \mathbf{x}_i, \mathbf{x}_j\rangle|$, $Y = \frac{\beta}{\sqrt{m}}\|\mathbf{x}_i\|_2$, and $\phi = \frac{2.3\beta}{\sqrt{m}}$ to get (80). We next bound the first term in the sum in (80). Since $\mathbf{x}_j \sim \mathcal{N}(\mathbf{0}, (1/m)\mathbf{I}_m)$, we have that, conditioned on $\mathbf{x}_i$, $\langle \mathbf{x}_j, \mathbf{x}_i\rangle \sim \mathcal{N}(0, \|\mathbf{x}_i\|_2^2/m)$. Hence, with $\beta = \sqrt{6\log N}$, it follows from (89) that

$$\mathrm{P}\left[|\langle \mathbf{x}_i, \mathbf{x}_j\rangle| \ge \frac{\beta}{\sqrt{m}}\|\mathbf{x}_i\|_2\right] \le 2e^{-\frac{\beta^2}{2}} = \frac{2}{N^3}. \qquad (81)$$

We next bound the second term in the sum in (80) and treat the cases where $\mathbf{x}_i$ is an inlier and where it is an outlier separately. First, suppose that $\mathbf{x}_i$ is an inlier. Since $\frac{1+2\sigma}{\sqrt{1+\sigma^2}} \le 2.3$ for $\sigma \ge 0$, we have

$$\begin{aligned}
\mathrm{P}[\|\mathbf{x}_i\|_2 \ge 2.3] &\le \mathrm{P}\left[\frac{1}{\sqrt{1+\sigma^2}}\left\|\mathbf{U}^{(\ell)}\mathbf{a}_i^{(\ell)} + \mathbf{e}_i^{(\ell)}\right\|_2 \ge \frac{1+2\sigma}{\sqrt{1+\sigma^2}}\right] \\
&\le \mathrm{P}\left[\left\|\mathbf{U}^{(\ell)}\mathbf{a}_i^{(\ell)}\right\|_2 + \left\|\mathbf{e}_i^{(\ell)}\right\|_2 \ge 1+2\sigma\right] \qquad (82) \\
&= \mathrm{P}\left[\left\|\mathbf{e}_i^{(\ell)}\right\|_2 \ge 2\sigma\right] \\
&\le \frac{1}{N^3} \qquad (83)
\end{aligned}$$

where (82) follows from the triangle inequality, and for (83) we applied (51) with $\beta = \sqrt{6 \log N}$ and used that $\beta \leq \sqrt{m}$, by assumption.

Next, suppose that $\mathbf{x}_i$ is an outlier. Applying (51) with $\sigma = 1$ (again using that $\beta \leq \sqrt{m}$, by assumption), we have

$$\mathrm{P}[\|\mathbf{x}_i\|_2 \geq 2.3] \leq \frac{1}{N^3}. \tag{84}$$

Finally, combining (81), (83) (for $\mathbf{x}_i$ an inlier), and (84) (for $\mathbf{x}_i$ an outlier) in (80) yields

$$\mathrm{P}\left[\max_{i \in [N] \setminus \{j\}} |\langle \mathbf{x}_i, \mathbf{x}_j \rangle| > \frac{2.3\sqrt{6 \log N}}{\sqrt{m}}\right] \leq \sum_{i \in [N] \setminus \{j\}} \left(\frac{2}{N^3} + \frac{1}{N^3}\right) \leq \frac{3}{N^2}.$$

Taking the union bound over all N_0 outliers yields the desired result.

Next, we bound the probability of our outlier detection scheme misclassifying a given inlier as an outlier. Consider the inlier $\mathbf{x}_j \in \mathcal{X}_\ell$. Then, we have

$$\begin{aligned}
\max_{i \in [N] \setminus \{j\}} |\langle \mathbf{x}_i, \mathbf{x}_j \rangle| &\geq \max_{i \in [n_\ell] \setminus \{j\}} \left|\left\langle \mathbf{x}_i^{(\ell)}, \mathbf{x}_j^{(\ell)} \right\rangle\right| \\
&\geq \max_{i \in [n_\ell] \setminus \{j\}} \frac{1}{1+\sigma^2} \left|\left\langle \mathbf{a}_i^{(\ell)}, \mathbf{a}_j^{(\ell)} \right\rangle + e_i^{(\ell)}\right| \\
&\geq \max_{i \in [n_\ell] \setminus \{j\}} \frac{1}{1+\sigma^2} (\tilde{z}_i^{(\ell)} - |e_i^{(\ell)}|)
\end{aligned}$$

