Dimensionality-reduced subspace clustering

Reinhard Heckel, Michael Tschannen, and Helmut Bölcskei

December 14, 2015

Abstract

Subspace clustering refers to the problem of clustering unlabeled high-dimensional data points into a union of low-dimensional linear subspaces, whose number, orientations, and dimensions are all unknown. In practice one may have access to dimensionality-reduced observations of the data only, resulting, e.g., from undersampling due to complexity and speed constraints on the acquisition device or mechanism. More pertinently, even if the high-dimensional data set is available it is often desirable to first project the data points into a lower-dimensional space and to perform clustering there; this reduces storage requirements and computational cost. The purpose of this paper is to quantify the impact of dimensionality reduction through random projection on the performance of three subspace clustering algorithms, all of which are based on principles from sparse signal recovery. Specifically, we analyze the thresholding based subspace clustering (TSC) algorithm, the sparse subspace clustering (SSC) algorithm, and an orthogonal matching pursuit variant thereof (SSC-OMP). We find, for all three algorithms, that dimensionality reduction down to the order of the subspace dimensions is possible without incurring significant performance degradation. Moreover, these results are order-wise optimal in the sense that reducing the dimensionality further leads to a fundamentally ill-posed clustering problem. Our findings carry over to the noisy case as illustrated through analytical results for TSC and simulations for SSC and SSC-OMP. Extensive experiments on synthetic and real data complement our theoretical findings.

1 Introduction

One of the major challenges in modern data analysis is to find low-dimensional structure in large high-dimensional data sets. A prevalent low-dimensional structure is that of data points lying in a union of (low-dimensional) subspaces. The problem of extracting such a structure from a given data set can be formalized as follows. Consider the (high-dimensional) set $\mathcal{Y}$ of points in $\mathbb{R}^m$ and assume that $\mathcal{Y} = \mathcal{Y}_1 \cup \ldots \cup \mathcal{Y}_L$, where the points in $\mathcal{Y}_\ell$ lie in a linear subspace $\mathcal{S}_\ell$ of $\mathbb{R}^m$. The association of the data points to the sets $\mathcal{Y}_\ell$, the orientations, dimensions, and the number of the subspaces $\mathcal{S}_\ell$ are all unknown. The problem of identifying the assignments of the points in $\mathcal{Y}$ to the $\mathcal{Y}_\ell$ is referred to as subspace clustering [34] or hybrid linear modeling and has applications, inter alia, in unsupervised learning, image representation and segmentation, computer vision, and disease detection.

R. Heckel was with the Department of Information Technology and Electrical Engineering, ETH Zurich, Switzerland, and is now with IBM Research, Zurich, Switzerland (e-mail: reinhard.heckel@gmail.com). M. Tschannen and H. Bölcskei are with the Department of Information Technology and Electrical Engineering, ETH Zurich, Switzerland (e-mail: michaelt@nari.ee.ethz.ch; boelcskei@nari.ee.ethz.ch).

Part of this paper was presented at the 2014 IEEE International Symposium on Information Theory (ISIT) [17].

In practice one may have access to dimensionality-reduced observations of $\mathcal{Y}$ only, resulting, e.g., from "undersampling" due to complexity and speed constraints on the acquisition device or mechanism. More pertinently, even if the data points in $\mathcal{Y}$ are directly accessible, it is often desirable to work on a dimensionality-reduced version of $\mathcal{Y}$ as this reduces data storage cost and leads to computational complexity savings. The idea of reducing computational complexity through dimensionality reduction appears, e.g., in [32] in a general context, and for subspace clustering in the experiments reported in [38, 9]. Dimensionality reduction also has a privacy-enhancing effect in the sense that no access to the original data is needed for processing [25].

Dimensionality reduction will, in general, come at the cost of clustering performance. The purpose of this paper is to analytically characterize this performance degradation for three subspace clustering algorithms, namely thresholding-based subspace clustering (TSC) [16], sparse subspace clustering (SSC) [8, 9], and SSC-orthogonal matching pursuit (SSC-OMP) [7]. The common theme underlying these three algorithms is that they apply spectral clustering to an adjacency matrix constructed from sparse representations of the data points, obtained through a nearest neighbor search in the case of TSC, through ℓ_1-minimization for SSC, and through OMP in the case of SSC-OMP. While there are numerous further approaches to subspace clustering (see [34] for an overview), we chose to study TSC, SSC, and SSC-OMP, as they belong to the small group of subspace clustering algorithms that are computationally tractable and succeed provably under nonrestrictive conditions [28, 29, 9, 7, 37, 16]. Specifically, the results in [16] for TSC, and in [28, 29] for SSC show that TSC and SSC can succeed even when the subspaces $\mathcal{S}_\ell$ intersect. The corresponding proof techniques, together with analytical performance guarantees for SSC-OMP developed in this paper, form the basis for our analytical characterization of the impact of dimensionality reduction on subspace clustering performance.

Formal problem statement and contributions. Consider a set of N data points $\mathcal{Y} \in \mathbb{R}^m$, and assume that $\mathcal{Y} = \mathcal{Y}_1 \cup \ldots \cup \mathcal{Y}_L$, where the points $\mathbf{y}_i^{(\ell)} \in \mathcal{Y}_\ell, i \in \{1, \ldots, n_\ell\}$, lie in a d_ℓ-dimensional linear subspace of $\mathbb{R}^m$, denoted by $\mathcal{S}_\ell$. Neither the assignments of the points in $\mathcal{Y}$ to the sets $\mathcal{Y}_\ell$ nor the subspaces $\mathcal{S}_\ell$ or the number of subspaces L are known. Traditional subspace clustering operates on the data $\mathcal{Y}$ with the goal of segmenting it into the sets $\mathcal{Y}_\ell$. Here, we assume, however, that clustering is performed on a dimensionality-reduced version of the points in $\mathcal{Y}$. Specifically, we employ the random projection method [32] by first applying the (same) realization of a random projection matrix $\mathbf{\Phi} \in \mathbb{R}^{p \times m}$ (typically $p \ll m$) to each point in $\mathcal{Y}$ to obtain the set of dimensionality-reduced data points $\mathcal{X}$. Then, we declare the segmentation obtained by operating on $\mathcal{X}$ to be the segmentation of the data points in $\mathcal{Y}$. The realization of $\mathbf{\Phi}$ does not need to be known. There are two error sources that determine the performance of this approach, first, the error that would be obtained even if clustering was performed on the high-dimensional data set $\mathcal{Y}$ directly, second, and more pertinently, the error incurred by operating on dimensionality-reduced data. The former is quantified for TSC in [16], for SSC in [28, 29], and for SSC-OMP this paper develops corresponding new results. Analytically characterizing the error incurred by dimensionality reduction is the main contribution of this paper.

While it is conceivable that TSC, which is based on thresholding inner products, exhibits graceful performance degradation as the data set's dimensionality is reduced through random projection, this is far from obvious for the ℓ_1-minimization based SSC algorithm and the iterative SSC-OMP algorithm. We prove our main results by first deriving conditions for TSC, SSC, and SSC-OMP to ensure correct clustering of dimensionality-reduced data. While these conditions are general, they only become amenable to insightful interpretations once particularized for a random data model, also used in [28, 16], that takes the subspace structure of the data set into account. The resulting

clustering conditions make the impact of dimensionality reduction explicit and reveal a tradeoff between the affinity of the subspaces $\mathcal{S}_\ell$ and the amount of dimensionality reduction possible. Specifically, we find that all three algorithms succeed provably under quite generous conditions on the relative orientations of the subspaces $\mathcal{S}_\ell$, provided that the dimensionality is reduced no more than down to the largest subspace dimension $d_{\max} = \max_\ell d_\ell$. As the computational complexity associated with the construction of the adjacency matrix is essentially linear in the dimension of the ambient space, m, for all three algorithms, random projection reduces the complexity of this step by a factor of $m/d_{\max}$. These complexity savings translate into, possibly significant, run-time savings for the overall clustering algorithms (which include the spectral clustering step), in particular when m is sufficiently large relative to N.

We study the impact of noise—added to the high-dimensional data points—on clustering performance. For TSC, we derive a clustering condition which quantifies the tradeoff between the affinity of the subspaces $\mathcal{S}_\ell$ and the amount of dimensionality reduction possible, as a function of noise variance. Specifically, this condition allows us to conclude that TSC succeeds provably provided that—as in the noiseless case—the dimensionality is reduced to no more than down to the largest subspace dimension $d_{\max}$, and the noise variance is sufficiently small. An approach akin to that used for TSC can be applied to establish a similar clustering condition for SSC-OMP. The corresponding technical details are, however, significantly more involved and cumbersome. We therefore decided not to state the formal result. Regarding SSC, we remark that Wang et al. [36] reported deterministic clustering conditions for the Lasso-version of SSC [29] applied to dimensionality-reduced noisy data. However, the corresponding results [36, Lem. 16, Thm. 18] make the critical assumption of the signal part of the *projected* noisy data being normalized, whereas the noise component remains un-normalized. It is difficult to see how one would realize this in practice, unless the noise realization is known perfectly, in which case the noise component could be removed which would take us back to the noiseless case. The results in [36] for noisy data therefore appear to be of limited practicality. While the statements in [36] may be particularized to the noiseless case, we note that corresponding results appeared in the conference version [17] of this paper before the publication of [36].

We note that our results, both for the noiseless and the noisy case, apply even when the subspaces $\mathcal{S}_\ell$ span the ambient space $\mathbb{R}^m$. This follows from our clustering conditions depending on the *pairwise* affinities between subspaces only, and pairwise affinities changing only moderately if the dimensionality is reduced down to no more than the order of the individual subspace dimensions.

Another popular dimensionality reduction method is principal component analysis (PCA). However, when used in the context of subspace clustering, PCA allows dimensionality reduction down to the dimension of the overall span of the subspaces only, in general; this results in no dimensionality reduction at all when the subspaces $\mathcal{S}_\ell$ span the ambient space. To see this, consider the L subspaces of dimension 1 that correspond to the standard basis in $\mathbb{R}^m$, i.e., the ℓ-th subspace is spanned by the vector $\mathbf{e}_\ell$ given by $[\mathbf{e}_\ell]_\ell = 1$ and $[\mathbf{e}_\ell]_i = 0$, for $i \neq \ell$. Assuming that each of the data points in the data set under consideration, denoted by $\mathbf{Y} \in \mathbb{R}^{m \times N}$, lies in one of these L subspaces, the corresponding sample covariance matrix $\mathbf{Y}\mathbf{Y}^T$ has non-zero entries only in its first L main diagonal entries. The first L principal components are therefore given by the vectors $\mathbf{e}_\ell$. Reducing the dimensionality of the data set to below L will result in certain data points being mapped to zero (owing to the orthogonality of the $\mathbf{e}_\ell$). Moreover, PCA has computational complexity $O(Nm^2 + m^3)$ while random projection through Gaussian matrices and fast random projection matrices [1] has complexity $O(pmN)$ and $O(\log(m)mN)$, respectively, and is therefore computationally much less demanding. This is an important aspect as computational complexity is a major motivation for dimensionality reduction.

Notation. We use lowercase boldface letters to denote (column) vectors and uppercase boldface letters to designate matrices. The superscript ^{T} stands for transposition. For the vector $\mathbf{x}$, x_q denotes its qth entry and $\mathbf{x}_S$ is the subvector of $\mathbf{x}$ with entries corresponding to the indices in the set S. For the matrix $\mathbf{A}$, $\mathbf{A}_{ij}$ designates the entry in its ith row and jth column, $\mathbf{A}_S$ the matrix containing the columns of $\mathbf{A}$ with indices in the set S, $\|\mathbf{A}\|_{2\to 2} \coloneqq \max_{\|\mathbf{v}\|_2=1} \|\mathbf{A}\mathbf{v}\|_2$ its spectral norm, $\sigma_{\min}(\mathbf{A})$ its minimum singular value, and $\|\mathbf{A}\|_F \coloneqq (\sum_{i,j} |\mathbf{A}_{ij}|^2)^{1/2}$ its Frobenius norm. If $\mathbf{A}$ has full column rank $\mathbf{A}^\dagger \coloneqq (\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T$ stands for its (left) pseudoinverse, and for $\mathbf{A}$ with full row rank, $\mathbf{A}^\dagger \coloneqq \mathbf{A}^T(\mathbf{A}\mathbf{A}^T)^{-1}$ is the (right) pseudoinverse. The identity matrix is denoted by $\mathbf{I}$. $\log(\cdot)$ refers to the natural logarithm, $\arccos(\cdot)$ is the inverse function of $\cos(\cdot)$, and $x \wedge y$ denotes the minimum of x and y. The set $\{1, \ldots, N\}$ is written as $[N]$. The cardinality of the set S is designated by $|S|$ and its complement is $\overline{S}$. $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ stands for the distribution of a real Gaussian random vector with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. We write $X \sim Y$ to indicate that the random variables X and Y are equally distributed. For notational convenience, we use the following shorthands: $\max_\ell$ for $\max_{\ell\in[L]}$, $\max_{k\neq\ell}$ for $\max_{k\in[L]\colon k\neq\ell}$, and $\max_{k,\ell\colon k\neq\ell}$ for $\max_{k,\ell\in[L]\colon k\neq\ell}$. The unit sphere in $\mathbb{R}^m$ is $\mathbb{S}^{m-1} \coloneqq \{\mathbf{x} \in \mathbb{R}^m \colon \|\mathbf{x}\|_2 = 1\}$. A subgraph H of a graph G is said to be connected if every pair of nodes in H can be joined by a path along edges with nodes exclusively in H. A subgraph H of G is called a connected component of G if H is connected and if there are no edges between nodes in H and the remaining nodes in G.

2 A brief review of TSC, SSC, and SSC-OMP

We next briefly summarize the TSC [16], SSC [8, 9], and SSC-OMP [7] algorithms. All three algorithms apply normalized spectral clustering [35] to an adjacency matrix $\mathbf{A}$ built by finding a sparse representation of each data point in terms of the other data points. Specifically, TSC is based on least-squares representations in terms of nearest neighbors while SSC and SSC-OMP construct $\mathbf{A}$ by finding sparse representations via ℓ_1-minimization and OMP, respectively. Note that the focus in [16] is on a version of TSC that uses a spherical distance measure between data points instead of least-squares regression coefficients to determine the entries of $\mathbf{A}$. The analytical results presented here apply to both versions of TSC. We decided, however, to work with the least-squares version as this formulation better elucidates the sparsity aspect and thereby the relationship to SSC and SSC-OMP.

In order to emphasize that we consider all three algorithms applied to dimensionality-reduced data, their descriptions will be in terms of the dimensionality-reduced data set $\mathcal{X} \subset \mathbb{R}^p$. We furthermore assume that an estimate $\hat{L}$ of the number of subspaces L is available. The estimation of L from $\mathcal{X}$ is discussed later. We also note that the formulations of the TSC and SSC-OMP algorithms below assume that the data points in $\mathcal{X}$ are of comparable ℓ_2-norm. This assumption is relevant for Step 1 in both cases and is not restrictive as the data points can be normalized prior to clustering.

The TSC algorithm: Given a set of N data points $\mathcal{X}$ in $\mathbb{R}^p$, an estimate of the number of subspaces $\hat{L}$, and the parameter q, perform the following steps:

Step 1: For every $\mathbf{x}_j \in \mathcal{X}$, find the set $S_j \subset [N]\backslash\{j\}$ of cardinality q defined by

$$|\langle \mathbf{x}_j, \mathbf{x}_i \rangle| \geq |\langle \mathbf{x}_j, \mathbf{x}_k \rangle|, \text{ for all } i \in S_j \text{ and all } k \notin S_j,$$

and let $\mathbf{z}_j$ be the coefficient vector corresponding to the minimum least-squares representation of $\mathbf{x}_j$ in terms of $\mathbf{x}_i, i \in S_j$. Specifically, set $(\mathbf{z}_j)_{S_j} = \arg\min_{\mathbf{z}} \left\|\mathbf{x}_j - \mathbf{X}_{S_j}\mathbf{z}\right\|_2$ (if multiple solutions exist, choose, e.g., the $\mathbf{z}$ with minimum ℓ_2-norm), and $(\mathbf{z}_j)_{\overline{S}_j} = \mathbf{0}$. Construct the adjacency matrix

$\mathbf{A}$ according to $\mathbf{A} = \mathbf{Z} + \mathbf{Z}^T$, where $\mathbf{Z} = \text{abs}([\mathbf{z}_1 \ \dots \ \mathbf{z}_N])$ and $\text{abs}(\cdot)$ takes absolute values element-wise.

Step 2: Apply normalized spectral clustering [26, 35] to $(\mathbf{A}, \hat{L})$.

The SSC algorithm: Given a set of N data points $\mathcal{X}$ in $\mathbb{R}^p$ and an estimate of the number of subspaces $\hat{L}$, perform the following steps:

Step 1: Let $\mathbf{X} \in \mathbb{R}^{p \times N}$ be the matrix whose columns are the points in $\mathcal{X}$. For every $\mathbf{x}_j \in \mathcal{X}$ determine $\mathbf{z}_j$ as a solution of

$$\underset{\mathbf{z}}{\text{minimize}} \ \|\mathbf{z}\|_1 \text{ subject to } \mathbf{x}_j = \mathbf{X}\mathbf{z} \text{ and } z_j = 0. \tag{1}$$

Construct the adjacency matrix $\mathbf{A}$ according to $\mathbf{A} = \mathbf{Z} + \mathbf{Z}^T$, where $\mathbf{Z} = \text{abs}([\mathbf{z}_1 \ \dots \ \mathbf{z}_N])$.

Step 2: Apply normalized spectral clustering [26, 35] to $(\mathbf{A}, \hat{L})$.

The SSC-OMP algorithm: Given a set of N data points $\mathcal{X}$ in $\mathbb{R}^p$, an estimate of the number of subspaces $\hat{L}$, and a maximum number of OMP iterations $s_{\max}$, perform the following steps:

Step 1: For every $\mathbf{x}_j \in \mathcal{X}$, find a sparse representation of $\mathbf{x}_j$ in terms of $\mathcal{X} \backslash \{\mathbf{x}_j\}$ using OMP as follows: Initialize the iteration counter $s = 0$, the residual $\mathbf{r}_0 = \mathbf{x}_j$, and the set of selected indices $\Lambda_0 = \emptyset$. For $s = 1, 2, \dots$ perform updates according to

$$\Lambda_s = \Lambda_{s-1} \cup \underset{i \in [N] \colon i \neq j}{\text{argmax}} \ |\langle \mathbf{x}_i, \mathbf{r}_{s-1} \rangle| \tag{2}$$

$$\mathbf{r}_s = (\mathbf{I} - \mathbf{X}_{\Lambda_s} \mathbf{X}_{\Lambda_s}^{\dagger}) \mathbf{x}_j \tag{3}$$

until $\mathbf{r}_s = \mathbf{0}$ or $s = s_{\max}$ (when the maximizer in (2) is not unique, select any of the solutions). With the number of OMP iterations actually performed denoted by s_{end}, set $(\mathbf{z}_j)_{\Lambda_{s_{\text{end}}}} = \mathbf{X}_{\Lambda_{s_{\text{end}}}}^{\dagger} \mathbf{x}_j$, $(\mathbf{z}_j)_{\overline{\Lambda}_{s_{\text{end}}}} = \mathbf{0}$, and construct the adjacency matrix $\mathbf{A}$ according to $\mathbf{A} = \mathbf{Z} + \mathbf{Z}^T$, where $\mathbf{Z} = \text{abs}([\mathbf{z}_1 \ \dots \ \mathbf{z}_N])$.

Step 2: Apply normalized spectral clustering [26, 35] to $(\mathbf{A}, \hat{L})$.

For all three algorithms the number of subspaces L can be estimated based on the insight that the number of zero eigenvalues of the normalized Laplacian of the graph G with adjacency matrix $\mathbf{A}$, henceforth simply referred to as "the graph G", is equal to the number of connected components of G [30]. A robust estimator for L is the *eigengap heuristic* described in [35].

Let the oracle segmentation of $\mathcal{X}$ be given by $\mathcal{X} = \mathcal{X}_1 \cup \ldots \cup \mathcal{X}_L$. If each connected component in the graph G corresponds exclusively to points from one of the sets $\mathcal{X}_\ell$, spectral clustering will deliver the oracle segmentation [35, Prop. 4] and the clustering error, i.e., the fraction of misclustered points, will be zero. Since conditions guaranteeing zero clustering error are inherently hard to obtain, we will work with an intermediate, albeit sensible, performance measure, also employed in [16, 28, 29, 7]. Specifically, this measure, termed the no-false connections property, declares success if the graph G has no false connections, i.e., if each $\mathbf{x}_j \in \mathcal{X}_\ell$ is connected to points in $\mathcal{X}_\ell$ only, for all ℓ. Guaranteeing the absence of false connections, does, however, not guarantee that the connected components of G correspond to the $\mathcal{X}_\ell$, as the points in a given set $\mathcal{X}_\ell$ may form two (or more) distinct connected components in G.

To counter this problem sufficiently many entries in each row/column of the adjacency matrix $\mathbf{A}$ have to be non-zero. Specifically, for the subgraphs of G corresponding to the $\mathcal{X}_\ell$ to be connected, each row/column of $\mathbf{A}$ corresponding to a point in $\mathcal{X}_\ell$ needs to have between $O(\log n_\ell)$ and $O(n_\ell)$ non-zero entries. As the solutions $\mathbf{z}$ to $\arg\min_{\mathbf{z}} \left\|\mathbf{x}_j - \mathbf{X}_{S_j}\mathbf{z}\right\|_2$ are typically dense, TSC is likely to select a representation of $\mathbf{x}_j$ in terms of points in $\mathcal{X}_\ell \backslash \{\mathbf{x}_j\}$ with on the order of q non-zero coefficients.

Choosing q large enough therefore ensures sufficient connectivity of the graph G generated by TSC. On the other hand, taking q to be large increases the probability of false connections. The performance guarantee we obtain for TSC therefore requires q to be sufficiently small relative to the n_ℓ.

