Journal of Machine Learning Research 22 (2021) 1-37 Submitted 11/18; Revised 1/21; Published 2/21

Subspace Clustering through Sub-Clusters

Weiwei Li WEIWEILI@LIVE.UNC.EDU
Department of Statistics and Operations Research
University of North Carolina at Chapel Hill
Chapel Hill, NC 27514, USA

Jan Hannig JAN.HANNIG@UNC.EDU
Department of Statistics and Operations Research
University of North Carolina at Chapel Hill
Chapel Hill, NC 27514, USA

Sayan Mukherjee SAYAN@STAT.DUKE.EDU
Department of Statistical Science
Mathematics, Computer Science, Biostatistics & Bioinformatics
Duke University
Durham, NC 27708, USA

Editor: David Blei

Abstract

The problem of dimension reduction is of increasing importance in modern data analysis. In this paper, we consider modeling the collection of points in a high dimensional space as a union of low dimensional subspaces. In particular we propose a highly scalable sampling based algorithm that clusters the entire data via first spectral clustering of a small random sample followed by classifying or labeling the remaining out-of-sample points. The key idea is that this random subset borrows information across the entire dataset and that the problem of clustering points can be replaced with the more efficient problem of "clustering sub-clusters". We provide theoretical guarantees for our procedure. The numerical results indicate that for large datasets the proposed algorithm outperforms other state-of-the-art subspace clustering algorithms with respect to accuracy and speed.

Keywords: dimension reduction, subspace clustering, sub-cluster, random sampling, scalability, handwritten digits, spectral clustering

1. Introduction

In data analysis, researchers are often given datasets with large volume and high dimensionality. To reduce the computational complexity arising in these settings, researchers resort to dimension reduction techniques. To this end, traditional methods like PCA Hotelling (1933) use few principal components to represent the original dataset; factor analysis (Cattell, 1952) seeks to get linear combinations of latent factors; subsequent works of PCA include kernel PCA (Schölkopf et al., 1998), generalized PCA (Vidal et al., 2005); manifold learning (Belkin and Niyogi, 2003) assumes data points collected from a high dimensional ambient space lie around a low dimensional manifold, and muli-manifold learning (Liu et al.,

2011) considers the setting of a mixture of manifolds. In this paper, we focus on one of the simplest manifold, a subspace, and consider the subspace clustering problem. Specifically, we approximate the original dataset as an union of subspaces. Representing the data as a union of subspaces allows for more computationally efficient downstream analysis on various problems such as motion segmentation (Elhamifar and Vidal, 2009), handwritten digits recognition (You et al., 2016a), and image compression (Hong et al., 2006).

1.1 Related Work

Many techniques have been developed for subspace clustering, see Elhamifar and Vidal (2013) for a review. The mainstream methods usually include two phases: (1) calculating the affinity matrix; (2) applying spectral clustering (Ng et al., 2002) to the affinity matrix to compute a label for each data point. For phase (1), the property of self-representation is often used to calculate the affinity matrix: self-representation states that a point can be represented by a linear combination of other points in the same subspace. Specifically, Elhamifar and Vidal (2009) proposed the sparse subspace clustering (SSC) algorithm which solves the lasso minimization problem N times, where N is the total number of data points, the theoretical property of SSC was further studied in Soltanolkotabi et al. (2012). Similarly, Rahmani and Atia (2017) proposed the direction search algorithm (DSC) which uses ℓ_1 minimization to find the "optimal direction" for each data point, these directions are then used to cluster the data points. One of the main drawbacks of SSC and DSC is their computational complexity of $O(N^2)$ in both time and space, which limits its application to large datasets. To address this limitation, a variety of methods have been proposed to avoid solving complicated optimization problems in constructing the affinity matrix. Heckel and Bölcskei (2015) used inner products with thresholding (TSC) to calculate the affinity between each pair of points, Park et al. (2014) used a greedy algorithm to find for each point the linear space spanned by its neighbors, similarly Dyer et al. (2013) and You et al. (2016c) used orthogonal matching pursuit (OMP), You et al. (2016b) used elastic the net for subspace clustering (ENSC) and proposed an efficient solver by active set method. However, these methods require running spectral clustering on the full $N \times N$ affinity matrix. A Bayesian mixture model was proposed for subspace clustering in Thomas et al. (2014), but its parameter inference is not scalable to large dataset. Zhou et al. (2018) used a deep learning based method which does not have theoretical guarantee.

