Convergence and Recovery Guarantees of the K-Subspaces Method for Subspace Clustering

Peng Wang [1] **Huikang Liu** [2] **Anthony Man-Cho So** [3] **Laura Balzano** [1]

Abstract

The K-subspaces (KSS) method is a generalization of the K-means method for subspace clustering. In this work, we present local convergence analysis and a recovery guarantee for KSS, assuming data are generated by the semi-random union of subspaces model, where N points are randomly sampled from $K \geq 2$ overlapping subspaces. We show that if the initial assignment of the KSS method lies within a neighborhood of a true clustering, it converges at a superlinear rate and finds the correct clustering within $\Theta(\log\log N)$ iterations with high probability. Moreover, we propose a thresholding inner-product based spectral method for initialization and prove that it produces a point in this neighborhood. We also present numerical results of the studied method to support our theoretical developments.

1. Introduction

Subspace clustering (SC) is a fundamental problem in unsupervised learning, which can be applied to do dimensionality reduction and data analysis. It has found wide applications in diverse fields, such as computer vision (Ho et al., 2003; Vidal et al., 2008), gene expression analysis (Jiang et al., 2004; Ucar et al., 2011), and image segmentation (Hong et al., 2006), to name a few. In research on SC, the union of subspace (UoS) model, which assumes that data points lie in one of multiple underlying subspaces, is a typical model for studying SC. In particular, substantial advances have been made recently on designing algorithms for solving the SC problem and on establishing theoretical foundations in the UoS model; see, e.g., Vidal (2011); Vidal et al. (2016); Meng et al. (2018) and the references therein.

In the UoS model, the goal of SC is to recover the underlying subspaces and cluster the unlabeled data points into the corresponding subspaces. To achieve this goal, many algorithms have been proposed in the past two decades, such as sparse subspace clustering methods (Elhamifar & Vidal, 2013; Wang & Xu, 2013), low-rank representation-based methods (Liu et al., 2012), thresholding-based methods (Heckel & Bölcskei, 2015; Li & Gu, 2021), and K-subspaces (KSS) method (Bradley & Mangasarian, 2000). In these methods, the KSS method, which is known as a generalization of the K-means method, can handle clusters in subspaces. In particular, it is conceptually simple and has linear complexity per iteration. This computational benefits render it suitable to handle large-scale datasets in practice. However, a complete theoretical understanding of its convergence behavior and recovery performance is not found in the literature, to the best of our knowledge. This is due in part to its alternating and discrete nature, as well as the fact that as with the K-means, KSS can easily get stuck in bad local minima without a good initialization. Consequently, it remains a major challenge to provide the theoretical foundations for KSS. In this work, we provide guarantees for the convergence behavior and recovery performance of the KSS method. We also develop a simple initialization method with provable guarantees for the KSS method. It is worth mentioning that our results improve on state-of-the-art theory with respect to allowable affinity between subspaces, and support the algorithm's competitive performance in our numerical evaluation.

1.1. Related Works

Over the past years, a substantial body of literature explores algorithmic development and theoretical analysis of SC. One of the most well-studied methods is arguably sparse subspace clustering (SSC), which is motivated by representing each data point as a sparse linear combination of the remaining ones. A seminal work by Elhamifar & Vidal

[1]Department of Electrical Engineering and Computer Science, University of Michigan, Ann Arbor [2]Research Institute for Interdisciplinary Sciences, School of Information Management and Engineering, Shanghai University of Finance and Economics, Shanghai [3]Department of Systems Engineering and Engineering Management, The Chinese University of Hong Kong, Hong Kong. Correspondence to: Peng Wang <peng8wang@gmail.com>.

Proceedings of the 39^{th} International Conference on Machine Learning, Baltimore, Maryland, USA, PMLR 162, 2022.

Table 1. Comparison of affinity requirement and recovery results of the surveyed methods in the noiseless semi-random UoS model with *overlapping* subspaces ($K \geq 2$).

References	Methods	Affinity[a]	Results
Soltanolkotabi et al. (2012)	SSC	$O(\frac{1}{\sqrt{\log N}})$	SDP
Wang & Xu (2013)	SSC	$O(\frac{1}{\sqrt{\log N}})$	SDP
Tschannen & Bölcskei (2018)	(O)MP	$O(\frac{1}{\sqrt{\log N}})$	SDP
Wang et al. (2016)	SSC	$O(\frac{1}{\sqrt{\log N}})$	CC
Heckel & Bölcskei (2015)	TSC	$O(\frac{1}{\sqrt{\log N}})$	CC
Park et al. (2014)	GSR	$O(\frac{1}{\sqrt{\log N}})^b$	CC
Lipor et al. (2021)	EKSS	$O(\frac{1}{\sqrt{\log N}})$	CC
Ours	**KSS**	$O(1)$	CC

[a] We use the notion (15) to measure the subspace affinity. [b] This is obtained by taking $\delta = 1/N$ in Park et al. (2014, Theorem 3.2).

(2013) proposed and studied this method. The algorithm proceeds by solving a convex sparse optimization problem, followed by applying spectral clustering to the graph constructed by a solution of this convex problem. In particular, they showed that when the data points are drawn from the *disjoint* subspaces in the noiseless setting, the solution is non-trivial and no edges in the constructed graph connect two points in different subspaces. This is referred to as the *subspace detection property* (SDP) in literature; see, e.g., Wang & Xu (2013); Soltanolkotabi et al. (2012; 2014). We should point out that SDP does not imply *correct clustering* (CC) of data points as mentioned in Wang et al. (2016); Li & Gu (2021). Following this line of work, theoretical results on the SSC method in various contexts have been established, and many variants and extensions of the SSC method have been proposed. For example, Soltanolkotabi et al. (2012) developed a unified analysis framework of the SSC method, which showed that the SDP holds even when subspaces can be *overlapping* in the noiseless setting. Later, Soltanolkotabi et al. (2014) extended their analysis and results to the noisy setting. Meanwhile, an independent work by Wang & Xu (2013) also studied the behavior of SSC based on the SDP in the noisy setting. In spite of the solid theoretical guarantees and great empirical performance, SSC suffers from high computational cost. To tackle this issue, Dyer et al. (2013) applied an orthogonal matching pursuit (OMP) method to SSC. Then, Tschannen & Bölcskei (2018) analyzed the performance of this method in the noisy setting and also introduced and studied the matching pursuit (MP) method for SSC. Recently, more and more variants and extensions for solving SSC have been proposed; see, e.g., Ding et al. (2021); Wang et al. (2019); Wu et al. (2020); Chen et al. (2020); Matsushima & Brbic (2019); Traganitis & Giannakis (2017); You et al. (2016).

As for other methods for SC, Liu et al. (2012) proposed a low-rank representation (LRR) method by minimizing a nuclear norm regularized problem. In particular, they showed that the proposed method can recover the row space of the data points. Later, Shen et al. (2016) developed an online version of the LLR method, which reduces its computational cost significantly. Another notable approach for SC is thresholding-based methods, which exploit the correlation between data points. For example, Heckel & Bölcskei (2015) proposed a thresholding-based subspace clustering (TSC) method, which applies spectral clustering to a weight matrix with entries depending on spherical distances of each data point to its nearest neighbors. They showed that TSC can achieve correct clustering by proving that the formed graph has no false connection and K connected subgraphs. Li & Gu (2021) proposed a thresholding inner-product (TIP) method for SC, which constructs an adjacency matrix by thresholding magnitudes of inner products between data points. In particular, they provided an explicit bound on the error rate of the TIP method when there are only two subspaces of the same dimension. Moreover, Park et al. (2014) proposed a greedy subspace clustering (GSC) method that constructs a neighborhood matrix using a nearest subspace neighbor method and then recovers subspaces by a greedy algorithm. They showed that their approach can guarantee correct clustering. However, they assumed that the dimension of each subspace is known and same and the number of data points in each subspace is also same. Actually, there are still numerous other popular methods using different techniques for SC, such as matrix factorization-based method (Boult & Brown, 1991; Pimentel-Alarcón et al., 2016; Fan, 2021) and principal component analysis type methods (Vidal et al., 2005; McWilliams & Montana, 2014).

In contrast to the above methods, the KSS method (Bradley & Mangasarian, 2000; Agarwal & Mustafa, 2004; Tseng, 2000) is essentially a generalization of the *k-means* method for SC, which minimizes the sum of distances of each point to its projection onto the assigned subspace, i.e.,

$$\min_{\mathcal{C},\boldsymbol{U}} \sum_{k=1}^{K} \sum_{i \in \mathcal{C}_k} \|\boldsymbol{z}_i - \boldsymbol{U}_k \boldsymbol{U}_k^T \boldsymbol{z}_i\|^2, \tag{1}$$

where $\{\boldsymbol{z}_i\}_{i=1}^N \subseteq \mathbb{R}^n$ denotes the set of N data points, $\mathcal{C} = \{\mathcal{C}_k\}_{k=1}^K$ denotes the set of $K \geq 2$ estimated clusters, and $\boldsymbol{U} = \begin{bmatrix}\boldsymbol{U}_1 & \dots & \boldsymbol{U}_K\end{bmatrix}$ with $\boldsymbol{U}_k$ being an orthonormal basis of the corresponding cluster. Similar to the k-means method, the KSS method proceeds by alternating between the *subspace update step* and the *cluster assignment step*. As a local search algorithm, it is conceptually simple and has linear complexity as a function of the number of data points, while many popular methods based on self-expression property, such as the surveyed SSC, OMP-based SSC, and LLR, have at least quadratic complexity. This computational advantage renders it more suitable to

handle large-scale datasets than these self-expression property based-methods. However, due to its non-convex nature, it suffers from sensitivity to initialization and lack of theoretical understanding. To fix the former issue, some heuristics for good initialization have been proposed; see, e.g., He et al. (2016); Zhang et al. (2009). To improve the performance of the KSS method, Gitlin et al. (2018) employed a coherence pursuit algorithm. Recently, Lipor et al. (2021) applied an ensembles approach to the KSS method with random initialization and showed that it achieves correct clustering based on the argument in Heckel & Bölcskei (2015). However, their analysis can only tackle one KSS iteration. Generally, it remains open to propose a provable initialization scheme for the KSS method and fully understand its convergence behavior and recovery performance. Due to this, the KSS method has mostly been superseded by convex methods based on self-expression property, which are widely studied and have solid theoretical results. Moreover, establishing theoretical foundations for the KSS method may open the door for the study of various non-convex methods for SC.

1.2. Our Contributions

In this work, we study the KSS method for SC in the semi-random UoS model. First, we provide theoretical guarantees for the convergence behavior and recovery performance of the KSS method. Specifically, we prove the existence of a *basin of attraction*, whose radius is as large as $O(\sqrt{N})$ around the true clustering of the data points, when the cluster sample sizes are on the same order and the subspace dimensions are also on the same order. If the initial assignment of the KSS method lies within this basin, the algorithm is guaranteed to converge to the true clustering at a *superlinear* rate. In particular, once the number of iterations reaches $\Theta(\log\log N)$, the KSS method yields the true clustering with the corresponding orthonormal bases exactly. It is worth emphasizing that these results are obtained under the condition that the normalized affinity between pairwise subspaces is $O(1)$, which is generally milder than those in the existing literature; see the comparison in Table 1. Second, we propose a thresholding inner-product based spectral method for initialization of the KSS method. We show that it can generate a point lying in the basin of attraction of the KSS method by deriving its clustering error rate. Our core argument is to derive a spectral bound for a random adjacency matrix without independence structure, which could be of independent interest. In conclusion, our work demystifies the computational efficiency of the KSS method and provides a provable initialization scheme for it, thus bridging the gap between theory and practice. From a broader perspective, our work also contributes to the literature on simple and scalable non-convex methods with provable guarantees; see, e.g., Wang et al. (2021a;b); Zhang et al. (2020); Gao & Zhang (2019); Boumal (2016); Ling (2022).

Notation. Let $\mathbb{R}^n$ be the n-dimensional Euclidean space and $\|\cdot\|$ be the Euclidean norm. Given a matrix $\boldsymbol{A}$, we use $\|\boldsymbol{A}\|$ to denote its spectral norm, $\sigma_i(\boldsymbol{A})$ its i-th largest singular value, $\|\boldsymbol{A}\|_F$ its Frobenius norm, and a_{ij} its (i,j)-th element. Given a vector $\boldsymbol{a}\in\mathbb{R}^n$, we denote by a_i its i-th element. Given a positive integer n, we denote by $[n]$ the set $\{1,\dots,n\}$. Given $d_1,\dots,d_K$, let $d_{\min}=\min\{d_k : k\in[K]\}$ and $d_{\max}=\max\{d_k : k\in[K]\}$. Given a discrete set S, we denote by $|S|$ its cardinality. Given two sets $A,B\subseteq[n]$, the set difference between A and B denoted by $A\setminus B$ is defined by $A\setminus B=\{x\in A : x\notin B\}$. We use $\mathcal{O}^{n\times d}$ to denote the set of all $n\times d$ matrices that have orthonormal columns (in particular, $\mathcal{O}^d$ denotes the set of all $d\times d$ orthogonal matrices) and Π_K to denote the set of all $K\times K$ permutation matrices. Let $\pi:[K]\to[K]$ denote a permutation of the elements in $[K]$. Each π corresponds to a $\boldsymbol{Q}_\pi=\{q_{ij}\}_{1\le i,j\le K}\in\Pi_K$ such that $q_{ij}=1$ if $j=\pi(i)$ and $q_{ij}=0$ otherwise for all $i\in[K]$. The converse also holds. Moreover, for any $\boldsymbol{U},\boldsymbol{V}\in\mathcal{O}^{n\times d}$, we denote by $d(\boldsymbol{U},\boldsymbol{V})=\left\|\boldsymbol{U}\boldsymbol{U}^T-\boldsymbol{V}\boldsymbol{V}^T\right\|$ the distance between the subspaces spanned by $\boldsymbol{U}$ and $\boldsymbol{V}$. We define $\mathbb{S}^{d-1}=\left\{\boldsymbol{a}\in\mathbb{R}^d : \|\boldsymbol{a}\|=1\right\}$ and denote by $\mathrm{Unif}(\mathbb{S}^{d-1})$ a uniform distribution over the sphere in $\mathbb{R}^d$. For non-negative sequences $\{a_k\}$ and $\{b_k\}$, we write $a_k\gtrsim b_k$ if there exists a universal constant $C>0$ such that $a_k\ge Cb_k$ for all k.

2. Preliminaries and Main Results

In this section, we formally set up the SC problem in the semi-random UoS model, introduce the KSS method for tackling the SC problem, propose an initialization scheme for the KSS method, and give a summary of our main results.

2.1. Semi-Random UoS Model

Definition 1. *Suppose that a family of sets $\{\mathcal{C}_k\}_{k=1}^K$ is a partition of $[N]$. We say that $\boldsymbol{H}\in\mathbb{R}^{N\times K}$ is a membership matrix if $h_{ik}=1$ if $i\in\mathcal{C}_k$ and $h_{ik}=0$ otherwise. For simplicity, we use $\mathcal{M}^{N\times K}$ to denote the collections of all such $N\times K$ membership matrices.*

Given an $\boldsymbol{H}\in\mathcal{M}^{N\times K}$, each row of it has exactly one 1 and $(K-1)$ 0's. Besides, $\boldsymbol{H}\boldsymbol{Q}$ for any $\boldsymbol{Q}\in\Pi_K$ represents the same partition as $\boldsymbol{H}$ up to a permutation of the cluster labels. We define the distance between two membership matrices $\boldsymbol{H},\boldsymbol{H}'\in\mathcal{M}^{N\times K}$ by

$$d_F(\boldsymbol{H},\boldsymbol{H}')=\min_{\boldsymbol{Q}\in\Pi_K}\|\boldsymbol{H}-\boldsymbol{H}'\boldsymbol{Q}\|_F.$$

Then, one can verify that the number of misclassified points in $\boldsymbol{H}$ with respect to $\boldsymbol{H}'$ is $d_F^2(\boldsymbol{H},\boldsymbol{H}')/2$.

Definition 2 (Semi-Random UoS Model[1]). *Let S_k^* denote a subspace of $\mathbb{R}^n$ of dimension d_k with $\boldsymbol{U}_k^* \in \mathcal{O}^{n\times d_k}$ being its orthonormal basis for all $k \in [K]$. Let $\boldsymbol{H}^* \in \mathcal{M}^{N\times K}$ represent a partition of $[N]$ into K clusters, each of which is of size N_k for all $k \in [K]$. Then, we say that a collection of $N \geq 2$ points $\{\boldsymbol{z}_i\}_{i=1}^N$ is generated according to the semi-random UoS model with parameters $\left(N, K, \{\boldsymbol{U}_k^*\}_{k=1}^K, \boldsymbol{H}^*\right)$ if*

$$\boldsymbol{z}_i = \boldsymbol{U}_k^* \boldsymbol{a}_i, \tag{2}$$

where $k \in [K]$ satisfies $h_{ik}^* = 1$ and $\boldsymbol{a}_i \overset{i.i.d.}{\sim} \mathrm{Unif}(\mathbb{S}^{d_k-1})$ for all $i \in [N]$.

Intuitively, given an unknown partition encoded by $\boldsymbol{H}^*$, this model generates a collection of unlabeled observations. Given these observations, the goal of SC is to design an algorithm that finds the true partition, i.e., $\boldsymbol{H}^*\boldsymbol{Q}$ for some $\boldsymbol{Q} \in \Pi_K$. We should point out that the subspace dimensions $d_1, \ldots, d_K$ are also all unknown.

2.2. The KSS Method

In this subsection, we introduce the KSS method by interpreting it as an application of the alternating minimization method to Problem (1). Such an interpretation is similar to that in Bradley & Mangasarian (2000). By introducing $\boldsymbol{H} \in \mathcal{M}^{N\times K}$, we can reformulate Problem (1) as

$$\begin{aligned} \min \ & \sum_{i=1}^N \sum_{k=1}^K h_{ik} \left(\|\boldsymbol{z}_i\|^2 - \|\boldsymbol{U}_k^T \boldsymbol{z}_i\|^2\right) \\ \text{s.t. } & \boldsymbol{H} \in \mathcal{M}^{N\times K},\ \boldsymbol{U}_k \in \mathcal{O}^{n\times \hat{d}_k},\ \text{for all } k \in [K], \end{aligned} \tag{3}$$

where $\hat{d}_k$ for all $k \in [K]$ are candidate subspace dimensions. Observe that this problem is in a form that is amenable to the alternating minimization method (see, e.g., Ghosh & Kannan (2020); Hardt (2014); Zhang (2020)). Given the current iterate $(\boldsymbol{H}^t, \boldsymbol{U}_1^t, \ldots, \boldsymbol{U}_K^t) \in \mathcal{M}^{N\times K} \times \mathcal{O}^{n\times \hat{d}_1} \times \cdots \times \mathcal{O}^{n\times \hat{d}_K}$, the method generates the next iterate via

$$\boldsymbol{U}_k^{t+1} \in \underset{\boldsymbol{U}_k \in \mathcal{O}^{n\times \hat{d}_k}}{\arg\min} \sum_{i=1}^N h_{ik}^t \left(\|\boldsymbol{z}_i\|^2 - \|\boldsymbol{U}_k^T \boldsymbol{z}_i\|^2\right) \tag{4}$$

for all $k \in [K]$ and

$$\boldsymbol{H}^{t+1} \in \mathcal{T}\left(\boldsymbol{G}_{\boldsymbol{H}}(\boldsymbol{U}^{t+1})\right), \tag{5}$$

where the (i,k)-th element of $\boldsymbol{G}_{\boldsymbol{H}}(\boldsymbol{U}) \in \mathbb{R}^{N\times K}$ is $\|\boldsymbol{z}_i\|^2 - \|\boldsymbol{U}_k^T \boldsymbol{z}_i\|^2$ and $\mathcal{T}$ denotes the operator that for any $\boldsymbol{G} \in \mathbb{R}^{N\times K}$,

$$\mathcal{T}(\boldsymbol{G}) = \arg\min \left\{\langle \boldsymbol{G}, \boldsymbol{H}\rangle : \boldsymbol{H} \in \mathcal{M}^{N\times K}\right\}. \tag{6}$$

[1]This model is called semi-random UoS because the subspaces are arbitrary but data points are randomly generated.

It is worth noting that the updates (4) and (5) both admit closed-form solutions, which respectively correspond to the *subspace update step* and the *cluster assignment step* of the KSS method. Indeed, the update (4) is typically a PCA problem and its solution is given by

$$\boldsymbol{U}_k^{t+1} = \mathrm{PCA}\left(\sum_{i=1}^N h_{ik}^t \boldsymbol{z}_i \boldsymbol{z}_i^T, \hat{d}_k\right), \tag{7}$$

where $\mathrm{PCA}(\boldsymbol{A}, d) : \mathbb{S}^n \times \mathbb{R} \to \mathbb{R}^{n\times d}$ is the operator that computes the eigenvectors associated with the d leading eigenvalues of $\boldsymbol{A}$. Moreover, the update (5) is a special assignment problem, whose solution is given by

$$h_{ik}^{t+1} = \begin{cases} 1, & \text{if } k = I_i, \\ 0, & \text{otherwise}, \end{cases} \tag{8}$$

where $I_i \in [K]$ satisfies $\|\boldsymbol{U}_{I_i}^{t+1^T} \boldsymbol{z}_i\| \geq \|\boldsymbol{U}_k^{t+1^T} \boldsymbol{z}_i\|$ for all $k \neq I_i$; see Lemma 5.

A natural question arising in the update (7) is how to choose $\hat{d}_k$ for all $k \in [K]$. Generally, the KSS method assumes that the subspace dimensions $d_1, \ldots, d_K$ are known beforehand (Vidal, 2011), which is not practical in applications. Even if $d_1, \ldots, d_K$ are known but unequal, it is still unknown how to find an one-to-one mapping between $\{d_k\}_{k=1}^K$ and $\{\hat{d}_k\}_{k=1}^K$ due to the fact that the permutation of clusters is unknown. To fix this issue, we propose an adaptive strategy to choose $\hat{d}_k$ for all $k \in [K]$. Specifically, let $\lambda_{k1}^t \geq \cdots \geq \lambda_{kd}^t$ be the d leading eigenvalues of $\sum_{i=1}^N h_{ik}^t \boldsymbol{z}_i \boldsymbol{z}_i^T$ for all $k \in [K]$, where d is an input parameter satisfying $d > d_{\max}$. Then for all $k \in [K]$, we set

$$\hat{d}_k^{t+1} = \underset{i\in[d-1]}{\arg\max} \left(\lambda_{ki}^t - \lambda_{k(i+1)}^t\right) \tag{9}$$

and replace (7) by

$$\boldsymbol{U}_k^{t+1} = \mathrm{PCA}\left(\sum_{i=1}^N h_{ik}^t \boldsymbol{z}_i \boldsymbol{z}_i^T, \hat{d}_k^{t+1}\right). \tag{10}$$

2.3. Initialization Method

A key ingredient in our approach is to identify a proper initial point $\boldsymbol{H}^0$ that may guarantee rapid convergence of the KSS method for solving Problem (3). Motivated by the thresholding inner-product based scheme in Li & Gu (2021), we propose a thresholding inner-product based spectral method (TIPS) for initialization. Specifically, given a thresholding parameter $\tau > 0$, a graph G with adjacency matrix $\boldsymbol{A} \in \mathbb{R}^{N\times N}$ is generated by

$$a_{ij} = \begin{cases} 1, & \text{if } |\langle \boldsymbol{z}_i, \boldsymbol{z}_j\rangle| \geq \tau \text{ and } i \neq j, \\ 0, & \text{otherwise}, \end{cases} \tag{11}$$

for all $1 \le i \le j \le N$. Then, the initial cluster assignment $\boldsymbol{H}^0$ is obtained by applying the *k-means* to the matrix $\boldsymbol{V}$ formed by the eigenvectors associated with the K leading eigenvalues of $\boldsymbol{A}$. Although it is NP-hard in the worst case to compute a global minimizer of the k-means problem (see, e.g., Aloise et al. (2009)), some polynomial-time algorithms have been proposed for finding an approximate solution whose value is within a constant fraction of the optimal value (see, e.g., Kumar & Kannan (2010)), i.e.,

$$(\boldsymbol{H}^0, \hat{\boldsymbol{X}}) \in \mathcal{M}^{N\times K} \times \mathbb{R}^{K\times K} \quad \text{s.t. } \|\boldsymbol{H}^0\hat{\boldsymbol{X}} - \boldsymbol{V}\|_F^2 \le (1+\theta) \min_{(\boldsymbol{H},\boldsymbol{X})\in\mathcal{M}^{N\times K}\times\mathbb{R}^{K\times K}} \|\boldsymbol{H}\boldsymbol{X} - \boldsymbol{V}\|_F^2, \quad (12)$$

where $\theta > 0$ is a constant. We assume that we can find such an approximate solution. We now summarize the proposed method in Algorithm 1.

Algorithm 1 The TIPS initialized KSS method

Input: samples $\{\boldsymbol{z}_i\}_{i=1}^N$, $\tau > 0$, $\theta > 0$, $d, T, K \in \mathbb{Z}_+$
`/* The TIPS initialization */`
construct an adjacency matrix $\boldsymbol{A} \in \mathbb{R}^{N\times N}$ by (11)
calculate $\boldsymbol{V} \in \mathbb{R}^{N\times K}$ formed by the eigenvectors associated with the K leading eigenvalues of $\boldsymbol{A}$
let $(\boldsymbol{H}^0, \hat{\boldsymbol{X}})$ be a $(1+\theta)$-approximate solution to the k-means problem (12) with K clusters and input $\boldsymbol{V}$
`/* The KSS method */`
for $t = 0, 1, \ldots, T$ **do**
  `/* subspace update step */`
  for $k = 1, \ldots, K$ **do**
    Compute $\hat{d}_k^{t+1}$ via (9) and $\boldsymbol{U}_k^{t+1}$ via (10)
  end for
  `/* cluster assignment step */`
  compute $\boldsymbol{H}^{t+1}$ via (8)
end for

2.4. Main Theorems

Before we proceed, we introduce a definition to capture notions of affinity between pairwise subspaces and impose an assumption on the affinity.