For SSC and SSC-OMP, the number of non-zero entries in each row/column of $\mathbf{A}$ turns out to be tied to d_ℓ, rather than n_ℓ. To see this, suppose that both algorithms exclusively select data points from $\mathcal{X}_\ell \backslash \{\mathbf{x}_j\}$ to represent $\mathbf{x}_j$. Moreover, assume that the $\mathcal{X}_\ell$ are non-degenerate in the sense that, indeed, d_ℓ points are needed to represent $\mathbf{x}_j \in \mathcal{X}_\ell$ through points in $\mathcal{X}_\ell \backslash \{\mathbf{x}_j\}$; this precludes, e.g., that $\mathcal{X}_\ell$ contains multiple copies of the same data point. The OMP algorithm in SSC-OMP then terminates after $\min(d_\ell, s_{\max})$ (recall that $d_\ell = \dim(\mathcal{S}_\ell)$) iterations for $\mathbf{x}_j \in \mathcal{X}_\ell$ and hence results in exactly $\min(d_\ell, s_{\max})$ non-zero entries in the corresponding column of $\mathbf{Z}$ (recall that $\mathbf{A} = \mathbf{Z} + \mathbf{Z}^T$). For SSC, we simply note that d_ℓ points are enough to represent $\mathbf{x}_j \in \mathcal{X}_\ell$ through other points in $\mathcal{X}_\ell$ and we cannot guarantee more than d_ℓ non-zero entries in the corresponding column of $\mathbf{Z}$, in general. This will lead to insufficient connectivity for SSC and SSC-OMP when d_ℓ is not in the range $O(\log n_\ell)$–$O(n_\ell)$. The problem is exacerbated when the data set is degenerate. To counter insufficient connectivity in SSC a modification which adds an ℓ_2-penalty to the cost function in (1) was proposed in [9, Sec. 5]. Such a modification is not known for SSC-OMP, and this may be considered a limitation of SSC-OMP.

We finally remark that TSC and SSC-OMP can be made essentially parameterless, like SSC. Specifically, a procedure for choosing the TSC parameter q in a data-driven fashion is described in [15], and for SSC-OMP we can get rid of the parameter $s_{\max}$ by stopping the OMP step once the ℓ_2-norm of the residual $\mathbf{r}_s$ falls below a threshold value.

3 Main results

We start by specifying the statistical data model used throughout the paper. The subspaces $\mathcal{S}_\ell$ are taken to be deterministic and the points within the $\mathcal{S}_\ell$ are chosen randomly. Specifically, the elements of the set $\mathcal{Y}_\ell$ in $\mathcal{Y} = \mathcal{Y}_1 \cup \ldots \cup \mathcal{Y}_L$ are obtained by choosing n_ℓ points at random according to $\mathbf{y}_j^{(\ell)} = \mathbf{U}^{(\ell)} \mathbf{a}_j^{(\ell)}, j \in [n_\ell]$, where the columns of $\mathbf{U}^{(\ell)} \in \mathbb{R}^{m \times d_\ell}$ form an orthonormal basis for the d_ℓ-dimensional subspace $\mathcal{S}_\ell$, and the $\mathbf{a}_j^{(\ell)}$ are i.i.d. uniform on $\mathbb{S}^{d_\ell - 1}$. As the $\mathbf{U}^{(\ell)}$ are orthonormal, the data points $\mathbf{y}_j^{(\ell)}$ are distributed uniformly on the set $\{\mathbf{y} \in \mathcal{S}_\ell \colon \|\mathbf{y}\|_2 = 1\} = \mathcal{S}_\ell \cap \mathbb{S}^{m-1}$, which avoids degenerate situations where the data points lie in preferred directions. To see why such degeneracies can lead to ambiguous results, consider a two-dimensional subspace and assume that the data points in this subspace 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.

The dimensionality-reduced data set $\mathcal{X} \subset \mathbb{R}^p$ is obtained by applying the (same) realization of a random matrix $\mathbf{\Phi} \in \mathbb{R}^{p \times m}$ ($p \geq \max_\ell d_\ell$) to each point in $\mathcal{Y}$. The elements of the sets $\mathcal{X}_\ell$ in $\mathcal{X} = \mathcal{X}_1 \cup \ldots \cup \mathcal{X}_L$ are hence given by $\mathbf{x}_j^{(\ell)} = \mathbf{\Phi} \mathbf{y}_j^{(\ell)}, j \in [n_\ell]$. We take $\mathbf{\Phi}$ as a random matrix satisfying the following concentration inequality

$$\mathrm{P}\left[\left| \|\mathbf{\Phi}\mathbf{x}\|_2^2 - \|\mathbf{x}\|_2^2 \right| \geq t \|\mathbf{x}\|_2^2 \right] \leq 2e^{-\tilde{c} t^2 p}, \quad \forall\, t > 0, \forall \mathbf{x} \in \mathbb{R}^m, \tag{4}$$

where $\tilde{c}$ is either a numerical constant or a parameter mildly depending on m. Random matrices satisfying (4) realize, with high probability, linear embeddings in the sense of the Johnson-Lindenstrauss (JL) Lemma, see e.g., [32], [10, Sec. 9.5]. The JL Lemma says that every set

of N points in Euclidean space can be embedded in an $O(\epsilon^{-2}\log N)$-dimensional space without perturbing the pairwise Euclidean distances between the points by more than a factor of $1\pm\epsilon$.

A similar statement on random projection preserving affinities between subspaces–as defined in (5)–is used in our proofs. Specifically, we show that randomly projecting a set of d-dimensional subspaces into p-dimensional space does not increase their pairwise affinities by more than const.$\sqrt{d/p}$, with high probability (cf. (45)). The concentration inequality (4) holds, inter alia, for matrices with i.i.d. subgaussian[1] entries [10, Lem. 9.8]; this includes $\mathcal{N}(0,1/p)$ entries and entries that are uniformly distributed on $\{-1/\sqrt{p},1/\sqrt{p}\}$. Such matrices may, however, be costly to generate, store, and apply to high-dimensional data points. In order to reduce these costs structured random matrices satisfying (4) (with $\tilde{c}$ possibly mildly dependent on m) were proposed in [1, 21]. For example, the structured random matrix proposed in [1] (and described in detail in Section 5) satisfies (4) with $\tilde{c}=c_2\log^{-4}(m)$, where c_2 is a numerical constant [21, Prop. 3.2], and can be applied in time $O(m\log m)$ as opposed to time $O(mp)$ for the realizations of general subgaussian random matrices.

The clustering performance guarantees we obtain below are all in terms of the affinity between the subspaces $\mathcal{S}_k$ and $\mathcal{S}_\ell$ defined as [28, Def. 2.6], [29, Def. 1.2]

$$\operatorname{aff}(\mathcal{S}_k,\mathcal{S}_\ell)\coloneqq\frac{1}{\sqrt{d_k\wedge d_\ell}}\big\|\mathbf{U}^{(k)^T}\mathbf{U}^{(\ell)}\big\|_F. \tag{5}$$

Note that $0\leq\operatorname{aff}(\mathcal{S}_k,\mathcal{S}_\ell)\leq 1$, with $\operatorname{aff}(\mathcal{S}_k,\mathcal{S}_\ell)=1$ if $\mathcal{S}_k\subseteq\mathcal{S}_\ell$ or $\mathcal{S}_\ell\subseteq\mathcal{S}_k$ and $\operatorname{aff}(\mathcal{S}_k,\mathcal{S}_\ell)=0$ if $\mathcal{S}_k$ and $\mathcal{S}_\ell$ are orthogonal to each other. Moreover, we have

$$\operatorname{aff}(\mathcal{S}_k,\mathcal{S}_\ell)=\sqrt{\cos^2(\theta_1)+\ldots+\cos^2(\theta_{d_k\wedge d_\ell})}/\sqrt{d_k\wedge d_\ell}, \tag{6}$$

where $\theta_1\leq\ldots\leq\theta_{d_k\wedge d_\ell}$ are the principal angles between $\mathcal{S}_k$ and $\mathcal{S}_\ell$ [12, Sec. 6.3.4]. If $\mathcal{S}_k$ and $\mathcal{S}_\ell$ intersect in t dimensions, i.e., if $\mathcal{S}_k\cap\mathcal{S}_\ell$ is t-dimensional, then $\cos(\theta_1)=\ldots=\cos(\theta_t)=1$ and hence $\operatorname{aff}(\mathcal{S}_k,\mathcal{S}_\ell)\geq\sqrt{t/(d_k\wedge d_\ell)}$. The affinity between subspaces plays an important role in subspace classification [27] as well, see [19, Thms. 2 and 3].

We start with our main result for TSC.

Theorem 1. *Choose q such that $q\leq\min_\ell n_\ell/6$. If*

$$\max_{k,\ell:\,k\neq\ell}\operatorname{aff}(\mathcal{S}_k,\mathcal{S}_\ell)+\frac{\sqrt{11}}{\sqrt{3\tilde{c}}}\frac{\sqrt{d_{\max}}}{\sqrt{p}}\leq\frac{1}{15\log N}, \tag{7}$$

where $d_{\max}=\max_\ell d_\ell$ and $\tilde{c}$ is the constant in the concentration inequality (4), then the graph G obtained by applying TSC to $\mathcal{X}$ has no false connections with probability at least $1-7N^{-1}-\sum_{\ell=1}^L n_\ell e^{-c(n_\ell-1)}$, where $c>1/20$ is a numerical constant.

Our main result for SSC is the following.

Theorem 2. *Let $\rho_\ell\coloneqq(n_\ell-1)/d_\ell$, $\ell\in[L]$, $\rho_{\min}\coloneqq\min_\ell\rho_\ell\geq\rho_0$, where $\rho_0>1$ is a numerical constant, and pick any $\tau>0$. Set $d_{\max}=\max_\ell d_\ell$ and suppose that*

$$\max_{k,\ell:\,k\neq\ell}\operatorname{aff}(\mathcal{S}_k,\mathcal{S}_\ell)+\sqrt{\frac{28d_{\max}+8\log L+2\tau}{3\tilde{c}p}}\leq\frac{\sqrt{\log\rho_{\min}}}{65\log N}, \tag{8}$$

where $\tilde{c}$ is the constant in (4). Then, the graph G obtained by applying SSC to $\mathcal{X}$ has no false connections with probability at least $1-4e^{-\tau/2}-N^{-1}-\sum_{\ell=1}^L n_\ell e^{-\sqrt{\rho_\ell}d_\ell}$.

[1] A random variable x is subgaussian [10, Sec. 7.4] if its tail probability satisfies $\mathrm{P}[|x|>t]\leq c_1e^{-c_2t^2}$ for constants $c_1,c_2>0$.

Finally, for SSC-OMP we obtain the following statement.

Theorem 3. *Let $\rho_\ell \coloneqq (n_\ell - 1)/d_\ell$, $\ell \in [L]$, $\rho_{\min} \coloneqq \min_\ell \rho_\ell \geq \rho_0$, where $\rho_0 > 1$ is a numerical constant, and pick any $\tau > 0$. Set $d_{\min} \coloneqq \min_\ell d_\ell$, $d_{\max} \coloneqq \max_\ell d_\ell$, and suppose that $\mathbf{\Phi}$ has (in addition to satisfying the concentration inequality (4)) a rotationally invariant distribution, i.e., $\mathbf{\Phi}\mathbf{V} \sim \mathbf{\Phi}$ for all unitary matrices $\mathbf{V} \in \mathbb{R}^{m\times m}$. If*

$$\max_{k,\ell:\, k\neq\ell} \operatorname{aff}(\mathcal{S}_k, \mathcal{S}_\ell) + \sqrt{\frac{28 d_{\max} + 8\log L + 2\tau}{12\tilde{c}p}} \sqrt{\frac{d_{\max}}{d_{\min}}} \leq \frac{3}{200} \frac{\sqrt{\log \rho_{\min}}}{\log N}, \tag{9}$$

where $\tilde{c}$ is the constant in (4), then, irrespectively[2] of the choice of the maximum number of OMP-iterations $s_{\max}$, the graph G obtained by applying SSC-OMP to $\mathcal{X}$ has no false connections with probability at least $1 - 4e^{-\tau/2} - 4N^{-1} - \sum_{\ell=1}^{L} n_\ell e^{-\sqrt{\rho_\ell} d_\ell}$.

The proofs of Theorems 1, 2, and 3 are provided in Appendices A, B, and C, respectively, and are established by first deriving deterministic clustering conditions that are then evaluated for our statistical data model.

Theorems 1 and 2 essentially say that even when p is on the order of $d_{\max}$, TSC and SSC succeed with high probability if the affinities between the subspaces $\mathcal{S}_\ell$ are sufficiently small and if $\mathcal{X}$ contains sufficiently many points from each subspace. The same conclusion applies to SSC-OMP provided that the term $\sqrt{d_{\max}/d_{\min}}$ is not too large, which is the case if the dimensions d_ℓ, $\ell \in [L]$, of the subspaces are of the same order. This condition is satisfied in many practical applications, such as, e.g., for the face clustering and the handwritten digit clustering problems described in Section 5. We believe the occurrence of the factor $\sqrt{d_{\max}/d_{\min}}$ in (9) to be an artifact of our proof technique. Also note that Theorem 3 imposes more restrictive conditions on $\mathbf{\Phi}$ than Theorems 1 and 2, namely the distribution of $\mathbf{\Phi}$ has to be rotationally invariant. This is a technical condition and it is not implied by (4). Examples of rotationally invariant matrices satisfying (4) include matrices with i.i.d. $\mathcal{N}(0, 1/p)$ entries.

Theorems 1–3 apply even when the subspaces $\mathcal{S}_\ell$ span the ambient space $\mathbb{R}^m$. This follows by virtue of our clustering conditions depending only on the pairwise affinities between subspaces, and pairwise affinities changing only moderately if the dimensionality is reduced down to no more than the order of the individual subspace dimensions.

Theorems 1–3 show that for all three algorithms, p may be taken to be linear (up to log-factors) in $d_{\max}$. We can therefore conclude that the dimensionality of the data set $\mathcal{Y}$ can be reduced down to the order of the largest subspace dimension without affecting clustering performance significantly. This has important practical ramifications as, for all three algorithms considered, the computational complexity associated with the construction of the adjacency matrix is essentially linear in the dimension of the ambient space the data points "live in". To get an idea of the resulting overall complexity savings, let us consider the TSC algorithm and assume that the (high-dimensional) data set $\mathcal{Y} \subset \mathbb{R}^m$ is projected down to $\mathbb{R}^p$, with $p = O(d_{\max}\log^2(N))$, via a Gaussian random projection; this choice of p guarantees, by Theorem 1, that clustering performance is not affected significantly by dimensionality reduction. The complexity associated with the construction of the adjacency matrix for TSC is given by the cost of computing the inner products between all pairs of data points, and is therefore $O(mN^2)$ for the original data set $\mathcal{Y} \subset \mathbb{R}^m$ and $O(pN^2)$ for the projected data set $\mathcal{X} \subset \mathbb{R}^p$. Adding the cost for applying the Gaussian random projection results in an overall cost of $O(pN^2) + O(pNm) = O(d_{\max}\log^2(N)N(N+m))$ for building the

[2]While the statement holds irrespectively of $s_{\max}$, recall from Section 2 that choosing $s_{\max}$ too small may result in too few non-zeros in the adjacency matrix $\mathbf{A}$ for successful clustering.

adjacency matrix associated with $\mathcal{X}$. The resulting complexity savings for TSC are therefore given by $O(\min(m, N)/(d_{\max} \log^2(N))$. The absolute run-time savings are even more pronounced for SSC-OMP and SSC, as the corresponding costs for building the adjacency matrix is larger than $O(mN^2)$. Further gains can be obtained by employing fast random projections [1].

Dimensionality reduction affects the computational cost associated with the construction of the adjacency matrix only. The spectral clustering step, which when naïvely implemented has complexity $O(N^3)$, may be the dominating factor in the overall computational cost, in particular when m is small relative to N^3. Notwithstanding, dimensionality reduction can still lead to significant total run-time savings. Our numerical results in Section 5 demonstrate this for SSC. To see savings on the same order for SSC-OMP and TSC, we would have to consider problems with N smaller relative to m.

The probability lower bounds in Theorems 1–3 are independent of p and m and require the total number of data points N to be large in absolute terms in order to ensure a success probability close to one.

Theorems 1–3 are order-optimal in the following sense. If dimensionality is reduced to below $d_{\max}$, then, in general, there are points from different subspaces that are projected into the same lower-dimensional subspace, which renders the resulting clustering problem fundamentally ill-posed. To see this, take $d_\ell = d$, for all ℓ, and assume that $p \leq d$. Next, note that the (randomly projected) points $\mathcal{X}_\ell$ lie in the column span of $\mathbf{\Phi}\mathbf{U}^{(\ell)}$. As $\mathbf{U}^{(\ell)}$ is a basis for the d_ℓ-dimensional subspace $\mathcal{S}_\ell \subset \mathbb{R}^m$, the span of $\mathbf{\Phi}\mathbf{U}^{(\ell)}$ is $\mathbb{R}^p$, for all ℓ, and therefore all points in the projected data set $\mathcal{X} = \mathcal{X}_1 \cup ... \cup \mathcal{X}_L$ lie in the same p-dimensional subspace, which renders the clustering problem ill-posed.

We next compare the clustering conditions (7), (8), and (9) in Theorems 1, 2, and 3 with their counterparts for clustering of the original, high-dimensional data set $\mathcal{Y}$. Specifically, such reference conditions can be found in [16, Thm. 2] for TSC and in [28, Thm. 2.8] for SSC, but do not seem to be available for SSC-OMP for the statistical data model considered in this paper. However, setting $\mathbf{\Phi} = \mathbf{I}$ in the proof of Theorem 3, we can easily get a reference condition for SSC-OMP. Rather than providing the details of this simple modification, we refer the reader to the proof in [31, Chap. 4].

Corollary 1. *Let $\rho_\ell \coloneqq (n_\ell - 1)/d_\ell$, $\ell \in [L]$, and suppose that $\rho_{\min} \coloneqq \min_\ell \rho_\ell \geq \rho_0$, where $\rho_0 > 1$ is a numerical constant. If*

$$\max_{k,\ell:\, k \neq \ell} \operatorname{aff}(\mathcal{S}_k, \mathcal{S}_\ell) \leq \frac{\sqrt{\log \rho_{\min}}}{64 \log N}, \tag{10}$$

then the graph G obtained by applying SSC-OMP to the original, high-dimensional data set $\mathcal{Y}$ has no false connections with probability at least $1 - 2N^{-1} - \sum_{\ell=1}^{L} n_\ell e^{-\sqrt{\rho_\ell} d_\ell}$.

We conclude that for all three algorithms the impact of dimensionality reduction is essentially quantified through a term proportional to $\sqrt{d_{\max}/p}$ that adds to the maximum affinity between the subspaces $\mathcal{S}_\ell$ in the clustering conditions (7), (8), and (9). These clustering conditions nicely reflect the intuition that the smaller the affinities between the subspaces $\mathcal{S}_\ell$, the more aggressively we can reduce the dimensionality of the data set without compromising clustering performance.

As the result in Corollary 1 is new, a few comments on its relation to existing results, specifically those in [7] and [37], are in order. Corollary 1 imposes less restrictive conditions on the relative orientations of the subspaces than [7, Thm. 3], [37, Thm. 2, Cor. 1], but makes stronger assumptions on the data model. The result in [37, Thm. 3] applies to subspaces with random orientations, and therefore does not allow for statements involving subspace affinities. We refer the reader to the thesis [31, Sec. 4.1] for a more detailed comparison of Corollary 1 above to [7, Thm. 3]. Finally,

numerical results corroborating the fundamental nature of the clustering condition (10) can be found in [31, Sec. 5.1].

4 Impact of noise

In many practical applications the data points to be clustered are corrupted by noise, typically modeled as additive Gaussian noise. In this section, we study the interplay between dimensionality reduction and additive noise for the TSC algorithm. Specifically, we let the high-dimensional data points be corrupted by Gaussian noise according to

$$\tilde{\mathbf{y}}_i^{(\ell)} = \mathbf{y}_i^{(\ell)} + \mathbf{e}_i^{(\ell)},$$

where $\mathbf{e}_i^{(\ell)} \sim \mathcal{N}(0, (\sigma^2/m)\mathbf{I})$, and assume, as before, that $\mathbf{y}_i^{(\ell)}$ is drawn i.i.d. uniformly from the intersection of the d_ℓ-dimensional subspace $\mathcal{S}_\ell$ with the unit sphere. The dimensionality-reduced noisy data set $\tilde{\mathcal{X}} \subset \mathbb{R}^p$ is obtained by applying the same realization of the random projection matrix $\mathbf{\Phi} \in \mathbb{R}^{p\times m}$ to all (noisy) data points $\tilde{\mathbf{y}}_i^{(\ell)}$. The elements of the sets $\tilde{\mathcal{X}}_\ell$ in $\tilde{\mathcal{X}} = \tilde{\mathcal{X}}_1 \cup \ldots \cup \tilde{\mathcal{X}}_L$ are hence given by

$$\tilde{\mathbf{x}}_j^{(\ell)} = \mathbf{\Phi}(\mathbf{y}_i^{(\ell)} + \mathbf{e}_i^{(\ell)}), \quad j \in [n_\ell]. \tag{11}$$

Theorem 4. *Choose q such that $q \leq \min_\ell n_\ell/6$, and let $m \geq 6 \log N$. If*

$$\max_{k,\ell:\, k\neq\ell} \operatorname{aff}(\mathcal{S}_k, \mathcal{S}_\ell) + \frac{\sqrt{11}}{\sqrt{3\tilde{c}}} \frac{\sqrt{d_{\max}}}{\sqrt{p}} + \frac{\sigma(1+\sigma)\sqrt{6}}{\sqrt{\bar{c}\log N}} \frac{\sqrt{d_{\max}}}{\sqrt{p}} \leq \frac{1}{15 \log N}, \tag{12}$$

where $d_{\max} = \max_\ell d_\ell$ and $\bar{c} = \min(6, \tilde{c})$ with $\tilde{c}$ the constant in the concentration inequality (4), then the graph G obtained by applying TSC to $\tilde{\mathcal{X}}$ has no false connections with probability at least $1 - 14N^{-1} - 2Ne^{-m} - \sum_{\ell=1}^{L} n_\ell e^{-c(n_\ell - 1)}$, where $c > 1/20$ is a numerical constant.

Theorem 4 states that in the noisy case—just as in the noiseless case—TSC succeeds for p as small as $d_{\max}$, order-wise, provided that the affinities between the subspaces $\mathcal{S}_\ell$ are sufficiently small and $\tilde{\mathcal{X}}$ contains sufficiently many points from each subspace. More specifically, comparing the noiseless clustering condition (7) to (12), we can see that the impact of noise is simply to add the offset $\frac{\sigma(1+\sigma)\sqrt{6}}{\sqrt{\bar{c}\log N}} \frac{\sqrt{d_{\max}}}{\sqrt{p}}$ to the LHS of the clustering condition. For fixed σ, owing to the factor $\sqrt{d_{\max}/p}$, the impact of noise on the effective affinity as quantified by the LHS of (12) becomes more pronounced when the dimensionality is reduced more aggressively.

Theorem 4 continues to hold (with $\bar{c}$ in the term $\frac{\sigma(1+\sigma)\sqrt{6}}{\sqrt{\bar{c}\log N}} \frac{\sqrt{d_{\max}}}{\sqrt{p}}$ replaced by a numerical constant, and e^{-m} in the success probability replaced by e^{-m}), if noise $\tilde{\mathbf{e}}_i^{(\ell)} \sim \mathcal{N}(0, (\sigma^2/p)\mathbf{I})$ is added *after* random projection according to $\tilde{\mathbf{x}}_j^{(\ell)} = \mathbf{\Phi}\mathbf{y}_i^{(\ell)} + \tilde{\mathbf{e}}_i^{(\ell)}$. This is not surprising, as the absolute amount of noise injected remains the same, i.e., $\mathbb{E}\left[\left\|\tilde{\mathbf{e}}_i^{(\ell)}\right\|_2^2\right] = \mathbb{E}\left[\left\|\mathbf{e}_i^{(\ell)}\right\|_2^2\right] = \sigma^2$.