Recently, there have been two methods that increase the scalability of sparse subspace clustering. The SSSC algorithm and its varieties (Peng et al., 2015) clusters a random subset of the whole dataset and then uses this clustering to classify or label the out-of-sample data points. This method scales well when the random subset is small, however a great deal of information is discarded as only the information in the subset is used. In You et al. (2016a) a divide and conquer strategy is used for SSC—the dataset is split into several small subsets on which SSC is run, and clustering results are merged. This method cannot reduce the computational complexity of the SSC by an order of magnitude so is limited in its ability to scale to large dataset.

1.2 Contribution

In this paper, we propose a novel, efficient sampling based algorithm with provable guarantees that extends the ideas in previous scalable methods (Peng et al., 2015; You et al., 2016a). The motivation for using sampling based algorithm is twofold. From the theoretical perspective, Luxburg et al. (2005) showed that under certain assumptions, the spectral clustering results on the sampled subset will converge to the results on the whole dataset. This gives us the insight that as the size of the sampled subset increases properly, the subset becomes almost as informative as the whole dataset. From the computational perspective, traditional spectral clustering based algorithms need to build a "neighborhood" for each data point. Thus the complexity (both in time and memory) is usually at least quadratic in the total number of data points, while sampling based algorithms need to find neighboring points only within the subset. This greatly reduces the computational resources needed incurring some loss of information.

Our algorithm seeks to combine strengths from both approaches. In particular, for each point in the subset we find its nearest neighbors in the complete dataset and use these points to construct a sub-cluster, these sub-clusters contain information from the entire dataset and not just the random sample. Finding neighboring points among the whole dataset makes it possible to get a neighborhood with big enough size and few false connections for each sampled point. The affinity matrix for the subset is then constructed from these sub-clusters. The idea is that we change the problem from "clustering of data points" to "clustering of sub-clusters", which integrates information across the dataset and should deliver better clustering results.

We provide theoretical guarantees for our procedure in Section 3. The analysis reveals that under mild conditions, the subspaces can share arbitrarily many intersections as long as most of their principal angles are larger than a certain threshold. While our algorithm for finding neighboring points is similar to that of Heckel and Bölcskei (2015), the data generation model and assumptions underlying our theorems are different—we take into account the fact that after normalization the noisy terms will no longer follow a multivariate Gaussian distribution. While our work is originally designed for linear subspace clustering problems. The idea of clustering through sub-clusters can be easily extended to general clustering problems.

Finally, we study empirical properties of the proposed algorithm on both synthetic and real-world datasets selected to have diverse sizes. We show that the clustering through sub-clusters algorithm is highly scalable and can significantly boost the clustering accuracy on both the subset and whole dataset. The advantage of our algorithm over other state-of-the-art algorithms changes from marginal to significant as the size of the dataset increases.

1.3 Paper Organization

The rest of this paper is organized as follows: in Section 2, we describe the implementation of our clustering procedure, in Section 3 we state the model setting and theoretical guarantees for our procedure and explain in some details the geometric and distributional intuitions underlying our procedure. The detailed proofs can be found in Appendix C. In Section 4 we present numerical experiments and compare our method with other state-of-the-art methods, a comprehensive report of the numerical results can be found in Appendix E.

1.4 Notation

Unless specified otherwise, we use capital bold letter to denote data matrix, and corresponding lower bold letter to denote the columns of it. In this paper, we are given a dataset $\mathbf{Y}$ with N data points in $\mathbb{R}^D$. We use both $\mathbf{y}_i$ and $[\mathbf{Y}]_i$ to denoted the i-th column of $\mathbf{Y}$, and $\mathbf{Y}_{-i}$ is the matrix $\mathbf{Y}$ with the i-th column removed. Similarly, we write $\mathbf{y}_{-j}$ as vector $\mathbf{y}$ with the j-th entry removed. The ij-th entry of a matrix $\mathbf{Y}$ is denoted as $[\mathbf{Y}]_{ij}$. The complement of event $\mathcal{E}$ is denoted by $\mathcal{E}^{\complement}$. The cardinality of $\mathcal{E}$ is denoted by $\mathbf{card}(\mathcal{E})$, and the mode of $\mathcal{E}$ is $\mathbf{mode}(\mathcal{E})$. We use subscript with parenthesis to represent the order statistics of entries in a vector, for example $\mathbf{a}_{(i)}$ is the i-th smallest entry in vector $\mathbf{a}$, while without ambiguity both $\mathbf{a}(i)$ and a_i refer to the i-th element of vector $\mathbf{a}$. The unit sphere in $\mathbb{R}^d$ is denoted by $\mathbb{S}^{d-1}$. We assume each data point of $\mathbf{Y}$ lies on one of K linear subspaces denoted by $\{\mathcal{S}_k\}_{k=1}^K$. Here K is a known constant and $\mathcal{S}_k$ is the k-th linear subspace. The subspace clustering problem aims assigning to each point in $\mathbf{Y}$ the membership to a subspace (cluster) $\mathcal{S}_k$.