Definition 3. *The affinity between two subspaces S_k and S_ℓ is defined by*

$$\mathrm{aff}(S_k, S_\ell) = \sqrt{\sum_{i=1}^{\min\{d_k,d_\ell\}} \left(\sigma_{k\ell}^{(i)}\right)^2}, \quad (13)$$

where $\sigma_{k\ell}^{(1)} \ge \cdots \ge \sigma_{k\ell}^{(\min\{d_k,d_\ell\})} \ge 0$ are the singular vaules of $\boldsymbol{U}_k^T\boldsymbol{U}_\ell \in \mathbb{R}^{d_k\times d_\ell}$ with $\boldsymbol{U}_k, \boldsymbol{U}_\ell$ being respectively orthonormal bases of S_k and S_ℓ. The normalized affinity between two subspaces S_k and S_ℓ is defined by

$$\overline{\mathrm{aff}}(S_k, S_\ell) = \frac{\mathrm{aff}(S_k, S_\ell)}{\min\{\sqrt{d_k}, \sqrt{d_\ell}\}}. \quad (14)$$

For ease of exposition, we define the maximum of the normalized affinities as

$$\kappa = \max_{1\le k\ne\ell\le K} \overline{\mathrm{aff}}(S_k^*, S_\ell^*) \quad (15)$$

and define

$$\kappa_d = \frac{d_{\max}}{d_{\min}},\ \kappa_N = \frac{N_{\max}}{N_{\min}}. \quad (16)$$

Assumption 1. *The affinity between pairwise subspaces in the UoS model satisfies $\kappa \in (0, 1/2]$.*

We remark that this affinity condition is milder than those in the related literature. Because this assumption allows the affinity $\mathrm{aff}(S_k, S_\ell)$ to be as large as $\min\{\sqrt{d_k}, \sqrt{d_\ell}\}/2$, while those in the literature require $\mathrm{aff}(S_k, S_\ell) \lesssim \min\{\sqrt{d_k}, \sqrt{d_\ell}\}/\sqrt{\log N}$ for all $1 \le k \ne \ell \le K$. Please see the comparison in Table 1. We next present a main theorem of this work, which shows that the KSS method converges superlinearly and achieves the correct clustering under Assumption 1.

Theorem 1. *Let $\{\boldsymbol{z}_i\}_{i=1}^N$ be data points generated according to the semi-random UoS model with parameters $(N, K, \{\boldsymbol{U}_k^*\}_{k=1}^K, \boldsymbol{H}^*)$. Suppose that Assumption 1 holds, $N_{\min} \gtrsim d_k \gtrsim \log N$ for all $k \in [K]$, and the initial point $\boldsymbol{H}^0 \in \mathcal{M}^{N\times K}$ satisfies*

$$d_F(\boldsymbol{H}^0, \boldsymbol{H}^*) \le \frac{(1-\kappa)N_{\min}}{5\kappa_d\sqrt{N}}. \quad (17)$$

Set $T = \Theta(\log\log N)$ and $d \in \mathbb{Z}_+$ satisfying $d > d_{\max}$ in Algorithm 1. Then, the following statements hold with probability at least $1 - N^{-\Omega(1)}$:
(i) For all $t \in [T]$, it holds that

$$d_F(\boldsymbol{H}^t, \boldsymbol{H}^*) \le \kappa_1^{2^t-1} d_F\left(\boldsymbol{H}^0, \boldsymbol{H}^*\right), \quad (18)$$

where $\kappa_1 \in (0,1)$ is an absolute constant.
(ii) It holds for a permutation $\pi : [K] \to [K]$ that

$$\boldsymbol{H}^T = \boldsymbol{H}^*\boldsymbol{Q}_\pi \quad (19)$$

and $\hat{d}_{\pi(k)}^{T+1} = d_k$ for all $k \in [K]$,

$$\boldsymbol{U}_{\pi(k)}^{T+1} = \boldsymbol{U}_k^*\boldsymbol{O}_k,\ \boldsymbol{O}_k \in \mathcal{O}^{d_k} \text{ for all } k \in [K]. \quad (20)$$

Before we proceed, some remarks are in order. First, while Problem (1) is NP-hard in the worst case (Gitlin et al., 2018), the assumption that the data points are generated by the semi-random UoS model allows us to conduct an average-case analysis of the KSS method. Second, a neighborhood of size $O\left(\frac{N_{\min}}{\kappa_d\sqrt{N}}\right)$ around each true cluster forms a *basin of attraction* in the UoS model, in which the KSS method converges superlinearly. In particular, if κ_d, κ_N

are both constant, we can see that the size of this basin is $O(\sqrt{N})$, which is rather large. Provided that the initial point $\boldsymbol{H}^0$ lies within this basin, the subsequent iterates are guaranteed to converge to ground truth at a *superlinear* rate. Third, if the number of iterations reaches $\Theta(\log\log N)$, the KSS method can not only find correct clustering, but also exactly recovers the orthonormal basis of each subspace. This demonstrates the efficacy of the KSS method. Finally, any method that can return a point satisfying (17) is qualified as an initialization scheme for the KSS method. In this work, we design a simple initialization scheme in the first stage of Algorithm 1 that can provably generate a point in the basin of attraction under the following assumption. Before we proceed, let $\boldsymbol{B} \in \mathbb{R}^{K\times K}$ be a symmetric matrix whose elements are given by

$$b_{k\ell} = 2 - 2\Phi\left(\frac{\tau\sqrt{d_k d_\ell}}{\mathrm{aff}(S_k^*, S_\ell^*)}\right),\ \forall\, 1 \le k,\ell \le K, \quad (21)$$

where $\Phi(x) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{x}\exp\left(-\frac{t^2}{2}\right)dt$ denotes the cumulative distribution function of the standard normal distribution. It is worth noting that $b_{k\ell}$ is an approximation of the probability of $a_{ij} = 1$ if $\boldsymbol{z}_i \in S_k^*$ and $\boldsymbol{z}_j \in S_\ell^*$ for all $1 \le k,\ell \le K$, where a_{ij} is given in (11); see Lemma 1.

Assumption 2. *The thresholding parameter is set as*

$$\tau = \frac{\sqrt{c}}{\sqrt{d_{\max}}}, \quad (22)$$

where $c > 0$ is a constant. The parameter κ_d is a constant and the maximum of the normalized affinities satisfying

$$\kappa \in \left(0, \frac{\sqrt{c}}{\sqrt{\kappa_d}\Phi^{-1}\left(1 - \frac{1-\Phi(\sqrt{c})}{2(K-1)}\right)}\right) \quad (23)$$

is also a constant. Moreover, the affinity between pairwise subspaces satisfies

$$\mathrm{aff}(S_k^*, S_\ell^*) \gtrsim \log N,\ \forall\, 1 \le k \ne \ell \le K \quad (24)$$

and the subspace dimension satisfies

$$d_{\min} \gtrsim \log^3 N\ . \quad (25)$$

We will use this assumption in the following theorem, restricting our result to the high affinity case. In general, the clustering becomes harder as the affinity increases; see, e.g., Soltanolkotabi et al. (2014, Section 1.3.1). Then, it is natural to assume that κ is a constant and (24) holds. We want to also highlight that (24) implies that our subspaces are of generally moderate dimension, which is made precise in (25) of the assumption. While this is slightly restrictive, it is in line with theoretical results in other subspace clustering literature, and it also simplifies our theoretical analysis. We leave an analysis of the low-to-moderate affinity settings and low-rank subspaces to future work.

Theorem 2. *Let $\{\boldsymbol{z}_i\}_{i=1}^N$ be data points generated according to the semi-random UoS model with parameters $(N, K, \{\boldsymbol{U}_k^*\}_{k=1}^K, \boldsymbol{H}^*)$. Suppose that Assumption 2 holds, $\kappa_d \le \sqrt{\log N}$, and $\kappa_d\kappa_N^2 \lesssim \sqrt{\log N}$. It holds with probability at least $1 - N^{-\Omega(1)}$ that*

$$d_F(\boldsymbol{H}^0, \boldsymbol{H}^*) \lesssim \sqrt{\kappa_d}\kappa_N \frac{\sqrt{N_{\max}}}{\sqrt[4]{\log N}}. \quad (26)$$

In particular, if both κ_d and κ_N are constants and N is sufficiently large, $\boldsymbol{H}^0$ satisfies (17) with probability at least $1 - N^{-\Omega(1)}$.

To put the above results in perspective, we make some remarks. First, according to the fact that $d_F^2(\boldsymbol{H}^0, \boldsymbol{H}^*)/2$ denotes the number of misclassified data points, the bound (26) implies that the TIPS method only misclassifies $O(N/\sqrt{\log N})$ points when κ_d, κ_N are constants and the normalized subspace affinity is $O(1)$. This automatically satisfies (17), which requires the number of misclassified points to be $O(N)$ when κ_d, κ_N are constants. Second, we believe that the recovery error bound (26) can be improved by enhancing the spectral bound in Proposition 1. This is left for future research.

3. Proofs of Main Results

In this section, we sketch the proofs of the theorems in Section 2. The complete proofs can be found in Sections B, C, and D of the appendix.

3.1. Analysis of Initialization Method

In this subsection, our goal is to establish a recovery error bound of the TIPS method. To begin, we estimate the connection probability of data points (i.e., the probability of $a_{ij} = 1$ in (11)) according to their memberships after the thresholding procedure (11). Moreover, we show that the connection probability of data points in the same subspace is larger than that of data points in different subspaces.

Lemma 1. *Consider the setting in Theorem 2. Let $p_{k\ell} \in \mathbb{R}$ denote the connection probability between any pair of data points that respectively belong to the subspaces S_k^* and S_ℓ^* for all $1 \le k,\ell \le K$. Then, it holds for all $1 \le k,\ell \le K$ that*

$$|p_{k\ell} - b_{k\ell}| \lesssim \kappa_d/\sqrt{\log N}, \quad (27)$$

where $b_{k\ell}$ is defined in (21), and

$$p_{kk} - p_{k\ell} \gtrsim 1/\sqrt{\kappa_d}. \quad (28)$$

Under Assumption 2, we can show that the approximate connection matrix $\boldsymbol{B}$ is non-degenerate, which is crucial for the analysis of the k-means error bound.

Lemma 2. *Consider the setting in Theorem 2. The matrix $\boldsymbol{B}$ defined in (21) is of full rank and its smallest singular value γ satisfies $\gamma \geq 1 - \Phi(\sqrt{c})$, where c is the constant in Assumption 2.*

Next, we present a spectral bound on the deviation of $\boldsymbol{A}$ from its mean.

Proposition 1. *Consider the setting in Theorem 2. Then, it holds with probability at least $1 - 6K^2N^{-1}$ that*

$$\|\boldsymbol{A} - \mathbb{E}[\boldsymbol{A}]\| \lesssim \frac{\sqrt{\kappa_d}N}{\sqrt[4]{\log N}}. \tag{29}$$

Despite that this bound seems large, it is sufficient for proving (26). A key observation is that the entries in the i-th column of $\boldsymbol{A}$ are independent conditioned on $\boldsymbol{z}_i$, while they are dependent. This plays an important role in our analysis. Compared to the results in Lei et al. (2015, Theorem 5.2), this lemma provides a spectral bound for an adjacency matrix without independence structure, which could be of independent interest.

Equipped with Proposition 1, Assumption 2, Lemma 2, and Lei et al. (2015, Lemmas 5.1, 5.3), we can prove Theorem 2. The complete proof is provided in Section B.4 of the appendix.

3.2. Analysis of Subspace Update Step

In this subsection, we analyze convergence behavior of the subspace update step in the KSS iterations. For ease of exposition, let us introduce some further notation. Given an $\boldsymbol{H} \in \mathcal{M}^{N\times K}$, let $\mathcal{C}_k = \{i \in [N] : h_{ik} = 1\}$ and $n_k = |\mathcal{C}_k|$ for all $k \in [K]$. Given $\mathcal{C}_1, \ldots, \mathcal{C}_K$, let

$$n_{k\ell} = |\mathcal{C}_k \cap \mathcal{C}^*_\ell|, \quad \boldsymbol{\Psi}_{k\ell} = \frac{1}{n_{k\ell}} \sum_{i \in \mathcal{C}_k \cap \mathcal{C}^*_\ell} \boldsymbol{a}_i \boldsymbol{a}_i^T \tag{30}$$

for all $k, \ell \in [K]$, where $\boldsymbol{a}_i$ for all $i \in [N]$ are given in the UoS model. Given a permutation $\pi : [K] \to [K]$ and a partition $\{\mathcal{C}_1, \ldots, \mathcal{C}_K\}$ of $[N]$ represented by $\boldsymbol{H} \in \mathcal{M}^{N\times K}$, we define the maximum of the number of misclassified points in $\mathcal{C}_k$ w.r.t. $\mathcal{C}^*_{\pi^{-1}(k)}$ and that in $\mathcal{C}^*_k$ w.r.t. $\mathcal{C}_{\pi(k)}$ as

$$W_k(\boldsymbol{H}) = \max\left\{|\mathcal{C}_k \setminus \mathcal{C}^*_{\pi^{-1}(k)}|, |\mathcal{C}^*_k \setminus \mathcal{C}_{\pi(k)}|\right\}. \tag{31}$$

To begin, we present a lemma that estimates the singular values of $\boldsymbol{\Psi}_{k\ell}$ for all $1 \leq k \neq \ell \leq K$.

Lemma 3. *Suppose that $\pi : [K] \to [K]$ is a permutation, $N_k \gtrsim d_k \gtrsim \log N$ for all $k \in [K]$, and $\boldsymbol{H} \in \mathcal{M}^{N\times K}$ satisfies*

$$W_k(\boldsymbol{H}) \leq \frac{1}{8} N_{\min} \text{ for all } k \in [K]. \tag{32}$$

It holds with probability at least $1 - 2K/(d^2_{\min}N)$ for all $1 \leq k \neq \ell \leq K$ that

$$\left|\sigma_i\left(\boldsymbol{\Psi}_{\pi(k)k}\right) - \frac{1}{d_k}\right| \leq \frac{1}{32d_k} \text{ for all } i \in [d_k], \tag{33}$$

$$\sigma_1\left(\boldsymbol{\Psi}_{\pi(k)\ell}\right) \leq \frac{1}{d_\ell} + \frac{5c_1}{4d_\ell}\left(\sqrt{\frac{d_\ell}{n_{\pi(k)\ell}}} + \frac{d_\ell}{n_{\pi(k)\ell}}\right), \tag{34}$$

where $c_1 > 0$ is an absolute constant.

Armed with this lemma, we are now ready to show that the distance from the subspaces generated by the update steps to the true ones can be bounded by the number of misclassfied data points.

Lemma 4. *Let $\boldsymbol{G}_{\boldsymbol{U}_k}(\boldsymbol{H}) = \sum_{i=1}^N h_{ik}\boldsymbol{z}_i\boldsymbol{z}_i^T$ for some $\boldsymbol{H} \in \mathcal{M}^{N\times K}$ and $\lambda_{k1} \geq \cdots \geq \lambda_{kd}$ be the d leading eigenvalues of $\boldsymbol{G}_{\boldsymbol{U}_k}(\boldsymbol{H})$ for all $k \in [K]$. Suppose that for all $k \in [K]$,*

$$\hat{d}_k = \operatorname*{arg\,max}_{i\in[d-1]} \left(\lambda_{ki} - \lambda_{k(i+1)}\right) \tag{35}$$

and

$$\boldsymbol{U}_k = \mathrm{PCA}(\boldsymbol{G}_{\boldsymbol{U}_k}(\boldsymbol{H}), \hat{d}_k). \tag{36}$$

Suppose in addition that $\pi : [K] \to [K]$ is a permutation, $N_{\min} \gtrsim d_k \gtrsim \log N$ for all $k \in [K]$, and $\varepsilon \in (0, 1/(8\kappa_d)]$ is a constant such that

$$W_k(\boldsymbol{H}) \leq \varepsilon N_{\min} \text{ for all } k \in [K]. \tag{37}$$

Then, it holds with probability at least $1 - 2K/(d^2_{\min}N)$ that

$$\hat{d}_{\pi(k)} = d_k \text{ for all } k \in [K], \tag{38}$$

$$\sum_{k=1}^K d(\boldsymbol{U}_{\pi(k)}, \boldsymbol{U}^*_k) \leq \frac{2d_{\max}}{N_{\min}} \max\left\{\frac{1}{d_{\min}}\|\boldsymbol{H} - \boldsymbol{H}^*\boldsymbol{Q}_\pi\|_F^2, 2K^2(c_1+1)c_1\right\}, \tag{39}$$

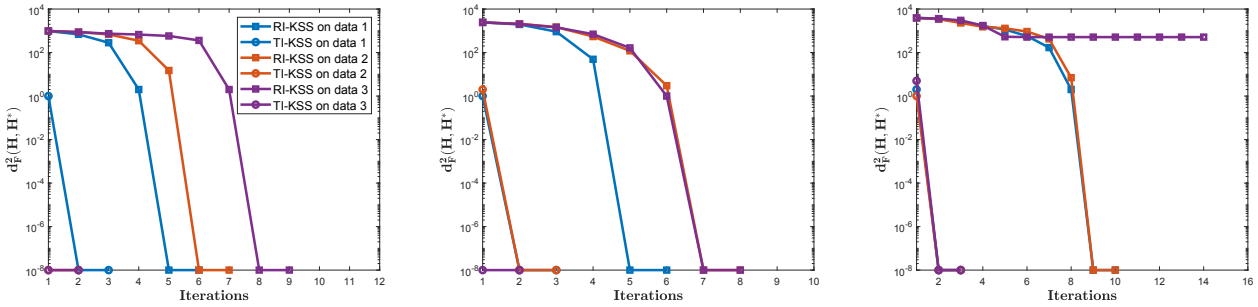
where c_1 is the constant in Lemma 3.

3.3. Analysis of Cluster Assignment Step

In this subsection, we turn to study convergence behavior of the cluster assignment step in the KSS iterations. Observe that Problem (6) is row-separable, and thus we can solve it by dividing it into N subproblems. Specifically, for a row of $\boldsymbol{G}$ denoted by $\boldsymbol{g} \in \mathbb{R}^K$, it suffices to consider

$$\mathcal{T}(\boldsymbol{g}) = \arg\min\left\{\langle \boldsymbol{g}, \boldsymbol{h}\rangle : \boldsymbol{h}^T\boldsymbol{1}_K = 1,\ \boldsymbol{h} \in \{0,1\}^K\right\}.$$

Then, we can show that this problem admits a closed-form solution, which may be not unique.



(a) $K = 3$ (b) $K = 6$ (c) $K = 9$

Figure 1. Convergence performance of KSS: The x-axis is number of iterations and the y-axis is the distance from an iterate to a ground truth, i.e., $d_F^2(\boldsymbol{H}^k, \boldsymbol{H}^*) + 10^{-8}$, where $\boldsymbol{H}^k$ is the k-th iterate generated by KSS.

Lemma 5. *For any $\boldsymbol{g} \in \mathbb{R}^K$, it holds that $\boldsymbol{v} \in \mathcal{T}(\boldsymbol{g})$ if and only if $v_k = 1$ and $v_\ell = 0$ for all $\ell \neq k$, where $k \in [K]$ satisfies $g_k \leq g_\ell$ for all $\ell \neq k$. Moreover, $\boldsymbol{v} \in \mathcal{T}(\boldsymbol{g})$ if and only if $\boldsymbol{Q}\boldsymbol{v} \in \mathcal{T}(\boldsymbol{Q}\boldsymbol{g})$ for $\boldsymbol{Q} \in \Pi_K$.*

Based on the above result, we can prove that the operator $\mathcal{T}$ possesses a Lipschitz-like property.

Lemma 6. *Suppose that $\boldsymbol{g} \in \mathbb{R}^K$ is arbitrary and $\delta > 0$ is a constant such that $g_\ell - g_k \geq \delta$ for some $k \in [K]$ and all $\ell \neq k$. Then, for any $\boldsymbol{v} \in \mathcal{T}(\boldsymbol{g})$, $\boldsymbol{g}' \in \mathbb{R}^K$, and $\boldsymbol{v}' \in \mathcal{T}(\boldsymbol{g}')$, it holds that*

$$\|\boldsymbol{v} - \boldsymbol{v}'\| \leq \frac{2\|\boldsymbol{g} - \boldsymbol{g}'\|}{\delta}. \tag{40}$$

We are now ready to show that the number of misclassified points is bounded by the subspace distance.

Lemma 7. *Let $\pi : [K] \to [K]$ be a permutation such that $\boldsymbol{U} = (\boldsymbol{U}_1, \ldots, \boldsymbol{U}_K)$ with $\boldsymbol{U}_{\pi(k)} \in \mathcal{O}^{n \times d_k}$ for all $k \in [K]$. Suppose that $\bar{\boldsymbol{H}} \in \mathcal{T}(\boldsymbol{G}_{\boldsymbol{H}}(\boldsymbol{U}))$, where the (i,k)-th element of $\boldsymbol{G}_{\boldsymbol{H}}(\boldsymbol{U}) \in \mathbb{R}^{N \times K}$ is $\|\boldsymbol{z}_i\|^2 - \|\boldsymbol{U}_k^T \boldsymbol{z}_i\|^2$. Then, it holds for all $i \in [N]$ that*

$$\|\bar{\boldsymbol{h}}_i - \boldsymbol{h}_i^* \boldsymbol{Q}_\pi\| \leq \frac{2\sqrt{\sum_{k=1}^K d(\boldsymbol{U}_{\pi(k)}, \boldsymbol{U}_k^*)}}{1 - \max_{\ell \neq I_i} \|\boldsymbol{U}_\ell^{*T} \boldsymbol{z}_i\|^2}, \tag{41}$$

where the row vectors $\bar{\boldsymbol{h}}_i, \boldsymbol{h}_i^* \in \mathbb{R}^K$ respectively denote the i-th row of $\bar{\boldsymbol{H}}$ and $\boldsymbol{H}^*$, and $I_i \in [K]$ satisfies $h_{iI_i}^* = 1$ for all $i \in [N]$.

The following lemma indicates that the KSS iterations directly converge to ground truth once the distance from the current iterate to ground truth is small enough. This implies finite termination of the KKS method.

Lemma 8. *Suppose that Assumption 1 holds, $N_{\min} \gtrsim d_k \gtrsim \log N$ for all $k \in [K]$, and $\boldsymbol{H}^t \in \mathcal{M}^{N \times K}$ satisfies*

$$d_F^2(\boldsymbol{H}^t, \boldsymbol{H}^*) \leq 2K^2(c_1 + 1)c_1 d_{\min}, \tag{42}$$

where c_1 is the constant in Lemma 3. Then, it holds with probability at least $1 - 2K/(d_{\min}^2 N) - 5K^2/N$ that

$$\boldsymbol{H}^{t+1} = \boldsymbol{H}^* \boldsymbol{Q}_\pi$$

for some $\boldsymbol{Q}_\pi \in \Pi_K$.

Equipped with the results in Sections 3.2 and 3.3, we can prove Theorem 1. The complete proof can be found in Section D.5 of the appendix.

4. Experiment Results

In this section, we report the convergence behavior, recovery performance, and numerical efficiency of the KSS method for SC on both synthetic and real datasets. All of our experiments are implemented in MATLAB R2020a on the Great Lakes HPC Cluster of the University of Michigan with 180GB memory and 16 cores. Our code is available at https://github.com/peng8wang/ICML2022-K-Subspaces.

4.1. Convergence Behavior and Recovery Performance

We first conduct 3 sets of numerical tests, which correspond to $K \in \{3, 6, 9\}$, to examine the convergence behavior and recovery performance of the KSS method in the semi-random UoS model (see Definition 2). We generate K overlapping subspaces as follows. First, we set $n = 300$, $\overline{d} = 30$, $\underline{d} = 25$, and uniformly at random select $d_k \in [\underline{d}, \overline{d}]$ for all $k \in [K]$. Second, we arbitrarily generate an orthogonal matrix $\boldsymbol{U} = [\boldsymbol{u}_1, \ldots, \boldsymbol{u}_n] \in \mathcal{O}^n$ and set the shared basis as $\bar{\boldsymbol{U}} = [\boldsymbol{u}_{n-s+1}, \ldots, \boldsymbol{u}_n]$ for an integer $s \in [0, \underline{d}]$. Next, we generate $\boldsymbol{V}_k$ by randomly picking up $(d_k - s)$ columns, which are not repeated, from the first $n - s$ columns of $\boldsymbol{U}$. Finally, we form $\boldsymbol{U}_k^* = \begin{bmatrix} \boldsymbol{V}_k & \bar{\boldsymbol{U}} \end{bmatrix}$ for all $k \in [K]$, which ensures that the intersection between S_k and S_ℓ is at least of dimension s for all $1 \leq k \neq \ell \leq K$. In each test, we generate 3 datasets by setting $s = 6$ and $N_k = 500$ for all $k \in [K]$ and respectively run the KSS method with random initialization (denoted by RI-KSS)

and TIPS initialization (denoted by TI-KSS) by setting $\tau = 2/\sqrt{\overline{d}}$ on them. Then, we plot the distance of the iterates to ground truth, i.e., $d_F^2(\boldsymbol{H}^k, \boldsymbol{H}^*) + 10^{-8}$, against the iteration numbers in Figure 1. It can be observed that with a proper initialization, the KSS method converges so quickly that it finds the correct clustering within 10 iterations. This supports the result in Theorem 1. Additionally, it exhibits a finite termination phenomenon that corroborates the result in Lemma 8. Moreover, it is observed in Figure 1(c) that RI-KSS gets stuck at a local minimum while TI-KSS does not on data 3.

Table 2. Average CPU time (in seconds) and the best clustering accuracy of the tested methods on real datasets.

Accuracy	*COIL100*	*YaleB*	*USPS*	*MNIST*
KSS	**0.8117**	0.7154	**0.8172**	**0.9780**
SSC	0.6732	**0.8277**	0.6583	–
OMP	0.3393	**0.8268**	0.2109	0.5749
TSC	**0.7343**	0.4878	0.6693	**0.8514**
GSC	0.6550	0.7071	**0.9522**	0.6306
LRR	0.5500	0.6828	0.7129	–
LRSSC	0.5200	0.7088	0.6443	–
Time (s)	***COIL100***	***YaleB***	***USPS***	***MNIST***
KSS	53.53	6.90	**8.85**	**30.53**
SSC	912.25	136.36	1217.88	–
OMP	**12.12**	**1.02**	31.12	398.37
TSC	**29.78**	**3.06**	**2.66**	**154.46**
GSC	178.15	24.22	105.59	1800.00
LRR	144.25	63.31	112.56	–
LRSSC	1800.00	444.28	1800.00	–

"–" denotes out of memory.