where we used the reverse triangle inequality, and $\tilde{z}_i^{(\ell)}$ and $e_i^{(\ell)}$ were defined in (36) and (38), respectively. Thus, for $\epsilon \geq 0$, under the assumption

$$\frac{1}{1+\sigma^2}\left(\frac{1}{\sqrt{d_\ell}} - \epsilon\right) \geq \frac{2.3\sqrt{6 \log N}}{\sqrt{m}} \tag{85}$$

resolved below, we have

$$\begin{aligned}
\mathrm{P}\left[\max_{i \in [N] \setminus \{j\}} |\langle \mathbf{x}_i, \mathbf{x}_j \rangle| \leq \frac{2.3\sqrt{6 \log N}}{\sqrt{m}}\right] &\leq \mathrm{P}\left[\max_{i \in [n_\ell] \setminus \{j\}} \frac{1}{1+\sigma^2}(\tilde{z}_i^{(\ell)} - |e_i^{(\ell)}|) \leq \frac{1}{1+\sigma^2}\left(\frac{1}{\sqrt{d_\ell}} - \epsilon\right)\right] \\
&\leq \mathrm{P}\left[\max_{i \in [n_\ell] \setminus \{j\}} \tilde{z}_i^{(\ell)} - \max_{i \in [n_\ell] \setminus \{j\}} |e_i^{(\ell)}| \leq \frac{1}{\sqrt{d_\ell}} - \epsilon\right] \\
&\leq \mathrm{P}\left[\max_{i \in [n_\ell] \setminus \{j\}} \tilde{z}_i^{(\ell)} \leq \frac{1}{\sqrt{d_\ell}}\right] + \mathrm{P}\left[\epsilon \leq \max_{i \in [n_\ell] \setminus \{j\}} |e_i^{(\ell)}|\right] \qquad (86)
\end{aligned}$$

where (86) follows from (41) with $X = \max_{i \in [n_\ell] \setminus \{j\}} \tilde{z}_i^{(\ell)} - \frac{1}{\sqrt{d_\ell}}$, $Y = \max_{i \in [n_\ell] \setminus \{j\}} |e_i^{(\ell)}| - \epsilon$, and $\phi = \varphi = 0$. Next, note that (59) with $\nu = 1$ yields

$$\mathrm{P}\left[\max_{i \in [n_\ell] \setminus \{j\}} \tilde{z}_i^{(\ell)} \leq \frac{1}{\sqrt{d_\ell}}\right] = \prod_{i \in [n_\ell] \setminus \{j\}} \mathrm{P}\left[\tilde{z}_i^{(\ell)} \leq \frac{1}{\sqrt{d_\ell}}\right] \leq \prod_{i \in [n_\ell] \setminus \{j\}} \sqrt{\frac{2}{\pi}} = e^{-\frac{1}{2}\log\left(\frac{\pi}{2}\right)(n_\ell - 1)}. \tag{87}$$

Application of (87) and (42) with $\epsilon = \frac{2\sigma(1+\sigma)}{\sqrt{m}}\sqrt{6 \log N}$ (using that $\beta \leq \sqrt{m}$, as verified below) to (86) yields

$$\mathrm{P}\left[\max_{i \in [N] \setminus \{j\}} |\langle \mathbf{x}_i, \mathbf{x}_j \rangle| \leq \frac{2.3\sqrt{6 \log N}}{\sqrt{m}}\right] \leq e^{-\frac{1}{2}\log\left(\frac{\pi}{2}\right)(n_\ell - 1)} + n_\ell \frac{7}{N^3}. \tag{88}$$

We next show that choosing c_1 sufficiently small, specifically $c_1 \leq 1/6$, guarantees that $\beta \leq \sqrt{m}$. To this end simply note that (13) implies

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

and take $c_1 \leq 1/6$.