We finally note that an approach similar to that used for TSC can be applied to extend our result for SSC-OMP to the noisy case resulting in clustering conditions analogous to those for TSC. The corresponding technical details are, however, significantly more involved and cumbersome. We therefore decided not to state the formal result. We expect that a similar result can be proven for (a robust version of) SSC as our simulation results in Section 5.1.2 indicate that the qualitative behavior of all three algorithms in the presence of noise is essentially identical, and, in addition, is qualitatively accurately predicted by Theorem 4.

5 Numerical Results

We evaluate the impact of dimensionality reduction on the clustering error (CE), i.e., the fraction of misclustered points, for TSC, SSC, and SSC-OMP applied to synthetic data as well as to publicly available standard data sets widely used in the subspace clustering literature. Specifically, we consider the problems of clustering faces, handwritten digits, and gene expression data. All three algorithms, TSC, SSC, and SSC-OMP, were observed to tolerate massive dimensionality reduction in all experiments. The performance ranking of the three algorithms according to CE varies considerably across data sets. Specifically, in order to demonstrate that none of the algorithms uniformly outperforms the others, we chose to report the results for all three data sets. We also compare the algorithms in terms of their running times on a PC with 32 GB RAM and 8-core Intel Core i7-3770K CPU clocked at 3.50 GHz.

TSC and SSC-OMP were implemented in Matlab following the specifications in Section 2. For SSC, we used the Matlab implementation provided in [9], which is based on Lasso (instead of ℓ_1-minimization) and uses the Alternating Direction Method of Multipliers (ADMM). Code to reproduce the experiments in this section is available at `http://www.nari.ee.ethz.ch/commth/research/`. Information on the number of Monte Carlo runs used in our experiments is contained in this Matlab code.

Unless stated otherwise, we select the Lasso parameter λ in SSC from the set $\{0.001, 0.002, 0.004, 0.008, 0.01, 0.02, 0.04, 0.08, 0.1, 0.2\}$ such that the lowest clustering error is obtained on the original high-dimensional data set $\mathcal{Y}$. The parameters q and $s_{\max}$ for TSC and SSC-OMP, respectively, are chosen analogously from the set $\{2, 4, \ldots, 18\}$. Although these parameter selection procedures may not yield the optimum parameters for the projected data set $\mathcal{X}$ for all realizations of $\mathbf{\Phi}$, we desist from selecting the parameters for every realization of $\mathbf{\Phi}$ individually as this may lead to overly optimistic results.

As projection matrices we consider i.i.d. $\mathcal{N}(0, 1/p)$ Gaussian random matrices (referred to as GRP) and fast random projection (FRP) matrices [1] given by the real part of $\mathbf{FD} \in \mathbb{C}^{p \times m}$, where $\mathbf{D} \in \mathbb{R}^{m \times m}$ is diagonal with main diagonal elements drawn i.i.d. uniformly from $\{-1, 1\}$, and $\mathbf{F} \in \mathbb{C}^{p \times m}$ is obtained by choosing a set of p rows uniformly at random from the rows of an $m \times m$ discrete Fourier transform (DFT) matrix. In all experiments the dimensionality-reduced data set $\mathcal{X}$ is obtained by applying the (same) realization of either a GRP or an FRP matrix to all data points in $\mathcal{Y}$. The FRP can be implemented efficiently by premultiplying $\mathcal{Y}$ by $\mathbf{D}$ and then applying the FFT to each data point. With regards to storage space, we note that the FRP only requires the storage of a binary m-dimensional vector (namely the diagonal entries of $\mathbf{D}$), in contrast to mp real numbers for GRPs.

5.1 Synthetic data

5.1.1 Comparison of TSC, SSC, and SSC-OMP

We use the data model described in Section 3 with $m = 2^{15} = 32768$ and generate $L = 3$ subspaces $\mathcal{S}_\ell$ of $\mathbb{R}^m$ of dimension $d = 20$ at random such that every pair of subspaces intersects in at least r dimensions; this implies $\operatorname{aff}(\mathcal{S}_k, \mathcal{S}_\ell) \geq \sqrt{r/d}$, for all $k, \ell \in [L], k \neq \ell$. More specifically, we take the basis matrices to be given by $\mathbf{U}^{(\ell)} = [\mathbf{U} \;\; \tilde{\mathbf{U}}^{(\ell)}]$, where $\mathbf{U} \in \mathbb{R}^{m \times r}$ and the $\tilde{\mathbf{U}}^{(\ell)} \in \mathbb{R}^{m \times (d-r)}$, $\ell \in [L]$, are chosen uniformly at random among all orthonormal matrices of dimensions $m \times r$ and $m \times (d - r)$, respectively. We sample $n_\ell = 80$ data points, for each $\ell \in [L]$, resulting in a total of $N = 240$ data points.

In Figure 1, we plot the CE as a function of p for TSC, SSC, and SSC-OMP applied to the

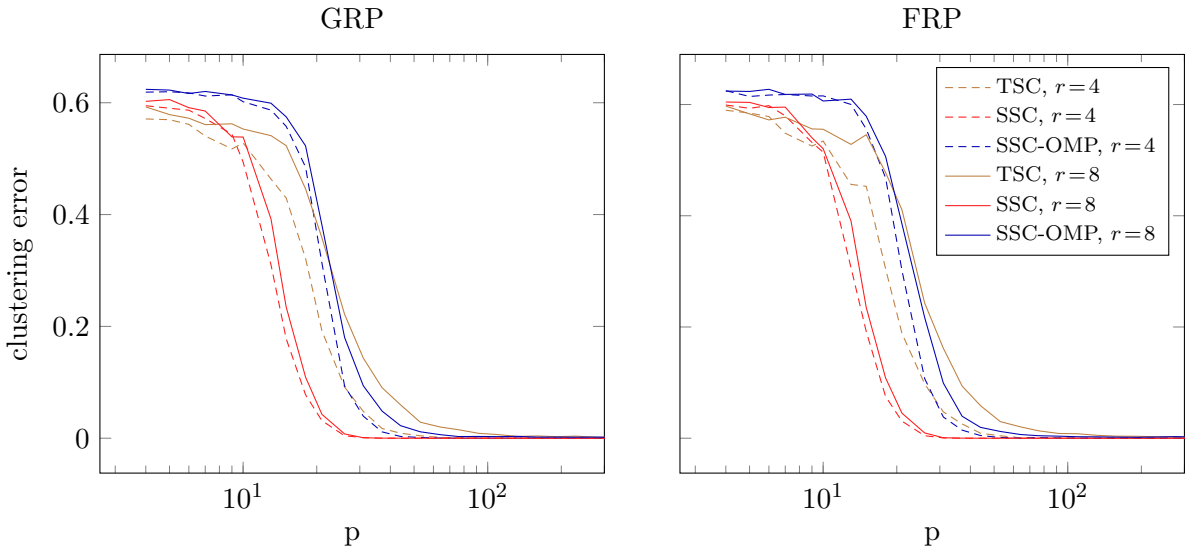

Figure 1: Clustering error for synthetic data as a function of p using GRP (left) and FRP (right). Recall that for $r = 4$ and $r = 8$ we have $\mathrm{aff}(\mathcal{S}_k, \mathcal{S}_\ell) \geq \sqrt{1/5}$ and $\mathrm{aff}(\mathcal{S}_k, \mathcal{S}_\ell) \geq \sqrt{2/5}$, respectively, for all $k, \ell \in [L]$, $k \neq \ell$.

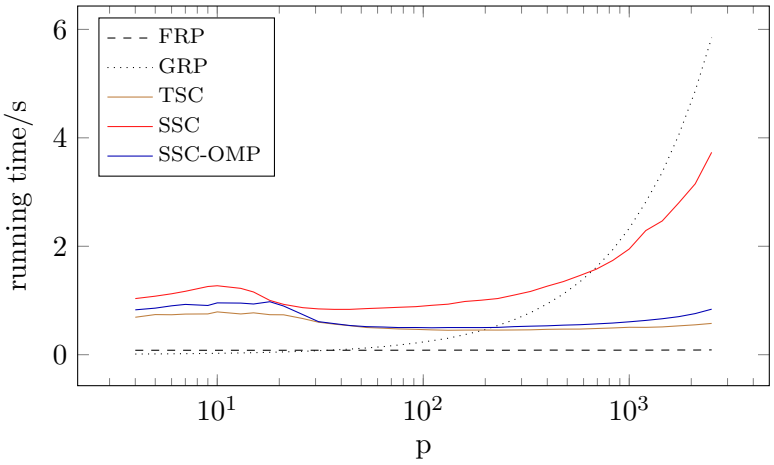

Figure 2: Running times (in seconds) for clustering synthetic data.

dimensionality-reduced data set $\mathcal{X}$ with $r = 4$ and $r = 8$. Figure 2 shows the running times corresponding to the application of the FRP and the GRP matrix to the (entire) data set $\mathcal{Y}$ along with the running times of the clustering algorithms alone.

The results show, as predicted by Theorems 1–3, that TSC, SSC, and SSC-OMP, indeed, succeed provided that $\sqrt{d/p}$ is sufficiently small. Specifically, we observe a transition to CE ≈ 0 for p between 20 and 100. As the subspaces $\mathcal{S}_\ell$ are of dimension 20 this corroborates the fact that the dimensionality of the data can be reduced down to the dimension of the subspaces without compromising clustering performance significantly. Equivalently, we accomplish a dimensionality reduction by a factor of about 1600–320.

For all three algorithms the numerical results further confirm the tradeoff between the affinities

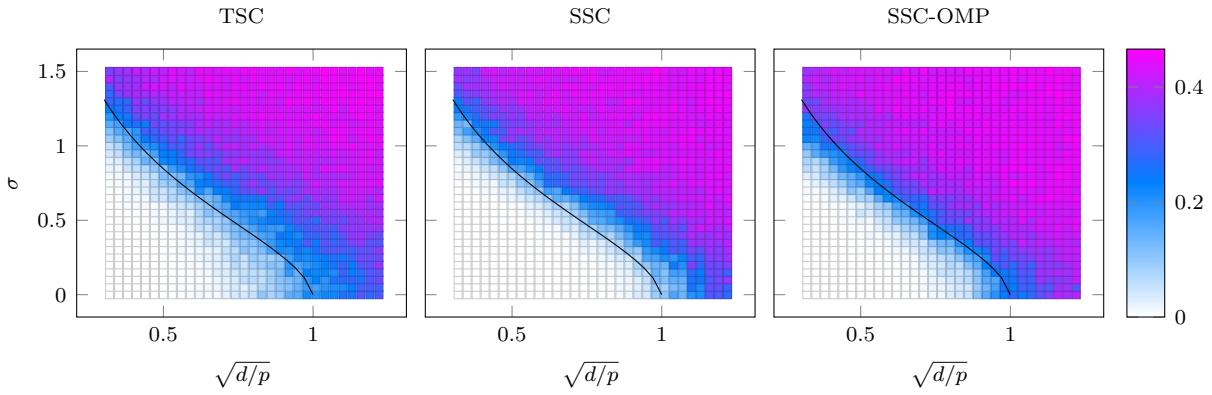

Figure 3: CE (color coded) as a function of $\sqrt{d/p}$ and σ for $L = 2$ orthogonal subspaces of $\mathbb{R}^{100}$. The black lines correspond to the curve $\sqrt{d/p}(0.8 + \sigma(0.1 + \sigma)) = 0.8$, and roughly separate the regimes where clustering succeeds from that where it fails.

of the $\mathcal{S}_\ell$ and the amount of dimensionality reduction possible as quantified by the clustering conditions (7), (8), and (9). Specifically, the CE increases as r and hence $\text{aff}(\mathcal{S}_k, \mathcal{S}_\ell)$ increases. In this example, SSC consistently outperforms TSC and SSC-OMP, albeit at the cost of significantly longer running time (see Figure 2). While the running time of SSC exhibits very pronounced increasing behavior in p, that of SSC-OMP shows much less pronounced increases, and that of TSC does not increase notably in p. It is furthermore interesting to see that the clustering performance is essentially identical for FRP and GRP. This is remarkable as the application of FRP requires only $O(m \log m)$ operations (per data point) and therefore its running time does not depend on p. Application of the GRP, in contrast, requires $O(mp)$ operations (per data point), which results in a running time that is linear in p.

5.1.2 Impact of noise

In the next experiment we study the interplay between noise and dimensionality reduction. We use the data model described in Section 3 with $m = 100$ and generate $L = 2$ orthogonal subspaces $\mathcal{S}_\ell$ of $\mathbb{R}^m$ of dimension $d = 10$. This ensures that the affinity between the subspaces equals 0 (fixing the affinity to some other constant would not change the qualitative conclusions). We generate the noisy data set $\tilde{\mathcal{Y}}$ by sampling $n_\ell = 30$ points from each of the two subspaces and adding $\mathcal{N}(0, (\sigma^2/m)\mathbf{I})$ noise. Figure 3 shows the CE as a function of $\sqrt{d/p}$ and σ for dimensionality reduction via GRP.

The clustering condition in Theorem 4 guarantees that TSC succeeds as long as $\sqrt{d/p}(c_1 + \sigma(c_2 + \sigma)) \leq c_3$, where c_1, c_2, c_3 are independent of d, p, m, and σ^2. In order to find out whether this sufficient condition predicts the fundamental clustering behavior qualitatively correctly, we test whether a phase transition, separating the region where clustering succeeds from that where it fails, indeed, occurs at

$$\sqrt{d/p}(c_1 + \sigma(c_2 + \sigma)) = c_3. \tag{13}$$

To this end, we fit (13)—by choosing c_1, c_2, c_3—into the plots in Figure 3 and observe that the answer is in the affirmative. Moreover, our numerical results show that the phase transition behavior of SSC and SSC-OMP is essentially identical to that of TSC, which provides evidence for SSC and SSC-OMP behaving similarly to TSC in the noisy case.

5.1.3 Dimensionality reduction when the subspaces span the ambient space

As noted in Section 3, Theorems 1–4 indicate that dimensionality reduction down to the order of the subspace dimensions is possible even when the subspaces $\mathcal{S}_\ell$ span the ambient space $\mathbb{R}^m$. To verify this observation empirically, we perform the following experiment. We draw a random Gaussian matrix $\mathbf{V} \in \mathbb{R}^{200\times 200}$. With probability one, the columns of $\mathbf{V}$ span $\mathbb{R}^{200}$. We then extract the 200×20 matrices $\mathbf{V}^{(\ell)}$ from $\mathbf{V}$ according to $[\mathbf{V}^{(1)} \ \ldots \ \mathbf{V}^{(10)}] = \mathbf{V}$, and let the subspace $\mathcal{S}_\ell$ be given by the span of $\mathbf{V}^{(\ell)}, \ell = 1, ..., 10$. This guarantees that the union of the $\mathcal{S}_\ell$ span $\mathbb{R}^{200}$. Note, however, that the affinities between pairs of the resulting subspaces will be small with high probability. We again use the data model described in Section 3 and sample $n_\ell = 60$ points on $\mathcal{S}_\ell \cap \mathbb{S}^{d_\ell - 1}$, for all $\ell \in [L]$, to obtain a data set $\mathcal{Y}$ with a total of $N = 600$ points. We select the values for q, λ, and $s_{\max}$ that yield the lowest CE for the majority of values for p.

Figure 4 shows the CE as a function of p for TSC, SSC, and SSC-OMP. The CE starts to be non-zero for $p < 60$ for TSC and SSC-OMP, and for $p < 40$ for SSC. We therefore conclude that the dimensionality can, indeed, be reduced, quite significantly, even when the subspaces span the ambient space, as indicated by Theorems 1–4.

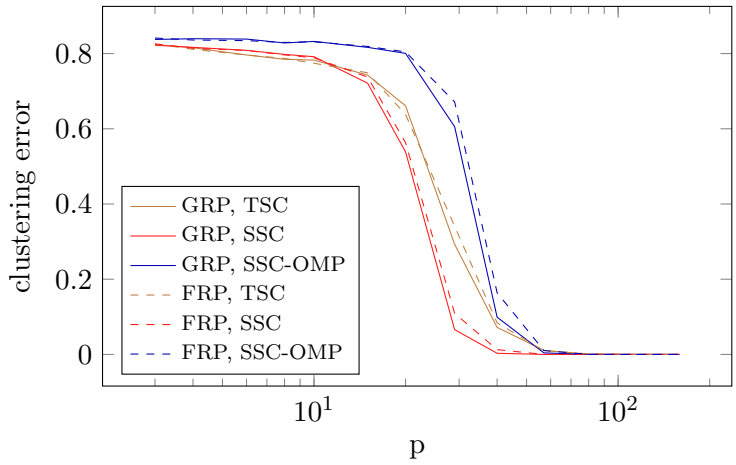

Figure 4: CE as a function of p for $L = 10$ subspaces that collectively span the ambient space $\mathbb{R}^{200}$.

5.2 Clustering faces

We next evaluate the impact of dimensionality reduction in the problem of clustering face images taken from the Extended Yale B data set [11, 24], which contains 192×168 pixel ($m = 32256$) frontal face images of 38 individuals, with 64 images per individual, each acquired under different illumination conditions. The motivation for applying subspace clustering algorithms 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 $\mathcal{S}_\ell$ would then contain the images corresponding to a given person.

We generate $\mathcal{Y}$ by first selecting a subset of $L = 2$ individuals uniformly at random from the set of all $\binom{38}{2}$ pairs and then collecting all images corresponding to the two selected individuals. In Figure 5, we plot the corresponding CE and the running times as a function of p. Again, for each p, the CE and the running times are obtained by averaging over 500 problem instances generated by randomly choosing 100 instances of $\mathcal{Y}$ and 5 realizations of the projection matrix per chosen data set $\mathcal{Y}$. In contrast to the preceding experiment, here, SSC-OMP consistently outperforms TSC and

SSC. For all three algorithms the dimensionality of the data can be reduced by a factor of about 100 without notably increasing the CE. Note, however, that in this experiment the dimensionality cannot be reduced as aggressively as in the preceding synthetic data experiment. Specifically, here the data points lie in 9-dimensional subspaces and dimensionality reduction by a factor of 100 corresponds to $p \approx 322$. One possible explanation for this observation is that the principal angles between the subspaces spanned by the face images of different subjects are typically small (see [9, Sec. 7]), which means that the subspace affinities in this data set are large. The conclusions regarding running times and choice of the random projection matrix are analogous to those reported for synthetic data above.

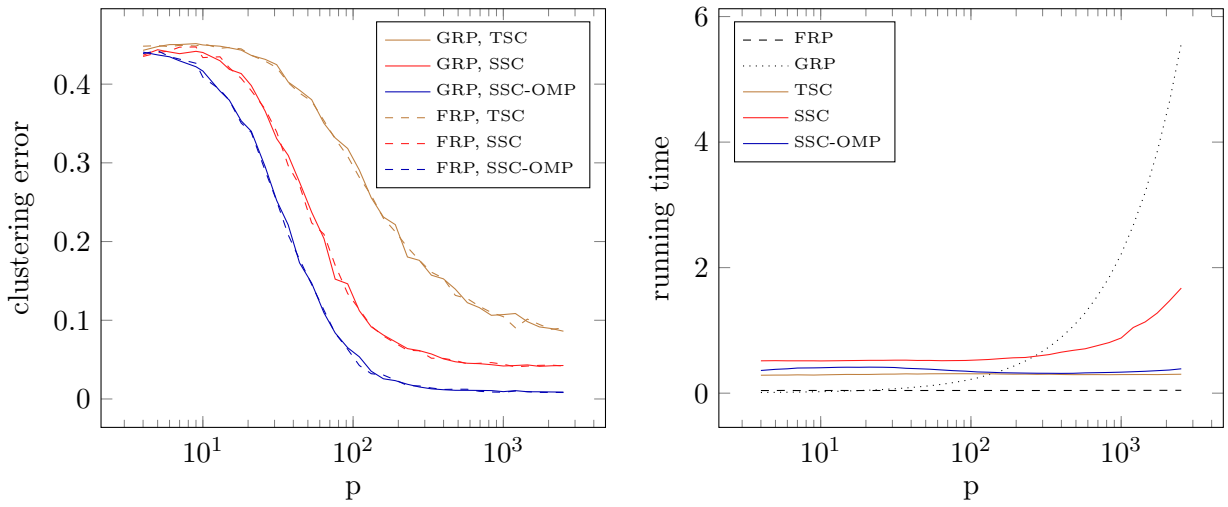

Figure 5: Clustering error and running times (in seconds) for clustering $L = 2$ faces from the Extended Yale B data set.

5.3 Clustering handwritten digits

In this experiment, we investigate the impact of dimensionality reduction in the context of clustering images of handwritten digits. We use the MNIST data set [23] containing $10,000$ images of (horizontally and vertically) aligned handwritten digits of size 28×28 pixels ($m = 784$). The motivation for employing subspace clustering in this context stems from the observation that vectorized images of different handwritten versions of the same digit tend to lie near a low-dimensional subspace [14].

We generate the data sets $\mathcal{Y}$ by selecting 250 images (out of 1000) uniformly at random from each of the sets corresponding to the digits 2, 4, and 8. There is no specific reason for our choice of the digits 2, 4, and 8; other combinations of three digits yield similar results. However, some combinations of digits are more difficult to cluster than others, e.g., 1 and 7 are "closer" (in terms of the affinities between the subspaces the corresponding images approximately lie in) than 1 and 8; clustering 1 and 7 therefore typically results in a larger error than clustering 1 and 8. The results depicted in Figure 6 show that the dimensionality of the data set can be reduced from $m = 784$ to $p = 200$, i.e., by a factor of 3.9, without notably increasing the CE incurred by TSC and SSC. For sufficiently large p, TSC yields a slightly lower clustering error than SSC. SSC-OMP is outperformed considerably by the other two algorithms.

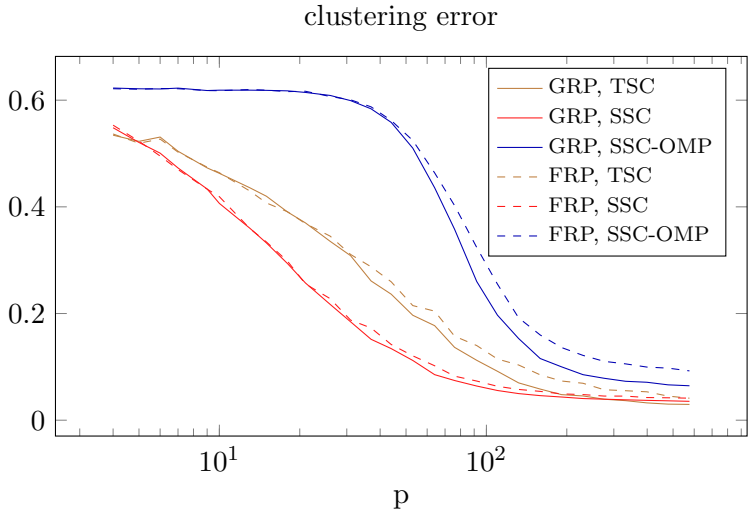

Figure 6: Clustering error for handwritten digits 2,4, and 8 from the MNIST data set.