4.2. Numerical Efficiency and Accuracy on Real Data

We now conduct experiments to examine the computational efficiency and recovery accuracy of the KSS method on real datasets. We also compare it with several state-of-the-art methods: SSC in Elhamifar & Vidal (2013), SSC solved by OMP in You et al. (2016), TSC in Heckel & Bölcskei (2015), GSC in Park et al. (2014), LRR in Liu et al. (2012), and LRSSC in Wang et al. (2019). In the implementations of SSC, OMP, LRR, and LRSSC, we use the source codes provided by their authors. We use the real datasets *COIL-100* (S. A. Nene & Murase, 1996a), the cropped extended *Yale B* (Georghiades et al., 2001), *USPS* (Hull, 1994), and *MNIST* (LeCun, 1998).[2] The stopping criteria for the tested methods are given as follows. For KSS, we terminate it when the norm of two consecutive iterates is less than 10^{-2}. For SSC, LRR, and LRSSC, we use the stopping criteria in their source codes. No stopping criterion is needed for TSC and GSC due to their one-shot nature. We set the maximum iteration number of KSS, SSC, LRR, and LRSSC as 200. We set the maximum running time of all tested algorithms as 1800 seconds. For the implementation of KSS, we used the TIPS initialization except for on *MNIST*, where we use random initialization in Algorithm 1. More details, including data processing, parameter settings, and test results, can be found in Section F of the appendix. Then, we run each method 10 times. Note that if the algorithms are initialized deterministically, the only randomness is from the initialization for k-means in spectral clustering. To compare the computational efficiency and recovery accuracy of the tested methods, we report the average running time and best clustering accuracy for all runs of each method in Table 2. More experiment results can be also found in Section F of the appendix. It can be observed that the KSS method is in the top three in terms of *both* accuracy and computational efficiency for every dataset. This demonstrates the efficiency and efficacy of the KSS method for SC.

[2]The datasets *COIL-100*, *Yale B*, and *USPS* are downloaded from http://www.cad.zju.edu.cn/home/dengcai/Data/data.html. The dataset *MNIST* is downloaded from LIBSVM (Chang & Lin, 2011) at https://www.csie.ntu.edu.tw/ cjlin/libsvmtools/datasets/.

5. Concluding Remarks

In this work, we analyzed the KSS method for subspace clustering and provided a TIPS method for its initialization in the semi-random UoS model. We showed that provided an initial assignment satisfying a partial recovery condition, the KSS method converges superlinearly and achieves correct clustering within $\Theta(\log\log N)$ iterations, even when the normalized affinity between pairwise subspaces is $O(1)$. Moreover, we proved that the proposed initialization method can return a qualified initial point. All these results are demonstrated by the numerical results. A natural future direction is to study the convergence behavior and recovery performance of the KSS method in the noisy UoS model; see, e.g., Heckel & Bölcskei (2015); Soltanolkotabi et al. (2014); Tschannen & Bölcskei (2018); Wang & Xu (2013).

Acknowledgements

The first and last authors are supported in part by ARO YIP award W911NF1910027, in part by AFOSR YIP award FA9550-19-1-0026, and in party by NSF CAREER award CCF-1845076. The second author is supported by the National Natural Science Foundation of China (NSFC) Grant 72192832. The third author is supported in part by the Hong Kong Research Grants Council (RGC) General Research Fund (GRF) Project CUHK 14205421. The authors thank the reviewers for their insightful comments. They also thank John Lipor for sharing some data and code.

References

Agarwal, P. K. and Mustafa, N. H. K-means projective clustering. In *Proceedings of the 23rd ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems*, pp. 155–165, 2004.

Aloise, D., Deshpande, A., Hansen, P., and Popat, P. NP-hardness of euclidean sum-of-squares clustering. *Machine Learning*, 75(2):245–248, 2009.

Boult, T. E. and Brown, L. G. Factorization-based segmentation of motions. In *Proceedings of the IEEE Workshop on Visual Motion*, pp. 179–180. IEEE Computer Society, 1991.

Boumal, N. Nonconvex phase synchronization. *SIAM Journal on Optimization*, 26(4):2355–2377, 2016.

Bradley, P. S. and Mangasarian, O. L. K-plane clustering. *Journal of Global optimization*, 16(1):23–32, 2000.

Bruna, J. and Mallat, S. Invariant scattering convolution networks. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1872–1886, 2013.

Chang, C.-C. and Lin, C.-J. Libsvm: a library for support vector machines. *ACM transactions on intelligent systems and technology (TIST)*, 2(3):1–27, 2011.

Chen, Y., Li, C.-G., and You, C. Stochastic sparse subspace clustering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4155–4164, 2020.

Ding, T., Zhu, Z., Tsakiris, M., Vidal, R., and Robinson, D. Dual principal component pursuit for learning a union of hyperplanes: Theory and algorithms. In *International Conference on Artificial Intelligence and Statistics*, pp. 2944–2952. PMLR, 2021.

Dyer, E. L., Sankaranarayanan, A. C., and Baraniuk, R. G. Greedy feature selection for subspace clustering. *The Journal of Machine Learning Research*, 14(1):2487–2517, 2013.

Elhamifar, E. and Vidal, R. Sparse subspace clustering: Algorithm, theory, and applications. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(11): 2765–2781, 2013.

Fan, J. Large-scale subspace clustering via k-factorization. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, pp. 342–352, 2021.

Gao, C. and Zhang, A. Y. Iterative algorithm for discrete structure recovery. *arXiv preprint arXiv:1911.01018*, 2019.

Georghiades, A., Belhumeur, P., and Kriegman, D. 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.

Ghosh, A. and Kannan, R. Alternating minimization converges super-linearly for mixed linear regression. In *International Conference on Artificial Intelligence and Statistics*, pp. 1093–1103. PMLR, 2020.

Gitlin, A., Tao, B., Balzano, L., and Lipor, J. Improving K-subspaces via coherence pursuit. *IEEE Journal of Selected Topics in Signal Processing*, 12(6):1575–1588, 2018.

Golub, G. H. and Van Loan, C. F. *Matrix computations*, volume 3. JHU press, 2013.

Hardt, M. Understanding alternating minimization for matrix completion. In *2014 IEEE 55th Annual Symposium on Foundations of Computer Science*, pp. 651–660. IEEE, 2014.

He, J., Zhang, Y., Wang, J., Zeng, N., and Hao, H. Robust k-subspaces recovery with combinatorial initialization. In *2016 IEEE International Conference on Big Data (Big Data)*, pp. 3573–3582. IEEE, 2016.

Heckel, R. and Bölcskei, H. Robust subspace clustering via thresholding. *IEEE Transactions on Information Theory*, 61(11):6320–6342, 2015.

Ho, J., Yang, M.-H., Lim, J., Lee, K.-C., and Kriegman, D. Clustering appearances of objects under varying illumination conditions. In *2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2003. Proceedings.*, volume 1, pp. I–I. IEEE, 2003.

Hong, W., Wright, J., Huang, K., and Ma, Y. Multiscale hybrid linear models for lossy image representation. *IEEE Transactions on Image Processing*, 15(12):3655–3671, 2006.

Hull, J. J. A database for handwritten text recognition research. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 16(5):550–554, 1994.

Jiang, D., Tang, C., and Zhang, A. Cluster analysis for gene expression data: a survey. *IEEE Transactions on Knowledge and Data Engineering*, 16(11):1370–1386, 2004.

Kumar, A. and Kannan, R. Clustering with spectral norm and the k-means algorithm. In *2010 IEEE 51st Annual Symposium on Foundations of Computer Science*, pp. 299–308. IEEE, 2010.

LeCun, Y. The mnist database of handwritten digits. *http://yann.lecun.com/exdb/mnist/*, 1998.

Lei, J., Rinaldo, A., et al. Consistency of spectral clustering in stochastic block models. *The Annals of Statistics*, 43 (1):215–237, 2015.

Li, G. and Gu, Y. Theory of spectral method for union of subspaces-based random geometry graph. *arXiv preprint arXiv:1907.10906*, 2019.

Li, G. and Gu, Y. Theory of spectral method for union of subspaces-based random geometry graph. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 6337–6345. PMLR, 2021.

Ling, S. Improved performance guarantees for orthogonal group synchronization via generalized power method. *SIAM Journal on Optimization*, 32(2):1018–1048, 2022.

Lipor, J., Hong, D., Tan, Y. S., and Balzano, L. Subspace clustering using ensembles of K-subspaces. *Information and Inference: A Journal of the IMA*, 10(1):73–107, 2021.

Liu, G., Lin, Z., Yan, S., Sun, J., Yu, Y., and Ma, Y. Robust recovery of subspace structures by low-rank representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(1):171–184, 2012.

Matsushima, S. and Brbic, M. Selective sampling-based scalable sparse subspace clustering. *Advances in Neural Information Processing Systems*, 32:12416–12425, 2019.

McWilliams, B. and Montana, G. Subspace clustering of high-dimensional data: a predictive approach. *Data Mining and Knowledge Discovery*, 28(3):736–772, 2014.

Meng, L., Li, G., Yan, J., and Gu, Y. A general framework for understanding compressed subspace clustering algorithms. *IEEE Journal of Selected Topics in Signal Processing*, 12(6):1504–1519, 2018.

Park, D., Caramanis, C., and Sanghavi, S. Greedy subspace clustering. *Advances in Neural Information Processing Systems*, 27:2753–2761, 2014.

Pimentel-Alarcón, D., Balzano, L., Marcia, R., Nowak, R., and Willett, R. Group-sparse subspace clustering with missing data. In *2016 IEEE Statistical Signal Processing Workshop (SSP)*, pp. 1–5. IEEE, 2016.

S. A. Nene, S. K. N. and Murase, H. Columbia object image library (COIL-100). *Technical Report CUCS-006-96*, 1996a.

S. A. Nene, S. K. N. and Murase, H. Columbia object image library (COIL-20). *Technical Report CUCS-005-96*, 1996b.

Shen, J., Li, P., and Xu, H. Online low-rank subspace clustering by basis dictionary pursuit. In *International Conference on Machine Learning*, pp. 622–631. PMLR, 2016.

Soltanolkotabi, M., Candes, E. J., et al. A geometric analysis of subspace clustering with outliers. *The Annals of Statistics*, 40(4):2195–2238, 2012.

Soltanolkotabi, M., Elhamifar, E., Candes, E. J., et al. Robust subspace clustering. *Annals of Statistics*, 42(2):669–699, 2014.

Traganitis, P. A. and Giannakis, G. B. Sketched subspace clustering. *IEEE Transactions on Signal Processing*, 66 (7):1663–1675, 2017.

Tschannen, M. and Bölcskei, H. Noisy subspace clustering via matching pursuits. *IEEE Transactions on Information Theory*, 64(6):4081–4104, 2018.

Tseng, P. Nearest q-flat to m points. *Journal of Optimization Theory and Applications*, 105(1):249–252, 2000.

Ucar, D., Hu, Q., and Tan, K. Combinatorial chromatin modification patterns in the human genome revealed by subspace clustering. *Nucleic Acids Research*, 39(10): 4063–4075, 2011.

Vershynin, R. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge University Press, 2018.

Vidal, R. Subspace clustering. *IEEE Signal Processing Magazine*, 28(2):52–68, 2011.

Vidal, R., Ma, Y., and Sastry, S. Generalized principal component analysis (GPCA). *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(12):1945–1959, 2005.

Vidal, R., Tron, R., and Hartley, R. Multiframe motion segmentation with missing data using powerfactorization and GPCA. *International Journal of Computer Vision*, 79(1):85–105, 2008.

Vidal, R., Ma, Y., and Sastry, S. S. *Generalized principal component analysis*, volume 5. Springer, 2016.

Wang, P., Liu, H., Zhou, Z., and So, A. M.-C. Optimal non-convex exact recovery in stochastic block model via projected power method. In *International Conference on Machine Learning*, pp. 10828–10838. PMLR, 2021a.

Wang, P., Zhou, Z., and So, A. M.-C. Non-convex exact community recovery in stochastic block model. *Mathematical Programming*, pp. 1–37, 2021b.

Wang, Y., Wang, Y.-X., and Singh, A. Graph connectivity in noisy sparse subspace clustering. In *Artificial Intelligence and Statistics*, pp. 538–546. PMLR, 2016.

Wang, Y.-X. and Xu, H. Noisy sparse subspace clustering. In *International Conference on Machine Learning*, pp. 89–97. PMLR, 2013.

Wang, Y.-X., Xu, H., and Leng, C. Provable subspace clustering: When LRR meets SSC. *IEEE Transactions on Information Theory*, 65(9):5406–5432, 2019.

Wu, J.-Y., Huang, L.-C., Yang, M.-H., and Liu, C.-H. Sparse subspace clustering via two-step reweighted L1-minimization: Algorithm and provable neighbor recovery rates. *IEEE Transactions on Information Theory*, 2020.

You, C., Robinson, D., and Vidal, R. Scalable sparse subspace clustering by orthogonal matching pursuit. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3918–3927, 2016.

Zhang, T. Phase retrieval by alternating minimization with random initialization. *IEEE Transactions on Information Theory*, 66(7):4563–4573, 2020.

Zhang, T., Szlam, A., and Lerman, G. Median k-flats for hybrid linear modeling with many outliers. In *2009 IEEE 12th International Conference on Computer Vision Workshops, ICCV Workshops*, pp. 234–241. IEEE, 2009.

Zhang, Y., Qu, Q., and Wright, J. From symmetry to geometry: Tractable nonconvex problems. *arXiv preprint arXiv:2007.06753*, 2020.

Appendix

In the appendix, we provide proofs of the technical results presented in Sections 2 and 3. To proceed, we introduce some further notations. Given a vector $\boldsymbol{a} \in \mathbb{R}^n$, we denote by $\mathrm{diag}(\boldsymbol{a}) \in \mathbb{R}^{n\times n}$ the diagonal matrix with $\boldsymbol{a}$ on its diagonal. Given a symmetric matrix $\boldsymbol{A}$, we use $\lambda_{\min}(\boldsymbol{A})$ to denote its smallest eigenvalue. We respectively use $\mathbf{1}_n$, $\boldsymbol{E}_n$, and $\boldsymbol{I}_d$ to denote the n-dimensional all-one vector, $n \times n$ all-one matrix, and $d \times d$ identity matrix, and simply write $\mathbf{1}$, $\boldsymbol{E}$, and $\boldsymbol{I}$ when their dimension can be inferred from the context. Given two random variables X and Y, we write $X \overset{d}{=} Y$ if X and Y are equal in distribution. We use $\boldsymbol{e}_i$ to denote a standard basis with a 1 in the i-th coordinate and 0's elsewhere. For a vector $\boldsymbol{x} \in \mathbb{R}^n$, we denote by $\boldsymbol{x}_S$ its subvector consisting of the elements indexed by the set S. We denote the cumulative distribution function of the standard normal distribution by

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{t^2}{2}} dt.$$

For any random vector $\boldsymbol{a} \sim \mathrm{Unif}(\mathbb{S}^{d-1})$, it is known that there exists a standard normal random vector such that $\boldsymbol{a}$ is its normalization. We denote such vector by $\bar{\boldsymbol{a}}$. Thus, it holds that

$$\bar{\boldsymbol{a}} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}_d),\ \boldsymbol{a} = \frac{\bar{\boldsymbol{a}}}{\|\bar{\boldsymbol{a}}\|}. \tag{43}$$

Moreover, let

$$\boldsymbol{U}_k^{*T} \boldsymbol{U}_\ell^* = \boldsymbol{U}_{k\ell}^* \boldsymbol{\Sigma}_{k\ell}^* \boldsymbol{V}_{k\ell}^{*T} \tag{44}$$

be a singular value decomposition (SVD) of $\boldsymbol{U}_k^{*T} \boldsymbol{U}_\ell^*$, where $\sigma_{k\ell}^{(1)} \geq \cdots \geq \sigma_{k\ell}^{(\min\{d_k, d_\ell\})} \geq 0$ are the singular values of $\boldsymbol{U}_k^{*T} \boldsymbol{U}_\ell^*$ and $\boldsymbol{U}_{k\ell}^* \in \mathcal{O}^{d_k}, \boldsymbol{V}_{k\ell}^* \in \mathcal{O}^{d_\ell}$. Suppose that $d_k \geq d_\ell$. We have

$$\boldsymbol{U}_k^{*T} \boldsymbol{U}_\ell^* = \boldsymbol{U}_{k\ell}^* \begin{bmatrix} \bar{\boldsymbol{\Sigma}}_{k\ell}^* \\ \mathbf{0} \end{bmatrix} \boldsymbol{V}_{k\ell}^{*T}, \tag{45}$$

where $\bar{\boldsymbol{\Sigma}}_{k\ell}^* = \mathrm{diag}\left(\sigma_{k\ell}^{(1)}, \ldots, \sigma_{k\ell}^{(d_\ell)}\right)$. Suppose to the contrary that $d_k < d_\ell$. Then, we have

$$\boldsymbol{U}_k^{*T} \boldsymbol{U}_\ell^* = \boldsymbol{U}_{k\ell}^* \begin{bmatrix} \bar{\boldsymbol{\Sigma}}_{k\ell}^* & \mathbf{0} \end{bmatrix} \boldsymbol{V}_{k\ell}^{*T}, \tag{46}$$

where $\bar{\boldsymbol{\Sigma}}_{k\ell}^* = \mathrm{diag}\left(\sigma_{k\ell}^{(1)}, \ldots, \sigma_{k\ell}^{(d_k)}\right)$. According to (13) and (44), one can verify that

$$\mathrm{aff}(S_k^*, S_\ell^*) = \|\boldsymbol{\Sigma}_{k\ell}^*\|_F. \tag{47}$$

Recall that for any $\boldsymbol{U}, \boldsymbol{V} \in \mathcal{O}^{n\times d}$, we use $d(\boldsymbol{U}, \boldsymbol{V}) = \left\|\boldsymbol{U}\boldsymbol{U}^T - \boldsymbol{V}\boldsymbol{V}^T\right\|$ to denote the distance between the subspaces respectively spanned by $\boldsymbol{U}$ and $\boldsymbol{V}$. Then, one can verify

$$d(\boldsymbol{U}, \boldsymbol{V}) = \left\|(\boldsymbol{I} - \boldsymbol{U}\boldsymbol{U}^T)\boldsymbol{V}\right\| = \left\|(\boldsymbol{I} - \boldsymbol{V}\boldsymbol{V}^T)\boldsymbol{U}\right\| = \sqrt{1 - \sigma_{\min}^2(\boldsymbol{U}^T\boldsymbol{V})}. \tag{48}$$

A. Concentration Inequalities

In this section, we present some concentration inequalities for random vectors. These inequalities play an important role in the analysis of the proposed method. We first introduce a spectral bound on the covariance estimation for random vectors generated by a uniform distribution over the sphere. It is a direct consequence of Vershynin (2018, Theorem 4.7.1) and thus we omit its proof.

Lemma 9. *Suppose that $\boldsymbol{a}_1, \ldots, \boldsymbol{a}_m \in \mathbb{R}^d$ are i.i.d. uniformly distributed over the unit sphere. Then, it holds with probability at least $1 - 2e^{-u}$ that*

$$\left\| \frac{1}{m} \sum_{i=1}^{m} \boldsymbol{a}_i \boldsymbol{a}_i^T - \frac{1}{d} \boldsymbol{I}_d \right\| \leq \frac{c_1}{d} \left(\sqrt{\frac{d+u}{m}} + \frac{d+u}{m} \right),$$

where $c_1 > 0$ is an absolute constant.

We next present a bound on the deviation of the weighted sum of standard normal random variables from its mean. This is an extension of Li & Gu (2019, Lemma 7).

Lemma 10. *Let $\boldsymbol{x} \in \mathbb{R}^d$ be a normal random vector such that $\boldsymbol{x} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \boldsymbol{I}_d)$. It holds for $\lambda_1, \ldots, \lambda_d \in [0,1]$ with $\sum_{i=1}^d \lambda_i^2 \geq 4$ and $t > 0$ that*

$$\mathbb{P}\left(\left|\sqrt{\sum_{i=1}^d \lambda_i^2 x_i^2} - \sigma\sqrt{\sum_{i=1}^d \lambda_i^2}\right| \geq t + 2\sigma\right) \leq 2\exp\left(-\frac{t^2}{2\sigma^2}\right).$$

Proof of Lemma 10. We define

$$f(\boldsymbol{x}) = \sqrt{\sum_{i=1}^d \lambda_i^2 x_i^2}.$$

By calculation, we obtain

$$\|\nabla f(\boldsymbol{x})\| = \sqrt{\frac{\sum_{i=1}^d \lambda_i^4 x_i^2}{\sum_{i=1}^d \lambda_i^2 x_i^2}} \leq 1.$$

Applying the concentration inequality for Lipschitz functions (see, e.g., Li & Gu (2019, Lemma 6))) to $f(\boldsymbol{x})$ yields that

$$\mathbb{P}\left(|f(\boldsymbol{x}) - \mathbb{E}[f(\boldsymbol{x})]| \geq t\right) \leq 2\exp\left(-\frac{t^2}{2\sigma^2}\right). \tag{49}$$

We first note that

$$\mathbb{E}[f(\boldsymbol{x})] \leq \sqrt{\mathbb{E}[f^2(\boldsymbol{x})]} = \sqrt{\mathbb{E}\left[\sum_{i=1}^d \lambda_i^2 x_i^2\right]} = \sigma\sqrt{\sum_{i=1}^d \lambda_i^2}. \tag{50}$$

By letting $X = f(\boldsymbol{x}) \geq 0$ and $\mu = \mathbb{E}[X]$, we can compute

$$\begin{aligned}\mathrm{Var}(X) = \mathbb{E}\left[(X-\mu)^2\right] &= \int_0^\infty t^2 d\mathbb{P}\left(|X-\mu| \leq t\right) = -\int_0^\infty t^2 d\mathbb{P}\left(|X-\mu| > t\right) \\ &= \int_0^\infty 2t\mathbb{P}\left(|X-\mu| > t\right) dt \leq \int_0^\infty 4t\exp\left(-\frac{t^2}{2\sigma^2}\right) dt = 4\sigma^2,\end{aligned}$$

where the forth equality and the last one follow from integration by parts and the inequality is due to (49). Thus, we have

$$\mathbb{E}^2[f(\boldsymbol{x})] = \mathbb{E}[f^2(\boldsymbol{x})] - \mathrm{Var}\left(f(\boldsymbol{x})\right) = \mathbb{E}\left[\sum_{i=1}^d \lambda_i^2 x_i^2\right] - \mathrm{Var}\left(f(\boldsymbol{x})\right) \geq \sigma^2\left(\sum_{i=1}^d \lambda_i^2 - 4\right).$$

This, together with $\sum_{i=1}^d \lambda_i^2 \geq 4$, implies

$$\mathbb{E}[f(\boldsymbol{x})] \geq \sigma\sqrt{\sum_{i=1}^d \lambda_i^2 - 4} \geq \sigma\left(\sqrt{\sum_{i=1}^d \lambda_i^2} - 2\right). \tag{51}$$

Plugging (50) and (51) into (49) yields the desired result. □

Equipped with the above results, we are ready to present a lemma that characterizes the properties of a uniform distribution over the sphere. This plays an important role in the subsequent analysis.

Lemma 11. *Suppose that* $\|\mathbf{\Sigma}_{k\ell}^*\|_F \geq 2$ *for all* $1 \leq k \neq \ell \leq K$ *and* $\boldsymbol{a}_i \overset{i.i.d.}{\sim} \mathrm{Unif}(\mathbb{S}^{d_\ell - 1})$ *for all* $i \in [N]$. *Then, it holds with probability at least* $1 - 4N^{-2}$ *that for some* $i \in [N]$ *and* $1 \leq k \neq \ell \leq K$,

$$\left| \|\bar{\boldsymbol{a}}_i\| - \sqrt{d_\ell} \right| \leq \alpha, \quad |\|\mathbf{\Sigma}_{k\ell}^* \bar{\boldsymbol{a}}_i\| - \|\mathbf{\Sigma}_{k\ell}^*\|_F| \leq \alpha, \tag{52}$$

and

$$\frac{\|\mathbf{\Sigma}_{k\ell}^*\|_F - \alpha}{\sqrt{d_\ell} + \alpha} \leq \|\mathbf{\Sigma}_{k\ell}^* \boldsymbol{a}_i\| \leq \frac{\|\mathbf{\Sigma}_{k\ell}^*\|_F + \alpha}{\sqrt{d_\ell} - \alpha}, \tag{53}$$

where $\alpha = 2\sqrt{\log N} + 2$.

Proof of Lemma 11. We first prove (52). Applying Lemma 10 with $t = 2\sqrt{\log N}$ and $\lambda_j = 1$ for all $j \in [d_\ell]$ to $\bar{\boldsymbol{a}}_i \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_{d_\ell})$ yields that

$$\mathbb{P}\left(\left| \|\bar{\boldsymbol{a}}_i\| - \sqrt{d_\ell} \right| \geq \alpha \right) \leq 2N^{-2}. \tag{54}$$

Suppose that $d_k \geq d_\ell$. According to (45), we have $\mathbf{\Sigma}_{k\ell}^* = \begin{bmatrix} \bar{\mathbf{\Sigma}}_{k\ell}^* \\ \boldsymbol{0} \end{bmatrix}$. Applying Lemma 10 with $t = 2\sqrt{\log N}$ and $\lambda_j = \sigma_{k\ell}^{(j)}$ for all $j \in [d_\ell]$ to $\bar{\boldsymbol{a}}_i \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_{d_\ell})$ yields

$$\mathbb{P}\left(\left| \|\bar{\mathbf{\Sigma}}_{k\ell}^* \bar{\boldsymbol{a}}_i\| - \|\bar{\mathbf{\Sigma}}_{k\ell}^*\|_F \right| \geq \alpha \right) \leq 2N^{-2}.$$

This, together with $\|\mathbf{\Sigma}_{k\ell}^* \bar{\boldsymbol{a}}_i\| = \|\bar{\mathbf{\Sigma}}_{k\ell}^* \bar{\boldsymbol{a}}_i\|$ and $\|\mathbf{\Sigma}_{k\ell}^*\|_F = \|\bar{\mathbf{\Sigma}}_{k\ell}^*\|_F$, implies

$$\mathbb{P}\left(|\|\mathbf{\Sigma}_{k\ell}^* \bar{\boldsymbol{a}}_i\| - \|\mathbf{\Sigma}_{k\ell}^*\|_F| \geq \alpha \right) \leq 2N^{-2}. \tag{55}$$

Suppose to the contrary that $d_k < d_\ell$. According to (46), we have $\mathbf{\Sigma}_{k\ell}^* \bar{\boldsymbol{a}}_i = \left(\sigma_{k\ell}^{(1)} \bar{a}_{i1}, \ldots, \sigma_{k\ell}^{(d_k)} \bar{a}_{id_k} \right)$. Applying Lemma 10 with $t = 2\sqrt{\log N}$ and $\lambda_j = \sigma_{k\ell}^{(j)}$ for all $j \in [d_k]$ to $[\bar{\boldsymbol{a}}_i]_S \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_{d_k})$ with $S = [d_k]$ yields

$$\mathbb{P}\left(|\|\mathbf{\Sigma}_{k\ell}^* \bar{\boldsymbol{a}}_i\| - \|\mathbf{\Sigma}_{k\ell}^*\|_F| \geq \alpha \right) \leq 2N^{-2}.$$

This, together with (54), (55), and the union bound, implies (52).