We write d_k as the dimension of subspace $\mathcal{S}_k$ and $\mathbf{U}_k \in \mathbb{R}^{D\times d_k}$ as its corresponding orthogonal base. The number of points belong to cluster $\mathcal{S}_k$ is N_k. We use $\mathbf{y_i}^{(k)} \in \mathbb{R}^D$ to represent the i-th point from the k-th cluster, the set $\{\mathbf{y_1}^{(k)}, ..., \mathbf{y_{N_k}}^{(k)}\}$ contains all points that belong to $\mathcal{S}_k$. Finally, we write $F_{m,n}$ as the F-distribution with parameters (m, n), $Dir(\boldsymbol{\alpha})$ as the Dirichlet distribution with parameter vector $\boldsymbol{\alpha}$, $\beta(a, b)$ as the beta distribution with parameters (a, b), $\mathcal{N}(\mu, \sigma)$ as the Gaussian distribution with mean (vector) μ and variance (covariance matrix) σ^2, χ^2_d as the chi-square distribution with d degrees of freedom, and $U(\mathbb{S}^{d-1})$ as the uniform distribution on the surface of unit sphere $\mathbb{S}^{d-1}$.

2. The Algorithm for Sampling Based Subspace Clustering

In this section, we introduce our sampling based algorithm for subspace clustering (SBSC). In Appendix A we will discuss issues regarding hyper-parameters. Throughout this section, we assume the columns of $\mathbf{Y}$ have unit ℓ_2 norm.

Our main algorithm takes the raw dataset $\mathbf{Y}$ that has N observations and several parameters as inputs and outputs the clustering assignment for each point in the dataset, it proceeds in two stages (see the matched steps in Algorithm 1 for further details):

- Stage 1: In-sample clustering
    1. Draw a subset $\tilde{\mathbf{Y}}$ of $n \ll N$ points.
    2. For each point $\tilde{\mathbf{y}}_i \in \tilde{\mathbf{Y}}$, find its $(d_{\max} + 1)$ nearest neighbor points in $\mathbf{Y}$ and use $\mathcal{C}_i$ to denote the index set of these points. We call $\mathbf{Y}_{\mathcal{C}_i}$ the sub-cluster of $\tilde{\mathbf{y}}_i$.
    3. Compute the affinity matrix $\mathbf{D}$ where each element $[\mathbf{D}]_{ij}$ is the similarity calculated between $\mathbf{Y}_{\mathcal{C}_i}$ and $\mathbf{Y}_{\mathcal{C}_j}$.
    4. Sparsify the affinity matrix by removing possible spurious connections.
    5. Conduct spectral clustering on $\tilde{\mathbf{Y}}$ with the sparsified affinity matrix.
- Stage 2: Out-of-sample classification
    6. Fit a classifier to the clustered points in $\tilde{\mathbf{Y}}$ and classify the points in $\mathbf{Y} \setminus \tilde{\mathbf{Y}}$.

In Stage 1, we formulated n sub-clusters out of the sampled dataset. These n sub-clusters are then further grouped into K clusters. An algorithm based on a similar idea was proposed for Gaussian mixture models by Aragam et al. (2020), they first divide the dataset into $L(\gg K)$ mixtures, and then cluster these mixtures into K mixtures. In our paper the linear structure of subspaces is central to clustering while in their paper the distributional properties of mixture models play the key role in grouping, this is a significant difference. On a high level, we transfer the problem from "clustering points" to "clustering sub-clusters".

Step 2 computes the neighborhood of points around each sampled points by thresholding a similarity score based on inner products, this method was also used in Heckel and Bölcskei (2015). The intuitive reason for this step is that for normalized data, two vectors are more likely to lie in the same linear subspace if the absolute magnitude of the inner product between the points is large. One may use other measure of similarities in Step 2 to find the neighboring points. In addition to the standard algorithm, we also present experimental results based on other similarity measures in Section 4 and Appendix E.

The idea of using distance between the sub-clusters to construct an affinity matrix in Step 3 relies on the self-representative property of linear subspaces—see Theorem 2 for technical details. Please note that each entry of the affinity matrix measures the closeness between data points, hence it is a decreasing function of distance. There is both theoretical and empirical evidence that sparsification of an affinity matrix by setting smaller elements to zero improves clustering results (Belkin and Niyogi, 2003; Von Luxburg, 2007). For this reason in Step 4 we threshold the affinity matrix.

In Stage 2, the remaining points are labeled via a classifier where a regression model is fitted on the clustered data, specifically a residual minimization ridge regression procedure. If both n, d_{max} and D are linear in $\log N$, the complexity of our algorithm is $O(N \log N)$.