5.4 Clustering gene expression data

Finally, we consider clustering of gene expression level data–originating from different types of cancer cells–according to cancer type. This problem is of significant practical relevance as it helps, inter alia, to identify genes that are involved in the same cellular process [20]. The use of subspace clustering in this context was suggested in [20]. We use the publicly available Novartis multi-tissue data set from the Broad Institute Cancer Program database [5]. This data set contains the 1000-dimensional gene expression level data of $n = 103$ tissue samples taken from $L = 4$ different cancer types. In order to illustrate that the gene expression level vectors of a single cancer type, indeed, lie near a low-dimensional subspace, we plot, in Figure 7, the singular values of the data matrices corresponding to a single cancer type. We observe that the singular values decay rapidly and for every cancer type, more than 94% of the energy of the corresponding data vectors is concentrated in a 6-dimensional subspace of the 1000-dimensional ambient space.

We cluster all $n = 103$ available samples. The CE obtained by averaging, for each p, over 200 realizations of the random projection matrix is shown in Figure 7. For $p \approx 100$, which corresponds to dimensionality reduction by a factor of 10, the CEs of TSC and SSC are comparable to those obtained when operating on the original high-dimensional data set. SSC is seen to consistently (across p) perform best, followed by TSC and SSC-OMP. As in previous experiments the CEs observed for GRP and FRP, for each of the three algorithms, are virtually identical.

Acknowledgments

The authors would like to thank Robert Calderbank for very helpful comments on an earlier version of this manuscript.

A Proof of Theorems 1 and 4

The proof idea for Theorem 1 is to turn the effect of the random projection into an additive perturbation and to show that this perturbation is small for all values of p down to the order of

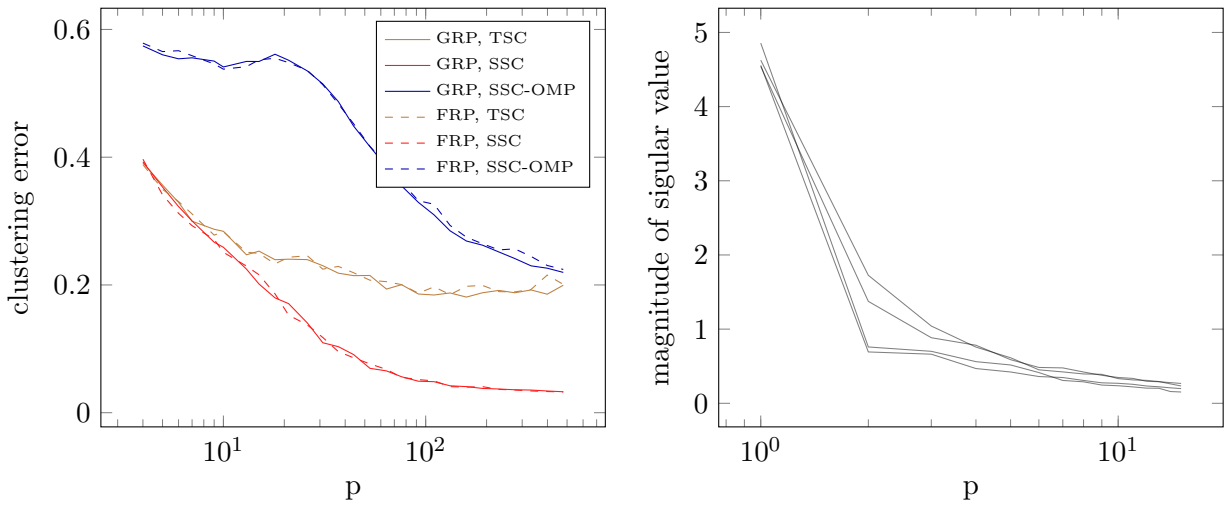

Figure 7: Clustering error for gene expression level data of $L = 4$ cancer types (left). Singular values of data matrices corresponding to a single cancer type (right).

$d_{\max}$. In the noisy case, addressed by Theorem 4, we have an additional perturbation due to noise. We detail the proof of the more general Theorem 4 below, and explain in Appendix A.3 the simple changes that yield Theorem 1. The proof of Theorem 4 follows closely that of [16, Thm. 3], which quantifies the performance of TSC under additive Gaussian noise alone. We therefore elaborate only on the steps that are new relative to [16] and encourage the interested reader to consult [16] for the arguments not repeated here.

The graph G obtained by applying TSC to the dimensionality-reduced noisy data set $\tilde{\mathcal{X}}$ has no false connections, i.e., each $\tilde{\mathbf{x}}_i^{(\ell)}$ is connected to points in $\tilde{\mathcal{X}}_\ell$ only, if for each $\tilde{\mathbf{x}}_i^{(\ell)} \in \tilde{\mathcal{X}}_\ell$ the associated set S_i corresponds to points in $\tilde{\mathcal{X}}_\ell$ only, for all ℓ. This is the case if

$$z_{(n_\ell - q)}^{(\ell)} > \max_{k \neq \ell, j} z_j^{(k)}, \tag{14}$$

where $z_j^{(k)} := \left|\langle \tilde{\mathbf{x}}_j^{(k)}, \tilde{\mathbf{x}}_i^{(\ell)} \rangle\right|$ and $z_{(1)}^{(\ell)} \leq z_{(2)}^{(\ell)} \leq \ldots \leq z_{(n_\ell - 1)}^{(\ell)}$ are the order statistics of $\{z_j^{(\ell)}\}_{j \in [n_\ell] \setminus \{i\}}$ and $\max_{k \neq \ell, j}$ denotes maximization over $k \in [L]$, $k \neq \ell$, and over the indices j of the corresponding points $\tilde{\mathbf{x}}_j^{(k)} \in \tilde{\mathcal{X}}_k$. Note that, for simplicity of exposition, the notation $z_j^{(k)}$ does not reflect dependence on $\tilde{\mathbf{x}}_i^{(\ell)}$. The proof is established by upper-bounding the probability of (14) being violated for a given data point $\tilde{\mathbf{x}}_i^{(\ell)}$. A union bound over all N points $\tilde{\mathbf{x}}_i^{(\ell)}, i \in [n_\ell], \ell \in [L]$, then yields the final result. We start by setting $\bar{z}_j^{(k)} := \left|\left\langle \mathbf{y}_j^{(k)}, \mathbf{y}_i^{(\ell)} \right\rangle\right|$, where $\mathbf{y}_j^{(k)} = \mathbf{U}^{(k)} \mathbf{a}_j^{(k)}$ are the original data points in the (high-dimensional) space $\mathbb{R}^m$, and noting that $z_j^{(k)} = \left|\left\langle \tilde{\mathbf{x}}_j^{(k)}, \tilde{\mathbf{x}}_i^{(\ell)} \right\rangle\right| = \left|\left\langle \mathbf{y}_j^{(k)}, \mathbf{y}_i^{(\ell)} \right\rangle + e_j^{(k)}\right|$,

where we defined the "distortion"

$$\begin{aligned} e_j^{(k)} &\coloneqq \left\langle \boldsymbol{\Phi}\tilde{\mathbf{y}}_j^{(k)}, \boldsymbol{\Phi}\tilde{\mathbf{y}}_i^{(\ell)} \right\rangle - \left\langle \mathbf{y}_j^{(k)}, \mathbf{y}_i^{(\ell)} \right\rangle \\ &= \left\langle (\boldsymbol{\Phi}^T\boldsymbol{\Phi} - \mathbf{I})\mathbf{y}_j^{(k)}, \mathbf{y}_i^{(\ell)} \right\rangle + \left\langle \boldsymbol{\Phi}\mathbf{y}_j^{(k)}, \boldsymbol{\Phi}\mathbf{e}_i^{(\ell)} \right\rangle + \left\langle \boldsymbol{\Phi}\mathbf{e}_j^{(k)}, \boldsymbol{\Phi}\mathbf{y}_i^{(\ell)} \right\rangle + \left\langle \boldsymbol{\Phi}\mathbf{e}_j^{(k)}, \boldsymbol{\Phi}\mathbf{e}_i^{(\ell)} \right\rangle \\ &= \underbrace{\left\langle {\mathbf{U}^{(\ell)}}^T(\boldsymbol{\Phi}^T\boldsymbol{\Phi} - \mathbf{I})\mathbf{U}^{(k)}\mathbf{a}_j^{(k)}, \mathbf{a}_i^{(\ell)} \right\rangle}_{\bar{e}_j^{(k)}} \\ &\quad + \underbrace{\left\langle \boldsymbol{\Phi}^T\boldsymbol{\Phi}\mathbf{y}_j^{(k)}, \mathbf{e}_i^{(\ell)} \right\rangle + \left\langle \mathbf{e}_j^{(k)}, \boldsymbol{\Phi}^T\boldsymbol{\Phi}\mathbf{y}_i^{(\ell)} \right\rangle + \left\langle \boldsymbol{\Phi}^T\boldsymbol{\Phi}\mathbf{e}_j^{(k)}, \mathbf{e}_i^{(\ell)} \right\rangle}_{\tilde{e}_j^{(k)}}. \end{aligned} \tag{15}$$

Here, the term $\bar{e}_j^{(k)}$ accounts for the perturbation caused by random projection, and $\tilde{e}_j^{(k)}$ corresponds to the perturbation caused by noise. The probability of (14) being violated can now be upper-bounded according to

$$\begin{aligned} \mathrm{P}\left[z_{(n_\ell - q)}^{(\ell)} \leq \max_{k\neq\ell, j} z_j^{(k)} \right] &\leq \mathrm{P}\left[\bar{z}_{(n_\ell - q)}^{(\ell)} \leq \frac{2}{3\sqrt{d_\ell}} \right] \\ &\quad + \mathrm{P}\left[\max_{k\neq\ell, j} \bar{z}_j^{(k)} \geq \alpha \right] + \mathrm{P}\left[\max_{(j,k)\neq(i,\ell)} \left| e_j^{(k)} \right| \geq \epsilon \right], \end{aligned} \tag{16}$$

where we set

$$\alpha \coloneqq \frac{4\sqrt{6}\log N}{\sqrt{d_\ell}} \max_{k\neq\ell} \frac{1}{\sqrt{d_k}} \left\| {\mathbf{U}^{(k)}}^T \mathbf{U}^{(\ell)} \right\|_F, \quad \epsilon \coloneqq \frac{\beta}{\sqrt{d_\ell}}\delta + \beta\frac{2\sigma(1+\sigma)}{\sqrt{m}}\delta' \tag{17}$$

with $\beta \coloneqq \sqrt{6\log N}$, $\delta \coloneqq \frac{\sqrt{28 d_{\max} + 8\log L + 8\log N}}{\sqrt{3\bar{c}p}}$, and $\delta' \coloneqq \sqrt{\frac{6m}{\bar{c}p}}$, and assumed

$$\alpha + 2\epsilon \leq \frac{2}{3\sqrt{d_\ell}}. \tag{18}$$

We refer the reader to [16, Proof of Thm. 3, Eq. (40)] for an explanation of the steps leading to (16) (while [16, Eq. (40)] is not completely equivalent to (16), the steps leading to (16) are essentially identical). Resolving the assumption (18) leads to

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

which is implied by (12) (using that $\sqrt{28 d_{\max} + 8\log L + 8\log N}/\sqrt{\log N} \leq \sqrt{44 d_{\max}}$ because $\log L/\log N \leq 1$, $d_{\max} \geq 1$, and $\log N > 1$ for $N \geq 3$). With ϵ as defined in (17), and the triangle inequality, it follows that $\max_{(j,k)\neq(i,\ell)} \left| e_j^{(k)} \right| \geq \epsilon$ implies that either $\max_{(j,k)\neq(i,\ell)} \left| \bar{e}_j^{(k)} \right| \geq \frac{\beta}{\sqrt{d_\ell}}\delta$ or $\max_{(j,k)\neq(i,\ell)} \left| \tilde{e}_j^{(k)} \right| \geq \beta\frac{2\sigma(1+\sigma)}{\sqrt{m}}\delta'$, or both. Therefore, by a union bound argument

$$\mathrm{P}\left[\max_{(j,k)\neq(i,\ell)} \left| e_j^{(k)} \right| \geq \epsilon \right] \leq \mathrm{P}\left[\max_{(j,k)\neq(i,\ell)} \left| \bar{e}_j^{(k)} \right| \geq \frac{\beta}{\sqrt{d_\ell}}\delta \right] + \mathrm{P}\left[\max_{(j,k)\neq(i,\ell)} \left| \tilde{e}_j^{(k)} \right| \geq \beta\frac{2\sigma(1+\sigma)}{\sqrt{m}}\delta' \right]. \tag{19}$$

Here, the first and second term on the RHS of (19) correspond to the perturbation caused by random projection and by noise, respectively. As established in Sections A.1 and A.2 these terms can be upper-bounded by $\frac{4}{N^2}$ and $2e^{-m} + \frac{7}{N^2}$, respectively, which yields

$$\mathrm{P}\left[\max_{(j,k)\neq(i,\ell)} \left|e_j^{(k)}\right| \geq \epsilon\right] \leq \frac{4}{N^2} + 2e^{-m} + \frac{7}{N^2}. \tag{20}$$

The remaining terms on the RHS of (16) are upper-bounded as shown in Steps 3 and 2 in [16, Proof of Thm. 3], respectively, using standard concentration of measure results, according to

$$\mathrm{P}\left[\bar{z}_{(n_\ell - q)}^{(\ell)} \leq \frac{2}{3\sqrt{d_\ell}}\right] \leq e^{-c(n_\ell - 1)}$$

and

$$\mathrm{P}\left[\max_{k\neq\ell, j} \bar{z}_j^{(k)} \geq \alpha\right] \leq 3N^{-2},$$

where $c > 1/20$ is a numerical constant, and we employed the assumption $n_\ell \geq 6q$, for all ℓ.

With (20) we thus get that (14) is violated with probability at most $e^{-c(n_\ell - 1)} + 2e^{-m} + 14N^{-2}$. Taking the union bound over all points $\mathbf{x}_i^{(\ell)}, i \in [n_\ell], \ell \in [L]$, finishes the proof.

A.1 Perturbation caused by random projection

We next show that the first term on the RHS of (19) is upper-bounded by $4/N^2$. This term corresponds to the perturbation caused by random projection. For notational convenience, we set $\mathbf{B}_{k,\ell} = \mathbf{U}^{(\ell)^T}(\mathbf{\Phi}^T\mathbf{\Phi} - \mathbf{I})\mathbf{U}^{(k)}$ and note that

$$\begin{aligned}
&\mathrm{P}\left[\max_{(j,k)\neq(i,\ell)} \left|\bar{e}_j^{(k)}\right| \geq \frac{\beta}{\sqrt{d_\ell}}\delta\right] = \mathrm{P}\left[\max_{(j,k)\neq(i,\ell)} \left|\left\langle \mathbf{B}_{k,\ell}\mathbf{a}_j^{(k)}, \mathbf{a}_i^{(\ell)}\right\rangle\right| \geq \frac{\beta}{\sqrt{d_\ell}}\delta\right] \\
&= \mathrm{P}\left[\bigcup_{(j,k)\neq(i,\ell)} \left\{\left|\left\langle \mathbf{B}_{k,\ell}\mathbf{a}_j^{(k)}, \mathbf{a}_i^{(\ell)}\right\rangle\right| \geq \frac{\beta}{\sqrt{d_\ell}}\delta\right\}\right] \\
&\leq \mathrm{P}\left[\bigcup_{(j,k)\neq(i,\ell)} \left\{\left|\left\langle \mathbf{B}_{k,\ell}\mathbf{a}_j^{(k)}, \mathbf{a}_i^{(\ell)}\right\rangle\right| \geq \left\|\mathbf{B}_{k,\ell}\mathbf{a}_j^{(k)}\right\|_2 \frac{\beta}{\sqrt{d_\ell}}\right\} \cup \left\{\left\|\mathbf{B}_{k,\ell}\mathbf{a}_j^{(k)}\right\|_2 \geq \delta\right\}\right] \\
&\leq \mathrm{P}\left[\bigcup_{(j,k)\neq(i,\ell)} \left\{\left|\left\langle \mathbf{B}_{k,\ell}\mathbf{a}_j^{(k)}, \mathbf{a}_i^{(\ell)}\right\rangle\right| \geq \left\|\mathbf{B}_{k,\ell}\mathbf{a}_j^{(k)}\right\|_2 \frac{\beta}{\sqrt{d_\ell}}\right\} \cup \left\{\left\|\mathbf{B}_{k,\ell}\right\|_{2\to 2} \geq \delta\right\}\right] \qquad (21) \\
&\leq \mathrm{P}\left[\max_k \left\|\mathbf{B}_{k,\ell}\right\|_{2\to 2} \geq \delta\right] + \sum_{(j,k)\neq(i,\ell)} \mathrm{P}\left[\left|\left\langle \mathbf{B}_{k,\ell}\mathbf{a}_j^{(k)}, \mathbf{a}_i^{(\ell)}\right\rangle\right| \geq \left\|\mathbf{B}_{k,\ell}\mathbf{a}_j^{(k)}\right\|_2 \frac{\beta}{\sqrt{d_\ell}}\right] \qquad (22) \\
&\leq 2e^{-\tau/2} + N2e^{-\frac{6\log N}{2}} = \frac{4}{N^2}, \qquad (23)
\end{aligned}$$

where (21) follows from $\left\|\mathbf{B}_{k,\ell}\mathbf{a}_j^{(k)}\right\|_2 \leq \left\|\mathbf{B}_{k,\ell}\right\|_{2\to 2}$, (22) is by the union bound, and (23) follows from (45) in Appendix B with $\tau = 4\log N$ and Proposition 1 below with $\mathbf{a} = \mathbf{a}_i^{(\ell)}$, $\mathbf{b} = \mathbf{B}_{k,\ell}\mathbf{a}_j^{(k)}$, $d = d_\ell$, and $\beta = \sqrt{6\log N}$.

Proposition 1 (E.g., [33, Ex. 5.25]). *Let* $\mathbf{a}$ *be uniformly distributed on* $\mathbb{S}^{d-1}$ *and fix* $\mathbf{b} \in \mathbb{R}^d$. *Then, for* $\beta \geq 0$, *we have*

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

A.2 Perturbation caused by noise

In this section, we deal with the perturbation caused by noise. Specifically, we establish that the second term on the RHS of (19) satisfies

$$\mathrm{P}\left[\max_{(j,k)\neq(i,\ell)} \left|\tilde{e}_j^{(k)}\right| \geq \beta\frac{2\sigma(1+\sigma)}{\sqrt{m}}\delta'\right] \leq 2e^{-m} + \frac{7}{N^2}. \tag{24}$$

For notational convenience, we set $\bar{\mathbf{y}}_j^{(k)} = \mathbf{\Phi}^T\mathbf{\Phi}\mathbf{y}_j^{(k)}$ and drop the indices i and ℓ to write $\mathbf{y} = \mathbf{y}_i^{(\ell)}$, $\bar{\mathbf{y}} = \bar{\mathbf{y}}_i^{(\ell)}$, $\mathbf{e} = \mathbf{e}_i^{(\ell)}$. We first note that

$$\begin{aligned}
&\left\{\max_{(j,k)\neq(i,\ell)} \left|\tilde{e}_j^{(k)}\right| \geq \beta\frac{2\sigma(1+\sigma)}{\sqrt{m}}\delta'\right\} = \bigcup_{(j,k)\neq(i,\ell)} \left\{\left|\tilde{e}_j^{(k)}\right| \geq \beta\frac{2\sigma(1+\sigma)}{\sqrt{m}}\delta'\right\} \\
&\subseteq \bigcup_{(j,k)\neq(i,\ell)} \left\{\left|\left\langle \bar{\mathbf{y}}_j^{(k)}, \mathbf{e}\right\rangle\right| \geq \beta\frac{\sigma}{\sqrt{m}}\delta'\right\} \cup \left\{\left|\left\langle \mathbf{e}_j^{(k)}, \bar{\mathbf{y}}\right\rangle\right| \geq \beta\frac{\sigma}{\sqrt{m}}\delta'\right\} \cup \left\{\left|\left\langle \mathbf{\Phi}^T\mathbf{\Phi}\mathbf{e}_j^{(k)}, \mathbf{e}\right\rangle\right| \geq \beta\frac{2\sigma^2}{\sqrt{m}}\delta'\right\} \qquad (25) \\
&\subseteq \left\{\left\|\mathbf{\Phi}^T\mathbf{\Phi}\right\|_{2\to 2} \geq \delta'\right\} \cup \bigcup_{(j,k)\neq(i,\ell)} \left[\left\{\left|\left\langle \bar{\mathbf{y}}_j^{(k)}, \mathbf{e}\right\rangle\right| \geq \beta\frac{\sigma}{\sqrt{m}}\left\|\bar{\mathbf{y}}_j^{(k)}\right\|_2\right\} \cup \left\{\left|\left\langle \mathbf{e}_j^{(k)}, \bar{\mathbf{y}}\right\rangle\right| \geq \beta\frac{\sigma}{\sqrt{m}}\|\bar{\mathbf{y}}\|_2\right\}\right. \\
&\quad \left. \cup \left\{\left|\left\langle \mathbf{\Phi}^T\mathbf{\Phi}\mathbf{e}_j^{(k)}, \mathbf{e}\right\rangle\right| \geq \beta\frac{\sigma}{\sqrt{m}}\left\|\mathbf{\Phi}^T\mathbf{\Phi}\mathbf{e}_j^{(k)}\right\|_2\right\} \cup \left\{\left\|\mathbf{e}_j^{(k)}\right\|_2 \geq 2\sigma\right\}\right]. \qquad (26)
\end{aligned}$$

Here, (25) follows from the triangle inequality. To verify (26), consider the first event in (25) and note that

$$\left\{\left|\left\langle \bar{\mathbf{y}}_j^{(k)}, \mathbf{e}\right\rangle\right| \geq \beta\frac{\sigma}{\sqrt{m}}\delta'\right\} \subseteq \left\{\left\|\mathbf{\Phi}^T\mathbf{\Phi}\right\|_{2\to 2} \geq \delta'\right\} \cup \left\{\left|\left\langle \bar{\mathbf{y}}_j^{(k)}, \mathbf{e}\right\rangle\right| \geq \beta\frac{\sigma}{\sqrt{m}}\left\|\bar{\mathbf{y}}_j^{(k)}\right\|_2\right\}. \tag{27}$$