Taking a union bound over all inliers in $\mathcal{X}_\ell$ shows that the probability of the outlier detection scheme misclassifying one or more of the inliers in $\mathcal{X}_\ell$ as an outlier is at most

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

Finally, we resolve (85) by showing that it is implied, for all $\ell \in [L]$, by (13). Rewriting (85) yields

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

Since $\frac{\sigma(1+\sigma)}{1+\sigma^2} \leq 1.3$ for $\sigma \geq 0$, (85) is implied by

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

which equals (13) 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 6 ([51, Prop. 19.4.2]). *Let $x \sim \mathcal{N}(0,1)$. For $\beta \geq \frac{1}{\sqrt{2\pi}}$, we have*

$$\mathrm{P}[x \geq \beta] \leq e^{-\frac{\beta^2}{2}}. \tag{89}$$

Theorem 7 ([52, Eq. 1.6]). *Let f be Lipschitz on $\mathbb{R}^m$ with Lipschitz constant $L \in \mathbb{R}$, i.e., $|f(\mathbf{b}_1) - f(\mathbf{b}_2)| \leq L\|\mathbf{b}_1 - \mathbf{b}_2\|_2$, for all $\mathbf{b}_1, \mathbf{b}_2 \in \mathbb{R}^m$, and let $\mathbf{x} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_m)$. Then, for $\beta > 0$, we have*

$$\mathrm{P}[f(\mathbf{x}) \geq \mathbb{E}[f(\mathbf{x})] + \beta] \leq e^{-\frac{\beta^2}{2L^2}}.$$

Let $\mathbf{x} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_m)$. Applying the concentration inequality in Theorem 7 to $f(\mathbf{x}) = \|\mathbf{x}\|_2$ which has Lipschitz constant $L = 1$, and using Jensen's inequality to get $(\mathbb{E}[\|\mathbf{x}\|_2])^2 \leq \mathbb{E}\big[\|\mathbf{x}\|_2^2\big] = m$, we obtain

$$\mathrm{P}\big[\|\mathbf{x}\|_2 \geq \sqrt{m} + \beta\big] \leq e^{-\frac{\beta^2}{2}}. \tag{90}$$

Proposition 1 (E.g., [53, Ex. 5.25]). *Let $\mathbf{a}$ be uniformly distributed on $\mathbb{S}^{m-1}$ and fix $\mathbf{b} \in \mathbb{R}^m$. Then, for $\beta \geq 0$, we have*

$$\mathrm{P}\left[|\langle \mathbf{a}, \mathbf{b} \rangle| > \frac{\beta}{\sqrt{m}} \|\mathbf{b}\|_2 \right] \leq 2e^{-\frac{\beta^2}{2}}.$$

References

[1] R. Heckel and H. Bölcskei, "Subspace clustering via thresholding and spectral clustering," in *Proc. of IEEE Int. Conf. Acoust. Speech Sig. Proc.*, 2013, pp. 3263–3267.

[2] ——, "Noisy subspace clustering via thresholding," in *Proc. of IEEE Int. Symp. on Inf. Theory*, 2013, pp. 1382–1386.

[3] R. Basri and D. Jacobs, "Lambertian reflectance and linear subspaces," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 25, no. 2, pp. 218–233, 2003.

[4] R. Vidal, "Subspace clustering," *IEEE Signal Process. Mag.*, vol. 28, no. 2, pp. 52–68, 2011.

[5] T. Zhang, A. Szlam, Y. Wang, and G. Lerman, "Hybrid linear modeling via local best-fit flats," *Int. J. Comput. Vision*, vol. 100, pp. 217–240, 2012.

[6] J. Ho, M. Yang, J. Lim, K. Lee, and D. Kriegman, "Clustering appearances of objects under varying illumination conditions," in *Proc. of IEEE Conf. Comput. Vision Pattern Recogn.*, vol. 1, 2003, pp. 11–18.

[7] W. Hong, J. Wright, K. Huang, and Y. Ma, "Multiscale hybrid linear models for lossy image representation," *IEEE Trans. Image Process.*, vol. 15, no. 12, pp. 3655–3671, 2006.

[8] R. Vidal and R. Hartley, "Motion segmentation with missing data using PowerFactorization and GPCA," in *Proc. of IEEE Conf. Comput. Vision Pattern Recogn.*, vol. 2, 2004, pp. 310–316.