Note that any classifier can be used to do the out-of-sample classification. While ridge regression worked well for linear subspace clustering problems in this paper, we encourage users of Algorithm 1 to choose their own favorite classifier, e.g., svm, random forest, or even deep neural networks, based on their understandings of the data.

3. Clustering Accuracy

3.1 Model Specification

We assume all subspaces have the same dimension d and the data generating process is

$$\hat{\mathbf{y}_\mathbf{i}}^{(k)} = \zeta_i^{(k)} \mathbf{U_k} \mathbf{a}_i^{(k)} + \hat{\mathbf{e}}_i^{(k)}, \quad i = 1, \ldots, N_k, \quad k = 1, ..., K,$$

where $\mathbf{a}_i^{(k)} \in \mathbb{R}^d$ is sampled from the uniform distribution on the surface of $\mathbb{S}^{d-1}$, $\zeta_i^{(k)}$ is a random scalar such that $\zeta_i^{(k)2} \sim \chi_d^2$, and $\hat{\mathbf{e}}_i^{(k)} \sim \mathcal{N}(\mathbf{0}, d\sigma^2 \mathbf{I}_D)$. However $\hat{\mathbf{y}_\mathbf{i}}^{(k)}$ are unobserved and we only observe the normalized version $\mathbf{y}_i^{(k)} = \hat{\mathbf{y}_\mathbf{i}}^{(k)} / \|\hat{\mathbf{y}_\mathbf{i}}^{(k)}\|_2$. We then have

$$\mathbf{y}_i^{(k)} = \frac{\mathbf{U_k} \mathbf{a}_i^{(k)} + \sigma \mathbf{e}_i^{(k)}}{\left\| \mathbf{U}_k \mathbf{a}_i^{(k)} + \sigma \mathbf{e}_i^{(k)} \right\|_2}. \tag{1}$$

Consequently, each entry in $\mathbf{e}_i^{(k)} = \hat{\mathbf{e}}_i^{(k)} / (\sigma \zeta_i^{(k)})$ follows multivariate t-distribution with d degrees of freedom, and $\|\mathbf{e}_i^{(k)}{}_2\|_2^2 / D \sim F_{D,d}$. Numerically, the normalizing constant

input : Data $\mathbf{Y}$, number of subspaces K, sampling size n, neighbor threshold $d_{\max}$, regularization parameters λ_1 and λ_2, residual minimization parameter m, affinity threshold $t_{\max}$.

output: The label vector $\boldsymbol{\ell}$ of all points in $\mathbf{Y}$

1. Uniformly sample n points $\tilde{\mathbf{Y}}$ from $\mathbf{Y}$.
2. Construct the sub-clusters:

for $i = 1$ **to** n **do**
  $\mathbf{p} = |\langle \tilde{\mathbf{y}}_i, \mathbf{Y} \rangle|$;
  $\mathcal{C}_i := \{j : |\langle \tilde{\mathbf{y}}_i, \mathbf{y}_j \rangle| \geq \mathbf{p}_{(N-d_{\max})}\}$.
end

3. Construct affinity matrix $[\mathbf{D}]_{ij} = e^{-d(\mathbf{Y}_{\mathcal{C}_i}, \mathbf{Y}_{\mathcal{C}_j})/2}$ for $i \neq j \in \{1, ..., n\}$ and

$$\begin{aligned} d(\mathbf{Y}_{\mathcal{C}_i}, \mathbf{Y}_{\mathcal{C}_j}) &= ||\mathbf{Y}_{\mathcal{C}_i} - \mathbf{Y}_{\mathcal{C}_j}(\mathbf{Y}_{\mathcal{C}_j}^T \mathbf{Y}_{\mathcal{C}_j} + \lambda_1 \mathbf{I})^{-1} \mathbf{Y}_{\mathcal{C}_j}^T \mathbf{Y}_{\mathcal{C}_i}||_F \\ &\quad + ||\mathbf{Y}_{\mathcal{C}_j} - \mathbf{Y}_{\mathcal{C}_i}(\mathbf{Y}_{\mathcal{C}_i}^T \mathbf{Y}_{\mathcal{C}_i} + \lambda_1 \mathbf{I})^{-1} \mathbf{Y}_{\mathcal{C}_i}^T \mathbf{Y}_{\mathcal{C}_j}||_F. \end{aligned}$$

4. Sparsify the adjacency matrix:

for $j = 1$ **to** n **do**
  $\mathbf{v} := [\mathbf{D}]_j$;
  for $i = 1$ **to** n **do**
    if $[\mathbf{D}]_{ij} \leq \mathbf{v}_{(n-d_{\max})}$ **then**
      $[\mathbf{D}]_{ij} := 0$
    end
  end
end