To see this, simply take the complement of (27) according to

$$\left\{\left\|\mathbf{\Phi}^T\mathbf{\Phi}\right\|_{2\to 2} < \delta'\right\} \cap \left\{\left|\left\langle \bar{\mathbf{y}}_j^{(k)}, \mathbf{e}\right\rangle\right| < \beta\frac{\sigma}{\sqrt{m}}\left\|\bar{\mathbf{y}}_j^{(k)}\right\|_2\right\} \subseteq \left\{\left|\left\langle \bar{\mathbf{y}}_j^{(k)}, \mathbf{e}\right\rangle\right| < \beta\frac{\sigma}{\sqrt{m}}\delta'\right\}$$

where we used

$$\left\|\bar{\mathbf{y}}_j^{(k)}\right\|_2 = \left\|\mathbf{\Phi}^T\mathbf{\Phi}\mathbf{y}_j^{(k)}\right\|_2 \leq \left\|\mathbf{\Phi}^T\mathbf{\Phi}\right\|_{2\to 2}\left\|\mathbf{y}_j^{(k)}\right\|_2 = \left\|\mathbf{\Phi}^T\mathbf{\Phi}\right\|_{2\to 2}.$$

Treating the second and the third event in (25) similarly establishes (26). A union bound argument now yields

$$\mathrm{P}\left[\max_{(j,k)\neq(i,\ell)} \left|\tilde{e}_j^{(k)}\right| \geq \beta\frac{2\sigma(1+\sigma)}{\sqrt{m}}\delta'\right] \leq \mathrm{P}\left[\left\|\mathbf{\Phi}^T\mathbf{\Phi}\right\|_2 \geq \delta'\right] \tag{28}$$

$$+ \sum_{(j,k)\neq(i,\ell)} \mathrm{P}\left[\left|\left\langle \bar{\mathbf{y}}_j^{(k)}, \mathbf{e}\right\rangle\right| \geq \beta\frac{\sigma}{\sqrt{m}}\left\|\bar{\mathbf{y}}_j^{(k)}\right\|_2\right] \tag{29}$$

$$+ \sum_{(j,k)\neq(i,\ell)} \mathrm{P}\left[\left|\left\langle \mathbf{e}_j^{(k)}, \bar{\mathbf{y}}\right\rangle\right| \geq \beta\frac{\sigma}{\sqrt{m}}\|\bar{\mathbf{y}}\|_2\right] \tag{30}$$

$$+ \sum_{(j,k)\neq(i,\ell)} \mathrm{P}\left[\left|\left\langle \mathbf{\Phi}^T\mathbf{\Phi}\mathbf{e}_j^{(k)}, \mathbf{e}\right\rangle\right| \geq \beta\frac{\sigma}{\sqrt{m}}\left\|\mathbf{\Phi}^T\mathbf{\Phi}\mathbf{e}_j^{(k)}\right\|_2\right] \tag{31}$$

$$+ \sum_{(j,k)\neq(i,\ell)} \mathrm{P}\left[\left\|\mathbf{e}_j^{(k)}\right\|_2 \geq 2\sigma\right] \tag{32}$$

$$\leq 2e^{-m} + 6Ne^{-\frac{\beta^2}{2}} + Ne^{-\frac{\beta^2}{2}}. \tag{33}$$

To get (33) we upper-bounded the terms on the RHSs of (28)-(32) as follows. For the RHS of (28) we note that

$$\mathrm{P}\left[\left\|\mathbf{\Phi}^T\mathbf{\Phi}\right\|_2 \geq \delta'\right] \leq 2e^{-m},$$

which is a consequence of Theorem 6 stated in Appendix B below. Specifically, with $1 \leq \sqrt{\frac{6m}{\bar{c}p}}$, which follows from $\bar{c} = \min(6, \tilde{c}) \leq 6$ and $p \leq m$, both by assumption, we have

$$\mathrm{P}\left[\left\|\mathbf{\Phi}^T\mathbf{\Phi}\right\|_{2\to2} \geq \sqrt{\frac{24m}{\bar{c}p}}\right] \leq \mathrm{P}\left[\left\|\mathbf{\Phi}^T\mathbf{\Phi}\right\|_{2\to2} \geq 1 + \sqrt{\frac{6m}{\tilde{c}p}}\right]$$

$$\leq \mathrm{P}\left[\left\|\mathbf{\Phi}^T\mathbf{\Phi} - \mathbf{I}\right\|_{2\to2} \geq \sqrt{\frac{6m}{\tilde{c}p}}\right] \tag{34}$$

$$\leq 2e^{-m} \tag{35}$$

where (35) is by Theorem 6 (with $t = \sqrt{2m}$). To establish (34), first note that $\left\|\mathbf{\Phi}^T\mathbf{\Phi} - \mathbf{I}\right\|_{2\to2} \leq \delta'$ (with $\delta' = \sqrt{\frac{6m}{\bar{c}p}}$) implies $\sigma_{\max}\left(\mathbf{\Phi}^T\mathbf{\Phi}\right) \leq 1 + \delta'$, which in turn is equivalent to $\left\|\mathbf{\Phi}^T\mathbf{\Phi}\right\|_{2\to2} \leq 1 + \delta'$. We can therefore conclude that $\left\|\mathbf{\Phi}^T\mathbf{\Phi}\right\|_{2\to2} \geq 1 + \delta'$ implies $\left\|\mathbf{\Phi}^T\mathbf{\Phi} - \mathbf{I}\right\|_{2\to2} \geq \delta'$.

The terms inside the sums on the RHSs of (29), (30), and (31), were upper-bounded by applying Lemma 1, stated below. Specifically, we note that $\left\langle \bar{\mathbf{y}}_j^{(k)}, \mathbf{e}\right\rangle \sim \mathcal{N}(0, \sigma^2\left\|\bar{\mathbf{y}}_j^{(k)}\right\|_2^2)$, $\left\langle \mathbf{e}_j^{(k)}, \bar{\mathbf{y}}\right\rangle \sim \mathcal{N}(0, \sigma^2\left\|\bar{\mathbf{y}}\right\|_2^2)$, and $\left\langle \mathbf{\Phi}^T\mathbf{\Phi}\mathbf{e}_j^{(k)}, \mathbf{e}\right\rangle \sim \mathcal{N}(0, \sigma^2\left\|\mathbf{\Phi}^T\mathbf{\Phi}\mathbf{e}_j^{(k)}\right\|_2^2)$, where $\mathbf{y}_j^{(k)}$, $\bar{\mathbf{y}}$, and $\mathbf{\Phi}^T\mathbf{\Phi}\mathbf{e}_j^{(k)}$, respectively, can be regarded as fixed, and we used $\beta = \sqrt{6\log N} \geq \frac{1}{\sqrt{2\pi}}$, as $N \geq 1$.

Lemma 1 ([22, Prop. 19.4.2])**.** *Let $x \sim \mathcal{N}(0,1)$. For $\beta \geq \frac{1}{\sqrt{2\pi}}$, we have*

$$\mathrm{P}[x \geq \beta] \leq e^{-\frac{\beta^2}{2}}. \tag{36}$$

Finally, to upper-bound the terms inside the sum in (32), we used [16, Eq. (51)]

$$\mathrm{P}\left[\left\|\mathbf{e}_j^{(k)}\right\|_2 \geq 2\sigma\right] \leq e^{-\frac{\beta^2}{2}}. \tag{37}$$

A.3 Proof of Theorem 1

The proof of Theorem 1 is obtained from the proof of Theorem 4 by noting that in the noise-free case (i.e., $\sigma = 0$), the perturbation caused by noise satisfies $\tilde{e}_j^{(k)} = 0$, rendering the second term on the RHS of (19) void. Finally, we remark that the assumption $6 \log N \leq m$ is not needed in the noise-free case as it is involved only in establishing (24), which is void here.

B Proof of Theorem 2

We first note that the data points in $\mathcal{X}_\ell$ can be written as $\mathbf{x}_j^{(\ell)} = \mathbf{V}^{(\ell)} \mathbf{a}_j^{(\ell)}, j \in [n_\ell]$, where the $\mathbf{a}_j^{(\ell)}$ are i.i.d. uniform on $\mathbb{S}^{d_\ell - 1}$, and $\mathbf{V}^{(\ell)} \coloneqq \mathbf{\Phi} \mathbf{U}^{(\ell)}$ is a basis for the d_ℓ-dimensional subspace of $\mathbb{R}^p$ containing the points in $\mathcal{X}_\ell$ ($\mathbf{V}^{(\ell)}$ has full column rank with high probability, which follows from (44) as a consequence of the concentration inequality (4)). For the case where the $\mathbf{V}^{(\ell)}$ are orthonormal bases a sufficient condition for successful clustering was derived by Soltanolkotabi and Candès [28, Thm. 2.8]. However, owing to the projection $\mathbf{\Phi}$, the $\mathbf{V}^{(\ell)} = \mathbf{\Phi} \mathbf{U}^{(\ell)}$ will in general not be orthonormal. We will therefore need the following generalization of [28, Thm. 2.8] to arbitrary bases $\mathbf{V}^{(\ell)}$ for d_ℓ-dimensional subspaces of $\mathbb{R}^p$.

Theorem 5. *Suppose that the elements of the sets $\mathcal{X}_\ell$ in $\mathcal{X} = \mathcal{X}_1 \cup \ldots \cup \mathcal{X}_L$ are obtained by choosing n_ℓ points at random according to $\mathbf{x}_j^{(\ell)} = \mathbf{V}^{(\ell)} \mathbf{a}_j^{(\ell)}, j \in [n_\ell]$, where the $\mathbf{V}^{(\ell)} \in \mathbb{R}^{p \times d_\ell}$ have full rank, and the $\mathbf{a}_j^{(\ell)}$ are i.i.d. uniform on $\mathbb{S}^{d_\ell - 1}$. Assume that $\rho_\ell = (n_\ell - 1)/d_\ell \geq \rho_0$, for all ℓ, where $\rho_0 > 1$ is a numerical constant, and let $\rho_{\min} = \min_\ell \rho_\ell$. If*

$$\max_{k,\ell:\, k \neq \ell} \frac{1}{\sqrt{d_k}} \left\| \mathbf{V}^{(\ell)\dagger} \mathbf{V}^{(k)} \right\|_F \leq \frac{\sqrt{\log \rho_{\min}}}{64 \log N}, \tag{38}$$

where $\mathbf{V}^{(\ell)\dagger} = \left(\mathbf{V}^{(\ell)T} \mathbf{V}^{(\ell)} \right)^{-1} \mathbf{V}^{(\ell)T}$ is the pseudo-inverse of $\mathbf{V}^{(\ell)}$, then the graph G with adjacency matrix obtained by applying SSC to $\mathcal{X}$ has no false connections with probability at least $1 - N^{-1} - \sum_{\ell=1}^{L} n_\ell e^{-\sqrt{\rho_\ell} d_\ell}$.

Proof. See Appendix B.1. □

We now detail how Theorem 2 follows from Theorem 5. Specifically, we will show that (8) implies (38) with probability at least $1 - 4e^{-\tau/2}$, which, when combined with the probability bound in Theorem 5 via a union bound yields the final probability estimate in Theorem 2, and thereby concludes the proof.

We start filling in the details by showing how (8) implies (38). The LHS of (38) can be upper-

bounded as follows

$$\begin{aligned}
&\frac{1}{\sqrt{d_k}}\left\|\mathbf{V}^{(\ell)\dagger}\mathbf{V}^{(k)}\right\|_F = \frac{1}{\sqrt{d_k}}\left\|\left(\mathbf{V}^{(\ell)T}\mathbf{V}^{(\ell)}\right)^{-1}\mathbf{V}^{(\ell)T}\mathbf{V}^{(k)}\right\|_F \\
&\leq \left\|\left(\mathbf{V}^{(\ell)T}\mathbf{V}^{(\ell)}\right)^{-1}\right\|_{2\to 2}\frac{1}{\sqrt{d_k}}\left\|\mathbf{V}^{(\ell)T}\mathbf{V}^{(k)}\right\|_F \qquad (39)\\
&\leq \frac{\left\|\left(\mathbf{V}^{(\ell)T}\mathbf{V}^{(\ell)}\right)^{-1}\right\|_{2\to 2}}{\sqrt{d_k}}\left(\left\|\mathbf{U}^{(\ell)T}\mathbf{U}^{(k)}\right\|_F + \left\|\mathbf{U}^{(\ell)T}(\mathbf{\Phi}^T\mathbf{\Phi}-\mathbf{I})\mathbf{U}^{(k)}\right\|_F\right) \\
&\leq \left\|\left(\mathbf{V}^{(\ell)T}\mathbf{V}^{(\ell)}\right)^{-1}\right\|_{2\to 2}\left(\mathrm{aff}(\mathcal{S}_k,\mathcal{S}_\ell) + \left\|\mathbf{U}^{(\ell)T}(\mathbf{\Phi}^T\mathbf{\Phi}-\mathbf{I})\mathbf{U}^{(k)}\right\|_{2\to 2}\right) \qquad (40)\\
&\leq \frac{1}{1-\delta}(\mathrm{aff}(\mathcal{S}_k,\mathcal{S}_\ell)+\delta) \qquad (41)\\
&\leq \frac{65}{64}(\mathrm{aff}(\mathcal{S}_k,\mathcal{S}_\ell)+\delta) \leq \frac{\sqrt{\log\rho_{\min}}}{64\log N}, \qquad (42)
\end{aligned}$$

where (39) follows from $\|\mathbf{AB}\|_F^2 \leq \|\mathbf{A}\|_{2\to 2}^2\|\mathbf{B}\|_F^2$, (40) is a consequence of $\|\mathbf{B}\|_F \leq \sqrt{m\wedge n}\|\mathbf{B}\|_{2\to 2}$, for $\mathbf{B}\in\mathbb{R}^{m\times n}$ [18, Sec. 5.6, p. 365], and (41) holds with

$$\delta \coloneqq \sqrt{\frac{28d_{\max}+8\log L+2\tau}{3\tilde{c}p}}, \qquad (43)$$

with probability at least $1-4e^{-\tau/2}$ (here, $\tau>0$ is the numerical constant in the statement of Theorem 2). Eq. (41) holds with probability at least $1-4e^{-\tau/2}$ by

$$\mathrm{P}\left[\max_\ell \left\|\left(\mathbf{V}^{(\ell)T}\mathbf{V}^{(\ell)}\right)^{-1}\right\|_{2\to 2} \geq \frac{1}{1-\delta}\right] \leq 2e^{-\tau/2} \qquad (44)$$

and

$$\mathrm{P}\left[\max_{k,\ell}\left\|\mathbf{U}^{(\ell)T}(\mathbf{\Phi}^T\mathbf{\Phi}-\mathbf{I})\mathbf{U}^{(k)}\right\|_{2\to 2}\geq\delta\right] \leq 2e^{-\tau/2}, \qquad (45)$$

both proven below. Finally, to get (42) we invoked (8) twice, first we used $\mathrm{aff}(\mathcal{S}_k,\mathcal{S}_\ell)\geq 0$ and $\frac{\sqrt{\log\rho_{\min}}}{\log N} = \frac{\sqrt{\log\rho_{\min}}}{\log\left(\sum_{\ell=1}^L(\rho_\ell d_\ell+1)\right)} \leq 1$ in (8) to conclude that $\delta\leq 1/65$, i.e., $\frac{1}{1-\delta}\leq\frac{65}{64}$, and second, we applied (8) straight to upper-bound $\mathrm{aff}(\mathcal{S}_k,\mathcal{S}_\ell)$.

It remains to prove (44) and (45). For the special case of a Gaussian random matrix $\mathbf{\Phi}$, the probability bounds (44) and (45) can be obtained using standard results on the extremal singular values of Gaussian random matrices. For general $\mathbf{\Phi}$ satisfying the concentration inequality (4), the proofs of (44) and (45) rely on Theorem (6) below.

Theorem 6 ([10, Thm. 9.9, Rem. 9.10])**.** *Suppose that the random matrix $\mathbf{\Phi}\in\mathbb{R}^{p\times m}$ satisfies the concentration inequality* (4)*, i.e.,*

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

for all $t>0$ and for all $\mathbf{x}\in\mathbb{R}^m$, where $\tilde{c}$ is a constant. Then, for an orthonormal matrix $\mathbf{U}\in\mathbb{R}^{m\times d}$ and all $t>0$, we have

$$\mathrm{P}\left[\left\|\mathbf{U}^T\mathbf{\Phi}^T\mathbf{\Phi U}-\mathbf{I}\right\|_{2\to 2}\geq\sqrt{\frac{14d+2t^2}{3\tilde{c}p}}\right] \leq 2e^{-\frac{t^2}{2}}.$$

Additionally, for all $t > 0$, we have

$$\mathrm{P}\left[\left\|\mathbf{\Phi}^T\mathbf{\Phi} - \mathbf{I}\right\|_{2\to 2} \geq \sqrt{\frac{14m + 2t^2}{3\tilde{c}p}}\right] \leq 2e^{-\frac{t^2}{2}}.$$

Proof of (44)**:** By a union bound argument, we get

$$\mathrm{P}\left[\max_{\ell}\left\|\left(\mathbf{V}^{(\ell)^T}\mathbf{V}^{(\ell)}\right)^{-1}\right\|_{2\to 2} \geq \frac{1}{1-\delta}\right] \leq \sum_{\ell=1}^{L}\mathrm{P}\left[\left\|\left(\mathbf{V}^{(\ell)^T}\mathbf{V}^{(\ell)}\right)^{-1}\right\|_{2\to 2} \geq \frac{1}{1-\delta}\right]. \tag{46}$$

Note that $\left\|\mathbf{V}^{(\ell)^T}\mathbf{V}^{(\ell)} - \mathbf{I}\right\|_{2\to 2} \leq \delta$ implies that $\sigma_{\min}\left(\mathbf{V}^{(\ell)^T}\mathbf{V}^{(\ell)}\right) \geq 1-\delta$, which in turn implies $\left\|(\mathbf{V}^{(\ell)^T}\mathbf{V}^{(\ell)})^{-1}\right\|_{2\to 2} \leq \frac{1}{1-\delta}$. We can therefore conclude that $\left\|(\mathbf{V}^{(\ell)^T}\mathbf{V}^{(\ell)})^{-1}\right\|_{2\to 2} \geq \frac{1}{1-\delta}$ implies $\left\|\mathbf{V}^{(\ell)^T}\mathbf{V}^{(\ell)} - \mathbf{I}\right\|_{2\to 2} \geq \delta$, which can be formalized according to

$$\left\{\left\|(\mathbf{V}^{(\ell)^T}\mathbf{V}^{(\ell)})^{-1}\right\|_{2\to 2} \geq \frac{1}{1-\delta}\right\} \subseteq \left\{\left\|\mathbf{V}^{(\ell)^T}\mathbf{V}^{(\ell)} - \mathbf{I}\right\|_{2\to 2} \geq \delta\right\}.$$

Moreover, we have with δ as defined in (43) $\delta = \sqrt{\frac{28d_{\max}+2t^2}{3\tilde{c}p}} \geq \sqrt{\frac{14d_\ell+2t^2}{3\tilde{c}p}}$ ($2d_{\max} \geq d_{\max} \geq d_\ell$), with $t^2 = 4\log L + \tau$. Therefore, Theorem 6 (with $\mathbf{U} = \mathbf{U}^{(\ell)}$ and $t^2 = 4\log L + \tau$) yields

$$\mathrm{P}\left[\left\|\left(\mathbf{V}^{(\ell)^T}\mathbf{V}^{(\ell)}\right)^{-1}\right\|_{2\to 2} \geq \frac{1}{1-\delta}\right] \leq 2e^{-2\log L - \tau/2} = 2L^{-2}e^{-\tau/2} \leq 2L^{-1}e^{-\tau/2},$$

which when used on the RHS of (46) establishes (44).

Proof of (45)**:** Again, by a union bound argument, we get

$$\mathrm{P}\left[\max_{k,\ell}\left\|\mathbf{U}^{(\ell)^T}(\mathbf{\Phi}^T\mathbf{\Phi} - \mathbf{I})\mathbf{U}^{(k)}\right\|_{2\to 2} \geq \delta\right] \leq \sum_{k,\ell=1}^{L}\mathrm{P}\left[\left\|\mathbf{U}^{(\ell)^T}(\mathbf{\Phi}^T\mathbf{\Phi} - \mathbf{I})\mathbf{U}^{(k)}\right\|_{2\to 2} \geq \delta\right]. \tag{47}$$

We next upper-bound the probabilities on the RHS of (47). To this end, let $\tilde{\mathbf{U}} \in \mathbb{R}^{m\times\tilde{d}}$ be an orthonormal basis for the $\tilde{d}$-dimensional span of $[\mathbf{U}^{(\ell)}\ \mathbf{U}^{(k)}]$ ($\max(d_\ell, d_k) \leq \tilde{d} \leq d_\ell + d_k$). Since $\tilde{\mathbf{U}}\tilde{\mathbf{U}}^T$ is the orthogonal projection onto span($[\mathbf{U}^{(\ell)}\ \mathbf{U}^{(k)}]$), we have $\tilde{\mathbf{U}}\tilde{\mathbf{U}}^T\mathbf{U}^{(\ell)} = \mathbf{U}^{(\ell)}$ and $\tilde{\mathbf{U}}\tilde{\mathbf{U}}^T\mathbf{U}^{(k)} = \mathbf{U}^{(k)}$. Therefore, we get

$$\begin{aligned}
\left\|\mathbf{U}^{(\ell)^T}(\mathbf{\Phi}^T\mathbf{\Phi} - \mathbf{I})\mathbf{U}^{(k)}\right\|_{2\to 2} &= \left\|\mathbf{U}^{(\ell)^T}\tilde{\mathbf{U}}\tilde{\mathbf{U}}^T(\mathbf{\Phi}^T\mathbf{\Phi} - \mathbf{I})\tilde{\mathbf{U}}\tilde{\mathbf{U}}^T\mathbf{U}^{(k)}\right\|_{2\to 2} \\
&\leq \left\|\mathbf{U}^{(\ell)^T}\tilde{\mathbf{U}}\right\|_{2\to 2}\left\|\tilde{\mathbf{U}}^T\mathbf{\Phi}^T\mathbf{\Phi}\tilde{\mathbf{U}} - \mathbf{I}\right\|_{2\to 2}\left\|\tilde{\mathbf{U}}^T\mathbf{U}^{(k)}\right\|_{2\to 2} \\
&= \left\|\tilde{\mathbf{U}}^T\mathbf{\Phi}^T\mathbf{\Phi}\tilde{\mathbf{U}} - \mathbf{I}\right\|_{2\to 2},
\end{aligned}$$

where we used $\left\|\mathbf{U}^{(\ell)^T}\tilde{\mathbf{U}}\right\|_{2\to 2} = 1$, which holds since $\mathbf{U}^{(\ell)}$ is in the span of $\tilde{\mathbf{U}}$ and both $\mathbf{U}^{(\ell)}$ and $\tilde{\mathbf{U}}$ are orthonormal. This finally yields, with δ as defined in (43),

$$\begin{aligned}
\mathrm{P}\Big[\left\|\mathbf{U}^{(\ell)^T}(\boldsymbol{\Phi}^T\boldsymbol{\Phi}-\mathbf{I})\mathbf{U}^{(k)}\right\|_{2\to 2} \geq \delta\Big] &\leq \mathrm{P}\left[\left\|\tilde{\mathbf{U}}^T\boldsymbol{\Phi}^T\boldsymbol{\Phi}\tilde{\mathbf{U}}-\mathbf{I}\right\|_{2\to 2} \geq \sqrt{\frac{28d_{\max}+8\log L+2\tau}{3\tilde{c}p}}\right] \\
&\leq \mathrm{P}\left[\left\|\tilde{\mathbf{U}}^T\boldsymbol{\Phi}^T\boldsymbol{\Phi}\tilde{\mathbf{U}}-\mathbf{I}\right\|_{2\to 2} \geq \sqrt{\frac{14\tilde{d}+8\log L+2\tau}{3\tilde{c}p}}\right] \qquad (48) \\
&\leq 2e^{-\frac{4\log L+\tau}{2}} = 2L^{-2}e^{-\tau/2}, \qquad (49)
\end{aligned}$$

where (48) follows from $2d_{\max} \geq d_\ell + d_k \geq \tilde{d}$, and (49) is by application of Theorem 6 with $\mathbf{U} = \tilde{\mathbf{U}}$ and $t^2 = 4\log L + \tau$. The proof is concluded by using (49) on the RHS of (47).