[9] S. Rao, R. Tron, R. Vidal, and Y. Ma, "Motion segmentation via robust subspace separation in the presence of outlying, incomplete, or corrupted trajectories," in *Proc. of IEEE Conf. Comput. Vision Pattern Recogn.*, 2008, pp. 1–8.

[10] H. Kriegel, P. Kröger, and A. Zimek, "Clustering high-dimensional data: A survey on subspace clustering, pattern-based clustering, and correlation clustering," *ACM Trans. Knowl. Discov. Data*, vol. 3, no. 1, pp. 1–58, 2009.

[11] U. von Luxburg, "A tutorial on spectral clustering," *Stat. Comput.*, vol. 17, no. 4, pp. 395–416, 2007.

[12] M. Soltanolkotabi, E. Elhamifar, and E. J. Candès, "Robust subspace clustering," *Ann. Stat.*, vol. 42, no. 2, pp. 669–699, 2014.

[13] E. Elhamifar and R. Vidal, "Sparse subspace clustering," in *Proc. of IEEE Conf. Comput. Vision Pattern Recogn.*, 2009, pp. 2790–2797.

[14] ——, "Sparse subspace clustering: Algorithm, theory, and applications," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 35, no. 11, pp. 2765–2781, 2013.

[15] M. Soltanolkotabi and E. J. Candès, "A geometric analysis of subspace clustering with outliers," *Ann. Stat.*, vol. 40, no. 4, pp. 2195–2238, 2012.

[16] G. Liu, Z. Lin, and Y. Yu, "Robust subspace segmentation by low-rank representation," in *Proc. of 27th Int. Conf. on Mach. Learn.*, 2010, pp. 663–670.

[17] E. L. Dyer, A. C. Sankaranarayanan, and R. G. Baraniuk, "Greedy feature selection for subspace clustering," *Journal of Mach. Learn. Research*, vol. 14, pp. 2487–2517, 2013.

[18] Y. LeCun and C. Cortes, "The MNIST database," 2013, http://ya nn.lecun.com/exdb/mnist/.

[19] A. S. Georghiades, P. N. Belhumeur, and D. J. Kriegman, "From few to many: Illumination cone models for face recognition under variable lighting and pose," *IEEE Trans. Pattern Anal. Mach. Intelligence*, vol. 23, no. 6, pp. 643–660, 2001.

[20] K. C. Lee, J. Ho, and D. J. Kriegman, "Acquiring linear subspaces for face recognition under variable lighting," *IEEE Trans. Pattern Anal. Mach. Intelligence*, vol. 27, no. 5, pp. 684–698, 2005.

[21] F. Lauer and C. Schnorr, "Spectral clustering of linear subspaces for motion segmentation," in *Proc. of 12th IEEE Int. Conf. on Computer Vision*, 2009, pp. 678–685.

[22] J. Yan and M. Pollefeys, "A general framework for motion segmentation: Independent, articulated, rigid, non-rigid, degenerate and non-degenerate," in *European Conf. Computer Vision*, 2006, pp. 94–106.

[23] G. Lerman and T. Zhang, "Robust recovery of multiple subspaces by geometric ℓ_p minimization," *Ann. Statist.*, vol. 39, no. 5, pp. 2686–2715, 2011.

[24] G. Lerman, M. McCoy, J. A. Tropp, and T. Zhang, "Robust computation of linear models by convex relaxation," *Found. Comput. Math.*, vol. 15, no. 2, pp. 363–410, 2015.

[25] G. Chen and G. Lerman, "Spectral curvature clustering (SCC)," *Int. J. of Comput. Vision*, vol. 81, no. 3, pp. 317–330, Mar. 2009.

[26] ——, "Foundations of a multi-way spectral clustering framework for hybrid linear modeling," *Found. of Comput. Math.*, vol. 9, no. 5, pp. 517–558, Oct. 2009.

[27] A. Ng, I. M. Jordan, and W. Yair, "On spectral clustering: Analysis and an algorithm," in *Adv. Neural Inf. Process Syst.*, 2001, pp. 849–856.

[28] J. L. R. Kelley, *General Topology*. Springer, Berlin, Heidelberg, 1975.