5. Cluster $\tilde{\mathbf{Y}}$: set $\mathbf{D} := \mathbf{D} + \mathbf{D}^T$ and cluster the in-sample points in $\tilde{\mathbf{Y}}$ by applying spectral clustering on $\mathbf{D}$, use $\boldsymbol{\ell}_{in}$ to denote the labels of $\tilde{\mathbf{Y}}$.
6. Label the remaining points: use the Residual Minimization by Ridge Regression (RMRR) algorithm in Appendix D to classify the remaining points in $\mathbf{Y} \setminus \tilde{\mathbf{Y}}$, specifically for the out-of-sample label we have

$$\boldsymbol{\ell}_{out} = RMRR(\mathbf{Y} \setminus \tilde{\mathbf{Y}}, \tilde{\mathbf{Y}}, \boldsymbol{\ell}_{in}, \lambda_2, m)$$

7. Combine $\boldsymbol{\ell}_{in}$ and $\boldsymbol{\ell}_{out}$ to get $\boldsymbol{\ell}$, the label of the whole dataset $\mathbf{Y}$.

Algorithm 1: Sub-cluster Based Subspace Clustering (SBSC) algorithm.

$\|\mathbf{U}_k \mathbf{a}_i^{(k)} + \sigma \mathbf{e}_i^{(k)}\|_2$ will be approximately 1. In Heckel and Bölcskei (2015), the normalizing constants are treated directly as 1 and their noise vector is a multivariate Gaussian vector. In developing theoretical guarantees of this paper, we explicitly account for the normalizing constant $\|\mathbf{U}_k \mathbf{a}_i^{(k)} + \sigma \mathbf{e}_i^{(k)}\|_2$ and its effects.

Let $\lambda_1^{(ij)} \geq \lambda_2^{(ij)} \geq ... \geq \lambda_d^{(ij)}$ correspond to the cosine values of principal angles between $\mathcal{S}_i$ and $\mathcal{S}_j$, hence $\lambda_1^{(ij)} \leq 1$ and $\lambda_d^{(ij)} \geq 0$. Note that $\lambda_k^{(ij)} = \lambda_k^{(ji)}$ for $1 \leq k \leq d$ and $1 \leq i < j \leq K$. For each subspace $\mathcal{S}_k$, we define the uniformly maximal affinity vector to quantify its closeness with respect to all other subspaces.

Definition 1 *For each subspace $\mathcal{S}_k$, its uniformly maximal affinity vector with respect to the other subspaces is $[\lambda_1^{(k)}, ..., \lambda_d^{(k)}]$ with*

$$\lambda_i^{(k)} = \max_{j \neq k} \lambda_i^{(kj)}.$$

Definition 2 *The sub-cluster preserving property holds for an algorithm if each sub-cluster output contains only points from the same subspace.*

If the uniformly maximal affinity vectors have small entries, corresponding to large angles, we would expect that Algorithm 1 (SBSC) satisfies the sub-cluster preserving property.

In constructing the affinity matrix $\mathbf{D}$, we want the following property: two sub-clusters that belong to the same subspace have bigger affinities, hence smaller distances, than sub-clusters that belong to different subspaces.

Definition 3 *We say $\mathbf{Y}_{\mathcal{C}_i}$ has the correct neighborhood property with distance function $d(\cdot, \cdot)$ if*

$$d(\mathbf{Y}_{\mathcal{C}_i}, \mathbf{Y}_{\mathcal{C}_j}) < d(\mathbf{Y}_{\mathcal{C}_i}, \mathbf{Y}_{\mathcal{C}_k}),$$

for any $1 \leq j \neq k \leq n$ such that $\mathbf{Y}_{\mathcal{C}_i}$ and $\mathbf{Y}_{\mathcal{C}_j}$ belong to the same subspace, and $\mathbf{Y}_{\mathcal{C}_k}$ belongs to a different subspace than $\mathbf{Y}_{\mathcal{C}_i}$.

3.2 Theoretical Properties of SBSC

In this section we provide a theoretical analysis of Algorithm 1 providing conditions under which we have provable guarantees.

3.2.1 Assumptions

In this section, we list the assumptions used in the lemmas and theorems. On a high level, the theoretical properties developed in this paper require two groups of conditions. First, the subspaces need to be separated, this is Assumption $A2$. Second, the sub-clusters should be informative and carry enough information about the subspaces they belong to, this is $A3$. Notice both $A2$ and $A3$ include $A1$. Assumption $A4$ subsumes assumptions A1-A3 and adds slightly stronger conditions on amount of noise, this allows us to prove Theorem 2. The correct neighborhood property for sub-clusters stated in Theorem 2 is to the best of our knowledge novel for subspace clustering. An explanation of each assumption is provided at the end of this section.