We next prove (53) using (52). Using $\boldsymbol{a}_i = \bar{\boldsymbol{a}}_i / \|\bar{\boldsymbol{a}}_i\|$ and (52), we have

$$\|\mathbf{\Sigma}_{k\ell}^* \boldsymbol{a}_i\| = \frac{\|\mathbf{\Sigma}_{k\ell}^* \bar{\boldsymbol{a}}_i\|}{\|\bar{\boldsymbol{a}}_i\|} \leq \frac{\|\mathbf{\Sigma}_{k\ell}^*\|_F + \alpha}{\sqrt{d_\ell} - \alpha}$$

and

$$\|\mathbf{\Sigma}_{k\ell}^* \boldsymbol{a}_i\| = \frac{\|\mathbf{\Sigma}_{k\ell}^* \bar{\boldsymbol{a}}_i\|}{\|\bar{\boldsymbol{a}}_i\|} \geq \frac{\|\mathbf{\Sigma}_{k\ell}^*\|_F - \alpha}{\sqrt{d_\ell} + \alpha}.$$

Then, we complete the proof. □

Then, we present a lemma that estimates the magnitudes of some crucial parameters in our analysis.

Lemma 12. *Suppose that* $\boldsymbol{z}_i \in \mathbb{R}^N$ *are generated according to the semi-random UoS model such that* $\boldsymbol{z}_i \in S_\ell^*$. *Then, for any* $1 \leq i \neq j \leq N$ *and* $k \in [K]$, *it holds with probability at least* $1 - 5K^2/N$ *that*

$$\frac{\mathrm{aff}(S_k^*, S_\ell^*) - \alpha}{\sqrt{d_\ell} + \alpha} \leq \|\boldsymbol{U}_k^{*^T} \boldsymbol{z}_i\| \leq \frac{\mathrm{aff}(S_k^*, S_\ell^*) + \alpha}{\sqrt{d_\ell} - \alpha} \tag{56}$$

and

$$\left| \frac{\langle \boldsymbol{U}_k^{*^T} \boldsymbol{z}_i, \boldsymbol{U}_k^{*^T} \boldsymbol{z}_j \rangle}{\|\boldsymbol{U}_k^{*^T} \boldsymbol{z}_j\|} \right| \leq \frac{2\sqrt{\log N}}{\sqrt{d_\ell} - \alpha}, \tag{57}$$

where $\alpha = 2\sqrt{\log N} + 2$.

Proof of Lemma 12. Suppose that (52) and (53) hold for all $i \in [N]$ and $k, \ell \in [K]$, which happens with probability $1 - 4K^2N^{-1}$ according to Lemma 11 and the union bound. We first show (56). Since $\boldsymbol{z}_i \in S_\ell^*$ and a uniform distribution over the sphere is rotationally invariant, we have $\|\boldsymbol{U}_k^{*^T}\boldsymbol{z}_i\| = \|\boldsymbol{U}_k^{*^T}\boldsymbol{U}_\ell^*\boldsymbol{a}_i\| = \|\boldsymbol{U}_{k\ell}^*\boldsymbol{\Sigma}_{k\ell}^*\boldsymbol{V}_{k\ell}^{*^T}\boldsymbol{a}_i\| \sim \|\boldsymbol{\Sigma}_{k\ell}^*\boldsymbol{a}_i\|$. This, together with (53) and (47), implies that for any $j \in [N]$ and $\ell \in [K]$,

$$\frac{\mathrm{aff}(S_k^*, S_\ell^*) - \alpha}{\sqrt{d_\ell} + \alpha} \le \|\boldsymbol{U}_k^{*^T}\boldsymbol{z}_i\| \le \frac{\mathrm{aff}(S_k^*, S_\ell^*) + \alpha}{\sqrt{d_\ell} - \alpha}. \tag{58}$$

We next show (57). According to $\boldsymbol{z}_i \in S_\ell^*$ and (43), we have

$$\frac{\langle \boldsymbol{U}_k^{*^T}\boldsymbol{z}_i, \boldsymbol{U}_k^{*^T}\boldsymbol{z}_j\rangle}{\|\boldsymbol{U}_k^{*^T}\boldsymbol{z}_j\|} = \frac{\langle \boldsymbol{U}_k^{*^T}\boldsymbol{U}_\ell^*\boldsymbol{a}_i, \boldsymbol{U}_k^{*^T}\boldsymbol{z}_j\rangle}{\|\boldsymbol{U}_k^{*^T}\boldsymbol{z}_j\|} = \frac{\langle \boldsymbol{U}_k^{*^T}\boldsymbol{U}_\ell^*\bar{\boldsymbol{a}}_i, \boldsymbol{U}_k^{*^T}\boldsymbol{z}_j\rangle}{\|\bar{\boldsymbol{a}}_i\|\|\boldsymbol{U}_k^{*^T}\boldsymbol{z}_j\|}. \tag{59}$$

By letting X be a standard normal random variable, i.e., $X \sim N(0,1)$, we compute

$$\begin{aligned}\mathbb{P}\left(\left|\frac{\langle \boldsymbol{U}_k^{*^T}\boldsymbol{U}_\ell^*\bar{\boldsymbol{a}}_i, \boldsymbol{U}_k^{*^T}\boldsymbol{z}_j\rangle}{\|\boldsymbol{U}_k^{*^T}\boldsymbol{z}_j\|}\right| \le 2\sqrt{\log N} \;\middle|\; \boldsymbol{z}_j\right) &= \mathbb{P}\left(|X| \le \frac{2\|\boldsymbol{U}_k^{*^T}\boldsymbol{z}_j\|\sqrt{\log N}}{\|\boldsymbol{U}_\ell^{*^T}\boldsymbol{U}_k^*\boldsymbol{U}_k^{*^T}\boldsymbol{z}_j\|}\right) \\ &\ge \mathbb{P}\left(|X| \le 2\sqrt{\log N}\right) \ge 1 - \sqrt{\frac{2}{\pi}}\frac{N^{-2}}{\sqrt{\log N}},\end{aligned}$$

where the equality is due to $\langle \boldsymbol{U}_k^{*^T}\boldsymbol{U}_\ell^*\bar{\boldsymbol{a}}_i, \boldsymbol{U}_k^{*^T}\boldsymbol{z}_j\rangle/\|\boldsymbol{U}_k^{*^T}\boldsymbol{z}_j\| \sim \mathcal{N}\left(\boldsymbol{0}, \|\boldsymbol{U}_\ell^{*^T}\boldsymbol{U}_k^*\boldsymbol{U}_k^{*^T}\boldsymbol{z}_j\|^2/\|\boldsymbol{U}_k^{*^T}\boldsymbol{z}_j\|^2\right)$ and the first inequality is due to $\|\boldsymbol{U}_\ell^{*^T}\boldsymbol{U}_k^*\boldsymbol{U}_k^{*^T}\boldsymbol{z}_j\| \le \|\boldsymbol{U}_k^{*^T}\boldsymbol{z}_j\|$. This, together with (52), (59), and the union bound, implies that it holds with probability at least $1 - K^2N^{-1}/\sqrt{\log N}$ that for all $1 \le i \ne j \le N$ and $\ell \in [K]$,

$$\frac{|\langle \boldsymbol{U}_k^{*^T}\boldsymbol{z}_i, \boldsymbol{U}_k^{*^T}\boldsymbol{z}_j\rangle|}{\|\boldsymbol{U}_k^{*^T}\boldsymbol{z}_j\|} \le \frac{2\sqrt{\log N}}{\sqrt{d_\ell} - \alpha}.$$

This, together with (58) and the union bound, implies the desired results. □

B. Proofs in Section 3.1

According to Assumption 2, $d_{\min} \gtrsim \log^3 N$, and $\alpha = 2\sqrt{\log N} + 2$, there exists an $\varepsilon \lesssim 1/\sqrt{\log N}$ such that

$$\alpha \le \varepsilon \mathrm{aff}(S_k^*, S_\ell^*) \text{ for all } 1 \le k \ne \ell \le K, \quad \alpha \le \varepsilon\sqrt{d_k} \text{ for all } k \in [K]. \tag{60}$$

This result shall be used in the subsequent proofs again and again.

B.1. Proof of Lemma 1

Before we prove Lemma 1, we need the following lemma to estimate the probability of the event that $|\langle \boldsymbol{z}_i, \boldsymbol{z}_j\rangle| \ge \tau$ conditioned on $\boldsymbol{z}_j$. Recall that we denote the cumulative distribution function of the standard normal distribution by

$$\Phi(x) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{x} e^{-\frac{t^2}{2}}dt.$$

Lemma 13. *Suppose that $\boldsymbol{z}_i \in S_k^*$ for some $k \in [K]$. Then, it holds for any $1 \le i \ne j \le N$ that*

$$2 - 2\Phi\left(\frac{\tau(\sqrt{d_k} + \alpha)}{\|\boldsymbol{U}_k^{*^T}\boldsymbol{z}_j\|}\right) - 2N^{-2} \le \mathbb{P}\left(|\langle \boldsymbol{z}_i, \boldsymbol{z}_j\rangle| \ge \tau \;\middle|\; \boldsymbol{z}_j\right) \le 2 - 2\Phi\left(\frac{\tau(\sqrt{d_k} - \alpha)}{\|\boldsymbol{U}_k^{*^T}\boldsymbol{z}_j\|}\right) + 2N^{-2}, \tag{61}$$

where $\alpha = 2\sqrt{\log N} + 2$. In particular, we have

$$\mathbb{P}\left(|\langle \boldsymbol{z}_i, \boldsymbol{z}_j\rangle| \ge \frac{\|\boldsymbol{U}_k^{*^T}\boldsymbol{z}_j\|\sqrt{\log N}}{\sqrt{d_k}} \;\middle|\; \boldsymbol{z}_j\right) \le \sqrt{\frac{2}{\pi}}\frac{2\sqrt{d_k}}{(\sqrt{d_k} - \alpha)\sqrt{\log N}}N^{-\frac{(\sqrt{d_k}-\alpha)^2}{2d_k}} + 2N^{-2}. \tag{62}$$

Proof of Lemma 13. According to $\boldsymbol{z}_i \in S_k^*$ and (43), we have

$$\begin{aligned}\mathbb{P}\left(|\langle \boldsymbol{z}_i, \boldsymbol{z}_j\rangle| \geq \tau \;\middle|\; \boldsymbol{z}_j\right) &= 2\mathbb{P}\left(\langle \boldsymbol{z}_i, \boldsymbol{z}_j\rangle \geq \tau \;\middle|\; \boldsymbol{z}_j\right) = 2\mathbb{P}\left(\langle \boldsymbol{U}_k^* \boldsymbol{a}_i, \boldsymbol{z}_j\rangle \geq \tau \;\middle|\; \boldsymbol{z}_j\right) = 2\mathbb{P}\left(\langle \bar{\boldsymbol{a}}_i, \boldsymbol{U}_k^{*^T} \boldsymbol{z}_j\rangle \geq \tau\|\bar{\boldsymbol{a}}_i\| \;\middle|\; \boldsymbol{z}_j\right)\\ &\geq 2\mathbb{P}\left(\langle \bar{\boldsymbol{a}}_i, \boldsymbol{U}_k^{*^T} \boldsymbol{z}_j\rangle \geq \tau(\sqrt{d_k}+\alpha) \;\middle|\; \boldsymbol{z}_j\right) - 2\mathbb{P}\left(\|\bar{\boldsymbol{a}}_i\| \geq \sqrt{d_k}+\alpha\right)\\ &\geq 2 - 2\Phi\left(\frac{\tau(\sqrt{d_k}+\alpha)}{\|\boldsymbol{U}_k^{*^T}\boldsymbol{z}_j\|}\right) - 2N^{-2},\end{aligned}$$

where the first inequality is due to the union bound and the fact that $\boldsymbol{a}_i$ is independent of $\boldsymbol{a}_j$ and the second inequality follows from $\langle \bar{\boldsymbol{a}}_i, \boldsymbol{U}_k^{*^T}\boldsymbol{z}_j\rangle \sim \mathcal{N}(0, \|\boldsymbol{U}_k^{*^T}\boldsymbol{z}_j\|^2)$ and uses (54) for $\bar{\boldsymbol{a}}_i \in \text{Unif}\left(\mathbb{S}^{d_k-1}\right)$. By the same argument, we obtain

$$\begin{aligned}\mathbb{P}\left(|\langle \boldsymbol{z}_i, \boldsymbol{z}_j\rangle| \geq \tau \;\middle|\; \boldsymbol{z}_j\right) &= 2\mathbb{P}\left(\langle \bar{\boldsymbol{a}}_i, \boldsymbol{U}_k^{*^T} \boldsymbol{z}_j\rangle \geq \tau\|\bar{\boldsymbol{a}}_i\| \;\middle|\; \boldsymbol{z}_j\right)\\ &\leq 2\mathbb{P}\left(\langle \bar{\boldsymbol{a}}_i, \boldsymbol{U}_k^{*^T} \boldsymbol{z}_j\rangle \geq \tau(\sqrt{d_k}-\alpha) \;\middle|\; \boldsymbol{z}_j\right) + 2\mathbb{P}\left(\|\bar{\boldsymbol{a}}_i\| \leq \sqrt{d_k}-\alpha\right)\\ &\leq 2 - 2\Phi\left(\frac{\tau(\sqrt{d_k}-\alpha)}{\|\boldsymbol{U}_k^{*^T}\boldsymbol{z}_j\|}\right) + 2N^{-2}.\end{aligned}$$

This, together with $\tau = \|\boldsymbol{U}_k^{*^T}\boldsymbol{z}_j\|\sqrt{\log N}/\sqrt{d_k}$, yields

$$\begin{aligned}\mathbb{P}\left(|\langle \boldsymbol{z}_i, \boldsymbol{z}_j\rangle| \geq \frac{\|\boldsymbol{U}_k^{*^T}\boldsymbol{z}_j\|\sqrt{\log N}}{\sqrt{d_k}} \;\middle|\; \boldsymbol{z}_j\right) &\leq 2 - 2\Phi\left(\frac{\sqrt{\log N}(\sqrt{d_k}-\alpha)}{\sqrt{d_k}}\right) + 2N^{-2}\\ &\leq \sqrt{\frac{2}{\pi}}\int_{\frac{\sqrt{\log N}(\sqrt{d_k}-\alpha)}{\sqrt{d_k}}}^{\infty} \exp\left(-\frac{t}{2}\frac{\sqrt{\log N}(\sqrt{d_k}-\alpha)}{\sqrt{d_k}}\right) dt + 2N^{-2}\\ &= \sqrt{\frac{2}{\pi}}\frac{2\sqrt{d_k}}{(\sqrt{d_k}-\alpha)\sqrt{\log N}} N^{-\frac{(\sqrt{d_k}-\alpha)^2}{2d_k}} + 2N^{-2}.\end{aligned}$$

□

Proof of Lemma 1. First, we prove (27). Suppose that a pair of data points $\boldsymbol{z}_i, \boldsymbol{z}_j \in S_k^*$ for some $k \in [K]$. According to (61) in Lemma 13 and $\|\boldsymbol{U}_k^{*^T}\boldsymbol{z}_j\| = \|\boldsymbol{a}_j\| = 1$ due to the UoS model, we obtain

$$2 - 2\Phi\left(\tau(\sqrt{d_k}+\alpha)\right) - 2N^{-2} \leq \mathbb{P}\left(|\langle \boldsymbol{z}_i, \boldsymbol{z}_j\rangle| \geq \tau \;\middle|\; \boldsymbol{z}_j\right) \leq 2 - 2\Phi\left(\tau(\sqrt{d_k}-\alpha)\right) + 2N^{-2}.$$

This, together with $p_{kk} = \mathbb{E}\left[\mathbb{P}\left(|\langle \boldsymbol{z}_i, \boldsymbol{z}_j\rangle| \geq \tau \;\middle|\; \boldsymbol{z}_j\right)\right]$, implies

$$2 - 2\Phi\left(\tau(\sqrt{d_k}+\alpha)\right) - 2N^{-2} \leq p_{kk} \leq 2 - 2\Phi\left(\tau(\sqrt{d_k}-\alpha)\right) + 2N^{-2}. \tag{63}$$

This, together with (21), yields that for all $k \in [K]$,

$$p_{kk} - b_{kk} \leq 2\Phi\left(\tau\sqrt{d_k}\right) - 2\Phi\left(\tau(\sqrt{d_k}-\alpha)\right) + 2N^{-2} \leq \sqrt{\frac{2}{\pi}}\exp\left(-\frac{\tau^2(\sqrt{d_k}-\alpha)^2}{2}\right)\tau\alpha + 2N^{-2} \lesssim \frac{1}{\log N}, \tag{64}$$

where the last inequality is due to (22) and $d_{\min} \gtrsim \log^3 N$. By the same argument, we have $p_{kk} - b_{kk} \gtrsim -\frac{1}{\log N}$. This, together with (64), implies (27) for all $k = \ell$. Suppose that a pair of data points $\boldsymbol{z}_i \in S_k^*, \boldsymbol{z}_j \in S_\ell^*$ for some $1 \leq k \neq \ell \leq K$. Since a uniform distribution over the sphere is rotationally invariant, we have

$$\|\boldsymbol{U}_k^{*^T}\boldsymbol{z}_j\| = \|\boldsymbol{U}_k^{*^T}\boldsymbol{U}_\ell^*\boldsymbol{a}_j\| = \|\boldsymbol{U}_{k\ell}^*\boldsymbol{\Sigma}_{k\ell}^*\boldsymbol{V}_{k\ell}^{*^T}\boldsymbol{a}_j\| \sim \|\boldsymbol{\Sigma}_{k\ell}^*\boldsymbol{a}_j\|,$$

where the second equality is due to (44). This, together with (53) in Lemma 11 and (61) in Lemma 13, implies it holds with probability at least $1 - 2N^{-2}$ that

$$2 - 2\Phi\left(\frac{\tau(\sqrt{d_k}+\alpha)(\sqrt{d_\ell}+\alpha)}{\|\boldsymbol{\Sigma}_{k\ell}^*\|_F - \alpha}\right) - 2N^{-2} \leq \mathbb{P}\left(|\langle \boldsymbol{z}_i, \boldsymbol{z}_j\rangle| \geq \tau \;\middle|\; \boldsymbol{z}_j\right) \leq 2 - 2\Phi\left(\frac{\tau(\sqrt{d_k}-\alpha)(\sqrt{d_\ell}-\alpha)}{\|\boldsymbol{\Sigma}_{k\ell}^*\|_F + \alpha}\right) + 2N^{-2}.$$

This, together with $\mathrm{aff}(S_k^*, S_\ell^*) = \|\mathbf{\Sigma}_{k\ell}^*\|_F$ and $p_{k\ell} = \mathbb{E}\left[\mathbb{P}\left(|\langle \boldsymbol{z}_i, \boldsymbol{z}_j\rangle| \geq \tau \;\middle|\; \boldsymbol{z}_j\right)\right]$, further implies

$$2 - 2\Phi\left(\frac{\tau(\sqrt{d_k}+\alpha)(\sqrt{d_\ell}+\alpha)}{\mathrm{aff}(S_k^*, S_\ell^*) - \alpha}\right) - 4N^{-2} \leq p_{k\ell} \leq 2 - 2\Phi\left(\frac{\tau(\sqrt{d_k}-\alpha)(\sqrt{d_\ell}-\alpha)}{\mathrm{aff}(S_k^*, S_\ell^*) + \alpha}\right) + 4N^{-2}. \tag{65}$$

This, together with (21), yields that for all $1 \leq k \neq \ell \leq K$,

$$\begin{aligned}
p_{k\ell} - b_{k\ell} &\leq 2\Phi\left(\frac{\tau\sqrt{d_k d_\ell}}{\mathrm{aff}(S_k^*, S_\ell^*)}\right) - 2\Phi\left(\frac{\tau(\sqrt{d_k}-\alpha)(\sqrt{d_\ell}-\alpha)}{\mathrm{aff}(S_k^*, S_\ell^*) + \alpha}\right) + 4N^{-2} \\
&\leq \sqrt{\frac{2}{\pi}} \exp\left(-\frac{\tau^2(\sqrt{d_k}-\alpha)^2(\sqrt{d_\ell}-\alpha)^2}{2(\mathrm{aff}(S_k^*, S_\ell^*)+\alpha)^2}\right)\left(\frac{\tau\sqrt{d_k d_\ell}}{\mathrm{aff}(S_k^*, S_\ell^*)} - \frac{\tau(\sqrt{d_k}-\alpha)(\sqrt{d_\ell}-\alpha)}{\mathrm{aff}(S_k^*, S_\ell^*)+\alpha}\right) \\
&\leq \sqrt{\frac{2}{\pi}} \exp\left(-\frac{(1-\varepsilon)^4\tau^2 d_{\min}^2}{2(1+\varepsilon)^2 \mathrm{aff}^2(S_k^*, S_\ell^*)}\right)\left(1 - \frac{(1-\varepsilon)^2}{1+\varepsilon}\right)\frac{\tau\sqrt{d_k d_\ell}}{\mathrm{aff}(S_k^*, S_\ell^*)} \\
&= \sqrt{\frac{2}{\pi}} \frac{\varepsilon(3-\varepsilon)\tau\sqrt{d_k d_\ell}}{1+\varepsilon} \exp\left(-\frac{(1-\varepsilon)^4\tau^2 d_{\min}^2}{2(1+\varepsilon)^2 \mathrm{aff}^2(S_k^*, S_\ell^*)}\right)\frac{1}{\mathrm{aff}(S_k^*, S_\ell^*)} \\
&= \sqrt{\frac{2}{\pi}} \exp\left(-\frac{1}{2}\right)\frac{\varepsilon(3-\varepsilon)\sqrt{d_k d_\ell}}{(1-\varepsilon)^2 d_{\min}} \lesssim \frac{d_{\max}}{d_{\min}\sqrt{\log N}},
\end{aligned} \tag{66}$$

where the third inequality is due to (60) and the last inequality follows from $\varepsilon \lesssim 1/\sqrt{\log N}$ and the fact that $\exp\left(-\eta x^2/2\right) x$ attains the maximum at $x = 1/\sqrt{\eta}$ when $x \in (0, \infty)$. By the same argument, we have $p_{k\ell} - b_{k\ell} \gtrsim -\frac{d_{\max}^{3/2}}{d_{\min}^{3/2}\sqrt{\log N}}$. This, together with (64), implies (27) for all $1 \leq k \neq \ell \leq K$.

Next, we prove (28). Note that for any $1 \leq k \neq \ell \leq K$,

$$\begin{aligned}
(\sqrt{d_k}-\alpha)(\sqrt{d_\ell}-\alpha) - (\sqrt{d_k}+\alpha)\left(\mathrm{aff}(S_k^*, S_\ell^*)+\alpha\right) &= \sqrt{d_k}\left(\sqrt{d_\ell} - \mathrm{aff}(S_k^*, S_\ell^*)\right) - \alpha\left(2\sqrt{d_k} + \sqrt{d_\ell} + \mathrm{aff}(S_k^*, S_\ell^*)\right) \\
&\geq \sqrt{d_k}\left(\sqrt{d_\ell} - \mathrm{aff}(S_k^*, S_\ell^*)\right) - 2\alpha\left(\sqrt{d_k} + \sqrt{d_\ell}\right) \\
&\geq \frac{1}{10}\sqrt{d_k d_\ell},
\end{aligned} \tag{67}$$

where the first inequality follows from $\mathrm{aff}(S_k^*, S_\ell^*) \leq \sqrt{d_\ell}$ and the second inequality uses $\mathrm{aff}(S_k^*, S_\ell^*) \leq 4\sqrt{d_\ell}/5$ due to (15) and $d_{\min} \gtrsim \log^3 N$. For ease of exposition, let

$$x_{k\ell} = \frac{\tau(\sqrt{d_k}-\alpha)(\sqrt{d_\ell}-\alpha)}{\mathrm{aff}(S_k^*, S_\ell^*)+\alpha}, \quad y_k = \tau(\sqrt{d_k}+\alpha).$$

According to (67) and (22), we have $x_{k\ell} > y_k$ for any $1 \leq k \neq \ell \leq K$. For all $1 \leq k \neq \ell \leq K$ satisfying $\mathrm{aff}(S_k^*, S_\ell^*) \geq \tau(\sqrt{d_k}-\alpha)(\sqrt{d_\ell}-\alpha)/(2c) - \alpha$, we have

$$x_{k\ell} \leq 2c$$

and

$$x_{k\ell} - y_k \geq \frac{\tau\sqrt{d_k d_\ell}}{10\left(\mathrm{aff}(S_k^*, S_\ell^*)+\alpha\right)} \geq \frac{c\sqrt{d_k d_\ell}}{10\min\{\sqrt{d_k}, \sqrt{d_\ell}\}\sqrt{d_{\max}}} \gtrsim \frac{\sqrt{d_{\min}}}{\sqrt{d_{\max}}},$$

where the first inequality uses (67) and the second inequality follows from (60) and (15). This, together with (63), (65), and $x_{k\ell} > y_k$, yields that for all $1 \leq k \neq \ell \leq K$,

$$p_{kk} - p_{k\ell} \geq 2\left(\Phi\left(x_{k\ell}\right) - \Phi\left(y_k\right)\right) - 6N^{-2} \geq \sqrt{\frac{2}{\pi}} \exp\left(-\frac{x_{k\ell}^2}{2}\right)\left(x_{k\ell} - y_k\right) - 6N^{-2} \gtrsim \frac{\sqrt{d_{\min}}}{\sqrt{d_{\max}}}. \tag{68}$$

For all $1 \leq k \neq \ell \leq K$ satisfying $\mathrm{aff}(S_k^*, S_\ell^*) < \tau(\sqrt{d_k} - \alpha)(\sqrt{d_\ell} - \alpha)/(2c) - \alpha$, we have for all $1 \leq k \neq \ell \leq K$,

$$p_{kk} \geq 2 - 2\Phi\left(\frac{c(\sqrt{d_k} + \alpha)}{\sqrt{d_{\max}}}\right) - 2N^{-2} \geq 2 - 2\Phi\left((1+\varepsilon)c\right) - 2N^{-2},$$

where the first inequality is due to (63) and (22) and the second inequality uses (60), and

$$p_{k\ell} \leq 2 - 2\Phi\left(2c\right) + 4N^{-2}.$$

Then, we have for all $1 \leq k \neq \ell \leq K$,

$$p_{kk} - p_{k\ell} \geq 2\Phi\left(2c\right) - 2\Phi\left((1+\varepsilon)c\right) - 6N^{-2},$$

which is a constant due to the fact that $c > 0$ is a constant. This, together with (68) and $d_{\min} \gtrsim \log N$, implies (28). □

B.2. Proof of Lemma 2

Proof. According to (21) and (22), we have for all $1 \leq k \neq \ell \leq K$,

$$b_{kk} = 2 - 2\Phi\left(\frac{\sqrt{cd_k}}{\sqrt{d_{\max}}}\right) \geq 2 - 2\Phi(\sqrt{c}) \tag{69}$$

and

$$b_{k\ell} = 2 - 2\Phi\left(\frac{\sqrt{cd_k d_\ell}}{\sqrt{d_{\max}}\mathrm{aff}(S_k^*, S_\ell^*)}\right) \leq 2 - 2\Phi\left(\frac{\sqrt{cd_{\min}}}{\kappa\sqrt{d_{\max}}}\right) = 2 - 2\Phi\left(\frac{\sqrt{c}}{\kappa\sqrt{\kappa_d}}\right), \tag{70}$$

where the inequality is due to $\mathrm{aff}(S_k^*, S_\ell^*) \leq \kappa \min\{\sqrt{d_k}, \sqrt{d_\ell}\}$ for all $1 \leq k \neq \ell \leq K$ by Assumption 2 and (15). Then, we can decompose the symmetric matrix $\boldsymbol{B}$ into $\boldsymbol{B} = \boldsymbol{B}_1 + \boldsymbol{B}_2$, where

$$\boldsymbol{B}_1 = \begin{bmatrix} \frac{b_{11}}{2} & b_{12} & \dots & b_{1K} \\ b_{12} & \frac{b_{22}}{2} & \dots & b_{2K} \\ \vdots & \vdots & \ddots & \vdots \\ b_{1K} & b_{2K} & \dots & \frac{b_{KK}}{2} \end{bmatrix}, \quad \boldsymbol{B}_2 = \frac{1}{2}\begin{bmatrix} b_{11} & 0 & \dots & 0 \\ 0 & b_{22} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & b_{KK} \end{bmatrix}.$$

According to (69), (70), and (23), we can verify for all $k \in [K]$,

$$\frac{1}{2}|b_{kk}| - \left|\sum_{\ell:k\neq\ell} b_{k\ell}\right| \geq 1 - \Phi(\sqrt{c}) - 2(K-1)\left(1 - \Phi\left(\frac{\sqrt{c}}{\kappa\sqrt{\kappa_d}}\right)\right) \geq 0,$$

which implies that $\boldsymbol{B}_1$ is a symmetric diagonally dominant matrix (see Golub & Van Loan (2013, Section 4.1.1)). Using the result that a symmetric diagonally dominant matrix with real non-negative diagonal entries is positive semidefinite, we can conclude that $\boldsymbol{B}_1$ is positive semidefinite. On the other hand, we can see that $\boldsymbol{B}_2$ is a diagonal matrix with all the diagonal elements being larger than $1 - \Phi(\sqrt{c})$. Then, we have

$$\min_{\|\boldsymbol{x}\|=1} \boldsymbol{x}^T\boldsymbol{B}\boldsymbol{x} \geq \min_{\|\boldsymbol{x}\|=1} \boldsymbol{x}^T\boldsymbol{B}_1\boldsymbol{x} + \min_{\|\boldsymbol{x}\|=1} \boldsymbol{x}^T\boldsymbol{B}_2\boldsymbol{x} = \lambda_{\min}(\boldsymbol{B}_1) + \lambda_{\min}(\boldsymbol{B}_2) \geq 1 - \Phi(\sqrt{c}).$$

Then, we complete the proof. □

B.3. Proof of Proposition 1

Before we prove Proposition 1, we need estimate the covariance between the random variables a_{ik} and a_{jk} generated by the thresholding procedure (11) for all $1 \leq i \neq j \leq N$ and $k \in [N]$.