B.1 Proof of Theorem 5

Theorem 5 is a generalization of a result by Soltanolkotabi and Candès [28, Thm. 2.8] from orthonormal bases $\mathbf{V}^{(\ell)}$ for d_ℓ-dimensional subspaces of $\mathbb{R}^p$ to arbitrary bases $\mathbf{V}^{(\ell)}$ for d_ℓ-dimensional subspaces. The proof program essentially follows that of [28, Thm. 2.8]. However, some parts of the generalization are non-trivial. We only detail the arguments that are new relative to [28], and refer to [28] otherwise.

Throughout the proof, we use the following notation: Let $\mathbf{X}^{(\ell)} \in \mathbb{R}^{p\times n_\ell}$ be the matrix whose columns are the points in $\mathcal{X}_\ell$, and note that $\mathbf{X}^{(\ell)} = \mathbf{V}^{(\ell)}\mathbf{A}^{(\ell)}$, where $\mathbf{A}^{(\ell)} \in \mathbb{R}^{d_\ell\times n_\ell}$ is the matrix with columns $\mathbf{a}_i^{(\ell)}, i = 1,\ldots,n_\ell$. Set $\mathbf{X} = [\mathbf{X}^{(1)} \ \ldots \ \mathbf{X}^{(L)}] \in \mathbb{R}^{p\times N}$, and let $\mathbf{X}_{-i}$ be the matrix obtained by removing the ith column $\mathbf{x}_i$ from $\mathbf{X}$. $\mathcal{P}(\mathbf{X})$ denotes the symmetrized convex hull of the columns of $\mathbf{X}$ (i.e., the points in $\mathcal{X}$), that is, the convex hull of $\{\mathbf{x}_1, -\mathbf{x}_1, \ldots, \mathbf{x}_N, -\mathbf{x}_N\}$. For a convex body $\mathcal{P}$, its inradius $r(\mathcal{P})$ is defined as the radius of the largest Euclidean ball that can be inscribed in $\mathcal{P}$, and its circumradius $R(\mathcal{P})$ is defined as the radius of the smallest ball containing $\mathcal{P}$. Finally, the polar set of $\mathcal{K} \subset \mathbb{R}^n$ is defined as

$$\mathcal{K}^\circ = \{\mathbf{y} \in \mathbb{R}^n\colon \ \langle \mathbf{x}, \mathbf{y}\rangle \leq 1 \text{ for all } \mathbf{x} \in \mathcal{K}\}.$$

B.1.1 A deterministic clustering condition

We first establish a deterministic clustering condition. Specifically, in Theorem 7 below we present conditions guaranteeing that for $\mathbf{x}_i \in \mathcal{X}_\ell$ every solution of the problem

$$\underset{\mathbf{z}}{\text{minimize}} \ \|\mathbf{z}\|_1 \ \text{ subject to } \mathbf{X}_{-i}\mathbf{z} = \mathbf{x}_i \qquad (50)$$

has non-zero entries corresponding to columns of $\mathbf{X}^{(\ell)}$ only. The proof of Theorem 5 is then obtained by proving that these conditions are satisfied with high probability for the statistical data model in Theorem 5. We start by introducing terminology needed in the following. Define the primal optimization problem

$$P(\mathbf{y}, \mathbf{A})\colon \underset{\mathbf{z}}{\text{minimize}} \ \|\mathbf{z}\|_1 \text{ subject to } \mathbf{A}\mathbf{z} = \mathbf{y}$$

with the corresponding dual [4, Sec. 5.1.16]

$$D(\mathbf{y}, \mathbf{A})\colon \underset{\boldsymbol{\nu}}{\text{maximize}} \ \langle \mathbf{y}, \boldsymbol{\nu}\rangle \ \text{ subject to } \left\|\mathbf{A}^T\boldsymbol{\nu}\right\|_\infty \leq 1.$$

The problem (50) is then simply $P(\mathbf{x}_i, \mathbf{X}_{-i})$. The sets of optimal solutions of P and D are denoted by optsol$P(\mathbf{y}, \mathbf{A})$ and optsol$D(\mathbf{y}, \mathbf{A})$, respectively. A dual point $\boldsymbol{\lambda}(\mathbf{y}, \mathbf{A})$ is defined as a point in optsol$D(\mathbf{y}, \mathbf{A})$ of minimal Euclidean norm.

We are now ready to state the following generalization of [28, Thm. 2.5] from orthonormal bases $\mathbf{V}^{(\ell)}$ for d_ℓ-dimensional subspaces of $\mathbb{R}^p$ to arbitrary bases $\mathbf{V}^{(\ell)}$ for d_ℓ-dimensional subspaces.

Theorem 7. *Suppose that the elements of the sets $\mathcal{X}_\ell$ in $\mathcal{X} = \mathcal{X}_1 \cup \ldots \cup \mathcal{X}_L$ are obtained by choosing n_ℓ points according to $\mathbf{x}_i^{(\ell)} = \mathbf{V}^{(\ell)}\mathbf{a}_i^{(\ell)}, i \in [n_\ell]$, where the $\mathbf{a}_i^{(\ell)}$ are deterministic coefficient vectors and the $\mathbf{V}^{(\ell)} \in \mathbb{R}^{p \times d_\ell}$ are deterministic matrices of full column rank. Let $\mathbf{L} \in \mathbb{R}^{d_\ell \times (n_\ell - 1)}$ be the matrix whose columns are the normalized dual points $\tilde{\boldsymbol{\lambda}}(\mathbf{a}_i^{(\ell)}, \mathbf{A}_{-i}^{(\ell)}) = \boldsymbol{\lambda}(\mathbf{a}_i^{(\ell)}, \mathbf{A}_{-i}^{(\ell)}) / \big\|\boldsymbol{\lambda}(\mathbf{a}_i^{(\ell)}, \mathbf{A}_{-i}^{(\ell)})\big\|_2$, where $\mathbf{A}_{-i}^{(\ell)}$ is the matrix with columns $\mathbf{a}_j^{(\ell)}, j \in [n_\ell] \setminus \{i\}$. If*

$$\max_{k \neq \ell, j} \left\| \mathbf{L}^T \mathbf{V}^{(\ell)\dagger} \mathbf{V}^{(k)} \mathbf{a}_j^{(k)} \right\|_\infty < r(\mathcal{P}(\mathbf{A}_{-i}^{(\ell)})), \tag{51}$$

then the non-zero entries of all solutions of $P(\mathbf{x}_i^{(\ell)}, \mathbf{X}_{-(i,\ell)})$ correspond to points in $\mathcal{X}_\ell$ only (the columns of $\mathbf{X}_{-(i,\ell)}$ are the elements in $\mathcal{X} \setminus \{\mathbf{x}_i^{(\ell)}\}$).

Proof. The proof relies on the following lemma.

Lemma 2 ([28, Lem. 7.1], [6])**.** *Let T be a subset of the column indices of a given matrix $\mathbf{A}$. All solutions $\mathbf{c}^\star$ of $P(\mathbf{y}, \mathbf{A})$ satisfy $\mathbf{c}^\star_{\overline{T}} = \mathbf{0}$, if there exists a vector $\mathbf{c}$ such that $\mathbf{y} = \mathbf{A}\mathbf{c}$ with support $S \subseteq T$, and a (dual certificate) vector $\boldsymbol{\nu}$ obeying*

$$\mathbf{A}_S^T \boldsymbol{\nu} = \operatorname{sign}(\mathbf{c}_S) \tag{52}$$

$$\left\|\mathbf{A}_{T \cap \overline{S}}^T \boldsymbol{\nu}\right\|_\infty \leq 1 \tag{53}$$

$$\left\|\mathbf{A}_{\overline{T}}^T \boldsymbol{\nu}\right\|_\infty < 1. \tag{54}$$

We apply Lemma 2 with $\mathbf{A} = \mathbf{X}_{-(i,\ell)}$, $\mathbf{y} = \mathbf{x}_i^{(\ell)}$, and T the index set corresponding to the columns of $\mathbf{X}_{-i}^{(\ell)}$, and show that there exists a vector $\mathbf{c}$ supported on $S \subseteq T$ that obeys $\mathbf{x}_i^{(\ell)} = \mathbf{X}_{-(i,\ell)}\mathbf{c}$, and a corresponding vector $\boldsymbol{\nu}$ that satisfies (52)–(54). This then implies that the non-zero entries of all solutions of $P(\mathbf{x}_i^{(\ell)}, \mathbf{X}_{-(i,\ell)})$ correspond to points in $\mathcal{X}_\ell$ only, as desired.

We proceed with the explicit construction of the vector $\mathbf{c}$. Specifically, take $\mathbf{c}$ to be a vector that is zero on $\overline{T}$, and whose restriction to the index set T is given by $\mathbf{c}_T \in$ optsol$P(\mathbf{x}_i^{(\ell)}, \mathbf{X}_{-i}^{(\ell)})$. Let S be the support of $\mathbf{c}_T$, and let $\boldsymbol{\nu}_i^{(\ell)} = \left(\mathbf{V}^{(\ell)T}\right)^\dagger \boldsymbol{\lambda}_i^{(\ell)}$, where $\boldsymbol{\lambda}_i^{(\ell)}$ is taken to be a point of minimum ℓ_2-norm[3] in optsol$D(\mathbf{a}_i^{(\ell)}, \mathbf{A}_{-i}^{(\ell)})$. The next step is to show that $\boldsymbol{\nu}_i^{(\ell)} \in$ optsol$D(\mathbf{x}_i^{(\ell)}, \mathbf{X}_{-i}^{(\ell)})$, which will eventually allow us to establish that $\boldsymbol{\nu}_i^{(\ell)}$ satisfies the conditions of Lemma 2. To this end, we first

[3]For concreteness $\boldsymbol{\lambda}_i^{(\ell)}$ is taken to be a point of minimum ℓ_2-norm. Note, however, that for the proof to work we may let $\boldsymbol{\lambda}_i^{(\ell)}$ be an arbitrary point in optsol$P(\mathbf{a}_i^{(\ell)}, \mathbf{A}_{-i}^{(\ell)})$.

note that $\mathbf{x}_i^{(\ell)} = \mathbf{V}^{(\ell)}\mathbf{a}_i^{(\ell)}$ yields

$$\begin{aligned}
&\operatorname{optsol} D(\mathbf{x}_i^{(\ell)}, \mathbf{X}_{-i}^{(\ell)}) \\
&= \left\{ \arg\max_{\boldsymbol{\nu}} \left\langle \mathbf{a}_i^{(\ell)}, {\mathbf{V}^{(\ell)}}^T \boldsymbol{\nu} \right\rangle \text{ subject to } \left\| ({\mathbf{A}_{-i}^{(\ell)}})^T {\mathbf{V}^{(\ell)}}^T \boldsymbol{\nu} \right\|_\infty \leq 1 \right\} \\
&= \left\{ \boldsymbol{\nu} \colon \boldsymbol{\lambda} = {\mathbf{V}^{(\ell)}}^T \boldsymbol{\nu},\, \boldsymbol{\lambda} \in \operatorname{optsol} D(\mathbf{a}_i^{(\ell)}, \mathbf{A}_{-i}^{(\ell)}) \right\} \\
&\supseteq ({\mathbf{V}^{(\ell)}}^T)^\dagger \operatorname{optsol} D(\mathbf{a}_i^{(\ell)}, \mathbf{A}_{-i}^{(\ell)}),
\end{aligned}$$

where the inclusion holds as $({\mathbf{V}^{(\ell)}}^T)^\dagger \boldsymbol{\lambda}$ is the minimum norm solution to the linear system of equations $\boldsymbol{\lambda} = {\mathbf{V}^{(\ell)}}^T \boldsymbol{\nu}$, but in general not the only solution.

Since $P(\mathbf{y}, \mathbf{A})$ is a linear program, strong duality [4, Sec. 5.2.3] holds (provided that $P(\mathbf{y}, \mathbf{A})$ is feasible) and therefore the optimal objective values of $P(\mathbf{x}_i^{(\ell)}, \mathbf{X}_{-i}^{(\ell)})$ and $D(\mathbf{x}_i^{(\ell)}, \mathbf{X}_{-i}^{(\ell)})$ coincide. It therefore follows that

$$\|\mathbf{c}_T\|_1 = \left\langle \mathbf{x}_i^{(\ell)}, \boldsymbol{\nu}_i^{(\ell)} \right\rangle. \tag{55}$$

Since $\mathbf{c}_T \in \operatorname{optsol} P(\mathbf{x}_i^{(\ell)}, \mathbf{X}_{-i}^{(\ell)})$ and $\mathbf{c}_T$ is supported on S, both by assumption, we have $\mathbf{x}_i^{(\ell)} = \mathbf{X}_{-i}^{(\ell)}\mathbf{c}_T = (\mathbf{X}_{-i}^{(\ell)})_S \mathbf{c}_S$, and therefore (55) becomes

$$\langle \mathbf{c}_S, \operatorname{sign}(\mathbf{c}_S) \rangle = \left\langle (\mathbf{X}_{-i}^{(\ell)})_S \mathbf{c}_S, \boldsymbol{\nu}_i^{(\ell)} \right\rangle = \left\langle \mathbf{c}_S, (\mathbf{X}_{-i}^{(\ell)})_S^T \boldsymbol{\nu}_i^{(\ell)} \right\rangle. \tag{56}$$

On the other hand, as $\boldsymbol{\nu}_i^{(\ell)} \in \operatorname{optsol} D(\mathbf{x}_i^{(\ell)}, \mathbf{X}_{-i}^{(\ell)})$, we have $\left\| ({\mathbf{X}_{-i}^{(\ell)}})^T \boldsymbol{\nu}_i^{(\ell)} \right\|_\infty \leq 1$, which is equivalent to the following conditions (recall that the set T corresponds to the column indices of $\mathbf{X}_{-i}^{(\ell)}$):

$$\left\| ((\mathbf{X}_{-i}^{(\ell)})_S)^T \boldsymbol{\nu}_i^{(\ell)} \right\|_\infty \leq 1 \tag{57}$$

$$\left\| ((\mathbf{X}_{-i}^{(\ell)})_{T \cap \overline{S}})^T \boldsymbol{\nu}_i^{(\ell)} \right\|_\infty \leq 1. \tag{58}$$

As by (57), the entries of $(\mathbf{X}_{-i}^{(\ell)})_S^T \boldsymbol{\nu}_i^{(\ell)}$ are bounded in magnitude by 1 and the unique maximizer of $\max_{\mathbf{a}\colon \|\mathbf{a}\|_\infty \leq 1} \langle \mathbf{c}_S, \mathbf{a} \rangle$ is $\operatorname{sign}(\mathbf{c}_S)$, it follows from (56) that

$$(\mathbf{X}_{-i}^{(\ell)})_S^T \boldsymbol{\nu}_i^{(\ell)} = \operatorname{sign}(\mathbf{c}_S),$$

which establishes (52).

Thanks to (58), (53) is satisfied as well. It remains to verify (54), which here reads

$$\left| \left\langle \mathbf{x}_j^{(k)}, \boldsymbol{\nu}_i^{(\ell)} \right\rangle \right| < 1, \quad \text{for all } k \neq \ell, \text{ for all } j \in [n_k]. \tag{59}$$

With $\boldsymbol{\nu}_i^{(\ell)} = ({\mathbf{V}^{(\ell)}}^T)^\dagger \boldsymbol{\lambda}_i^{(\ell)}$, by definition, (59) becomes

$$\left| \left\langle \mathbf{x}_j^{(k)}, ({\mathbf{V}^{(\ell)}}^T)^\dagger \frac{\boldsymbol{\lambda}_i^{(\ell)}}{\left\| \boldsymbol{\lambda}_i^{(\ell)} \right\|_2} \right\rangle \right| < \frac{1}{\left\| \boldsymbol{\lambda}_i^{(\ell)} \right\|_2}, \text{ for all } k \neq \ell, \text{ for all } j \in [n_k]. \tag{60}$$

Since $((\mathbf{V}^{(\ell)^T})^\dagger)^T = \mathbf{V}^{(\ell)\dagger}$, and $\mathbf{x}_j^{(k)} = \mathbf{V}^{(k)}\mathbf{a}_j^{(k)}$, (60) is equivalent to

$$\left| \frac{\boldsymbol{\lambda}_i^{(\ell)^T}}{\left\|\boldsymbol{\lambda}_i^{(\ell)}\right\|_2} \mathbf{V}^{(\ell)\dagger}\mathbf{V}^{(k)}\mathbf{a}_j^{(k)} \right| < \frac{1}{\left\|\boldsymbol{\lambda}_i^{(\ell)}\right\|_2}, \text{ for all } k \neq \ell, \text{ for all } j \in [n_k]. \tag{61}$$

It now follows from $\boldsymbol{\lambda}_i^{(\ell)} \in \text{optsol}D(\mathbf{a}_i^{(\ell)}, \mathbf{A}_{-i}^{(\ell)})$ which holds by assumption, that

$$\left\|(\mathbf{A}_{-i}^{(\ell)})^T\boldsymbol{\lambda}_i^{(\ell)}\right\|_\infty \leq 1.$$

This, in turn, implies that $\boldsymbol{\lambda}_i^{(\ell)} \in \mathcal{P}^\circ(\mathbf{A}_{-i}^{(\ell)})$ where

$$\mathcal{P}^\circ(\mathbf{A}_{-i}^{(\ell)}) = \left\{\mathbf{z}\colon \left\|(\mathbf{A}_{-i}^{(\ell)})^T\mathbf{z}\right\|_\infty \leq 1\right\}$$

is the polar set of $\mathcal{P}(\mathbf{A}_{-i}^{(\ell)})$ (recall that $\mathcal{P}(\mathbf{A}_{-i}^{(\ell)})$ is the symmetrized convex hull of the columns in $\mathbf{A}_{-i}^{(\ell)}$). Since the inradius and the circumradius of a symmetric[4] convex body are related according to [13, Thm. 1.2]

$$r(\mathcal{P})R(\mathcal{P}^\circ) = 1,$$

we get from $\boldsymbol{\lambda}_i^{(\ell)} \in \mathcal{P}^\circ(\mathbf{A}_{-i}^{(\ell)})$ that

$$\left\|\boldsymbol{\lambda}_i^{(\ell)}\right\|_2 \leq R(\mathcal{P}^\circ(\mathbf{A}_{-i}^{(\ell)})) = \frac{1}{r(\mathcal{P}(\mathbf{A}_{-i}^{(\ell)}))}. \tag{62}$$

By (62), it follows that (61) holds if

$$\left| \frac{\boldsymbol{\lambda}_i^{(\ell)^T}}{\left\|\boldsymbol{\lambda}_i^{(\ell)}\right\|_2} \mathbf{V}^{(\ell)\dagger}\mathbf{V}^{(k)}\mathbf{a}_j^{(k)} \right| < r(\mathcal{P}(\mathbf{A}_{-i}^{(\ell)})), \text{ for all } k \neq \ell, \text{ for all } j \in [n_k],$$

which is implied by (51). This proves that (54) is satisfied as well, thereby concluding the proof. □

B.1.2 Evaluating the deterministic clustering condition for the statistical data model

Theorem 5 now follows from Theorem 7 by establishing that, for our statistical data model, the deterministic clustering condition (51) holds for all pairs (ℓ, i) with $\ell \in [L], i \in [n_\ell]$, with high probability. Specifically, by a union bound argument, we get

$$\begin{aligned}
&\text{P}[(51) \text{ is violated for at least one pair } (\ell, i)] \\
&\leq \sum_{(\ell,i)} \text{P}\left[\max_{k\neq\ell,j} \left\|\mathbf{L}^T\mathbf{V}^{(\ell)\dagger}\mathbf{V}^{(k)}\mathbf{a}_j^{(k)}\right\|_\infty \geq r(\mathcal{P}(\mathbf{A}_{-i}^{(\ell)}))\right] \\
&\leq \sum_{(\ell,i)} \left(\text{P}\left[\max_{k\neq\ell,j} \left\|\mathbf{L}^T\mathbf{V}^{(\ell)\dagger}\mathbf{V}^{(k)}\mathbf{a}_j^{(k)}\right\|_\infty \geq \frac{16\log N}{\sqrt{d_\ell d_k}} \left\|\mathbf{V}^{(\ell)\dagger}\mathbf{V}^{(k)}\right\|_F\right] + \text{P}\left[\frac{\sqrt{\log\rho_\ell}}{4\sqrt{d_\ell}} \geq r(\mathcal{P}(\mathbf{A}_{-i}^{(\ell)}))\right]\right) \qquad (63) \\
&\leq \sum_{\ell=1}^{L} n_\ell e^{-\sqrt{\rho_\ell} d_\ell} + N^{-1}. \qquad (64)
\end{aligned}$$

[4] A convex body $\mathcal{P}$ is called symmetric if $\mathbf{x} \in \mathcal{P}$ if and only if $-\mathbf{x} \in \mathcal{P}$.