A1. There exist positive constants T_l and ρ such that

$$T_l^2 \leq \min_{k=1,...,K} Q_{1-\frac{d_{max}}{N_k^{1-\rho}}}, \tag{2}$$

where Q_p denotes the p quantile of $\beta(\frac{1}{2}, \frac{d-1}{2})$.

A2. There exist positive constants $\{g_i\}_{i=1}^2$, η and $\rho \in (0,1)$, such that if we write $T = \frac{4g_2+2g_2^2}{1-g_2} + \frac{1+g_2}{1-g_2} g_1$, the following inequalities hold: (2) with T_l replaced by T, and

$$\sum_{i=1}^{d} \left(g_1^2 - \lambda_i^{(k)2}\right)_+^2 > \sum_{i=1}^{d} \left(g_1^2 - \lambda_i^{(k)2}\right)_-^2, \quad \sum_{i=1}^{d} \left(g_1^2 - \lambda_i^{(k)2}\right)_+ > \sum_{i=1}^{d} \left(g_1^2 - \lambda_i^{(k)2}\right)_-, \tag{3}$$

$$\frac{g_2^2}{D\sigma^2} > 3 + \frac{6}{\eta}, \quad \frac{d}{\log N} \geq (2+2\eta)^2. \tag{4}$$

A3. There exist positive constants T_l, q_0, ρ and t such that the following inequalities hold: (2), $d_{max} > d$ and

$$\frac{(T_l^2 d_{max} - C_2)C_2 - \frac{C_1^2}{4}}{T_l^2 d_{max}} \geq q_0, \tag{5}$$

where

$$C_1 = \left(2 + t\sqrt{\frac{\log N}{d-2}}\right)\sqrt{d_{max}}, \; C_2 = \left(\sqrt{\frac{2d_{max}}{\pi(d-1)}} - 2 - t\sqrt{\frac{\log N}{d-2}}\right)^2 \Big/ 2.$$

A4. There exist positive constants T_l, g, λ, η, q_0, ρ and t such that the following inequalities hold: (2), (3) with g_1 replaced by T_l, (4) with g_2 replaced by g, (5) and

$$\begin{aligned}
&f(d) := \frac{(2g-g^2)\,(d_{max}+1)}{2(1-g)} \cdot \sqrt{\frac{d(1+g)^4}{q_0^2} + \frac{D-d}{\lambda^2}} \leq \frac{1}{2}, \\
&\frac{f(d)\sqrt{d_{max}+1}}{1-f(d)} \cdot \sqrt{\frac{d(1+g)^4\lambda^2}{q_0^2} + D - d} \leq \frac{\lambda(1+g)^2\sqrt{d(d_{max}+1)}}{q_0(1-g)}, \\
&g\sqrt{\frac{d(1+g)^4\lambda^2}{q_0^2} + D - d} \leq \frac{\lambda(1+g)^2\sqrt{d(d_{max}+1)}}{q_0(1-g)}, \\
&\frac{6\lambda(1+g)^2\sqrt{d(d_{max}+1)}}{q_0(1-g)} \leq \sqrt{1-T_l^2}.
\end{aligned} \tag{6}$$

Assumption A1: The inner product is used to measure the distance between data points giving rise to the order statistics of a Beta distribution which is bounded in Assumption *A*1. The lower bound of the order statistics is used to control the separation between different subspaces. The upper bound controls the information carried in each sub-cluster. Mathematically, it implicitly controls the ratio between d and $\log N$. If we write $N_k = 10000$, $N = 10N_k$, $d_{max} = 3d$, $T_l^2 = 0.09$, and $\rho = 0.01$, then it suffices to have $\frac{d}{\log N_k} \leq 5$ for inequality (2). Please see the related derivation in Appendix C.

Assumption *A*2: This is the subspace separation assumption. We use it for the proof of Theorem 1. In Appendix C, we show that SBSC requires most of $\{\lambda_i^{(k)}\}_{i=1,k=1}^{d,\;\;K}$ to be smaller than g_1. This means large g_1 implies an easier clustering problem for SBSC, and vice versa. Throughout this paper we call g_1 the affinity threshold. Note that T is a upper bound

of the affinity threshold g_1, specifically if there was no noise $T = g_1$. From (2) we know that large $\frac{d}{\log N}$ implies a small T and g_1. Therefore, large d makes the clustering problem harder. This agrees with our intuition. Consider the extreme case where the subspaces are orthogonal, $\lambda_i^{(k)} = 0$ $(i = 1, \ldots, d, k = 1, \ldots, K)$, and Equations (3) are naturally true with any positive constant g_1. Finally, the constant g_2 in (4) controls the noise term. From the first condition in (4) we have $\sigma < \frac{g_2}{\sqrt{D}}$.