Lemma 14. *Suppose that $\boldsymbol{z}_i$, $\boldsymbol{z}_j$, and $\boldsymbol{z}_k$ are different points generated according to the semi-random UoS model such that $\boldsymbol{z}_k \in S_\ell$ for some $\ell \in [K]$. Suppose in addition that Assumption 2 holds, the thresholding parameter is set as in (22), $d_{\min} \gtrsim \log N$. Then, it holds for any $1 \leq i \neq j \leq N$ with probability at least $1 - 5K^2N^{-2}$ that*

$$|\mathbb{E}[a_{ik}a_{jk}|\boldsymbol{z}_i, \boldsymbol{z}_j] - \mathbb{E}[a_{ik}|\boldsymbol{z}_i]\mathbb{E}[a_{jk}|\boldsymbol{z}_j]| \lesssim \frac{d_{\max}}{d_{\min}\sqrt{\log N}}.$$

Proof of Lemma 14. Suppose that (56) and (57) hold, which happens with probability at least $1 - 5K^2/N$ according to Lemma 12. To simplify the notations, let

$$\boldsymbol{v}_i := \boldsymbol{U}_\ell^{*^T} \boldsymbol{z}_i,\ \tilde{\boldsymbol{v}}_i := \frac{\boldsymbol{v}_i}{\|\boldsymbol{v}_i\|},\ \text{for all } i \in [N]. \tag{71}$$

Besides, for any given $\boldsymbol{v}_i$ and $\boldsymbol{v}_j$, let

$$\beta_{ij} := \frac{\left(\tau + |\langle \boldsymbol{v}_i, \tilde{\boldsymbol{v}}_j\rangle| \sqrt{\log N}/\sqrt{d_\ell}\right)\left(\sqrt{d_\ell} + \alpha\right)}{\left(\|\boldsymbol{v}_i\| - |\langle \boldsymbol{v}_i, \tilde{\boldsymbol{v}}_j\rangle|\right)\sqrt{1 - (\log N)/d_\ell}},\ \beta'_{ij} := \frac{\left(\tau - |\langle \boldsymbol{v}_i, \tilde{\boldsymbol{v}}_j\rangle| \sqrt{\log N}/\sqrt{d_\ell}\right)\left(\sqrt{d_\ell} - \alpha\right)}{\|\boldsymbol{v}_i\|}. \tag{72}$$

In addition, suppose that the following inequalities hold:

$$\mathbb{E}[a_{ik}a_{jk}|\boldsymbol{z}_i, \boldsymbol{z}_j] \geq \left(2 - 2\Phi\left(\beta_{ij}\right) - 2N^{-2}\right)\mathbb{P}\left(\tau \leq |\langle \boldsymbol{z}_j, \boldsymbol{z}_k\rangle| \leq \frac{\|\boldsymbol{v}_j\|\sqrt{\log N}}{\sqrt{d_\ell}} \,\middle|\, \boldsymbol{z}_j\right) \tag{73}$$

$$\begin{aligned}\mathbb{E}[a_{ik}a_{jk}|\boldsymbol{z}_i, \boldsymbol{z}_j] \leq{}& \left(2 - 2\Phi\left(\beta'_{ij}\right) + 2N^{-2}\right)\mathbb{P}\left(\tau \leq |\langle \boldsymbol{z}_j, \boldsymbol{z}_k\rangle| \leq \frac{\|\boldsymbol{v}_j\|\sqrt{\log N}}{\sqrt{d_\ell}} \,\middle|\, \boldsymbol{z}_j\right) + \\ &\left(2 + 2N^{-2}\right)\mathbb{P}\left(\frac{\|\boldsymbol{v}_j\|\sqrt{\log N}}{\sqrt{d_\ell}} \leq |\langle \boldsymbol{z}_j, \boldsymbol{z}_k\rangle| \leq 1 \,\middle|\, \boldsymbol{z}_j\right).\end{aligned} \tag{74}$$

According to Lemma 13, we obtain

$$2 - 2\Phi\left(\frac{\tau(\sqrt{d_\ell} + \alpha)}{\|\boldsymbol{v}_i\|}\right) - 2N^{-2} \leq \mathbb{E}[a_{ik}|\boldsymbol{z}_i] \leq 2 - 2\Phi\left(\frac{\tau(\sqrt{d_\ell} - \alpha)}{\|\boldsymbol{v}_i\|}\right) + 2N^{-2}. \tag{75}$$

This, together with (73), yields

$$\begin{aligned}&\mathbb{E}[a_{ik}a_{jk}|\boldsymbol{z}_i, \boldsymbol{z}_j] - \mathbb{E}[a_{ik}|\boldsymbol{z}_i]\mathbb{E}[a_{jk}|\boldsymbol{z}_j] \\ &\geq 2\left(1 - \Phi\left(\beta_{ij}\right) - N^{-2}\right)\mathbb{P}\left(\tau \leq |\langle \boldsymbol{z}_j, \boldsymbol{z}_k\rangle| \leq \frac{\|\boldsymbol{v}_j\|\sqrt{\log N}}{\sqrt{d_\ell}} \,\middle|\, \boldsymbol{z}_j\right) - 2\left(1 - \Phi\left(\frac{\tau(\sqrt{d_\ell} - \alpha)}{\|\boldsymbol{v}_i\|}\right) + N^{-2}\right) \\ &\mathbb{P}\left(|\langle \boldsymbol{z}_j, \boldsymbol{z}_k\rangle| \geq \tau \,\middle|\, \boldsymbol{z}_j\right) = 2\left(\Phi\left(\frac{\tau(\sqrt{d_\ell} - \alpha)}{\|\boldsymbol{v}_i\|}\right) - \Phi\left(\beta_{ij}\right) - 2N^{-2}\right)\mathbb{P}\left(\tau \leq |\langle \boldsymbol{z}_j, \boldsymbol{z}_k\rangle| \leq \frac{\|\boldsymbol{v}_j\|\sqrt{\log N}}{\sqrt{d_\ell}} \,\middle|\, \boldsymbol{z}_j\right) \\ &- 2\left(1 - \Phi\left(\frac{\tau(\sqrt{d_\ell} - \alpha)}{\|\boldsymbol{v}_i\|}\right) + N^{-2}\right)\mathbb{P}\left(|\langle \boldsymbol{z}_j, \boldsymbol{z}_k\rangle| \geq \frac{\|\boldsymbol{v}_j\|\sqrt{\log N}}{\sqrt{d_\ell}} \,\middle|\, \boldsymbol{z}_j\right) \gtrsim -\frac{d_{\max}}{d_{\min}\sqrt{\log N}},\end{aligned} \tag{76}$$

where the last inequality is due to Lemma 16, Lemma 13, and (22). By the similar argument, according to (74) and (75), we have

$$\begin{aligned}\mathbb{E}[a_{ik}a_{jk}|\boldsymbol{z}_i, \boldsymbol{z}_j] - \mathbb{E}[a_{ik}|\boldsymbol{z}_i]\mathbb{E}[a_{jk}|\boldsymbol{z}_j] \leq{}& 2\left(\Phi\left(\frac{\tau(\sqrt{d_\ell} + \alpha)}{\|\boldsymbol{v}_i\|}\right) - \Phi(\beta'_{ij}) + 2N^{-2}\right)\mathbb{P}\left(\tau \leq |\langle \boldsymbol{z}_j, \boldsymbol{z}_k\rangle| \leq \frac{\sqrt{\log N}}{\sqrt{d_\ell}} \,\middle|\, \boldsymbol{z}_j\right) \\ &+ 2\left(\Phi\left(\frac{\tau(\sqrt{d_\ell} + \alpha)}{\|\boldsymbol{v}_i\|}\right) + 2N^{-2}\right)\mathbb{P}\left(|\langle \boldsymbol{z}_j, \boldsymbol{z}_k\rangle| \geq \frac{\sqrt{\log N}}{\sqrt{d_\ell}} \,\middle|\, \boldsymbol{z}_j\right) \lesssim \frac{d_{\max}}{d_{\min}\sqrt{\log N}}.\end{aligned}$$

According to this and (76), we complete the proof.

Then, the rest of the proof is devoted to proving (73) and (74). According to (11) and $\boldsymbol{z}_k = \boldsymbol{U}_\ell^* \boldsymbol{a}_k$, we have

$$\begin{aligned}\mathbb{E}[a_{ik}a_{jk}|\boldsymbol{z}_i, \boldsymbol{z}_j] &= \mathbb{P}\left(|\langle \boldsymbol{z}_i, \boldsymbol{z}_k\rangle| \geq \tau, |\langle \boldsymbol{z}_j, \boldsymbol{z}_k\rangle| \geq \tau \,\middle|\, \boldsymbol{z}_i, \boldsymbol{z}_j\right) \\ &= \int_{-\infty}^{\infty} \mathbb{P}\left(|\langle \boldsymbol{z}_i, \boldsymbol{z}_k\rangle| \geq \tau, |\langle \boldsymbol{z}_j, \boldsymbol{z}_k\rangle| \geq \tau \,\middle|\, |\langle \boldsymbol{z}_j, \boldsymbol{z}_k\rangle| = t, \boldsymbol{z}_i, \boldsymbol{z}_j\right) d\mathbb{P}\left(|\langle \boldsymbol{z}_j, \boldsymbol{z}_k\rangle| \leq t \,\middle|\, \boldsymbol{z}_j\right) \\ &= \int_{\tau}^{1} \mathbb{P}\left(|\langle \boldsymbol{z}_i, \boldsymbol{z}_k\rangle| \geq \tau \,\middle|\, |\langle \boldsymbol{z}_j, \boldsymbol{z}_k\rangle| = t, \boldsymbol{z}_i, \boldsymbol{z}_j\right) d\mathbb{P}\left(|\langle \boldsymbol{z}_j, \boldsymbol{z}_k\rangle| \leq t \,\middle|\, \boldsymbol{z}_j\right) \\ &= \int_{\tau}^{1} \mathbb{P}\left(|\langle \boldsymbol{v}_i, \boldsymbol{a}_k\rangle| \geq \tau \,\middle|\, |\langle \boldsymbol{v}_j, \boldsymbol{a}_k\rangle| = t, \boldsymbol{z}_i, \boldsymbol{z}_j\right) d\mathbb{P}\left(|\langle \boldsymbol{z}_j, \boldsymbol{z}_k\rangle| \leq t \,\middle|\, \boldsymbol{z}_j\right),\end{aligned} \tag{77}$$

where the second equality is due to the the law of total probability, the third equality uses $\tau \leq |\langle \boldsymbol{z}_j, \boldsymbol{z}_k \rangle| \leq 1$, and the last equality follows from $\boldsymbol{v}_i = \boldsymbol{U}_\ell^{*^T} \boldsymbol{z}_i$ for all $i \in [N]$. Due to $\|\boldsymbol{a}_k\| = 1$ and $\|\tilde{\boldsymbol{v}}_j\| = 1$, we can decompose $\boldsymbol{a}_k$ into two parts that are orthogonal:

$$\boldsymbol{a}_k = x\tilde{\boldsymbol{v}}_j + \sqrt{1-x^2}\boldsymbol{b}_k, \tag{78}$$

where $x \in \mathbb{R}$ and $\boldsymbol{b}_k \in \mathbb{R}^{d_\ell}$ satisfying $\langle \boldsymbol{b}_k, \tilde{\boldsymbol{v}}_j \rangle = 0$ and $\|\boldsymbol{b}_k\| = 1$. This, together with $|\langle \boldsymbol{v}_j, \boldsymbol{a}_k \rangle| = t$, implies

$$t = |x| \|\boldsymbol{v}_j\|. \tag{79}$$

Since $\|\tilde{\boldsymbol{v}}_j\| = 1$, there exists an orthogonal matrix $\boldsymbol{U} \in \mathcal{O}^{d_\ell}$ such that $\boldsymbol{U}\tilde{\boldsymbol{v}}_j = \boldsymbol{e}_1$. Let

$$\tilde{\boldsymbol{b}}_k := \frac{\boldsymbol{U}^T \boldsymbol{a}_k - x\boldsymbol{e}_1}{\sqrt{1-x^2}} \tag{80}$$

and $\boldsymbol{c}_k \in \mathbb{R}^{d_\ell - 1}$ such that $\boldsymbol{c}_k := (\tilde{b}_{k2}, \cdots, \tilde{b}_{kd_\ell})$. According to Lemma 15, we obtain

$$\boldsymbol{U}\boldsymbol{b}_k \sim \tilde{\boldsymbol{b}}_k \tag{81}$$

such that $\tilde{b}_{k1} = 0$ and $\boldsymbol{c}_k \sim \mathrm{Unif}(\mathbb{S}^{d_\ell - 2})$. Besides, let

$$\boldsymbol{w}_i := \boldsymbol{v}_i - \langle \boldsymbol{v}_i, \tilde{\boldsymbol{v}}_j \rangle \tilde{\boldsymbol{v}}_j. \tag{82}$$

According to (78), we have

$$\langle \boldsymbol{v}_i, \boldsymbol{a}_k \rangle = x\langle \boldsymbol{v}_i, \tilde{\boldsymbol{v}}_j \rangle + \sqrt{1-x^2}\langle \boldsymbol{v}_i, \boldsymbol{b}_k \rangle = x\langle \boldsymbol{v}_i, \tilde{\boldsymbol{v}}_j \rangle + \sqrt{1-x^2}\langle \boldsymbol{w}_i, \boldsymbol{b}_k \rangle, \tag{83}$$

where the second equality is due to (82) and $\langle \tilde{\boldsymbol{v}}_j, \boldsymbol{b}_k \rangle = 0$. According to $\langle \boldsymbol{w}_i, \tilde{\boldsymbol{v}}_j \rangle = 0$, we have $\langle \boldsymbol{U}^T \boldsymbol{w}_i, \boldsymbol{U}\tilde{\boldsymbol{v}}_j \rangle = \langle \boldsymbol{U}^T \boldsymbol{w}_i, \boldsymbol{e}_1 \rangle = 0$. This, together with letting $\boldsymbol{u}_i$ denote the i-th column of $\boldsymbol{U} \in \mathcal{O}^{d_\ell}$ and $\boldsymbol{d}_i := \left(\boldsymbol{u}_2^T \boldsymbol{w}_i, \ldots, \boldsymbol{u}_d^T \boldsymbol{w}_i\right) \in \mathbb{R}^{d_\ell - 1}$, we have $\boldsymbol{u}_1^T \boldsymbol{w}_i = 0$ and $\|\boldsymbol{d}_i\| = \|\boldsymbol{w}_i\|$. It follows from this and (81) that

$$|\langle \boldsymbol{w}_i, \boldsymbol{b}_k \rangle| = |\langle \boldsymbol{U}^T \boldsymbol{w}_i, \boldsymbol{U}\boldsymbol{b}_k \rangle| \sim |\langle \boldsymbol{U}^T \boldsymbol{w}_i, \tilde{\boldsymbol{b}}_k \rangle| = |\langle \boldsymbol{d}_i, \boldsymbol{c}_k \rangle|. \tag{84}$$

Now, we are ready to compute the lower bound of $\mathbb{E}[a_{ik}a_{jk}|\boldsymbol{z}_i, \boldsymbol{z}_j]$. According to (77), (79), (83), and (84), we have

$$\begin{aligned}
\mathbb{P}\left(|\langle \boldsymbol{v}_i, \boldsymbol{a}_k \rangle| \geq \tau \;\middle|\; |\langle \boldsymbol{v}_j, \boldsymbol{a}_k \rangle| = t, \boldsymbol{z}_i, \boldsymbol{z}_j\right) &\geq \mathbb{P}\left(|\langle \boldsymbol{w}_i, \boldsymbol{b}_k \rangle| \geq \frac{\tau + |x\langle \boldsymbol{v}_i, \tilde{\boldsymbol{v}}_j \rangle|}{\sqrt{1-x^2}} \;\middle|\; |x| = \frac{t}{\|\boldsymbol{v}_j\|}, \boldsymbol{z}_i, \boldsymbol{z}_j\right) \\
&= \mathbb{P}\left(|\langle \boldsymbol{d}_i, \boldsymbol{c}_k \rangle| \geq \frac{\tau\|\boldsymbol{v}_j\| + t|\langle \boldsymbol{v}_i, \tilde{\boldsymbol{v}}_j \rangle|}{\sqrt{\|\boldsymbol{v}_j\|^2 - t^2}} \;\middle|\; \boldsymbol{z}_i, \boldsymbol{z}_j\right).
\end{aligned} \tag{85}$$

Then, let

$$h(t) := \frac{\left(\tau\|\boldsymbol{v}_j\| + t|\langle \boldsymbol{v}_i, \tilde{\boldsymbol{v}}_j \rangle|\right)\left(\sqrt{d_\ell} + \alpha\right)}{\left(\|\boldsymbol{v}_i\| - |\langle \boldsymbol{v}_i, \tilde{\boldsymbol{v}}_j \rangle|\right)\sqrt{\|\boldsymbol{v}_j\|^2 - t^2}}.$$

According to the argument in Lemma 13 with $\alpha = 2\sqrt{\log N} + 2$, $\boldsymbol{c}_k \sim \mathrm{Unif}(\mathbb{S}^{d_\ell - 2})$, and $\|\boldsymbol{d}_i\| = \|\boldsymbol{w}_i\|$, we obtain

$$\begin{aligned}
\mathbb{P}\left(|\langle \boldsymbol{d}_i, \boldsymbol{c}_k \rangle| \geq \frac{\tau\|\boldsymbol{v}_j\| + t|\langle \boldsymbol{v}_i, \tilde{\boldsymbol{v}}_j \rangle|}{\sqrt{\|\boldsymbol{v}_j\|^2 - t^2}} \;\middle|\; \boldsymbol{z}_i, \boldsymbol{z}_j\right) &\geq 2 - 2\Phi\left(\frac{\left(\tau\|\boldsymbol{v}_j\| + t|\langle \boldsymbol{v}_i, \tilde{\boldsymbol{v}}_j \rangle|\right)\left(\sqrt{d_\ell} + \alpha\right)}{\|\boldsymbol{w}_i\|\sqrt{\|\boldsymbol{v}_j\|^2 - t^2}}\right) - 2N^{-2} \\
&\geq 2 - 2\Phi\left(h(t)\right) - 2N^{-2},
\end{aligned} \tag{86}$$

where the second inequality is due to $\|\boldsymbol{w}_i\| \geq \|\boldsymbol{v}_i\| - |\langle \boldsymbol{v}_i, \tilde{\boldsymbol{v}}_j \rangle| > 0$ according to (82) and the triangle inequality. According to (77), (85), and (86), we have

$$\begin{aligned}
\mathbb{E}[a_{ik}a_{jk}|\boldsymbol{z}_i, \boldsymbol{z}_j] &\geq \int_{\tau}^{1} \left(2 - 2\Phi\left(h(t)\right) - 2N^{-2}\right) d\mathbb{P}\left(|\langle \boldsymbol{z}_j, \boldsymbol{z}_k \rangle| \leq t \;\middle|\; \boldsymbol{z}_j\right) \\
&\geq \int_{\tau}^{\frac{\|\boldsymbol{v}_j\|\sqrt{\log N}}{\sqrt{d_\ell}}} \left(2 - 2\Phi\left(h(t)\right) - 2N^{-2}\right) d\mathbb{P}\left(|\langle \boldsymbol{z}_j, \boldsymbol{z}_k \rangle| \leq t \;\middle|\; \boldsymbol{z}_j\right) \\
&\geq \int_{\tau}^{\frac{\|\boldsymbol{v}_j\|\sqrt{\log N}}{\sqrt{d_\ell}}} \left(2 - 2\Phi\left(\beta_{ij}\right) - 2N^{-2}\right) d\mathbb{P}\left(|\langle \boldsymbol{z}_j, \boldsymbol{z}_k \rangle| \leq t \;\middle|\; \boldsymbol{z}_j\right) \\
&= \left(2 - 2\Phi\left(\beta_{ij}\right) - 2N^{-2}\right) \mathbb{P}\left(\tau \leq |\langle \boldsymbol{z}_j, \boldsymbol{z}_k \rangle| \leq \frac{\|\boldsymbol{v}_j\|\sqrt{\log N}}{\sqrt{d}} \;\middle|\; \boldsymbol{z}_j\right),
\end{aligned} \tag{87}$$

where the last inequality is because $h(t)$ is an increasing function. Next, by letting

$$g(t) := \frac{\left(\tau\|\boldsymbol{v}_j\| - t|\langle \boldsymbol{v}_i, \tilde{\boldsymbol{v}}_j \rangle|\right)\left(\sqrt{d_\ell} - \alpha\right)}{\|\boldsymbol{v}_i\|\|\boldsymbol{v}_j\|},$$

we can obtain the following inequality by the same argument as (85) and (86):

$$\begin{aligned}
\mathbb{P}\left(|\langle \boldsymbol{v}_i, \boldsymbol{a}_k \rangle| \geq \tau \;\middle|\; \langle \boldsymbol{v}_j, \boldsymbol{a}_k \rangle = t, \boldsymbol{z}_i, \boldsymbol{z}_j\right) &\leq 2 - 2\Phi\left(\frac{\left(\tau\|\boldsymbol{v}_j\| - t|\langle \boldsymbol{v}_i, \tilde{\boldsymbol{v}}_j \rangle|\right)\left(\sqrt{d_\ell} - \alpha\right)}{\|\boldsymbol{w}_i\|\sqrt{\|\boldsymbol{v}_j\|^2 - t^2}}\right) + 2N^{-2} \\
&\leq 2 - 2\Phi\left(g(t)\right) + 2N^{-2},
\end{aligned}$$

where the last inequality is due to $\|\boldsymbol{w}_i\| \leq \|\boldsymbol{v}_i\|$. Besides, it holds for $t \in (\tau, \|\boldsymbol{v}_j\|\sqrt{\log N}/\sqrt{d_\ell}]$ that

$$g(t) \geq \beta'_{ij}.$$

These, together with (77), imply

$$\begin{aligned}
\mathbb{E}[a_{ik}a_{jk}|\boldsymbol{z}_i, \boldsymbol{z}_j] &\leq \int_{\tau}^{1} \left(2 - 2\Phi\left(g(t)\right) + 2N^{-2}\right) d\mathbb{P}\left(|\langle \boldsymbol{z}_j, \boldsymbol{z}_k \rangle| \leq t \;\middle|\; \boldsymbol{z}_j\right) \\
&= \int_{\tau}^{\frac{\|\boldsymbol{v}_j\|\sqrt{\log N}}{\sqrt{d_\ell}}} \left(2 - 2\Phi\left(g(t)\right) + 2N^{-2}\right) d\mathbb{P}\left(|\langle \boldsymbol{z}_j, \boldsymbol{z}_k \rangle| \leq t \;\middle|\; \boldsymbol{z}_j\right) + \\
&\quad \int_{\frac{\|\boldsymbol{v}_j\|\sqrt{\log N}}{\sqrt{d_\ell}}}^{1} \left(2 - 2\Phi\left(g(t)\right) + 2N^{-2}\right) d\mathbb{P}\left(|\langle \boldsymbol{z}_j, \boldsymbol{z}_k \rangle| \leq t \;\middle|\; \boldsymbol{z}_j\right) \\
&\leq \left(2 - 2\Phi\left(\beta'_{ij}\right) + 2N^{-2}\right) \mathbb{P}\left(\tau \leq |\langle \boldsymbol{z}_j, \boldsymbol{z}_k \rangle| \leq \frac{\|\boldsymbol{v}_j\|\sqrt{\log N}}{\sqrt{d_\ell}} \;\middle|\; \boldsymbol{z}_j\right) + \\
&\quad \left(2 + 2N^{-2}\right) \mathbb{P}\left(\frac{\|\boldsymbol{v}_j\|\sqrt{\log N}}{\sqrt{d_\ell}} \leq |\langle \boldsymbol{z}_j, \boldsymbol{z}_k \rangle| \leq 1 \;\middle|\; \boldsymbol{z}_j\right).
\end{aligned} \tag{88}$$