In (63) we used that for random variables X and Y, possibly dependent, and constants ϕ and φ satisfying $\phi \leq \varphi$, we have

$$\begin{aligned} \mathrm{P}[X \geq Y] &\leq \mathrm{P}[\{X \geq \phi\} \cup \{\varphi \geq Y\}] \\ &\leq \mathrm{P}[X \geq \phi] + \mathrm{P}[\varphi \geq Y]\,. \end{aligned} \tag{65}$$

Specifically, we applied (65) with $\phi = \frac{16 \log N}{\sqrt{d_\ell d_k}} \left\| \mathbf{V}^{(\ell)\dagger} \mathbf{V}^{(k)} \right\|_F$ and $\varphi = \frac{\sqrt{\log \rho_\ell}}{4\sqrt{d_\ell}}$, which leads to the assumption

$$\frac{16 \log N}{\sqrt{d_\ell d_k}} \left\| \mathbf{V}^{(\ell)\dagger} \mathbf{V}^{(k)} \right\|_F \leq \frac{\sqrt{\log \rho_\ell}}{4\sqrt{d_\ell}}, \quad \text{for all } k, \ell \colon k \neq \ell,$$

implied by (38). To get (64) we used that, for all i,

$$\mathrm{P}\left[\frac{\sqrt{\log \rho_\ell}}{4\sqrt{d_\ell}} \geq r(\mathcal{P}(\mathbf{A}_{-i}^{(\ell)})) \right] \leq e^{-\sqrt{\rho_\ell} d_\ell} \tag{66}$$

and

$$\mathrm{P}\left[\max_{k \neq \ell, j} \left\| \mathbf{L}^T \mathbf{V}^{(\ell)\dagger} \mathbf{V}^{(k)} \mathbf{a}_j^{(k)} \right\|_\infty \geq \frac{16 \log N}{\sqrt{d_\ell d_k}} \left\| \mathbf{V}^{(\ell)\dagger} \mathbf{V}^{(k)} \right\|_F \right] \leq N^{-2}, \tag{67}$$

both of which are established next.

The upper bound (66) is an application of [28, Lem. 7.4], [2], and makes use of the assumption $(n_\ell - 1)/d_\ell = \rho_\ell \geq \rho_0 > 1$. Finally, (67) follows from a union bound argument and

$$\begin{aligned} \mathrm{P}\left[\left\| \mathbf{L}^T \mathbf{V}^{(\ell)\dagger} \mathbf{V}^{(k)} \mathbf{a}_j^{(k)} \right\|_\infty \geq \frac{16 \log N}{\sqrt{d_\ell d_k}} \left\| \mathbf{V}^{(\ell)\dagger} \mathbf{V}^{(k)} \right\|_F \right] \\ \leq (n_\ell + 1) e^{-4 \log N} \leq N^{-3}, \end{aligned} \tag{68}$$

which is a consequence of Lemma 3 below together with the fact that the normalized dual point $\tilde{\boldsymbol{\lambda}}_i^{(\ell)} = \boldsymbol{\lambda}_i^{(\ell)} / \|\boldsymbol{\lambda}_i^{(\ell)}\|_2$ is distributed uniformly on the unit sphere, as shown in [28, Sec. 7.2.2 Proof of Step 2].

Lemma 3 (Extracted from the proof of Lemma 7.5 in [28])**.** *Let the columns of $\mathbf{L} \in \mathbb{R}^{d_1 \times n_1}$ be i.i.d. uniform on $\mathbb{S}^{d_1 - 1}$, let $\mathbf{a}$ be uniform on $\mathbb{S}^{d_2 - 1}$, and let $\mathbf{B} \in \mathbb{R}^{d_1 \times d_2}$. Then, for $c \geq 12$, we have*

$$\mathrm{P}\left[\left\| \mathbf{L}^T \mathbf{B} \mathbf{a} \right\|_\infty \geq \frac{c}{\sqrt{d_1 d_2}} \|\mathbf{B}\|_F \right] \leq (n_1 + 1) e^{-\frac{c}{4}}.$$

C Proof of Theorem 3

The graph G obtained by SSC-OMP has no false connections if for each $\mathbf{x}_i^{(\ell)} \in \mathcal{X}_\ell$ the OMP algorithm as detailed in Section 2 selects points from $\mathcal{X}_\ell$ only, for all $\ell \in [L]$. This is the case if OMP selects points from $\mathcal{X}_\ell$ in all iterations $s \in [s_{\mathrm{end}}]$ (we explain below that OMP terminates after $s_{\mathrm{end}} = s_{\max} \wedge d_\ell$ iterations with high probability for our statistical data model). The OMP selection rule (2) implies that OMP selects a point from $\mathcal{X}_\ell$ in the $(s+1)$th iteration if

$$\max_{k \neq \ell, j} \left| \left\langle \mathbf{x}_j^{(k)}, \mathbf{r}_s \right\rangle \right| < \max_{j \in [n_\ell] \colon j \neq i} \left| \left\langle \mathbf{x}_j^{(\ell)}, \mathbf{r}_s \right\rangle \right|. \tag{69}$$

Hence, the graph G obtained by SSC-OMP has no false connections if the deterministic clustering condition (69) holds for all s_{end} OMP iterations, for all $\mathbf{x}_i^{(\ell)} \in \mathcal{X}_\ell, \ell \in [L]$. We will next establish that (69) is satisfied for our statistical data model with probability obeying the bound in Theorem 3.

As a vehicle for our analysis, we introduce the *reduced OMP algorithm* which, to compute sparse representations of the $\mathbf{x}_i^{(\ell)}$, has access to the corresponding *reduced data sets* $\mathcal{X}_\ell \backslash \{\mathbf{x}_i^{(\ell)}\}$ only, instead of the full data sets $\mathcal{X} \backslash \{\mathbf{x}_i^{(\ell)}\}$. If, for a given data set $\mathcal{X}$, the residuals computed by reduced OMP, henceforth denoted by $\mathbf{r}_s^{(\ell)}$, satisfy (69) for *all iterations*, then the reduced OMP algorithm and the original OMP algorithm (processing the same data set $\mathcal{X}$) select exactly the same data points in the same order and we have $\mathbf{r}_s = \mathbf{r}_s^{(\ell)}$ for all $s \in [s_{\max} \wedge d_\ell]$ by virtue of (3). We emphasize that for expositional convenience the notations $\mathbf{r}_s^{(\ell)}$ and $\tilde{\mathbf{r}}_s^{(\ell)}$ do not reflect dependence on i. The motivation for working with the reduced OMP algorithm is that $\mathbf{r}_s^{(\ell)}$ being a function of the data points in $\mathcal{X}_\ell$ only, conditionally on $\mathbf{\Phi}$, is statistically independent of the data points in $\mathcal{X} \backslash \mathcal{X}_\ell$. This will allow us to establish tail bounds for $|\langle \mathbf{x}_j^{(k)}, \mathbf{r}_s^{(\ell)} \rangle|$, $k \neq \ell$, $j \in [n_k]$, using standard concentration inequalities. We proceed to show that under the assumptions of Theorem 3 the reduced OMP residuals $\mathbf{r}_s^{(\ell)}$ indeed satisfy (69) for all $\ell \in [L]$, $i \in [n_\ell]$, and $s \in [s_{\max} \wedge d_\ell]$ with probability meeting the lower bound in Theorem 3.

Consider the reduced OMP algorithm for the data point $\mathbf{x}_i^{(\ell)}$ with fixed $\ell \in [L]$ and fixed $i \in [n_\ell]$. We start by noting that the reduced OMP index set Λ_s is a function of the data points in $\mathcal{X}_\ell$ only. After iteration s, with $\mathbf{x}_i^{(\ell)} = \mathbf{\Phi}\mathbf{U}^{(\ell)}\mathbf{a}_i^{(\ell)}$ and $\mathbf{X}_{\Lambda_s}^{(\ell)} = \mathbf{\Phi}\mathbf{U}^{(\ell)}\mathbf{A}_{\Lambda_s}^{(\ell)}$ inserted into (3), we get $\mathbf{r}_s^{(\ell)} = \mathbf{\Phi}\mathbf{U}^{(\ell)}\tilde{\mathbf{r}}_s^{(\ell)}$, where

$$\tilde{\mathbf{r}}_s^{(\ell)} \coloneqq (\mathbf{I} - \mathbf{A}_{\Lambda_s}^{(\ell)}(\mathbf{\Phi}\mathbf{U}^{(\ell)}\mathbf{A}_{\Lambda_s}^{(\ell)})^\dagger \mathbf{\Phi}\mathbf{U}^{(\ell)})\mathbf{a}_i^{(\ell)}.$$

We next establish a lower bound on the RHS of (69) and an upper bound on the LHS of (69). To isolate the impact of the different random quantities in the statistical data model, we will introduce events, upon the intersection of which (69) is implied by (9) via these bounds. A union bound on the probability of the intersection of these events then yields the final result.

We start by lower-bounding the RHS of (69) according to

$$\max_{j \in [n_\ell]\colon j \neq i} \left| \left\langle \mathbf{x}_j^{(\ell)}, \mathbf{r}_s^{(\ell)} \right\rangle \right| \geq \frac{1}{4}\sqrt{\frac{\log \rho_\ell}{d_\ell}}\sigma_{\min}\left(\mathbf{U}^{(\ell)^T}\mathbf{\Phi}^T\mathbf{\Phi}\mathbf{U}^{(\ell)}\right)\left\|\tilde{\mathbf{r}}_s^{(\ell)}\right\|_2 \tag{70}$$

$$\geq \frac{1}{4}\sqrt{\frac{\log \rho_\ell}{d_\ell}}(1-\delta)\left\|\tilde{\mathbf{r}}_s^{(\ell)}\right\|_2, \tag{71}$$

where (70) and (71) hold on the events

$$\mathcal{E}_1^{(\ell,i)} \coloneqq \left\{ \max_{j \in [n_\ell]\colon j \neq i} \left| \mathbf{x}_j^{(\ell)^T}\mathbf{\Phi}\mathbf{U}^{(\ell)}\mathbf{v} \right| > \frac{1}{4}\sqrt{\frac{\log \rho_\ell}{d_\ell}}\sigma_{\min}\left(\mathbf{U}^{(\ell)^T}\mathbf{\Phi}^T\mathbf{\Phi}\mathbf{U}^{(\ell)}\right)\|\mathbf{v}\|_2,\ \forall \mathbf{v} \in \mathbb{R}^{d_\ell} \right\}$$

and

$$\mathcal{E}_2 \coloneqq \left\{ \min_\ell \sigma_{\min}\left(\mathbf{U}^{(\ell)^T}\mathbf{\Phi}^T\mathbf{\Phi}\mathbf{U}^{(\ell)}\right) > 1-\delta \right\}, \qquad \delta \in (0,1),$$

respectively. Note that $\tilde{\mathbf{r}}_s^{(\ell)}$ in (70) not being statistically independent of the $\mathbf{x}_j^{(\ell)}$, $j \neq i$, is not an issue as we consider (70) on the event $\mathcal{E}_1^{(\ell,i)}$ and the inequality in the definition of $\mathcal{E}_1^{(\ell,i)}$ applies to *all* $\mathbf{v} \in \mathbb{R}^{d_\ell}$. Since $\mathbf{V}^{(\ell)} = \mathbf{\Phi}\mathbf{U}^{(\ell)}$ has full rank on $\mathcal{E}_2$, reduced OMP terminates after $s_{\max} \wedge d_\ell$

iterations. To see this, simply note that for $\mathbf{V}^{(\ell)}$ of full rank we need exactly d_ℓ points from $\mathcal{X}_\ell \backslash \{\mathbf{x}_i^{(\ell)}\}$ to represent $\mathbf{x}_i^{(\ell)} = \mathbf{V}^{(\ell)} \mathbf{a}_i^{(\ell)}$ (owing to the fact that the $\mathbf{a}_j^{(\ell)}$, $j \in [n_\ell]$, are i.i.d. uniform on $\mathbb{S}^{d_\ell - 1}$) and thus $\mathbf{r}_s^{(\ell)} = \mathbf{0}$ after exactly d_ℓ iterations.

We continue by upper-bounding the LHS of (69) according to

$$\begin{aligned}
\max_{k \neq \ell, j} \left| \left\langle \mathbf{x}_j^{(k)}, \mathbf{r}_s^{(\ell)} \right\rangle \right| &= \max_{k \neq \ell, j} \left| \mathbf{a}_j^{(k)^T} \mathbf{U}^{(k)^T} \mathbf{\Phi}^T \mathbf{\Phi} \mathbf{U}^{(\ell)} \tilde{\mathbf{r}}_s^{(\ell)} \right| \\
&= \max_{k \neq \ell, j} \left| \mathbf{a}_j^{(k)^T} \mathbf{U}^{(k)^T} \mathbf{U}^{(\ell)} \tilde{\mathbf{r}}_s^{(\ell)} + \mathbf{a}_j^{(k)^T} \mathbf{U}^{(k)^T} (\mathbf{\Phi}^T \mathbf{\Phi} - \mathbf{I}) \mathbf{U}^{(\ell)} \tilde{\mathbf{r}}_s^{(\ell)} \right| \\
&\leq \max_{k \neq \ell, j} \left| \mathbf{a}_j^{(k)^T} \mathbf{U}^{(k)^T} \mathbf{U}^{(\ell)} \tilde{\mathbf{r}}_s^{(\ell)} \right| + \max_{k \neq \ell, j} \left| \mathbf{a}_j^{(k)^T} \mathbf{U}^{(k)^T} (\mathbf{\Phi}^T \mathbf{\Phi} - \mathbf{I}) \mathbf{U}^{(\ell)} \tilde{\mathbf{r}}_s^{(\ell)} \right| \\
&< 4(3 \log N + \log s_{\max}) \max_{k \neq \ell} \frac{\left\| \mathbf{U}^{(k)^T} \mathbf{U}^{(\ell)} \right\|_F}{\sqrt{d_k} \sqrt{d_\ell}} \left\| \tilde{\mathbf{r}}_s^{(\ell)} \right\|_2 \\
&\quad + \sqrt{\frac{6 \log N + 2 \log s_{\max}}{d_{\min}}} \left\| \mathbf{U}^{(k)^T} (\mathbf{\Phi}^T \mathbf{\Phi} - \mathbf{I}) \mathbf{U}^{(\ell)} \right\|_{2 \to 2} \left\| \tilde{\mathbf{r}}_s^{(\ell)} \right\|_2 \qquad (72) \\
&\leq 4(3 \log N + \log s_{\max}) \max_{k \neq \ell} \frac{\left\| \mathbf{U}^{(k)^T} \mathbf{U}^{(\ell)} \right\|_F}{\sqrt{d_k} \sqrt{d_\ell}} \left\| \tilde{\mathbf{r}}_s^{(\ell)} \right\|_2 \\
&\quad + \sqrt{\frac{6 \log N + 2 \log s_{\max}}{d_{\min}}} \delta \left\| \tilde{\mathbf{r}}_s^{(\ell)} \right\|_2 \qquad (73) \\
&\leq \frac{1}{4} \sqrt{\frac{\log \rho_\ell}{d_\ell}} (1 - \delta) \left\| \tilde{\mathbf{r}}_s^{(\ell)} \right\|_2. \qquad (74)
\end{aligned}$$

Here, (72) holds on the intersection of the events

$$\begin{aligned}
\mathcal{E}_3^{(\ell,i,s)} &\coloneqq \left\{ \max_{k \neq \ell, j} \left| \mathbf{a}_j^{(k)^T} \mathbf{U}^{(k)^T} (\mathbf{\Phi}^T \mathbf{\Phi} - \mathbf{I}) \mathbf{U}^{(\ell)} \tilde{\mathbf{r}}_s^{(\ell)} \right| \right. \\
&\qquad \left. \leq \sqrt{\frac{6 \log N + 2 \log s_{\max}}{d_{\min}}} \left\| \mathbf{U}^{(k)^T} (\mathbf{\Phi}^T \mathbf{\Phi} - \mathbf{I}) \mathbf{U}^{(\ell)} \right\|_{2 \to 2} \left\| \tilde{\mathbf{r}}_s^{(\ell)} \right\|_2 \right\}, \\
\mathcal{E}_4^{(\ell,i,s)} &\coloneqq \left\{ \max_{k \neq \ell, j} \left| \mathbf{a}_j^{(k)^T} \mathbf{U}^{(k)^T} \mathbf{U}^{(\ell)} \tilde{\mathbf{r}}_s^{(\ell)} \right| < 4(3 \log N + \log s_{\max}) \max_{k \neq \ell} \frac{\left\| \mathbf{U}^{(k)^T} \mathbf{U}^{(\ell)} \right\|_F}{\sqrt{d_k} \sqrt{d_\ell}} \left\| \tilde{\mathbf{r}}_s^{(\ell)} \right\|_2 \right\}
\end{aligned}$$

and (73) holds on the event

$$\mathcal{E}_5 \coloneqq \left\{ \max_{k, \ell \colon k \neq \ell} \left\| \mathbf{U}^{(k)^T} (\mathbf{\Phi}^T \mathbf{\Phi} - \mathbf{I}) \mathbf{U}^{(\ell)} \right\|_{2 \to 2} < \delta \right\}.$$

Recall that the notation $\tilde{\mathbf{r}}_s^{(\ell)}$ does not reflect dependence on the index i. We do, however, make the dependence of $\mathcal{E}_3^{(\ell,i,s)}$ and $\mathcal{E}_4^{(\ell,i,s)}$ on i explicit.

Finally, setting $\delta \coloneqq \sqrt{\frac{28 d_{\max} + 8 \log L + 2\tau}{3 \tilde{c} p}}$ in (73), (74) follows from assumption (9). This is seen

as follows:

$$\max_{k\neq\ell}\frac{\left\|\mathbf{U}^{(k)^T}\mathbf{U}^{(\ell)}\right\|_F}{\sqrt{d_k}}+\frac{\delta\sqrt{\frac{d_\ell}{d_{\min}}}}{2\sqrt{6\log N+2\log s_{\max}}}\leq\max_{k,\ell:\ k\neq\ell}\operatorname{aff}(\mathcal{S}_k,\mathcal{S}_\ell)+\frac{\delta}{2}\sqrt{\frac{d_{\max}}{d_{\min}}} \tag{75}$$

$$\leq\frac{3}{200}\frac{\sqrt{\log\rho_{\min}}}{\log N} \tag{76}$$

$$\leq\frac{\sqrt{\log\rho_{\min}}}{50(\log N+(\log s_{\max})/3)} \tag{77}$$

$$\leq\frac{\sqrt{\log\rho_\ell}}{48(\log N+(\log s_{\max})/3)}(1-\delta), \tag{78}$$

where (76) is by (9) and (77) follows by noting that $(200/3)\log N = 50(\log N + (\log N)/3) > 50(\log N + (\log s_{\max})/3)$. Furthermore, we have

$$\frac{\sqrt{\log\rho_{\min}}}{\log N+(\log s_{\max})/3}\leq 1 \tag{79}$$

as a consequence of $\rho_{\min}=\min_\ell(n_\ell-1)/d_\ell < N/d_{\min}$, $N\geq 3$, and $d_{\min}\geq 1$. Next, (79) combined with $\sqrt{d_{\min}/d_{\max}}\leq 1$, $\max_{k,\ell:\ k\neq\ell}\operatorname{aff}(\mathcal{S}_k,\mathcal{S}_\ell)\geq 0$, and (77), implies that $\delta\leq\frac{2}{50}$, which yields $\frac{1}{50}\leq\frac{1}{48}(1-\delta)$ and hence establishes (78). Finally, (74) is obtained by rewriting the relation between the RHS of (75) and (78).

Note that the lower bound (71) on the RHS of (69) and the upper bound (74) on the LHS of (69) are equal; we have therefore established that, for fixed (i,ℓ), $\mathbf{r}_s^{(\ell)}$ obeys (69) on $\mathcal{E}_1^{(\ell,i)}\cap\mathcal{E}_2\cap\mathcal{E}_3^{(\ell,i,s)}\cap\mathcal{E}_4^{(\ell,i,s)}\cap\mathcal{E}_5$. It finally follows that on the event $\mathcal{E}_\star\coloneqq\bigcap_{\ell,i,s}\mathcal{E}_1^{(\ell,i)}\cap\mathcal{E}_2\cap\mathcal{E}_3^{(\ell,i,s)}\cap\mathcal{E}_4^{(\ell,i,s)}\cap\mathcal{E}_5$, the graph G obtained by SSC-OMP applied to the full data set $\mathcal{X}\setminus\{\mathbf{x}_i^{(\ell)}\}$ has no false connections. It remains to lower-bound $\mathrm{P}[\mathcal{E}_\star]$. Specifically, we have

$$\begin{aligned}\mathrm{P}[\mathcal{E}_\star]&=1-\mathrm{P}\left[\overline{\mathcal{E}_\star}\right]\\&\geq 1-\mathrm{P}\left[\overline{\mathcal{E}_2}\right]-\mathrm{P}\left[\overline{\mathcal{E}_5}\right]-\sum_{\ell\in[L],i\in[n_\ell]}\left(\mathrm{P}\left[\overline{\mathcal{E}}_1^{(\ell,i)}\right]+\sum_{s\in[s_{\max}\wedge d_\ell]}\left(\mathrm{P}\left[\overline{\mathcal{E}}_3^{(\ell,i,s)}\right]+\mathrm{P}\left[\overline{\mathcal{E}}_4^{(\ell,i,s)}\right]\right)\right)\\&\geq 1-4e^{-\tau/2}-\sum_{\ell\in[L]}n_\ell e^{-\sqrt{\rho_\ell}d_\ell}-\frac{4}{N},\end{aligned} \tag{80}$$

where the last inequality follows from

$$\mathrm{P}\left[\overline{\mathcal{E}}_1^{(\ell,i)}\right]\leq e^{-\sqrt{\rho_\ell}d_\ell} \tag{81}$$

$$\mathrm{P}\left[\overline{\mathcal{E}_2}\right]\leq 2e^{-\tau/2} \tag{82}$$

$$\mathrm{P}\left[\overline{\mathcal{E}}_3^{(\ell,i,s)}\right]\leq\frac{2}{s_{\max}N^2} \tag{83}$$

$$\mathrm{P}\left[\overline{\mathcal{E}}_4^{(\ell,i,s)}\right]\leq\frac{2}{s_{\max}N^2} \tag{84}$$

$$\mathrm{P}\left[\overline{\mathcal{E}_5}\right]\leq 2e^{-\tau/2}. \tag{85}$$

Here, (85) corresponds to (45), while the proofs of (81)–(84) are presented below.