Assumption $A3$: This guarantees the sub-clusters $\{\mathbf{Y}_{\mathcal{C}_i}\}_{i=1}^n$ are informative. We use it mainly for the proof of Lemma 5. Here C_1 and C_2 are closely related to the permeance statistics (Lerman et al., 2012), which measures how well a set of vectors is scattered across a space. Therefore a large $\frac{d_{max}}{d}$ implies that these vectors are well scattered. If T_l^2 equals to its upper bound in (2), $\rho = 0.01$, $N_k = 100000$, $N = 10N_k$, $d_{max} = 160d$, $t = 0.05$ and we want $q_0 \geq 0.5$, $A3$ requires $\frac{d}{\log N} \leq 5$ [1]

Assumption $A4$: Combines all previous assumptions, with slightly stronger conditions on subspace similarities and noise level; we use it for the proof of Theorem 2. The first three conditions in (6) essentially control the value of g, which in turn controls the magnitude of the norm of noise terms. The last condition in (6) controls the value of the regularization parameter in a distance function that will be defined latter.

3.2.2 Theoretical Properties of SBSC

Two theorems regarding Stage 1 of SBSC are stated and discussed in this section. Theorem 1 states a lower bound on the probability that SBSC satisfies the sub-cluster preserving property. Theorem 2 proves that SBSC has the correct-neighborhood property with high probability. Detailed proofs can be found in Appendix C.

Theorem 1 *Under Assumption A2, SBSC has sub-cluster preserving property with probability at least*

$$1 - \sum_{k=1}^{K} \frac{n_k(N_k - d_{\max})}{d_{\max}(N_k + 1)(N_k^{\rho} - 1)^2} - 2(K-1)ne^{-\epsilon_1^2} - \frac{2N}{N^{\left(1+\frac{\eta}{2+\eta}\right)^2}}, \tag{7}$$

where

$$\epsilon_1 = \min_k \frac{\sum_{i=1}^d \left(g_1^2 - \lambda_i^{(k)2}\right)_+ - \sum_{i=1}^d \left(g_1^2 - \lambda_i^{(k)2}\right)_-}{2\sqrt{\sum_{i=1}^d \left(g_1^2 - \lambda_i^{(k)2}\right)_+^2} + \sqrt{4\sum_{i=1}^d \left(g_1^2 - \lambda_i^{(k)2}\right)_+^2 + 2\sum_{i=1}^d \left(g_1^2 - \lambda_i^{(k)2}\right)_+}}. \tag{8}$$

If the subspaces are orthogonal with each other, i.e. $\{\lambda_i^{(k)}\}_{i=1,k=1}^{d,\ K} = 0$. Equation (8) reduces to

$$\epsilon_1 = \frac{\sqrt{d}}{2 + \sqrt{4 + \frac{2}{g_1^2}}}.$$

1. In this example, $\frac{d_{max}}{d}$ is fairly large. In the numerical section we found it is usually not necessary to choose large d_{max}. A better bound in Corollary 3 might be helpful the bridge the gap between numerical experiments and theoretical guarantee.

This shows ϵ_1 is linear in $\sqrt{d}$ and monotonically increasing in g_1^2. Appendix F establishes general conditions on g_1 and $\{\lambda_i^{(k)}\}_{i=1,k=1}^{d,\ K}$ under which ϵ_1 grows like $\sqrt{d}$. Combining this with Assumption $A2$, we observe that the third term of (7) is small for large N. When the number of data points in the sub-sample n is growing linearly in the log of the number of data points in the entire dataset $\log(N)$ (see Section 4.1.2), the second and fourth terms in Equation (7) go to 0 as N increase. This means (7) is close to 1 for large N.