□

Proof of Proposition 1. Suppose that (56) and (57) hold, which happens with probability at least $1 - 5K^2/N$ according to Lemma 12. For ease of exposition, let $\Delta := \boldsymbol{A} - \mathbb{E}[\boldsymbol{A}]$. Recall the definition of p_k and $q_{k\ell}$ for $1 \leq k \neq \ell \leq K$ in Lemma 1. It follows from (11) and Lemma 1 that the (i,j)-th element of $\boldsymbol{A}$ satisfies $a_{ij} \sim \mathbf{Bern}(p_{ij})$ such that

$$p_{ij} = \begin{cases} p_k, & \text{if } \boldsymbol{z}_i, \boldsymbol{z}_j \in S_k, \\ q_{k\ell}, & \text{if } \boldsymbol{z}_i \in S_k,\ \boldsymbol{z}_j \in S_\ell, \text{ and } k \neq \ell. \end{cases} \tag{89}$$

According to (63), (22), and $d_{\min} \gtrsim \log N$, one can verify that $p_k \in (0,1)$ is a constant for all $k \in [K]$. According to (65) and (22), we have for $1 \leq k \neq \ell \leq K$,

$$\begin{aligned} q_{k\ell} &\leq 2 - 2\Phi\left(\frac{\max_{k\neq\ell} \operatorname{aff}(S_k, S_\ell) + \alpha}{\operatorname{aff}(S_k, S_\ell) + \alpha} \frac{(\sqrt{d_k} - \alpha)(\sqrt{d_\ell} - \alpha)}{(\sqrt{d_{\max}} - \alpha)^2}\right) + 4N^{-2} \\ &\leq 2 - 2\Phi\left(\frac{(\sqrt{d_k} - \alpha)(\sqrt{d_\ell} - \alpha)}{(\sqrt{d_{\max}} - \alpha)^2}\right) + 4N^{-2}. \end{aligned} \tag{90}$$

Now, we are devoted to bounding $\|\Delta\|^2 = \|\Delta^2\|$. First, we consider the diagonal elements of Δ^2. According to (11), we note that a_{ij} for all $j \in [n]$ are mutually independent conditioned on $\boldsymbol{z}_i \in \mathbb{R}^N$. This, together with $\delta_{ij} = a_{ij} - \mathbb{E}[a_{ij}] \in \{1 - p_{ij}, -p_{ij}\}$ and the Hoeffding's inequality for general bounded random variables (see, e.g., Vershynin (2018, Theorem 2.2.6)), yields that

$$\mathbb{P}\left(\left|\sum_{j=1}^N \left(\delta_{ij}^2 - \mathbb{E}[\delta_{ij}^2]\right)\right| \geq \sqrt{N \log N} \;\middle|\; \boldsymbol{z}_i\right) \leq 2\exp\left(-\frac{2N\log N}{N}\right) = 2N^{-2}.$$

This, together with the union bound, yields that it holds with probability at least $1 - 2N^{-1}$ that for all $i \in [N]$,

$$\left|\sum_{j=1}^N \left(\delta_{ij}^2 - \mathbb{E}[\delta_{ij}^2]\right)\right| \leq \sqrt{N \log N}. \tag{91}$$

Due to the fact that $p_k \in (0,1)$ is a constant for all $k \in [K]$ and (90), we have for all $1 \leq i < j \leq N$,

$$\mathbb{E}[\delta_{ij}^2] = \mathbb{E}[a_{ij}^2] - \mathbb{E}^2[a_{ij}] = p_{ij}(1 - p_{ij})$$

is less than some constant. According to this and (91), it holds with probability at least $1 - 2N^{-1}$ that for all $i \in [N]$,

$$\left|(\Delta^2)_{ii}\right| = \left|\sum_{j=1}^N \delta_{ij}^2\right| \leq \left|\sum_{j=1}^N \mathbb{E}[\delta_{ij}^2]\right| + \sqrt{N \log N} \lesssim N. \tag{92}$$

Next, we consider the off-diagonal elements of Δ^2. According to (11), we note that $a_{ik}a_{jk}$ for all $k \neq i$ and $k \neq j$ are mutually independent conditioned on $\boldsymbol{z}_i, \boldsymbol{z}_j \in \mathbb{R}^N$ for all $1 \leq i \neq j \leq N$. This, together with $\delta_{ij} = a_{ij} - \mathbb{E}[a_{ij}]$ and the Hoeffding's inequality for general bounded random variables (see, e.g., Vershynin (2018, Theorem 2.2.6)), yields that

$$\mathbb{P}\left(\left|\sum_{k\neq i, k\neq j} (\delta_{ik}\delta_{jk} - \mathbb{E}[\delta_{ik}\delta_{jk}])\right| \geq \sqrt{2N \log N} \;\middle|\; \boldsymbol{z}_i, \boldsymbol{z}_j\right) \leq 2\exp\left(-\frac{4N\log N}{N}\right) = 2N^{-4}.$$

According to Jensen's inequality, Lemma 14, and $\mathbb{E}[\delta_{ik}\delta_{jk}|\boldsymbol{z}_i, \boldsymbol{z}_j] = \mathbb{E}[a_{ik}a_{jk}|\boldsymbol{z}_i, \boldsymbol{z}_j] - \mathbb{E}[a_{ik}|\boldsymbol{z}_i]\mathbb{E}[a_{jk}|\boldsymbol{z}_j]$, we obtain for $k \neq i, k \neq j$,

$$|\mathbb{E}[\delta_{ik}\delta_{jk}]| \leq \mathbb{E}\left[|\mathbb{E}[\delta_{ik}\delta_{jk}|\boldsymbol{z}_i, \boldsymbol{z}_j]|\right] \lesssim \frac{d_{\max}}{d_{\min}\sqrt{\log N}}.$$

These, together with the union bound, yields that it holds with probability at least $1 - 2N^{-2}$ that for all $1 \leq i \neq j \leq N$,

$$\left|\sum_{k\neq i, k\neq j} \delta_{ik}\delta_{jk}\right| \leq \left|\sum_{k\neq i, k\neq j} (\delta_{ik}\delta_{jk} - \mathbb{E}[\delta_{ik}\delta_{jk}])\right| + \left|\sum_{k\neq i, k\neq j} \mathbb{E}[\delta_{ik}\delta_{jk}]\right| \lesssim \sqrt{2N\log N} + \frac{d_{\max}N}{d_{\min}\sqrt{\log N}}.$$

As a result, it holds with probability at least $1 - 2N^{-2}$ that for any $1 \leq i \neq j \leq N$,

$$\left|(\Delta^2)_{ij}\right| = \left|\sum_{k=1}^N \delta_{ik}\delta_{jk}\right| \lesssim \frac{d_{\max}N}{d_{\min}\sqrt{\log N}}, \tag{93}$$

Applying the union bound to (92) and (93) yields that

$$\|\Delta^2\| \leq \max_{i\in[N]} \left|(\Delta^2)_{ii}\right| + \sqrt{\sum_{i\neq j} |(\Delta^2)_{ij}|^2} \lesssim N + \frac{d_{\max} N^2}{d_{\min}\sqrt{\log N}}$$

holds with probability at least $1 - 6K^2N^{-1}$. This further implies

$$\|\Delta\| \lesssim \sqrt{\frac{d_{\max}}{d_{\min}}} \frac{N}{\sqrt[4]{\log N}}.$$

□

B.4. Proof of Theorem 2

Proof. Suppose that (29) holds, which happens with probability at least $1 - 6K^2N^{-1}$ according to Proposition 1. Given $\boldsymbol{z}_i \in S_k^*$ and $\boldsymbol{z}_j \in S_\ell^*$, recall that $p_{k\ell} = \mathbb{P}\left(|\langle \boldsymbol{z}_i, \boldsymbol{z}_j\rangle| \geq \tau\right)$ denotes the connection probability between any pair of data points that respectively belong to the subspaces S_k^* and S_ℓ^* for all $1 \leq k, \ell \leq K$. Let $\boldsymbol{B} := \{b_{k\ell}\}_{1\leq k,\ell\leq K}$, $\boldsymbol{C} := \boldsymbol{H}^*\boldsymbol{B}\boldsymbol{H}^{*^T}$, $\boldsymbol{P} := \{p_{k\ell}\}_{1\leq k,\ell\leq K}$, and $\boldsymbol{D} := \boldsymbol{H}^*\boldsymbol{P}\boldsymbol{H}^{*^T}$, where $b_{k\ell}$ is defined in (21). In addition, let $\hat{\boldsymbol{U}}, \boldsymbol{U} \in \mathbb{R}^{n\times K}$ be respectively the eigenvectors of $\boldsymbol{A}$ and $\boldsymbol{C}$ associated with the K leading eigenvalues. According to (11), one can verify that

$$\mathbb{E}[\boldsymbol{A}] = \boldsymbol{D} - \mathrm{diag}(\boldsymbol{D}). \tag{94}$$

We claim that $\boldsymbol{C}$ is of rank K and its smallest singular value is larger than $N_{\min}\gamma$, where $\gamma \geq 1 - \Phi(c)$ is given in Lemma 2. Indeed, let $\boldsymbol{\Lambda} = \mathrm{diag}\left(\sqrt{N_1}, \ldots, \sqrt{N_K}\right)$. Then, we have

$$\boldsymbol{C} = \boldsymbol{H}^*\boldsymbol{\Lambda}^{-1}\boldsymbol{\Lambda}\boldsymbol{B}\boldsymbol{\Lambda}\left(\boldsymbol{H}^*\boldsymbol{\Lambda}^{-1}\right)^T.$$

One can verify that $\boldsymbol{H}^*\boldsymbol{\Lambda}^{-1}$ has orthonormal columns and

$$\sigma_{\min}(\boldsymbol{C}) \geq \sigma_{\min}\left(\boldsymbol{\Lambda}\boldsymbol{B}\boldsymbol{\Lambda}\right) \geq \sigma_{\min}^2\left(\boldsymbol{\Lambda}\right)\sigma_{\min}(\boldsymbol{B}) = N_{\min}\gamma.$$

According to (94) and Lemma 1, we have

$$\begin{aligned}\|\mathbb{E}[\boldsymbol{A}] - \boldsymbol{C}\| &= \|\boldsymbol{D} - \boldsymbol{C} - \mathrm{diag}(\boldsymbol{D})\| \leq \|\boldsymbol{D} - \boldsymbol{C}\| + \max_{1\leq k\leq K} p_{kk} \\ &\leq \|\boldsymbol{H}^*\|^2\|\boldsymbol{B} - \boldsymbol{P}\| + 1 \leq \|\boldsymbol{H}^{*^T}\boldsymbol{H}^*\|\|\boldsymbol{B} - \boldsymbol{P}\|_F + 1 \lesssim \frac{\kappa_d N_{\max}}{\sqrt{\log N}}.\end{aligned}$$

This, together with (29), yields that

$$\|\boldsymbol{A} - \boldsymbol{C}\| \leq \|\boldsymbol{A} - \mathbb{E}[\boldsymbol{A}]\| + \|\mathbb{E}[\boldsymbol{A}] - \boldsymbol{C}\| \lesssim \frac{\sqrt{\kappa_d}N}{\sqrt[4]{\log N}},$$

where the last inequality is due to $\kappa_d \lesssim \sqrt{\log N}$. This, together with Lei et al. (2015, Lemma 5.1), $\gamma \geq 1 - \Phi(c)$, and the fact that κ_d is a constant, yields that there exists a $\boldsymbol{Q} \in \mathcal{O}^K$ such that

$$\|\hat{\boldsymbol{U}} - \boldsymbol{U}\boldsymbol{Q}\|_F \leq \frac{2\sqrt{2K}}{N_{\min}\gamma}\|\boldsymbol{A} - \boldsymbol{C}\| \lesssim \frac{\sqrt{\kappa_d}N}{N_{\min}\sqrt[4]{\log N}}. \tag{95}$$

According to Lei et al. (2015, Lemma 2.1), we have $\boldsymbol{U} = \boldsymbol{H}^*\boldsymbol{X}$ for some $\boldsymbol{X} \in \mathbb{R}^{K\times K}$ with $\|\boldsymbol{x}_k - \boldsymbol{x}_\ell\| = \sqrt{1/N_k + 1/N_\ell}$, where $\boldsymbol{x}_k$ denotes k-th row of $\boldsymbol{X}$. By letting $\boldsymbol{X}' = \boldsymbol{X}\boldsymbol{Q}$, we obtain

$$\boldsymbol{U}\boldsymbol{Q} = \boldsymbol{H}^*\boldsymbol{X}',$$

where $\|\boldsymbol{x}_k' - \boldsymbol{x}_\ell'\| = \sqrt{1/N_k + 1/N_\ell}$. This, together with setting $\delta_k = 1/\sqrt{N_k}$ in Lei et al. (2015, Lemma 5.3) and (95), yields that

$$\min_{k\in[K]} N_k\delta_k^2 = 1 \gtrsim \frac{\kappa_d N^2}{N_{\min}^2\sqrt{\log N}} \gtrsim \|\hat{\boldsymbol{U}} - \boldsymbol{U}\boldsymbol{Q}\|_F^2$$

where the first inequality is due to $\kappa_N^2 \kappa_d \leq \sqrt{\log N}$. This implies that there exists a $\boldsymbol{Q} \in \mathcal{O}^K$ such that

$$\frac{\|\hat{\boldsymbol{U}} - \boldsymbol{U}\boldsymbol{Q}\|_F^2}{\delta_k^2} \lesssim N_k, \text{ for all } k \in [K]. \tag{96}$$

Let $\bar{\boldsymbol{U}} = \boldsymbol{H}^0 \hat{\boldsymbol{X}}$, where $(\boldsymbol{H}^0, \hat{\boldsymbol{X}})$ is an $(1+\varepsilon)$-approximate solution to Problem (12). Moreover, we define $T_k = \{i \in \mathcal{C}_k^* : \bar{\boldsymbol{u}}_k - \bar{\boldsymbol{u}}_\ell \geq \delta_k/2\}$ for all $k \in [K]$, where $\bar{\boldsymbol{u}}_i$ denote the i-th row of $\bar{\boldsymbol{U}}$. According to (96) and Lei et al. (2015, Lemma 5.3), we have

$$\sum_{k=1}^K \frac{|T_k|}{N_k} \lesssim \|\hat{\boldsymbol{U}} - \boldsymbol{U}\boldsymbol{Q}\|_F^2 \lesssim \frac{\kappa_d N^2}{N_{\min}^2 \sqrt{\log N}} \lesssim \frac{\kappa_d \kappa_N^2}{\sqrt{\log N}},$$

where the second inequality is due to (95). This implies

$$\sum_{k=1}^K |T_k| \lesssim \kappa_d \kappa_N^2 \frac{N_{\max}}{\sqrt{\log N}}$$

Note that Lei et al. (2015, Lemma 5.3) ensures that the membership is correctly recovered outside of $\cup_{k\in[K]} T_k$, then we have

$$d_F^2(\boldsymbol{H}, \boldsymbol{H}^*) \lesssim \kappa_d \kappa_N^2 \frac{N_{\max}}{\sqrt{\log N}},$$

which implies (26). Then, we complete the proof. □

C. Proofs in Section 3.2

Recall that given an $\boldsymbol{H} \in \mathcal{M}^{N\times K}$, $\mathcal{C}_k = \{i \in [N] : h_{ik} = 1\}$, $n_k = |\mathcal{C}_k|$ for all $k \in [K]$, and $n_{k\ell} = |\mathcal{C}_k \cap \mathcal{C}_\ell^*|$ for all $k, \ell \in [K]$. We can verify that the number of misclassified points in $\{\mathcal{C}_1, \ldots, \mathcal{C}_K\}$ represented by $\boldsymbol{H}$ with respect to $\{\mathcal{C}_1^*, \ldots, \mathcal{C}_K^*\}$ represented by $\boldsymbol{H}^*$ is $\|\boldsymbol{H} - \boldsymbol{H}^*\boldsymbol{Q}_{\pi^*}\|_F^2/2$, where $\boldsymbol{Q}_{\pi^*} \in \arg\min_{\boldsymbol{Q}\in\Pi_K} \|\boldsymbol{H} - \boldsymbol{H}^*\boldsymbol{Q}\|_F$. Moreover, we can verify that for a permutation $\pi : [K] \to [K]$,

$$\frac{1}{2}\|\boldsymbol{H} - \boldsymbol{H}^*\boldsymbol{Q}_\pi\|_F^2 = \sum_{k=1}^K \sum_{\ell \neq \pi^{-1}(k)} n_{k\ell} = \sum_{k=1}^K \sum_{\ell \neq \pi(k)} n_{\ell k}. \tag{97}$$

and

$$\sum_{\ell:\ell\neq\pi^{-1}(k)} n_{k\ell} = \frac{1}{2} \sum_{\ell:\ell\neq\pi^{-1}(k)} \sum_{i\in\mathcal{C}_k\cap\mathcal{C}_\ell^*} \|\boldsymbol{h}_i - \boldsymbol{h}_i^*\boldsymbol{Q}_\pi\|^2, \; W_k(\boldsymbol{H}) = \max\left\{ \sum_{\ell:\ell\neq\pi^{-1}(k)} n_{k\ell}, \sum_{\ell:\ell\neq\pi(k)} n_{\ell k} \right\}, \tag{98}$$

where $W_k(\boldsymbol{H})$ is defined in (31). Using Lemma 9, we can present a spectral bound on the deviation of the sample covariance of random vectors that follow a uniform distribution over the sphere from its mean.

Corollary 1. *Suppose that $d_k \geq 4\log\left(Kd_k^2N\right)$ for all $k \in [K]$. For all $k, \ell \in [K]$, it holds with probability at least $1 - 2K/(Nd_{\min}^2)$ that*

$$\left\| \boldsymbol{\Psi}_{k\ell} - \frac{1}{d_\ell}\boldsymbol{I}_{d_\ell} \right\| \leq \frac{5c_1}{4d_\ell} \left(\sqrt{\frac{d_\ell}{n_{k\ell}}} + \frac{d_\ell}{n_{k\ell}} \right), \tag{99}$$

where $\boldsymbol{\Psi}_{k\ell}$ is defined in (30) and $c_1 > 0$ is an absolute constant.

Proof of Corollary 1. Since $i \in \mathcal{C}_\ell^*$, we have $\boldsymbol{a}_i \in \mathrm{Unif}(\mathbb{S}^{d_\ell - 1})$ according to the UoS model in Definition 2. Applying Lemma 9 to (30) with $u = \log\left(Kd_\ell^2N\right)$ yields that it holds with probability at least $1 - 2/(Kd_\ell^2N)$ that

$$\left\| \boldsymbol{\Psi}_{k\ell} - \frac{1}{d_\ell}\boldsymbol{I}_{d_\ell} \right\| \leq \frac{c_1}{d_\ell} \left(\sqrt{\frac{d_\ell + \log(Kd_\ell^2N)}{n_{k\ell}}} + \frac{d_\ell + \log(Kd_\ell^2N)}{n_{k\ell}} \right) \leq \frac{5c_1}{4d_\ell} \left(\sqrt{\frac{d_\ell}{n_{k\ell}}} + \frac{d_\ell}{n_{k\ell}} \right),$$

where the second inequality is due to $d_\ell \geq 4\log(Kd_\ell^2N)$. This, together with the union bound, implies the desired result. □

C.1. Proof of Lemma 3

Proof of Lemma 3. Suppose that (99) holds for all $k, \ell \in [K]$, which happens with probability at least $1 - 2K/(Nd_{\min}^2)$ according to $N_k \gtrsim d_k \gtrsim \log N$ and Corollary 1. According to (30) and (32), we have for all $k \in [K]$,

$$n_{\pi(k)k} = |\mathcal{C}_{\pi(k)} \cap \mathcal{C}_k^*| = |\mathcal{C}_k^*| - \sum_{\ell:\ell\neq\pi(k)} |\mathcal{C}_\ell \cap \mathcal{C}_k^*| = N_k - \sum_{\ell:\ell\neq\pi(k)} n_{\ell k} \geq N_k - W_k(\boldsymbol{H}) \geq \frac{7}{8}N_k.$$

This, together with (99), yields that for all $k \in [K]$,

$$\left\|\boldsymbol{\Psi}_{\pi(k)k} - \frac{1}{d_k}\boldsymbol{I}_{d_k}\right\| \leq \frac{5c_1}{4d_k}\left(\sqrt{\frac{8d_k}{7N_k}} + \frac{8d_k}{7N_k}\right) \leq \frac{5c_1}{2d_k}\sqrt{\frac{8d_k}{7N_k}} \leq \frac{1}{32d_k},$$

where the third and last inequalities are due to $N_k \gtrsim d_k$ for all $k \in [K]$. This, together with Weyl's inequality, yields (33). Again, applying (99) to $\boldsymbol{\Psi}_{\pi(k)\ell}$ for all $\ell \neq k$ yields

$$\left\|\boldsymbol{\Psi}_{\pi(k)\ell} - \frac{1}{d_\ell}\boldsymbol{I}_{d_\ell}\right\| \leq \frac{5c_1}{4d_\ell}\left(\sqrt{\frac{d_\ell}{n_{\pi(k)\ell}}} + \frac{d_\ell}{n_{\pi(k)\ell}}\right).$$

This, together with Weyl's inequality, yields (34). Then, the proof is completed. □

C.2. Proof of Lemma 4

Proof of Lemma 4. Suppose that (33) and (34) hold, which happens with probability at least $1 - 2K/(d_{\min}^2 N)$ according to Lemma 3. Recall that

$$\boldsymbol{G}_{\boldsymbol{U}_k}(\boldsymbol{H}) = \sum_{i=1}^N h_{ik}\boldsymbol{z}_i\boldsymbol{z}_i^T \text{ for all } k \in [K]. \tag{100}$$

It follows from $\boldsymbol{H} \in \mathcal{M}^{N\times K}$ and $\mathcal{C}_k = \{i \in [N] : h_{ik} = 1\}$ that $h_{ik} = 1$ if $i \in \mathcal{C}_k$ and $h_{ik} = 0$ otherwise for all $i \in [N]$. Then, we note that

$$\boldsymbol{G}_{\boldsymbol{U}_{\pi(k)}}(\boldsymbol{H}) = \sum_{i\in\mathcal{C}_{\pi(k)}} \boldsymbol{z}_i\boldsymbol{z}_i^T = \sum_{\ell=1}^K \sum_{i\in\mathcal{C}_{\pi(k)}\cap\mathcal{C}_\ell^*} \boldsymbol{z}_i\boldsymbol{z}_i^T = \sum_{\ell=1}^K \sum_{i\in\mathcal{C}_{\pi(k)}\cap\mathcal{C}_\ell^*} \boldsymbol{U}_\ell^*\boldsymbol{a}_i\boldsymbol{a}_i^T\boldsymbol{U}_\ell^{*^T} = \sum_{\ell=1}^K n_{\pi(k)\ell}\boldsymbol{U}_\ell^*\boldsymbol{\Psi}_{\pi(k)\ell}\boldsymbol{U}_\ell^{*^T},$$

where the third equality is due to (2) in Definition 2 and the last equality follows from (30). To simplify the notations, we define

$$\boldsymbol{A}_\ell = \boldsymbol{U}_\ell^*\boldsymbol{\Psi}_{\pi(k)\ell}\boldsymbol{U}_\ell^{*^T} \text{ for all } \ell \in [K], \quad \delta_i = \sigma_i\left(\boldsymbol{G}_{\boldsymbol{U}_{\pi(k)}}(\boldsymbol{H})\right) - \sigma_{i+1}\left(\boldsymbol{G}_{\boldsymbol{U}_{\pi(k)}}(\boldsymbol{H})\right) \text{ for all } i \in [d-1].$$

On one hand, it follows from (33), $\sigma_{d_k}\left(\boldsymbol{\Psi}_{\pi(k)k}\right) \leq \sigma_{d_k}(\boldsymbol{A}_k)$, and $\sigma_1(\boldsymbol{A}_k) \leq \sigma_1\left(\boldsymbol{\Psi}_{\pi(k)k}\right)$ that

$$\frac{31}{32d_k} \leq \sigma_{d_k}(\boldsymbol{A}_k) \leq \cdots \leq \sigma_1(\boldsymbol{A}_k) \leq \frac{33}{32d_k}. \tag{101}$$

On the other hand, it follows from $\boldsymbol{U}_k^* \in \mathcal{O}^{n\times d_k}$ that

$$\sigma_{d_k+1}(\boldsymbol{A}_k) = \cdots = \sigma_n(\boldsymbol{A}_k) = 0. \tag{102}$$

According to (30), (31), and (37), we have for all $k \in [K]$,

$$n_{\pi(k)k} = |\mathcal{C}_k^*| - \sum_{\ell:\ell\neq\pi(k)} |\mathcal{C}_\ell \cap \mathcal{C}_k^*| \geq N_k - W_k(\boldsymbol{H}) \geq \frac{7}{8}N_k, \sum_{\ell:\ell\neq k} n_{\pi(k)\ell} \leq W_{\pi(k)}(\boldsymbol{H}) \leq \varepsilon N_{\min}. \tag{103}$$