Proof of (81)**:** Since $\mathbf{A}_{-i}^{(\ell)}$ has full column rank with probability 1, it follows from Lemma 4 below that

$$\begin{aligned}\left\|\mathbf{A}_{-i}^{(\ell)^T}\mathbf{U}^{(\ell)^T}\mathbf{\Phi}^T\mathbf{\Phi}\mathbf{U}^{(\ell)}\mathbf{v}\right\|_\infty &\geq r(\mathcal{P}(\mathbf{A}_{-i}^{(\ell)}))\left\|\mathbf{U}^{(\ell)^T}\mathbf{\Phi}^T\mathbf{\Phi}\mathbf{U}^{(\ell)}\mathbf{v}\right\|_2 \\ &\geq r(\mathcal{P}(\mathbf{A}_{-i}^{(\ell)}))\,\sigma_{\min}\left(\mathbf{U}^{(\ell)^T}\mathbf{\Phi}^T\mathbf{\Phi}\mathbf{U}^{(\ell)}\right)\|\mathbf{v}\|_2,\end{aligned}$$

for all $\mathbf{v}\in\mathbb{R}^{d_\ell}$. We therefore have

$$\begin{aligned}\mathrm{P}\left[\overline{\mathcal{E}}_1^{(\ell,i)}\right] &= \mathrm{P}\left[\left\|\mathbf{A}_{-i}^{(\ell)^T}\mathbf{U}^{(\ell)^T}\mathbf{\Phi}^T\mathbf{\Phi}\mathbf{U}^{(\ell)}\mathbf{v}\right\|_\infty \leq \frac{1}{4}\sqrt{\frac{\log\rho_\ell}{d_\ell}}\sigma_{\min}\left(\mathbf{U}^{(\ell)^T}\mathbf{\Phi}^T\mathbf{\Phi}\mathbf{U}^{(\ell)}\right)\|\mathbf{v}\|_2\right] \\ &\leq \mathrm{P}\left[r(\mathcal{P}(\mathbf{A}_{-i}^{(\ell)}))\leq\frac{1}{4}\sqrt{\frac{\log\rho_\ell}{d_\ell}}\right] \\ &\leq e^{-\sqrt{\rho_\ell}d_\ell}, \qquad (86)\end{aligned}$$

where (86) follows from (66), which uses the assumption $(n_\ell-1)/d_\ell=\rho_\ell\geq\rho_0>1$.

Lemma 4. *For a matrix $\mathbf{A}\in\mathbb{R}^{m\times n}$ of full column rank and $\mathbf{v}\in\mathbb{R}^m$, it holds that*

$$\left\|\mathbf{A}^T\mathbf{v}\right\|_\infty \geq r(\mathcal{P}(\mathbf{A}))\|\mathbf{v}\|_2, \qquad (87)$$

where $r(\mathcal{P}(\mathbf{A}))$ is the inradius of the symmetrized convex hull $\mathcal{P}(\mathbf{A})$ of the columns of $\mathbf{A}$.

Proof. The inequality (87) obviously holds for $\mathbf{v}=\mathbf{0}$. Pick any $\mathbf{v}\in\mathbb{R}^m\setminus\{\mathbf{0}\}$ and take $\epsilon\in(0,1)$. Let $\eta=\epsilon\|\mathbf{v}\|_2 r(\mathcal{P}(\mathbf{A}))$ and assume that $\mathbf{v}\in\eta\mathcal{P}^\circ(\mathbf{A})=\{\mathbf{z}\colon\left\|\mathbf{A}^T\mathbf{z}\right\|_\infty\leq\eta\}$, i.e., $\mathbf{v}$ is an element of the η-scaled version of the polar set $\mathcal{P}^\circ(\mathbf{A})$. Note that $\eta>0$ as $\|\mathbf{v}\|_2>0$, $\epsilon>0$, and $r(\mathcal{P}(\mathbf{A}))>0$ thanks to $\mathbf{A}$ having full column rank. It follows from [13, Thm. 1.2] that

$$\frac{\|\mathbf{v}\|_2}{\eta}\leq R(\mathcal{P}^\circ(\mathbf{A}))=\frac{1}{r(\mathcal{P}(\mathbf{A}))}. \qquad (88)$$

Now, owing to $\eta=\epsilon\|\mathbf{v}\|_2 r(\mathcal{P}(\mathbf{A}))$, (88) implies that $\epsilon\geq 1$, which contradicts $\epsilon\in(0,1)$. It therefore follows that $\mathbf{v}\in\mathbb{R}^m\setminus\{\eta\mathcal{P}^\circ(\mathbf{A})\}$ for all $\epsilon\in(0,1)$, which in turn implies that $\left\|\mathbf{A}^T\mathbf{v}\right\|_\infty>\eta=\epsilon\|\mathbf{v}\|_2 r(\mathcal{P}(\mathbf{A}))$ for all $\epsilon\in(0,1)$. In particular, letting $\epsilon\to 1$ yields $\left\|\mathbf{A}^T\mathbf{v}\right\|_\infty\geq\|\mathbf{v}\|_2 r(\mathcal{P}(\mathbf{A}))$ as desired. □

Proof of (82)**:** With $\sigma_{\min}(\mathbf{A})=\left\|\mathbf{A}^{-1}\right\|_{2\to2}^{-1}$ [33, Sec. 5.2.1] for a full rank matrix $\mathbf{A}\in\mathbb{R}^{m\times m}$ it follows that

$$\begin{aligned}\mathrm{P}\left[\overline{\mathcal{E}_2}\right] &= \mathrm{P}\left[\min_\ell\left\|\left(\mathbf{U}^{(\ell)^T}\mathbf{\Phi}^T\mathbf{\Phi}\mathbf{U}^{(\ell)}\right)^{-1}\right\|_{2\to2}^{-1}\leq 1-\delta\right] \\ &= \mathrm{P}\left[\max_\ell\left\|\left(\mathbf{U}^{(\ell)^T}\mathbf{\Phi}^T\mathbf{\Phi}\mathbf{U}^{(\ell)}\right)^{-1}\right\|_{2\to2}\geq\frac{1}{1-\delta}\right] \\ &\leq 2e^{-\tau/2},\end{aligned}$$

where $\tau>0$ is the numerical constant in Theorem 3 and the last inequality is thanks to (44).

Proof of (83)**:** By the union bound

$$\begin{aligned}
\mathrm{P}\Big[\overline{\mathcal{E}}_3^{(\ell,i,s)}\Big] &\leq \sum_{k\neq\ell,j} \mathrm{P}\Bigg[\Big|\mathbf{a}_j^{(k)^T}\mathbf{U}^{(k)^T}(\boldsymbol{\Phi}^T\boldsymbol{\Phi}-\mathbf{I})\mathbf{U}^{(\ell)}\tilde{\mathbf{r}}_s^{(\ell)}\Big| \\
&\qquad\qquad > \sqrt{\frac{6\log N + 2\log s_{\max}}{d_{\min}}}\Big\|\mathbf{U}^{(k)^T}(\boldsymbol{\Phi}^T\boldsymbol{\Phi}-\mathbf{I})\mathbf{U}^{(\ell)}\Big\|_{2\to 2}\Big\|\tilde{\mathbf{r}}_s^{(\ell)}\Big\|_2\Bigg] \\
&\leq \sum_{k\neq\ell,j} \mathrm{P}\Bigg[\Big|\mathbf{a}_j^{(k)^T}\mathbf{U}^{(k)^T}(\boldsymbol{\Phi}^T\boldsymbol{\Phi}-\mathbf{I})\mathbf{U}^{(\ell)}\tilde{\mathbf{r}}_s^{(\ell)}\Big| \\
&\qquad\qquad > \sqrt{\frac{6\log N + 2\log s_{\max}}{d_k}}\Big\|\mathbf{U}^{(k)^T}(\boldsymbol{\Phi}^T\boldsymbol{\Phi}-\mathbf{I})\mathbf{U}^{(\ell)}\tilde{\mathbf{r}}_s^{(\ell)}\Big\|_2\Bigg] \\
&\leq \sum_{k\neq\ell,j} \frac{2}{s_{\max}N^3} \leq \frac{2}{s_{\max}N^2}, \qquad (89)
\end{aligned}$$

where (89) follows from Proposition 1 with $\mathbf{a} = \mathbf{a}_j^{(k)}$, $\mathbf{b} = \mathbf{U}^{(k)^T}(\boldsymbol{\Phi}^T\boldsymbol{\Phi}-\mathbf{I})\mathbf{U}^{(\ell)}\tilde{\mathbf{r}}_s^{(\ell)}$, and $\beta = \sqrt{6\log N + 2\log s_{\max}}$.

Proof of (84)**:** We first show that $\tilde{\mathbf{r}}_s^{(\ell)}/\|\tilde{\mathbf{r}}_s^{(\ell)}\|_2$ is distributed uniformly at random on $\mathbb{S}^{d_\ell-1}$; (84) then follows by application of Lemma 3.

Recall that we consider reduced OMP, which computes a sparse representation of $\mathbf{x}_i^{(\ell)} = \boldsymbol{\Phi}\mathbf{U}^{(\ell)}\mathbf{a}_i^{(\ell)}$ using the columns of $\mathbf{X}_{-i}^{(\ell)} = \boldsymbol{\Phi}\mathbf{U}^{(\ell)}\mathbf{A}_{-i}^{(\ell)}$ as dictionary elements, i.e., Λ_s and $\tilde{\mathbf{r}}_s^{(\ell)}$ depend only on the random quantities $\boldsymbol{\Phi}\mathbf{U}^{(\ell)}$, $\mathbf{a}_i^{(\ell)}$, and $\mathbf{A}_{-i}^{(\ell)}$. In order to reflect these restricted dependencies, we write $\tilde{\mathbf{r}}_s^{(\ell)} = \tilde{\mathbf{r}}_s^{(\ell)}(\boldsymbol{\Phi}\mathbf{U}^{(\ell)}, \mathbf{a}_i^{(\ell)}, \mathbf{A}_{-i}^{(\ell)})$ and $\Lambda_s = \Lambda_s(\boldsymbol{\Phi}\mathbf{U}^{(\ell)}, \mathbf{a}_i^{(\ell)}, \mathbf{A}_{-i}^{(\ell)},)$. Here, the first argument specifies the basis matrix of the data points, the second argument corresponds to the coefficient vector of the data point (in the basis specified by the first argument) a sparse representation is to be computed for, and the third argument designates the coefficient matrix of the dictionary elements (again in the basis specified by the first argument).

We start by showing that the distribution of $\tilde{\mathbf{r}}_s^{(\ell)}$ is rotationally invariant. For a deterministic unitary matrix $\mathbf{W} \in \mathbb{R}^{d_\ell \times d_\ell}$, we have

$$\Lambda_s(\boldsymbol{\Phi}\mathbf{U}^{(\ell)}\mathbf{W}^T, \mathbf{W}\mathbf{a}_i^{(\ell)}, \mathbf{W}\mathbf{A}_{-i}^{(\ell)}) = \Lambda_s(\boldsymbol{\Phi}\mathbf{U}^{(\ell)}, \mathbf{a}_i^{(\ell)}, \mathbf{A}_{-i}^{(\ell)})$$

as the $\mathbf{x}_j^{(\ell)}$ can be written as $\mathbf{x}_j^{(\ell)} = \boldsymbol{\Phi}\mathbf{U}^{(\ell)}\mathbf{a}_j^{(\ell)} = \boldsymbol{\Phi}\mathbf{U}^{(\ell)}\mathbf{W}^T\mathbf{W}\mathbf{a}_j^{(\ell)}$.

Using the shorthand notation Λ_s' for $\Lambda_s(\boldsymbol{\Phi}\mathbf{U}^{(\ell)}\mathbf{W}^T, \mathbf{W}\mathbf{a}_i^{(\ell)}, \mathbf{W}\mathbf{A}_{-i}^{(\ell)})$ and recalling that $\tilde{\mathbf{r}}_s^{(\ell)} = (\mathbf{I} - \mathbf{A}_{\Lambda_s}^{(\ell)}(\boldsymbol{\Phi}\mathbf{U}^{(\ell)}\mathbf{A}_{\Lambda_s}^{(\ell)})^\dagger\boldsymbol{\Phi}\mathbf{U}^{(\ell)})\mathbf{a}_i^{(\ell)}$, it follows that

$$\begin{aligned}
\tilde{\mathbf{r}}_s^{(\ell)}(\boldsymbol{\Phi}\mathbf{U}^{(\ell)}\mathbf{W}^T, \mathbf{W}\mathbf{a}_i^{(\ell)}, \mathbf{W}\mathbf{A}_{-i}^{(\ell)}) &= \Big(\mathbf{I} - \mathbf{W}\mathbf{A}_{\Lambda_s'}^{(\ell)}\Big(\boldsymbol{\Phi}\mathbf{U}^{(\ell)}\mathbf{W}^T\mathbf{W}\mathbf{A}_{\Lambda_s'}^{(\ell)}\Big)^\dagger\boldsymbol{\Phi}\mathbf{U}^{(\ell)}\mathbf{W}^T\Big)\mathbf{W}\mathbf{a}_i^{(\ell)} \\
&= \Big(\mathbf{I} - \mathbf{W}\mathbf{A}_{\Lambda_s}^{(\ell)}\Big(\boldsymbol{\Phi}\mathbf{U}^{(\ell)}\mathbf{W}^T\mathbf{W}\mathbf{A}_{\Lambda_s}^{(\ell)}\Big)^\dagger\boldsymbol{\Phi}\mathbf{U}^{(\ell)}\mathbf{W}^T\Big)\mathbf{W}\mathbf{a}_i^{(\ell)} \\
&= \mathbf{W}\Big(\mathbf{I} - \mathbf{A}_{\Lambda_s}^{(\ell)}\Big(\boldsymbol{\Phi}\mathbf{U}^{(\ell)}\mathbf{A}_{\Lambda_s}^{(\ell)}\Big)^\dagger\boldsymbol{\Phi}\mathbf{U}^{(\ell)}\Big)\mathbf{a}_i^{(\ell)} \\
&= \mathbf{W}\tilde{\mathbf{r}}_s^{(\ell)}(\boldsymbol{\Phi}\mathbf{U}^{(\ell)}, \mathbf{a}_i^{(\ell)}, \mathbf{A}_{-i}^{(\ell)}).
\end{aligned}$$

By rotational invariance of the distributions of $\mathbf{a}_i^{(\ell)}, \mathbf{A}_{-i}^{(\ell)}$, and $\mathbf{\Phi}$ (by assumption in Theorem 3), we have $\mathbf{W}\mathbf{a}_i^{(\ell)} \sim \mathbf{a}_i^{(\ell)}, \mathbf{W}\mathbf{A}_{-i}^{(\ell)} \sim \mathbf{A}_{-i}^{(\ell)}$, and $\mathbf{\Phi}\mathbf{U}^{(\ell)}\mathbf{W}^T \sim \mathbf{\Phi}\mathbf{U}^{(\ell)}$ (because $\text{span}(\mathbf{U}^{(\ell)}\mathbf{W}^T) = \text{span}(\mathbf{U}^{(\ell)})$ and the columns of $\mathbf{U}^{(\ell)}\mathbf{W}^T$ are orthonormal). We therefore get

$$\begin{aligned}\tilde{\mathbf{r}}_s^{(\ell)}(\mathbf{\Phi}\mathbf{U}^{(\ell)}, \mathbf{a}_i^{(\ell)}, \mathbf{A}_{-i}^{(\ell)}) &\sim \tilde{\mathbf{r}}_s^{(\ell)}(\mathbf{\Phi}\mathbf{U}^{(\ell)}\mathbf{W}^T, \mathbf{W}\mathbf{a}_i^{(\ell)}, \mathbf{W}\mathbf{A}_{-i}^{(\ell)}) \\ &= \mathbf{W}\tilde{\mathbf{r}}_s^{(\ell)}(\mathbf{\Phi}\mathbf{U}^{(\ell)}, \mathbf{a}_i^{(\ell)}, \mathbf{A}_{-i}^{(\ell)}). \end{aligned} \tag{90}$$

Since (90) holds for all unitary matrices $\mathbf{W}$, the distribution of $\tilde{\mathbf{r}}_s^{(\ell)}$ is rotationally invariant and $\tilde{\mathbf{r}}_s^{(\ell)}/\|\tilde{\mathbf{r}}_s^{(\ell)}\|_2$ is, indeed, distributed uniformly on $\mathbb{S}^{d_\ell - 1}$. We finally exploit this property of $\tilde{\mathbf{r}}_s^{(\ell)}$ to upper-bound $\text{P}\left[\overline{\mathcal{E}}_4^{(\ell,i,s)}\right]$ as follows. A union bound over all k, $k \neq \ell$, yields

$$\begin{aligned}\text{P}\left[\overline{\mathcal{E}}_4^{(\ell,i,s)}\right] &\leq \sum_{k\neq\ell} \text{P}\left[\left\|\mathbf{A}^{(k)^T}\mathbf{U}^{(k)^T}\mathbf{U}^{(\ell)}\tilde{\mathbf{r}}_s^{(\ell)}\right\|_\infty \geq 4(3\log N + \log s_{\max})\frac{\left\|\mathbf{U}^{(k)^T}\mathbf{U}^{(\ell)}\right\|_F}{\sqrt{d_k}\sqrt{d_\ell}}\left\|\tilde{\mathbf{r}}_s^{(\ell)}\right\|_2\right] \\ &\leq \sum_{k\neq\ell} \frac{n_k + 1}{s_{\max}N^3} = \frac{N - n_\ell + L - 1}{s_{\max}N^3} < \frac{N+L}{s_{\max}N^3} \leq \frac{2}{s_{\max}N^2}, \end{aligned} \tag{91}$$

where (91) follows by application of Lemma 3 with $\mathbf{L} = \mathbf{A}^{(k)}$, $\mathbf{a} = \tilde{\mathbf{r}}_s^{(\ell)}/\|\tilde{\mathbf{r}}_s^{(\ell)}\|_2$, $\mathbf{B} = \mathbf{U}^{(k)^T}\mathbf{U}^{(\ell)}$, and $c = 4(3\log N + \log s_{\max})$.

References

[1] N. Ailon and E. Liberty. An almost optimal unrestricted fast Johnson-Lindenstrauss transform. *ACM Trans. Algorithms*, 9(3):1–12, 2013.

[2] David Alonso-Gutiérrez. On the isotropy constant of random convex sets. *Proc. Amer. Math. Soc.*, 136(9):3293–3300, 2008.

[3] R. Basri and D.W. Jacobs. Lambertian reflectance and linear subspaces. *IEEE Trans. Pattern Anal. Mach. Intell.*, 25(2):218–233, 2003.

[4] S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

[5] Broad-Institute. Cancer program data sets, 2013. `http://www.broadinstitute.org/cgi-bin/cancer/datasets.cgi`.

[6] E. J. Candes, J. Romberg, and T. Tao. Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information. *IEEE Trans. Inf. Theory*, 52(2):489–509, 2006.

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

[8] E. Elhamifar and R. Vidal. Sparse subspace clustering. In *Proc. of IEEE Conf. Comput. Vision Pattern Recogn.*, pages 2790–2797, 2009.

[9] E. Elhamifar and R. Vidal. Sparse subspace clustering: Algorithm, theory, and applications. *IEEE Trans. Pattern Anal. Machine Intell.*, 35(11):2765–2781, 2013.

[10] S. Foucart and H. Rauhut. *A Mathematical Introduction to Compressive Sensing*. Springer, Berlin, Heidelberg, 2013.

[11] 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*, 23(6):643–660, 2001.

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

[13] P. Gritzmann and V. Klee. Inner and outer j-radii of convex bodies in finite-dimensional normed spaces. *Discrete Comput. Geom.*, 7(1):255–280, 1992.

[14] T. Hastie and P. Y. Simard. Metrics and models for handwritten character recognition. *Stat. Sci.*, 13(1):54–65, 1998.

[15] R. Heckel, E. Agustsson, and H. Bölcskei. Neighborhood selection for thresholding based subspace clustering. In *Proc. of IEEE Int. Conf. Acoust. Speech Sig. Proc. (ICASSP)*, pages 6761–6765. IEEE, 2014.

[16] R. Heckel and H. Bölcskei. Robust subspace clustering via thresholding. *IEEE Trans. Inform. Theory*, 61(11):6320–6342, 2015.

[17] R. Heckel, M. Tschannen, and H. Bölcskei. Subspace clustering of dimensionality-reduced data. In *Proc. of IEEE Int. Symp. on Inf. Theory*, pages 2997–3001. IEEE, July 2014.

[18] R. A. Horn and C. R. Johnson. *Matrix analysis*. Cambridge University Press, 2012.

[19] Jiaji Huang, Qiang Qiu, and Robert Calderbank. The role of principal angles in subspace classification. *preprint, arXiv:1507.04230*, 2015.

[20] D. Jiang, C. Tang, and A. Zhang. Cluster analysis for gene expression data: A survey. *IEEE Trans. Knowl. Data Eng.*, 16(11):1370–1386, 2004.

[21] F. Krahmer and R. Ward. New and improved Johnson-Lindenstrauss embeddings via the restricted isometry property. *SIAM J. Math. Anal.*, 43(3):1269–1281, 2011.

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

[23] Y. LeCun and C. Cortes. The MNIST database, 2013. http://yann.lecun.com/exdb/mnist/.

[24] K. C. Lee, J. Ho, and D. J. Kriegman. Acquiring linear subspaces for face recognition under variable lighting. *IEEE Trans. Pattern Anal. Mach. Intell.*, 27(5):684–698, 2005.

[25] K. Liu, H. Kargupta, and J. Ryan. Random projection-based multiplicative data perturbation for privacy preserving distributed data mining. *IEEE Trans. Knowl. Data Eng.*, 18(1):92–106, 2006.

[26] A. Ng, I. M. Jordan, and W. Yair. On spectral clustering: Analysis and an algorithm. In *Advances in Neural Information Processing Systems*, pages 849–856, 2001.

[27] M. Nokleby, M. Rodrigues, and R. Calderbank. Discrimination on the Grassmann manifold: Fundamental limits of subspace classifiers. *IEEE Trans. Inform. Theory*, 61(4):2133–2147, 2015.

[28] M. Soltanolkotabi and E. J. Candès. A geometric analysis of subspace clustering with outliers. *Ann. Stat.*, 40(4):2195–2238, 2012.

[29] M. Soltanolkotabi, E. Elhamifar, and E. J. Candès. Robust subspace clustering. *Ann. Stat.*, 42(2):669–699, 2014.

[30] D. Spielman. Spectral graph theory. Lecture notes, 2012.

[31] M. Tschannen. Dimensionality reduction for sparse subspace clustering. MS thesis, ETH Zurich, March 2014.

[32] S. S. Vempala. *The Random Projection Method*. American Mathematical Society, 2005.

[33] R. Vershynin. Introduction to the non-asymptotic analysis of random matrices. In Y. C. Eldar and G. Kutyniok, editors, *Compressed sensing: Theory and applications*, pages 210–268. Cambridge University Press, 2012.

[34] R. Vidal. Subspace clustering. *IEEE Signal Process. Mag.*, 28(2):52–68, 2011.

[35] U. von Luxburg. A tutorial on spectral clustering. *Stat. Comput.*, 17(4):395–416, 2007.

[36] Y. Wang, Y. Wang, and A. Singh. A deterministic analysis of noisy sparse subspace clustering for dimensionality-reduced data. In *Proc. of Int. Conf. on Machine Learning (ICML)*, pages 1422–1431, 2015.

[37] C. You and R. Vidal. Sparse subspace clustering by orthogonal matching pursuit. July 2015.

[38] T. Zhang, A. Szlam, Yi Wang, and G. Lerman. Hybrid linear modeling via local best-fit flats. *Int. J. Comput. Vision*, 100:217–240, 2012.