[29] D. Spielman, "Spectral graph theory," 2012, lecture notes. [Online]. Available: http://www.cs.yale.edu/homes/spielman/561/

[30] G. H. Golub and C. F. Van Loan, *Matrix computations*. JHU Press, 1996.

[31] T. J. Hastie, R. J. Tibshirani, and J. J. H. Friedman, *The elements of statistical learning*. Springer, 2009.

[32] H. A. David and H. N. Nagaraja, *Order Statistics*. John Wiley & Sons, 2004.

[33] M. Brito, E. Chavez, A. Quiroz, and J. Yukich, "Connectivity of the mutual k-nearest-neighbor graph in clustering and outlier detection," *Stat. Probabil. Lett.*, vol. 35, no. 1, pp. 33–42, 1997.

[34] R. Heckel, E. Agustsson, and H. Bölcskei, "Neighborhood selection for thresholding based subspace clustering," in *Proc. of IEEE Int. Conf. Acoust. Speech Sig. Proc.*, 2014, pp. 6761–6765.

[35] R. Vidal, Y. Ma, and S. Sastry, "Generalized principal component analysis (GPCA)," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 12, pp. 1945–1959, 2005.

[36] P. Tseng, "Nearest q-flat to m points," *J. Optim. Theory Appl.*, vol. 105, no. 1, pp. 249–252, 2000.

[37] A. Topchy, M. Law, A. Jain, and A. Fred, "Analysis of consensus partition in cluster ensemble," in *Proc. of Fourth IEEE International Conf. on Data Mining*, 2004, pp. 225–232.

[38] T. Hastie and P. Y. Simard, "Metrics and models for handwritten character recognition," *Stat. Sci.*, vol. 13, no. 1, pp. 54–65, 1998.

[39] P. Favaro, R. Vidal, and A. Ravichandran, "A closed form solution to robust subspace estimation and clustering," in *IEEE Conf. on Computer Vision Pattern Recogn.*, 2011, pp. 1801–1807.

[40] G. Liu and S. Yan, "Latent low-rank representation for subspace segmentation and feature extraction," in *Proc. of IEEE Int. Conf. on Computer Vision*, 2011, pp. 1615–1622.

[41] T. H. Cormen, C. E. Leiserson, R. L. Rivest, and C. Stein, *Introduction to algorithms*. MIT Press, 2001.

[42] P. Balister, B. Bollobás, A. Sarkar, and M. Walters, "Connectivity of random k-nearest-neighbour graphs," *Adv. in Appl. Probab.*, vol. 37, no. 1, pp. 1–24, Mar. 2005.

[43] F. Xue and P. R. Kumar, "The number of neighbors needed for connectivity of wireless networks," *Wirel. Netw.*, vol. 10, no. 2, pp. 169–181, Mar. 2004.

[44] P. Leopardi, "A partition of the unit sphere into regions of equal area and small diameter," *Electron. Trans. Numer. Anal.*, vol. 25, pp. 309–327, 2006.

[45] ——, "Diameter bounds for equal area partitions of the unit sphere," *Electron. Trans. Numer. Anal.*, 2009.

[46] U. Feige and G. Schechtman, "On the optimality of the random hyperplane rounding technique for MAX CUT," *Random Struct. & Algor.*, vol. 20, no. 3, pp. 403–440, 2002.

[47] S. Janson, "On concentration of probability," in *Contemporary Combinatorics*. Springer, Berlin, Heidelberg, 2002.

[48] S. Foucart and H. Rauhut, *A mathematical introduction to compressive sensing*. Springer, Berlin, Heidelberg, 2013.

[49] W. B. Johnson and J. Lindenstrauss, "Extensions of Lipschitz mappings into a Hilbert space," *Contemp. Math.*, no. 26, pp. 189–206, 1984.

[50] R. A. Horn and C. R. Johnson, *Matrix analysis*. Cambridge University Press, 1986.

[51] A. Lapidoth, *A foundation in digital communication*. Cambridge University Press, 2009.

[52] M. Ledoux and M. Talagrand, *Probability in Banach spaces: Isoperimetry and processes*. Springer, Berlin, Heidelberg, 1991.

[53] R. Vershynin, "Introduction to the non-asymptotic analysis of random matrices," in *Compressed sensing: Theory and applications*. Cambridge University Press, 2012, pp. 210–268.