This, together with (34), yields that for all $k \in [K]$ and $\ell \neq k$,

$$n_{\pi(k)\ell}\sigma_1\left(\boldsymbol{\Psi}_{\pi(k)\ell}\right) \leq \frac{n_{\pi(k)\ell}}{d_\ell} + \frac{5c_1}{4}\sqrt{\frac{n_{\pi(k)\ell}}{d_\ell}} + \frac{5c_1}{4} \leq \frac{21n_{\pi(k)\ell}}{16d_\ell} + \frac{5c_1}{4}(c_1+1), \tag{104}$$

where the second inequality is due to $2\sqrt{\alpha\beta} \le \alpha/\rho + \rho\beta$ for any $\alpha, \beta \ge 0$ and $\rho > 0$. Summing up (104) for all $\ell \neq k$ yields that for all $k \in [K]$,

$$\sum_{\ell:\ell\neq k} n_{\pi(k)\ell}\sigma_1\left(\boldsymbol{\Psi}_{\pi(k)\ell}\right) \le \frac{21}{16}\sum_{\ell:\ell\neq k}\frac{n_{\pi(k)\ell}}{d_\ell} + \frac{5c_1}{4}K(c_1+1) \le \frac{21\varepsilon N_{\min}}{16d_{\min}} + \frac{3\varepsilon N_k}{16d_{\min}} \le \frac{3\varepsilon N_k}{2d_{\min}}, \tag{105}$$

where the second inequality is due to (103) and $N_k \gtrsim d_k$. We now show (38). According to $\boldsymbol{G}_{\boldsymbol{U}_{\pi(k)}}(\boldsymbol{H}) = n_{\pi(k)k}\boldsymbol{A}_k + \sum_{\ell\neq k} n_{\pi(k)\ell}\boldsymbol{A}_\ell$, we have for all $i \in [d-1]$,

$$\begin{aligned}
\delta_i &= \sigma_i\left(\boldsymbol{G}_{\boldsymbol{U}_{\pi(k)}}(\boldsymbol{H})\right) - \sigma_{i+1}\left(n_{\pi(k)k}\boldsymbol{A}_k\right) + \sigma_{i+1}\left(n_{\pi(k)k}\boldsymbol{A}_k\right) - \sigma_{i+1}\left(\boldsymbol{G}_{\boldsymbol{U}_{\pi(k)}}(\boldsymbol{H})\right)\\
&\le \sigma_i\left(n_{\pi(k)k}\boldsymbol{A}_k\right) - \sigma_{i+1}\left(n_{\pi(k)k}\boldsymbol{A}_k\right) + 2\sigma_1\left(\sum_{\ell:\ell\neq k} n_{\pi(k)\ell}\boldsymbol{A}_\ell\right),\\
&\le n_{\pi(k)k}\left(\sigma_i\left(\boldsymbol{A}_k\right) - \sigma_{i+1}\left(\boldsymbol{A}_k\right)\right) + 2\sum_{\ell:\ell\neq k} n_{\pi(k)\ell}\sigma_1\left(\boldsymbol{\Psi}_{\pi(k)\ell}\right),
\end{aligned} \tag{106}$$

where the first inequality is due to Weyl's inequality. Plugging (101), $n_{\pi(k)k} \le N_k$, and (105) into (106) yields that for all $i = 1, \ldots, d_k - 1$,

$$\delta_i \le \frac{N_k}{16d_k} + \frac{3\varepsilon N_k}{d_{\min}} \le \frac{7N_k}{16d_k}, \tag{107}$$

where the second inequality is due to $\varepsilon \le d_{\min}/(8d_k)$. Meanwhile, plugging (102) and (105) into (106) yields that for all $i = d_k + 1, \ldots, d$,

$$\delta_i \le \frac{3\varepsilon N_k}{d_{\min}} \le \frac{3N_k}{8d_k}, \tag{108}$$

where the second inequality is due to $\varepsilon \le d_{\min}/(8d_k)$. Note that $\boldsymbol{G}_{\boldsymbol{U}_{\pi(k)}}$ is a positive semidefinite matrix and satisfies

$$\begin{aligned}
\sigma_{d_k}\left(\boldsymbol{G}_{\boldsymbol{U}_{\pi(k)}}(\boldsymbol{H})\right) &\ge \sigma_{d_k}\left(\boldsymbol{U}_k^* n_{\pi(k)k}\boldsymbol{\Psi}_{\pi(k)k}\boldsymbol{U}_k^{*^T}\right) - \sigma_1\left(\sum_{\ell:\ell\neq k}\boldsymbol{U}_\ell^* n_{\pi(k)\ell}\boldsymbol{\Psi}_{\pi(k)\ell}\boldsymbol{U}_\ell^{*^T}\right)\\
&\ge n_{\pi(k)k}\sigma_{d_k}\left(\boldsymbol{\Psi}_{\pi(k)k}\right) - \sum_{\ell:\ell\neq k} n_{\pi(k)\ell}\sigma_1\left(\boldsymbol{\Psi}_{\pi(k)\ell}\right),
\end{aligned} \tag{109}$$

where the first inequality is due to Weyl's inequality and the second inequality follows from $\sigma_d(\boldsymbol{A}\boldsymbol{U}^T) \ge \sigma_d(\boldsymbol{A})$, $\sigma_1(\boldsymbol{A}\boldsymbol{U}^T) \le \sigma_1(\boldsymbol{A})$ for $\boldsymbol{A} \in \mathbb{R}^{d\times d}$ and $\boldsymbol{U} \in \mathcal{O}^{n\times d}$, and $\sigma_1(\boldsymbol{B}+\boldsymbol{C}) \le \sigma_1(\boldsymbol{B}) + \sigma_1(\boldsymbol{C})$ for $\boldsymbol{B}, \boldsymbol{C} \in \mathbb{R}^{m\times n}$. Plugging (33), (103), and (105) into (109) yields for all $k \in [K]$,

$$\sigma_{d_k}\left(\boldsymbol{G}_{\boldsymbol{U}_{\pi(k)}}(\boldsymbol{H})\right) \ge \frac{217N_k}{256d_k} - \frac{3\varepsilon N_k}{2d_{\min}} \ge \frac{217N_k}{256d_k} - \frac{3N_k}{16d_k} \ge \frac{169N_k}{256d_k}, \tag{110}$$

where the second inequality is due to $\varepsilon \le d_{\min}/(8d_k)$. Applying Weyl's inequality to $\boldsymbol{G}_{\boldsymbol{U}_{\pi(k)}}(\boldsymbol{H})$ gives

$$\sigma_{d_k+1}\left(\boldsymbol{G}_{\boldsymbol{U}_{\pi(k)}}(\boldsymbol{H})\right) \le n_{\pi(k)k}\sigma_{d_k+1}\left(\boldsymbol{A}_k\right) + \sigma_1\left(\sum_{\ell:\ell\neq k} n_{\pi(k)\ell}\boldsymbol{A}_\ell\right) \le \sum_{\ell:\ell\neq k} n_{\pi(k)\ell}\sigma_1\left(\boldsymbol{\Psi}_{\pi(k)\ell}\right) \le \frac{3N_k}{16d_k},$$

where the second inequality is due to (102) and the last inequality follows from (105) and $\varepsilon \le d_{\min}/(8d_k)$. This, together with (110), yields

$$\delta_{d_k} \ge \frac{169N_k}{256d_k} - \frac{3N_k}{16d_k} = \frac{121N_k}{256d_k}.$$

This, together with (35), $\lambda_{\pi(k)i} = \sigma_i\left(\boldsymbol{G}_{\boldsymbol{U}_{\pi(k)}}(\boldsymbol{H})\right)$ for all $i \in [d]$ due to the positive semidefiniteness of $\boldsymbol{G}_{\boldsymbol{U}_{\pi(k)}}(\boldsymbol{H})$, (107), and (108), implies (38).

We next show (39). Note that

$$\left\|(\boldsymbol{I}-\boldsymbol{U}_k^*\boldsymbol{U}_k^{*^T})\boldsymbol{G}_{\boldsymbol{U}_{\pi(k)}}(\boldsymbol{H})\right\| = \left\|\sum_{\ell:\ell\neq k}(\boldsymbol{I}-\boldsymbol{U}_k^*\boldsymbol{U}_k^{*^T})\boldsymbol{U}_\ell^* n_{\pi(k)\ell}\boldsymbol{\Psi}_{\pi(k)\ell}\boldsymbol{U}_\ell^{*^T}\right\|$$
$$\leq \sum_{\ell:\ell\neq k} n_{\pi(k)\ell}\left\|(\boldsymbol{I}-\boldsymbol{U}_k^*\boldsymbol{U}_k^{*^T})\boldsymbol{U}_\ell^*\right\|\|\boldsymbol{\Psi}_{\pi(k)\ell}\| \leq \sum_{\ell:\ell\neq k} n_{\pi(k)\ell}\sigma_1\left(\boldsymbol{\Psi}_{\pi(k)\ell}\right), \tag{111}$$

where the last inequality is due to $\left\|(\boldsymbol{I}-\boldsymbol{U}_k^*\boldsymbol{U}_k^{*^T})\boldsymbol{U}_\ell^*\right\| \leq 1$ for any $1\leq k\neq\ell\leq K$. Since $\boldsymbol{U}_k$ consists of eigenvectors associated with the top $\bar{d}_k$ eigenvalues of $\boldsymbol{G}_{\boldsymbol{U}_k}(\boldsymbol{H})$ that is positive semidefinite, then we have $\boldsymbol{G}_{\boldsymbol{U}_k}(\boldsymbol{H})\boldsymbol{U}_k = \boldsymbol{U}_k\boldsymbol{\Sigma}_k$ by the eigenvalue equation, where $\boldsymbol{\Sigma}_k$ is a diagonal matrix with the i-th diagonal element being $\sigma_i\left(\boldsymbol{G}_{\boldsymbol{U}_k}(\boldsymbol{H})\right)$ for all $i\in[\bar{d}_k]$. This, together with (110), implies $\boldsymbol{U}_k = \boldsymbol{G}_{\boldsymbol{U}_k}(\boldsymbol{H})\boldsymbol{U}_k\boldsymbol{\Sigma}_k^{-1}$ for all $k\in[K]$. According to (38), we have

$$d(\boldsymbol{U}_{\pi(k)},\boldsymbol{U}_k^*) = \left\|(\boldsymbol{I}-\boldsymbol{U}_k^*\boldsymbol{U}_k^{*^T})\boldsymbol{U}_{\pi(k)}\right\| = \left\|(\boldsymbol{I}-\boldsymbol{U}_k^*\boldsymbol{U}_k^{*^T})\boldsymbol{G}_{\boldsymbol{U}_{\pi(k)}}(\boldsymbol{H})\boldsymbol{U}_{\pi(k)}\boldsymbol{\Sigma}_{\pi(k)}^{-1}\right\|$$
$$\leq \left\|(\boldsymbol{I}-\boldsymbol{U}_k^*\boldsymbol{U}_k^{*^T})\boldsymbol{G}_{\boldsymbol{U}_{\pi(k)}}(\boldsymbol{H})\right\|\|\boldsymbol{\Sigma}_{\pi(k)}^{-1}\| \leq \frac{256d_k\sum_{\ell:\ell\neq k} n_{\pi(k)\ell}\sigma_1(\boldsymbol{\Psi}_{\pi(k)\ell})}{169N_k}, \tag{112}$$

where the first equality uses (48) and the last inequality is due to (110) and (111). Substituting (104) into the above inequality and summing up from $k=1$ to $k=K$ yield that for all $k\in[K]$,

$$\sum_{k=1}^K d(\boldsymbol{U}_{\pi(k)},\boldsymbol{U}_k^*) \leq \sum_{k=1}^K \frac{256d_k}{169N_k}\cdot\frac{21}{16}\left(K(c_1+1)c_1 + \frac{1}{d_{\min}}\sum_{\ell:\ell\neq k} n_{\pi(k)\ell}\right)$$
$$\leq \frac{2d_{\max}}{N_{\min}}\left(K^2(c_1+1)c_1 + \frac{1}{d_{\min}}\sum_{k=1}^K\sum_{\ell:\ell\neq k} n_{\pi(k)\ell}\right)$$
$$\leq \frac{2d_{\max}}{N_{\min}}\left(\frac{1}{2d_{\min}}\|\boldsymbol{H}-\boldsymbol{H}^*\boldsymbol{Q}_\pi\|_F^2 + K^2(c_1+1)c_1\right)$$
$$\leq \frac{2d_{\max}}{N_{\min}}\max\left\{\frac{1}{d_{\min}}\|\boldsymbol{H}-\boldsymbol{H}^*\boldsymbol{Q}_\pi\|_F^2, 2K^2(c_1+1)c_1\right\},$$

where the third inequality is due to (97) and $\sum_{k=1}^K\sum_{\ell\neq\pi^{-1}(k)} n_{k\ell} = \sum_{k=1}^K\sum_{\ell\neq k} n_{\pi(k)\ell}$. Then, we complete the proof. □

D. Proofs in Section 3.3

D.1. Proof of Lemma 5

Proof of Lemma 5. Note that $\boldsymbol{h}$ satisfying $\boldsymbol{h}^T\boldsymbol{1}_K = 1,\ \boldsymbol{h}\in\{0,1\}^K$ is a vector that has exactly one 1 and $(K-1)$ 0's. We can see that $\mathcal{T}(\boldsymbol{g})$ is to find the minimum element of $\boldsymbol{g}$. Then, the solution follows immediately. This also implies that for $\boldsymbol{Q}\in\Pi_K$, $\boldsymbol{v}\in\mathcal{T}(\boldsymbol{g})$ if and only if $\boldsymbol{Q}\boldsymbol{v}\in\mathcal{T}(\boldsymbol{Q}\boldsymbol{g})$. □

D.2. Proof of Lemma 6

Proof of Lemma 6. According to Lemma 5 and $g_\ell > g_k$ for a $k\in[K]$ and all $\ell\neq k$, we see that $\mathcal{T}(\boldsymbol{g})$ is a singleton and $\{\boldsymbol{v}\} = \mathcal{T}(\boldsymbol{g})$ satisfies $v_k = 1$ and $v_\ell = 0$ for all $\ell\neq k$. Let $\boldsymbol{g}'\in\mathbb{R}^K$ be arbitrary and $\boldsymbol{v}'\in\mathcal{T}(\boldsymbol{g}')$. It then follows from Lemma 5 that $v'_{k'} = 1$ and $v_{\ell'} = 0$ for some $k'\in[K]$ and all $\ell'\neq k'$ satisfying $g'_{k'}\leq g'_{\ell'}$. Suppose that $k' = k$. Then, we have $\|\boldsymbol{v}-\boldsymbol{v}'\| = 0$, and thus (40) holds trivially. Suppose to the contrary that $k'\neq k$. Then, we have $\|\boldsymbol{v}-\boldsymbol{v}'\| = \sqrt{2}$. Moreover, we can compute

$$\|\boldsymbol{g}-\boldsymbol{g}'\|^2 \geq (g_k-g'_k)^2 + (g_{k'}-g'_{k'})^2 \geq \frac{1}{2}(\underbrace{g_k-g_{k'}}_{\leq-\delta} + \underbrace{g'_{k'}-g'_k}_{\leq 0})^2 \geq \frac{1}{2}\delta^2.$$

This, together with $\|\boldsymbol{v}-\boldsymbol{v}'\| = \sqrt{2}$, implies the desired result (40). □

D.3. Proof of Lemma 7

Proof of Lemma 7. Let the row vectors $\boldsymbol{g}_i,\ \boldsymbol{g}_i^* \in \mathbb{R}^K$ denote the i-th row of $\boldsymbol{G}_{\boldsymbol{H}}(\boldsymbol{U})$ and $\boldsymbol{G}_{\boldsymbol{H}}(\boldsymbol{U}^*)$, respectively. For all $i \in [N]$, note that $I_i = \{k \in [K] : h_{ik}^* = 1\}$. According to the semi-random UoS model in Definition 2, we have for all $i \in [N]$ and $\ell \neq I_i$,

$$\begin{aligned} g_{i\ell}^* - g_{iI_i}^* &= \left(\|\boldsymbol{z}_i\|^2 - \|\boldsymbol{U}_\ell^{*^T}\boldsymbol{z}_i\|^2\right) - \left(\|\boldsymbol{z}_i\|^2 - \|\boldsymbol{U}_{I_i}^{*^T}\boldsymbol{z}_i\|^2\right) = \|\boldsymbol{U}_{I_i}^{*^T}\boldsymbol{U}_{I_i}^*\boldsymbol{a}_i\|^2 - \|\boldsymbol{U}_\ell^{*^T}\boldsymbol{z}_i\|^2 \\ &= 1 - \|\boldsymbol{U}_\ell^{*^T}\boldsymbol{z}_i\|^2 \geq 1 - \max_{\ell \neq I_i}\|\boldsymbol{U}_\ell^{*^T}\boldsymbol{z}_i\|^2, \end{aligned}$$

where the last equality is due to $\|\boldsymbol{a}_i\| = 1$. This, together with Lemma 5 and $\|\boldsymbol{U}_\ell^{*^T}\boldsymbol{z}_i\| < 1$ for all $\ell \neq I_i$, implies

$$\{\boldsymbol{H}^*\} = \mathcal{T}\left(\boldsymbol{G}_{\boldsymbol{H}}(\boldsymbol{U}^*)\right). \tag{113}$$

Besides, we note that for each $i \in [N]$ and $\boldsymbol{Q}_\pi \in \Pi_K$,

$$\|\boldsymbol{g}_i\boldsymbol{Q}_\pi^T - \boldsymbol{g}_i^*\|^2 = \sum_{k=1}^K \left(\|\boldsymbol{U}_{\pi(k)}^T\boldsymbol{z}_i\|^2 - \|\boldsymbol{U}_k^{*^T}\boldsymbol{z}_i\|^2\right)^2 \leq \sum_{k=1}^K \|\boldsymbol{U}_{\pi(k)}\boldsymbol{U}_{\pi(k)}^T - \boldsymbol{U}_k^*\boldsymbol{U}_k^{*^T}\|^2\|\boldsymbol{z}_i\|^4 = \sum_{k=1}^K d^2(\boldsymbol{U}_{\pi(k)}, \boldsymbol{U}_k^*),$$

where the first equality is due to $\boldsymbol{g}\boldsymbol{Q}_\pi^T = \begin{bmatrix} g_{\pi(1)} & \cdots & g_{\pi(K)} \end{bmatrix}$. This, together with Lemma 6 and (113), implies for all $i \in [N]$,

$$\|\bar{\boldsymbol{h}}_i - \boldsymbol{h}_i^*\boldsymbol{Q}_\pi\| = \|\bar{\boldsymbol{h}}_i\boldsymbol{Q}_\pi^T - \boldsymbol{h}_i^*\| \leq \frac{2\|\boldsymbol{g}_i\boldsymbol{Q}_\pi^T - \boldsymbol{g}_i^*\|}{1 - \max_{\ell \neq I_i}\|\boldsymbol{U}_\ell^{*^T}\boldsymbol{z}_i\|^2} \leq \frac{2\sqrt{\sum_{k=1}^K d^2(\boldsymbol{U}_{\pi(k)}, \boldsymbol{U}_k^*)}}{1 - \max_{\ell \neq I_i}\|\boldsymbol{U}_\ell^{*^T}\boldsymbol{z}_i\|^2},$$

where the first inequality uses the fact that $\boldsymbol{H}\boldsymbol{Q}^T \in \mathcal{T}(\boldsymbol{G}_{\boldsymbol{H}}(\boldsymbol{U})\boldsymbol{Q}^T)$ for $\boldsymbol{Q} \in \Pi_K$ if and only if $\boldsymbol{H} \in \mathcal{T}(\boldsymbol{G}_{\boldsymbol{H}}(\boldsymbol{U}))$ due to Lemma 5. ☐

With the preparations in Sections 3.2 and 3.3, we can analyze each iteration of the KSS method as follows.

Proposition 2. *Let $\varepsilon \in \left(0, \frac{d_{\min}}{8d_{\max}}\right]$ be a constant. Suppose that Assumption 1 holds, $N_{\min} \gtrsim d_k \gtrsim \log N$ for all $k \in [K]$, and $\boldsymbol{H}^t \in \mathcal{M}^{N\times K}$ satisfies*

$$\|\boldsymbol{H}^t - \boldsymbol{H}^*\boldsymbol{Q}_\pi\|_F^2 \leq 2\varepsilon N_{\min}, \tag{114}$$

where $\boldsymbol{Q}_\pi \in \arg\min_{\boldsymbol{Q}\in\Pi_K}\|\boldsymbol{H}^t - \boldsymbol{H}^*\boldsymbol{Q}\|_F$. Then, it holds with probability at least $1 - 2K/(d_{\min}^2 N) - 5K^2/N$ that $\hat{d}_{\pi(k)} = d_k$ for all $k \in [K]$,

$$\sum_{k=1}^K d(\boldsymbol{U}_{\pi(k)}^{t+1}, \boldsymbol{U}_k^*) \leq \frac{2d_{\max}}{N_{\min}} \max\left\{\frac{1}{d_{\min}}\|\boldsymbol{H}^t - \boldsymbol{H}^*\boldsymbol{Q}_\pi\|_F^2, 2K^2(c_1+1)c_1\right\}, \tag{115}$$

and for all $i \in [N]$,

$$\|\boldsymbol{h}_i^{t+1} - \boldsymbol{h}_i^*\boldsymbol{Q}_\pi\| \leq \frac{2}{1-\kappa}\sqrt{\sum_{k=1}^K d^2(\boldsymbol{U}_{\pi(k)}^{t+1}, \boldsymbol{U}_k^*)}, \tag{116}$$

where the row vectors $\boldsymbol{h}_i^{t+1}, \boldsymbol{h}_i^* \in \mathbb{R}^K$ respectively denote the i-th row of $\boldsymbol{H}^{t+1}$ and $\boldsymbol{H}^*$.

Proof of Proposition 2. Suppose that (33), (34), and (56) hold, which happens with probability at least $1 - 5K^2/N - 2K/(d_{\min}^2 N)$ according to Lemma 3, Lemma 12,and the union bound. According to (97), we have $W_k(\boldsymbol{H}^t) \leq \|\boldsymbol{H}^t - \boldsymbol{H}^*\boldsymbol{Q}_\pi\|_F^2/2 \leq \varepsilon N_{\min}$. It follows from this and Lemma 4 that $\hat{d}_{\pi(k)}^{t+1} = d_k$ for all $k \in [K]$ and (115). Next, note that $I_i = \{k \in [K] : h_{ik}^* = 1\}$ for all $i \in [N]$. This, together with (15) in Assumption 1 and (56), implies that for all $i \in [N]$ and $\ell \neq I_i$,

$$\|\boldsymbol{U}_\ell^{*^T}\boldsymbol{z}_i\|^2 \leq \left(\frac{\mathrm{aff}(S_{I_i}, S_\ell^*) + \alpha}{\sqrt{d_\ell} - \alpha}\right)^2 \leq \left(\frac{\kappa\sqrt{d_\ell} + \alpha}{\sqrt{d_\ell} - \alpha}\right)^2 = \left(\kappa + \frac{(1-\kappa)\alpha}{\sqrt{d_\ell} - \alpha}\right)^2 \leq 2\kappa^2 \leq \kappa, \tag{117}$$

where the third inequality is due to $d_k \gtrsim \log N$ for all $k \in [K]$ and the last inequality is due to $\kappa \leq 1/2$. Using this and Lemma 7 yields (116). ☐

D.4. Proof of Lemma 8

Proof of Lemma 8. Suppose that (115) and (116) hold, which happens with probability at least $1-2K/(d_{\min}^2 N)-5K^2/N$ according to (42), $N_{\min} \gtrsim d_k$ for all $k \in [K]$, and Proposition 2. According to (42) and (115), we obtain

$$\sum_{k=1}^{K} d(\boldsymbol{U}_{\pi(k)}^{t+1}, \boldsymbol{U}_k^*) \leq \frac{4d_{\max}}{N_{\min}} K^2(c_1+1)c_1.$$

This, together with (116), $\kappa \leq 1/2$, and $N_{\min} \gtrsim d_{\max}$, yields that for all $i \in [N]$,

$$\|\boldsymbol{h}_i^{t+1} - \boldsymbol{h}_i^* \boldsymbol{Q}_\pi\| \leq 4\sum_{k=1}^{K} d(\boldsymbol{U}_{\pi(k)}^{t+1}, \boldsymbol{U}_k^*) \leq \frac{16d_{\max}}{N_{\min}} K^2(c_1+1)c_1 < 1.$$

Since $\boldsymbol{h}_i^{t+1}, \boldsymbol{h}_i \in \{0,1\}^K$ for all $i \in [N]$, then we have $\boldsymbol{h}_i^{t+1} = \boldsymbol{h}_i^* \boldsymbol{Q}_\pi$ for all $i \in [N]$. Thus, $\boldsymbol{H}^{t+1} = \boldsymbol{H}^* \boldsymbol{Q}_\pi$. Moreover, due to the fact that $\min_{\boldsymbol{Q}\in\Pi_K} \|\boldsymbol{H}^{t+1} - \boldsymbol{H}^*\boldsymbol{Q}\|_F \leq \|\boldsymbol{H}^{t+1} - \boldsymbol{H}^*\boldsymbol{Q}_\pi\|_F = 0$, the desired result is implied. □

D.5. Proof of Theorem 1

We should point out that a technical issue occurred in our analysis is that we cannot infinitely use the result in Proposition 2 infinitely due to the union bound. Then, we study $T = \Theta(\log\log N)$ iterates.

Proof of Theorem 1. For ease of exposition, let $\pi^t : [K] \to [K]$ be a permutation such that

$$\boldsymbol{Q}_{\pi^t} \in \arg\min_{\boldsymbol{Q}\in\Pi_K} \|\boldsymbol{H}^t - \boldsymbol{H}^*\boldsymbol{Q}\|_F \tag{118}$$

and the row vector $\boldsymbol{h}_i \in \mathbb{R}^K$ denote the i-th row of $\boldsymbol{H} \in \mathcal{M}^{N\times K}$ for all $i \in [N]$. We first show (i). Suppose that $t \leq T$ is a positive integer such that

$$\|\boldsymbol{H}^t - \boldsymbol{H}^*\boldsymbol{Q}_{\pi^t}\|_F^2 \leq 2K^2(c_1+1)c_1 d_{\min}.$$

Using Lemma 8, it holds with probability at least $1-2K/(d_{\min}^2 N)-5K^2/N$ that

$$\boldsymbol{H}^{t+1} = \boldsymbol{H}^*\boldsymbol{Q}_{\pi^{t+1}}.$$

Then, it suffices to consider that for all $t \leq T$ such that

$$\|\boldsymbol{H}^t - \boldsymbol{H}^*\boldsymbol{Q}_{\pi^t}\|_F^2 > 2K^2(c_1+1)c_1 d_{\min}. \tag{119}$$

We first consider $t = 0$. According to (17), Proposition 2, and (119), it holds with probability at least $1-2K/(d_{\min}^2 N)-5K^2/N$ that $\hat{d}_{\pi^0(k)}^1 = d_k$ for all $k \in [K]$,

$$\sum_{k=1}^{K} d(\boldsymbol{U}_{\pi^0(k)}^1, \boldsymbol{U}_k^*) \leq \frac{2d_{\max}}{N_{\min} d_{\min}} \|\boldsymbol{H}^0 - \boldsymbol{H}^*\boldsymbol{Q}_{\pi^0}\|_F^2 \tag{120}$$

and for all $i \in [N]$,

$$\|\boldsymbol{h}_i^1 - \boldsymbol{h}_i^*\boldsymbol{Q}_{\pi^0}\| \leq \frac{2}{1-\kappa}\sqrt{\sum_{k=1}^{K} d^2(\boldsymbol{U}_{\pi^0(k)}^1, \boldsymbol{U}_k^*)}. \tag{121}$$

Summing up (121) from $i = 1$ to $i = N$ gives

$$\|\boldsymbol{H}^1 - \boldsymbol{H}^*\boldsymbol{Q}_{\pi^0}\|_F \leq \frac{2\sqrt{N}}{1-\kappa}\sqrt{\sum_{k=1}^{K} d^2(\boldsymbol{U}_{\pi^0(k)}^1, \boldsymbol{U}_k^*)} \leq \frac{2\sqrt{N}}{1-\kappa}\sum_{k=1}^{K} d(\boldsymbol{U}_{\pi^0(k)}^1, \boldsymbol{U}_k^*). \tag{122}$$

This, together with (120) and (17), yields that

$$\|\boldsymbol{H}^1-\boldsymbol{H}^*\boldsymbol{Q}_{\pi^0}\|_F\leq\frac{2\sqrt{N}}{1-\kappa}\frac{2d_{\max}}{N_{\min}d_{\min}}\|\boldsymbol{H}^0-\boldsymbol{H}^*\boldsymbol{Q}_{\pi^0}\|_F^2=\kappa_1\|\boldsymbol{H}^0-\boldsymbol{H}^*\boldsymbol{Q}_{\pi^0}\|_F, \tag{123}$$

where

$$\kappa_1:=\frac{2\sqrt{N}}{1-\kappa}\frac{2d_{\max}}{N_{\min}d_{\min}}\|\boldsymbol{H}^0-\boldsymbol{H}^*\boldsymbol{Q}_{\pi^0}\|_F\leq\frac{2\sqrt{N}}{1-\kappa}\frac{2d_{\max}}{N_{\min}d_{\min}}\frac{(1-\kappa)d_{\min}N_{\min}}{5d_{\max}\sqrt{N}}=\frac{4}{5}. \tag{124}$$

Now, we use mathematical induction to show that it holds for all $t\in[T]$ that

$$\sum_{k=1}^K d(\boldsymbol{U}^{t+1}_{\pi^t(k)},\boldsymbol{U}^*_k)\leq\kappa_1^{2^t}\sum_{k=1}^K d(\boldsymbol{U}^t_{\pi^{t-1}(k)},\boldsymbol{U}^*_k), \tag{125}$$

$$\|\boldsymbol{H}^{t+1}-\boldsymbol{H}^*\boldsymbol{Q}_{\pi^{t+1}}\|_F\leq\kappa_1^{2^t}\|\boldsymbol{H}^t-\boldsymbol{H}^*\boldsymbol{Q}_{\pi^t}\|_F, \tag{126}$$

$$\|\boldsymbol{H}^t-\boldsymbol{H}^*\boldsymbol{Q}_{\pi^{t-1}}\|_F\leq\frac{2\sqrt{N}}{1-\kappa}\sum_{k=1}^K d(\boldsymbol{U}^t_{\pi^{t-1}(k)},\boldsymbol{U}^*_k). \tag{127}$$

We first verify (125), (126), and (127) for $t=1$. Due to (118) and (123), we obtain

$$\|\boldsymbol{H}^1-\boldsymbol{H}^*\boldsymbol{Q}_{\pi^1}\|_F\leq\|\boldsymbol{H}^1-\boldsymbol{H}^*\boldsymbol{Q}_{\pi^0}\|_F\leq\kappa_1\|\boldsymbol{H}^0-\boldsymbol{H}^*\boldsymbol{Q}_{\pi^0}\|_F. \tag{128}$$

According to this, (124), Proposition 2, and (119), it holds with probability at least $1-2K/(d_{\min}^2N)-5K^2/N$ that $\hat{d}^2_{\pi^1(k)}=d_k$ for all $k\in[K]$,

$$\sum_{k=1}^K d(\boldsymbol{U}^2_{\pi^1(k)},\boldsymbol{U}^*_k)\leq\frac{2d_{\max}}{N_{\min}d_{\min}}\|\boldsymbol{H}^1-\boldsymbol{H}^*\boldsymbol{Q}_{\pi^1}\|_F^2 \tag{129}$$

and for all $i\in[N]$,

$$\|\boldsymbol{h}_i^2-\boldsymbol{h}_i^*\boldsymbol{Q}_{\pi^1}\|\leq\frac{2}{1-\kappa}\sqrt{\sum_{k=1}^K d^2(\boldsymbol{U}^2_{\pi^1(k)},\boldsymbol{U}^*_k)}. \tag{130}$$

Substituting (122) with the first inequality of (128) into (129) yields that

$$\begin{aligned}\sum_{k=1}^K d(\boldsymbol{U}^2_{\pi^1(k)},\boldsymbol{U}^*_k)&\leq\frac{2d_{\max}}{N_{\min}d_{\min}}\|\boldsymbol{H}^1-\boldsymbol{H}^*\boldsymbol{Q}_{\pi^1}\|_F\cdot\frac{2\sqrt{N}}{1-\kappa}\sum_{k=1}^K d(\boldsymbol{U}^1_{\pi^0(k)},\boldsymbol{U}^*_k)\\&\leq\frac{\kappa_1\sqrt{N}}{1-\kappa}\frac{4d_{\max}}{N_{\min}d_{\min}}\|\boldsymbol{H}^0-\boldsymbol{H}^*\boldsymbol{Q}_{\pi^0}\|_F\sum_{k=1}^K d(\boldsymbol{U}^1_{\pi^0(k)},\boldsymbol{U}^*_k)\\&=\kappa_1^2\sum_{k=1}^K d(\boldsymbol{U}^1_{\pi^0(k)},\boldsymbol{U}^*_k),\end{aligned}$$

where the second inequality follows from (128) and the equality is due to (124). Thus, (125) holds for $t=1$. According to (130), repeating the arguments in (122) and (123), we obtain

$$\|\boldsymbol{H}^2-\boldsymbol{H}^*\boldsymbol{Q}_{\pi^1}\|_F\leq\frac{2\sqrt{N}}{1-\kappa}\sum_{k=1}^K d(\boldsymbol{U}^2_{\pi^1(k)},\boldsymbol{U}^*_k)$$

and

$$\begin{aligned}\|\boldsymbol{H}^2-\boldsymbol{H}^*\boldsymbol{Q}_{\pi^1}\|_F&\leq\frac{2\sqrt{N}}{1-\kappa}\frac{2d_{\max}}{N_{\min}d_{\min}}\|\boldsymbol{H}^1-\boldsymbol{H}^*\boldsymbol{Q}_{\pi^1}\|_F^2\\&\leq\frac{2\sqrt{N}}{1-\kappa}\frac{2d_{\max}}{N_{\min}d_{\min}}\kappa_1\|\boldsymbol{H}^0-\boldsymbol{H}^*\boldsymbol{Q}_{\pi^0}\|_F\|\boldsymbol{H}^1-\boldsymbol{H}^*\boldsymbol{Q}_{\pi^1}\|_F=\kappa_1^2\|\boldsymbol{H}^1-\boldsymbol{H}^*\boldsymbol{Q}_{\pi^1}\|_F,\end{aligned}$$

where the second inequality is due to the equality in (123). This, together with (118), implies

$$\|\boldsymbol{H}^2 - \boldsymbol{H}^*\boldsymbol{Q}_{\pi^2}\|_F \leq \kappa_1^2 \|\boldsymbol{H}^1 - \boldsymbol{H}^*\boldsymbol{Q}_{\pi^1}\|_F.$$

Thus, (127) and (126) holds for $t = 1$. Next, suppose that (125), (126), and (127) hold for all $t \geq 1$. Then, we can show that (128),(129), and (130) also hold for $t+1$ using the same arguments as those of $t = 1$. Consequently, we can further show that (125), (126), and (127) hold for $t+1$ until $t = T$. Finally, we use mathematical induction and deduce that for all $t \in [T]$, the desired results (125) and (126) hold with probability at least $1 - (T+1)(2K/(d_{\min}^2 N) + 5K^2/N)$ according to the union bound. It follows from (123) and (126) that

$$d_F\left(\boldsymbol{H}^{t+1}, \boldsymbol{H}^*\right) \leq \kappa_1^{2^t}\kappa_1^{2^{t-1}} \ldots \kappa_1^{2^1} \|\boldsymbol{H}^1 - \boldsymbol{H}^*\boldsymbol{Q}_{\pi^1}\|_F \leq \kappa_1^{2^{t+1}-1} \|\boldsymbol{H}^0 - \boldsymbol{H}^*\boldsymbol{Q}_{\pi^0}\|_F = \kappa_1^{2^{t+1}-1} d_F\left(\boldsymbol{H}^0, \boldsymbol{H}^*\right).$$

We next show (ii). It follows from $T = \log_2\left(\frac{\log\left((1-\kappa)\sqrt{d_{\min}}N_{\min}\right) - \log(5\sqrt{2}Kc_1\kappa_1 d_{\max}\sqrt{N})}{\log(1/\kappa_1)}\right) + 1$ that

$$d_F\left(\boldsymbol{H}^{T-1}, \boldsymbol{H}^*\right) \leq \kappa_1^{2^{T-1}} \frac{(1-\kappa)d_{\min}N_{\min}}{5\kappa_1 d_{\max}\sqrt{N}} \leq Kc_1\sqrt{2d_{\min}}.$$

This, together with Lemma 8, yields (19). According to Proposition 2, we also have $\hat{d}_{\pi^T(k)}^{T+1} = d_k$ for all $k \in [K]$. This, together with (19) and (2) in Definition 2, yields

$$\boldsymbol{U}_{\pi(k)}^{T+1}\boldsymbol{U}_{\pi(k)}^{T+1^T} = \boldsymbol{U}_k^*\boldsymbol{U}_k^{*^T} \text{ for all } k \in [K]. \tag{131}$$

By letting $\boldsymbol{O}_k = \boldsymbol{U}_k^{*^T}\boldsymbol{U}_{\pi(k)}^{T+1}$, we have

$$\boldsymbol{O}_k^T\boldsymbol{O}_k = \boldsymbol{U}_{\pi(k)}^{T+1^T}\boldsymbol{U}_k^*\boldsymbol{U}_k^{*^T}\boldsymbol{U}_{\pi(k)}^{T+1} = \boldsymbol{I}_{d_k},$$

where the second equality is due to (131). This implies $\boldsymbol{Q}_k \in \mathcal{O}^{d_k}$ for all $k \in [K]$. Then, we prove (20). □

E. Auxiliary Lemmas

Lemma 15. *Suppose that $\boldsymbol{a} \sim \text{Unif}(\mathbb{S}^{d-1})$ and $\tilde{\boldsymbol{v}} \in \mathbb{R}^d$ is a fixed vector with $\|\tilde{\boldsymbol{v}}\| = 1$. Let $\boldsymbol{a}$ be decomposed as*

$$\boldsymbol{a} = x\tilde{\boldsymbol{v}} + \sqrt{1-x^2}\boldsymbol{b}, \tag{132}$$

where $x \in \mathbb{R}$ and $\boldsymbol{b} \in \mathbb{R}^d$ satisfying $\langle\tilde{\boldsymbol{v}}, \boldsymbol{b}\rangle = 0$ and $\|\boldsymbol{b}\| = 1$. There exists an orthogonal matrix $\boldsymbol{U} \in \mathcal{O}^d$ such that $\boldsymbol{U}\tilde{\boldsymbol{v}} = \boldsymbol{e}_1$. Let

$$\tilde{\boldsymbol{b}} = \frac{\boldsymbol{U}^T\boldsymbol{a} - x\boldsymbol{e}_1}{\sqrt{1-x^2}} \tag{133}$$

and $\boldsymbol{c} \in \mathbb{R}^{d-1}$ such that $\boldsymbol{c} = (\tilde{b}_2, \cdots, \tilde{b}_d)$. Then, it holds that $\boldsymbol{U}\boldsymbol{b} \sim \tilde{\boldsymbol{b}}$, where

$$\tilde{b}_1 = 0, \ \boldsymbol{c} \sim \text{Unif}(\mathbb{S}^{d-2}).$$

Proof of Lemma 15. According to (132) and the rotational invariance of a uniform distribution over sphere, we have

$$\boldsymbol{U}\boldsymbol{b} = \frac{\boldsymbol{U}\boldsymbol{a} - x\boldsymbol{e}_1}{\sqrt{1-x^2}} \sim \tilde{\boldsymbol{b}}.$$

Since $\langle\boldsymbol{a}, \tilde{\boldsymbol{v}}\rangle = x$, then $\langle\boldsymbol{U}^T\boldsymbol{a}, \boldsymbol{U}\tilde{\boldsymbol{v}}\rangle = \langle\boldsymbol{U}^T\boldsymbol{a}, \boldsymbol{e}_1\rangle = x$. This implies $\tilde{b}_1 = 0$. Moreover, since $\boldsymbol{a} \sim \text{Unif}(\mathbb{S}^{d-1})$, then $\boldsymbol{U}^T\boldsymbol{a} \sim \text{Unif}(\mathbb{S}^{d-1})$ due to the rotational invariance of a uniform distribution over sphere. Then, let $\boldsymbol{y} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_d)$ such that $\boldsymbol{U}^T\boldsymbol{a} = \boldsymbol{y}/\|\boldsymbol{y}\|$. This, together with $\langle\boldsymbol{U}^T\boldsymbol{a}, \boldsymbol{e}_1\rangle = x$, implies

$$y_1^2 = \frac{x^2\sum_{i\neq 1} y_i^2}{1-x^2}.$$

Then, we have

$$\|\boldsymbol{y}\|^2 = y_1^2 + \sum_{i\neq 1} y_i^2 = \frac{1}{1-x^2}\sum_{i\neq 1} y_i^2.$$

This, together with (133), implies that for any $i \neq 1$,

$$\tilde{b}_i = \frac{y_i}{\|\boldsymbol{y}\|\sqrt{1-x^2}} = \frac{y_i}{\sqrt{\sum_{i\neq 1} y_i^2}}.$$

Then, we complete the proof. □

Lemma 16. *Consider the setting in Lemma 14. Suppose that $\boldsymbol{v}_i$, $\tilde{\boldsymbol{v}}_i$ for $i \in [N]$ are defined in (71) and β_{ij} is defined as in (72) for some given $\boldsymbol{v}_i$ and $\tilde{\boldsymbol{v}}_j$. Then, it holds that*

$$\Phi\left(\beta_{ij}\right) - \Phi\left(\frac{\tau(\sqrt{d_\ell}-\alpha)}{\|\boldsymbol{v}_i\|}\right) \lesssim \frac{1}{\sqrt{\log N}}, \tag{134}$$

and

$$\Phi\left(\frac{\tau(\sqrt{d_\ell}+\alpha)}{\|\boldsymbol{v}_i\|}\right) - \Phi\left(\beta'_{ij}\right) \lesssim \frac{1}{\sqrt{\log N}}, \tag{135}$$

Proof of Lemma 16. Suppose that (56) and (57) hold, which happens with probability at least $1-5K^2N^{-2}$ according to Lemma 12. Let $k = \{\ell \in [K] : h^*_{i\ell} = 1\}$. It follows from (60) that

$$\|\boldsymbol{v}_i\| \geq \frac{\mathrm{aff}(S^*_k, S^*_\ell) - \alpha}{\sqrt{d_k}+\alpha} \geq \frac{(1-\varepsilon)\mathrm{aff}(S^*_k, S^*_\ell)}{(1+\varepsilon)\sqrt{d_k}}. \tag{136}$$

This, together with (57), yields that

$$|\langle \boldsymbol{v}_i, \tilde{\boldsymbol{v}}_j\rangle| \leq \frac{2\sqrt{\log N}}{\sqrt{d_k}-\alpha} \leq \frac{2\sqrt{\log N}}{(1-\varepsilon)\sqrt{d_k}} \leq \varepsilon\|\boldsymbol{v}_i\|. \tag{137}$$

According to (22) and (60), we have

$$\tau = \frac{\max_{k\neq\ell}\mathrm{aff}(S^*_k, S^*_\ell) + \alpha}{(\sqrt{d_{\max}}-\alpha)^2} \leq \frac{(1+\varepsilon)\max_{k\neq\ell}\mathrm{aff}(S^*_k, S^*_\ell)}{(1-\varepsilon)^2 d_{\max}} \leq \frac{1+\varepsilon}{(1-\varepsilon)^2\sqrt{d_{\max}}}. \tag{138}$$

We first compute

$$\begin{aligned}\frac{\tau(\sqrt{d_\ell}+\alpha)}{(\|\boldsymbol{v}_i\| - |\langle \boldsymbol{v}_i, \tilde{\boldsymbol{v}}_j\rangle|)\sqrt{1-(\log N)/d_\ell}} - \frac{\tau(\sqrt{d_\ell}-\alpha)}{\|\boldsymbol{v}_i\|} &\leq \frac{1+\varepsilon}{(1-\varepsilon)^2\sqrt{d_{\max}}}\left(\frac{1+\varepsilon}{(1-\varepsilon)^2} - (1-\varepsilon)\right)\frac{\sqrt{d_\ell}}{\|\boldsymbol{v}_i\|} \\ &\lesssim \frac{1}{\sqrt{\log N}\|\boldsymbol{v}_i\|},\end{aligned} \tag{139}$$

where the first inequality is due to (60) and (138). We next compute

$$\frac{|\langle \boldsymbol{v}_i, \tilde{\boldsymbol{v}}_j\rangle|\sqrt{(\log N)/d_\ell}(\sqrt{d_\ell}+\alpha)}{(\|\boldsymbol{v}_i\| - |\langle \boldsymbol{v}_i, \tilde{\boldsymbol{v}}_j\rangle|)\sqrt{1-(\log N)/d_\ell}} \leq \frac{4\sqrt{\log N}|\langle \boldsymbol{v}_i, \tilde{\boldsymbol{v}}_j\rangle|}{\|\boldsymbol{v}_i\|} \lesssim \frac{1}{\sqrt{\log N}\|\boldsymbol{v}_i\|}, \tag{140}$$

where the first inequality is due to (137) and $d_{\min} \gtrsim \log^3 N$ and the second one follows from the second inequality of (137) and $d_{\min} \gtrsim \log^3 N$. Then, we obtain

$$\begin{aligned}\beta_{ij} - \frac{\tau(\sqrt{d_\ell}-\alpha)}{\|\boldsymbol{v}_i\|} &= \frac{\tau(\sqrt{d_\ell}+\alpha)}{(\|\boldsymbol{v}_i\| - |\langle \boldsymbol{v}_i, \tilde{\boldsymbol{v}}_j\rangle|)\sqrt{1-(\log N)/d_\ell}} - \frac{\tau(\sqrt{d_\ell}-\alpha)}{\|\boldsymbol{v}_i\|} + \frac{|\langle \boldsymbol{v}_i, \tilde{\boldsymbol{v}}_j\rangle|\sqrt{(\log N)/d_\ell}(\sqrt{d}+\alpha)}{(\|\boldsymbol{v}_i\| - |\langle \boldsymbol{v}_i, \tilde{\boldsymbol{v}}_j\rangle|)\sqrt{1-(\log N)/d_\ell}} \\ &\lesssim \frac{1}{\sqrt{\log N}}\frac{1}{\|\boldsymbol{v}_i\|},\end{aligned} \tag{141}$$

where the first inequality is due to (139) and (140). Moreover, we have

$$\begin{aligned}\Phi\left(\beta_{ij}\right)-\Phi\left(\frac{\tau(\sqrt{d_\ell}-\alpha)}{\|\boldsymbol{v}_i\|}\right) &\leq \sqrt{\frac{1}{2\pi}}\exp\left(-\frac{\tau^2(\sqrt{d_\ell}-\alpha)^2}{2\|\boldsymbol{v}_i\|^2}\right)\left(\beta_{ij}-\frac{\tau(\sqrt{d_\ell}-\alpha)}{\|\boldsymbol{v}_i\|}\right)\\ &\leq \frac{1}{\sqrt{2\pi\log N}}\exp\left(-\frac{\left(\kappa\sqrt{d_{\min}}+\alpha\right)^2\left(\sqrt{d_\ell}-\alpha\right)^2}{2(\sqrt{d_{\max}}-\alpha)^4\|\boldsymbol{v}_i\|^2}\right)\frac{1}{\|\boldsymbol{v}_i\|}\\ &\lesssim \frac{d_{\max}}{d_{\min}\sqrt{\log N}},\end{aligned}$$

where the first inequality is due to the basic inequality for the integral, the second inequality uses the inequality of (138) and (141), and the last inequality follows from $d_{\min}\gtrsim\log N$ and the fact that $\exp\left(-cx^2/2\right)x$ attains the maximum at $x=1/\sqrt{c}$ when $x\in(0,\infty)$. The proof of (135) follows from the same argument as above.

□

F. Experiment Setups and Results in Section 4.2

In this section, we provide more implementation details and results for the experiments in Section 4.2. We use the real datasets *COIL-20* (S. A. Nene & Murase, 1996b), *COIL-100* (S. A. Nene & Murase, 1996a), the cropped extended *Yale B* (Georghiades et al., 2001), *USPS* (Hull, 1994), and *MNIST* (LeCun, 1998).[3] The information about the used real-world datasets can be found in Table 3. Before using these datasets in the experiments, we normalize them such that each sample has unit length. Note that the MNIST dataset contains 70000 images of handwritten digits 0-9. Following the preprocessing technique in You et al. (2016); Lipor et al. (2021), we represent each image by a feature vector of dimension 3472 using the scattering convolutional network (Bruna & Mallat, 2013) and reduce the dimension of each vector to 500 using PCA.

Table 3. The parameters for the real datasets: N is the number of samples, n is the dimension of samples, and K is the number of clusters.

Datasets	N	n	K
COIL20	1440	1024	20
COIL100	7200	1024	100
YaleB	2414	1024	38
USPS	9298	256	10
MNIST	70000	780	10

Since the data points in real datasets generally do not follow the semi-random UoS model in Definition 2, we cannot guarantee good clustering performance if we directly apply the TIPS method for initializing the KSS method. Therefore, in the implementation of the TIPS method, we improve the idea of the thresholding inner product to construct the weight matrix $\boldsymbol{A}=\{a_{ij}\}_{1\leq i,j\leq N}$ by

$$a_{ij}=\begin{cases}|\langle\boldsymbol{z}_i,\boldsymbol{z}_j\rangle|, & \text{if } |\langle\boldsymbol{z}_i,\boldsymbol{z}_j\rangle|\geq\tau \text{ or } j\in\mathcal{T}_i \text{ and } i\neq j,\\ 0, & \text{otherwise,}\end{cases}$$

where $\mathcal{T}_i\subseteq[N]\setminus\{i\}$ with $|\mathcal{T}_i|=2$ satisfies $|\langle\boldsymbol{z}_i,\boldsymbol{z}_j\rangle|\geq|\langle\boldsymbol{z}_i,\boldsymbol{z}_k\rangle|$ for all $j\in\mathcal{T}_i$ and $k\notin\mathcal{T}_i$. Introducing $\mathcal{T}_i$ is to ensure that each column of $\boldsymbol{A}$ contains at least two non-zero elements. For the implementation of the KSS method, we simply set $d_1=\cdots=d_K=d$, where d is given in Table 4. For all algorithms, we assume that K is known and given in Table 3. We present the parameters of all the tested methods in Table 4.

To complement the result of recovery accuracy in Table 2, we also report the running time and clustering accuracy for all runs of each method in Table 5.

[3]The datasets *COIL-20*, *COIL-100*, the cropped extended *Yale B*, and *USPS* are downloaded from http://www.cad.zju.edu.cn/home/dengcai/Data/data.html. The dataset *MNIST* is downloaded from https://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/.

Table 4. Parameters setting of the tested methods in the experiments .

	COIL20	*COIL100*	*YaleB*	*USPS*	*MNIST*
KSS	$(d,\tau)=(10,0.98)$	$(d,\tau)=(10,0.98)$	$(d,\tau)=(8,0.98)$	$(d,\tau)=(9,0.99)$	$(d,\tau)=(18,0.98)$
SSC	$(\alpha,\rho)=(10,0.8)$	$(\alpha,\rho)=(10,2)$	$(\alpha,\rho)=(10,1)$	$(\alpha,\rho)=(10,0.5)$	$(\alpha,\rho)=(10,0.8)$
TSC	$q=4$	$q=3$	$q=4$	$q=5$	$q=6$
GSC	$q=25$	$q=15$	$q=20$	$q=20$	$q=20$
LRR	$\lambda=10^{-2}$	$\lambda=10^{-3}$	$\lambda=0.1$	$\lambda=10^{-3}$	$\lambda=10^{-2}$
LRSSC	$(\sigma,\lambda)=(0.2,0.5)$	$(\sigma,\lambda)=(1,2)$	$(\sigma,\lambda)=(0.1,1)$	$(\sigma,\lambda)=(10,1)$	$(\sigma,\lambda)=(0.2,0.5)$
OMP	$q=2$	$q=2$	$q=5$	$q=25$	$q=20$

Table 5. CPU times (in seconds) and the clustering accuracy of the tested methods on real datasets over 10 runs.

Accuracy	*COIL20*	*COIL100*	*YaleB*	*USPS*	*MNIST*
KSS	**0.9187±0**	**0.8050±0.0040**	0.6715±0.0253	**0.8120±0.0164**	**0.8989±0.0796**
SSC	**0.9075±0.0164**	0.6542±0.0165	**0.8179±0.0074**	0.6582±0.0002	–
OMP	0.5012±0.0168	0.3273±0.0083	**0.7968±0.0216**	0.1967±0.0071	0.5749±0
TSC	0.8271±0	**0.7164±0.0093**	0.4700±0.0092	0.6688±0.0002	**0.8514±0**
GSC	0.7896±0	0.6445±0.0084	0.6852±0.0135	**0.9522±0.0001**	0.5411±0.0427
LRR	0.7161±0.0064	0.5403±0.0066	0.6534±0.0146	0.7129±0.0001	–
LRSSC	0.8194±0	0.5035±0.0101	0.6971±0.0097	0.6440±0.0005	–
Time (s)	*COIL20*	*COIL100*	*YaleB*	*USPS*	*MNIST*
KSS	1.32±0.08	53.53±6.78	5.94±0.84	**8.85±0.67**	**30.5287±13.15**
SSC	55.37±4.99	912.25±42.12	136.36±13.64	1217.88±27.21	–
OMP	**0.62**±0.04	**12.11±0.54**	**1.02±0.06**	31.12±0.29	398.37±8.14
TSC	**0.66**±0.03	**29.78±1.05**	**3.06±0.18**	**2.66±0.07**	**154.46±20.91**
GSC	11.73±0.54	178.15±7.93	24.22±0.85	105.59±7.22	1800.00±0
LRR	33.63±2.62	144.25±7.99	63.30±16.06	111.56±9.05	–
LRSSC	73.31±3.45	1800.00±0	444.28±37.95	1800.00±0	–

"–" denotes out of memory.