Next, we use the sub-cluster preserving property established in Theorem 1 to prove the theoretical guarantee for correct neighborhood property (see Definition 3). We use the distance function proposed in You et al. (2016a)

$$\begin{aligned} d(\mathbf{Y}_{\mathcal{C}_i}, \mathbf{Y}_{\mathcal{C}_j}) &= \|\mathbf{Y}_{\mathcal{C}_i} - \mathbf{Y}_{\mathcal{C}_j}(\mathbf{Y}_{\mathcal{C}_j}^T\mathbf{Y}_{\mathcal{C}_j} + \lambda\mathbf{I})^{-1}\mathbf{Y}_{\mathcal{C}_j}^T\mathbf{Y}_{\mathcal{C}_i}\|_F + \\ &\quad \|\mathbf{Y}_{\mathcal{C}_j} - \mathbf{Y}_{\mathcal{C}_i}(\mathbf{Y}_{\mathcal{C}_i}^T\mathbf{Y}_{\mathcal{C}_i} + \lambda\mathbf{I})^{-1}\mathbf{Y}_{\mathcal{C}_i}^T\mathbf{Y}_{\mathcal{C}_j}\|_F, \end{aligned} \tag{9}$$

where $\lambda > 0$ is a regularization parameter. This distance function is used to decide if two sets of points belong to the same cluster. Intuitively, the first term in Equation (9) computes the ℓ_2 norm of the residuals from a ridge regression where $\mathbf{Y}_{\mathcal{C}_i}$ is the response and $\mathbf{Y}_{\mathcal{C}_j}$ is the design matrix. The second term exchanges the roles of $\mathbf{Y}_{\mathcal{C}_i}$ and $\mathbf{Y}_{\mathcal{C}_j}$.

Theorem 2 *Assume sub-cluster preserving property is true for SBSC with probability at least $1 - p_s$, and Assumption A4 is satisfied. Then $\{\mathbf{Y}_{\mathcal{C}_i}\}_{i=1}^n$ have the correct neighborhood property with the distance function* (9) *with probability at least*

$$\begin{aligned} &1 - 4n(n-1)e^{-\epsilon_1^2} - \frac{2n}{N^{t^2/2}} - \sum_{k=1}^{K} n_k \left(\frac{N_k - d_{max}}{d_{max}(N_k+1)(N_k^\rho - 1)^2} + 2(N_k - 1)e^{-\epsilon_2^2} \right) \\ &- \frac{2N}{N^{\left(1+\frac{\eta}{2+\eta}\right)^2}} - p_s, \end{aligned} \tag{10}$$

where ϵ_1 is defined in (8) *with g_1 replaced by T_l and*

$$\epsilon_2 = \frac{\sqrt{d-1} - 1}{2 + \frac{1}{\sqrt{d-1}+1}}. \tag{11}$$

Similarly as before, one can show that Equation (10) goes to 1 with large N. Specifically, from Assumption A2 we know the term $n_k(N_k - 1)e^{-\epsilon_2^2}$ grows linearly in $\frac{n_k}{N^\eta}$.

4. Experimental Results

In this section, we test the performance of SBSC on both synthetic and benchmark datasets. In addition to Algorithm 1, we also consider two modifications of the SBSC algorithm. The SBSC-DSC algorithm uses the optimal direction search algorithm (Rahmani and Atia, 2017) instead of correlations to find neighboring points in Step 2 of Algorithm 1. The SBSC-SSC algorithm uses lasso minimization in Step 2 of Algorithm 1; its numerical results are reported in Appendix E.

The performance of the three versions of SBSC is compared to other state-of-the art algorithms. These include classic subspace clustering method: Sparse Subspace Clustering

(SSC, Elhamifar and Vidal, 2009; You et al., 2016a), Thresholding Subspace Clustering (TSC, Heckel and Bölcskei, 2015), Direction Search Subspace Clustering (DSC, Rahmani and Atia, 2017), Least Square Regression (LSR, Lu et al., 2012), Low-Rank Representation (LRR, Liu et al., 2010), Subspace Clustering by Orthogonal Matching Pursuit (SSC-OMP, You et al., 2016c), Elastic Net Subspace Clustering (ENSC, You et al., 2016b); and sampling based algorithms (Peng et al., 2013): Scalable Sparse Subspace Clustering (SSSC, reported in Appendix E), Scalable Thresholding Subspace Clustering (STSC), Scalable Direction Search (SDSC), Scalable Least Square Regression (SLSR), Scalable Low-Rank Representation (SLRR). To make fair comparisons, where possible we replicated results on our machine. Some results are copied from the original paper due to the unavailability of code.

Throughout this section, we use clustering accuracy, normalized mutual information (NMI) and running time as the metrics for performance evaluation. The formulas for clustering accuracy and NMI are presented in Appendix B. To demonstrate the advantage of using sub-clusters (i.e. borrowing information from the whole dataset) to cluster the data points in the subset. For sampling based algorithms we also report their clustering accuracy on the subset. In the rest of this paper, we call the clustering accuracy on the whole dataset as accuracy, and the clustering accuracy on the subset as accuracy-sub. For randomized algorithms, reported results are averaged over 10 trials. Additional numerical results are presented in Appendix E. The code used to generate these results can be found in the supplementary material.

4.1 Results on Synthetic Dataset

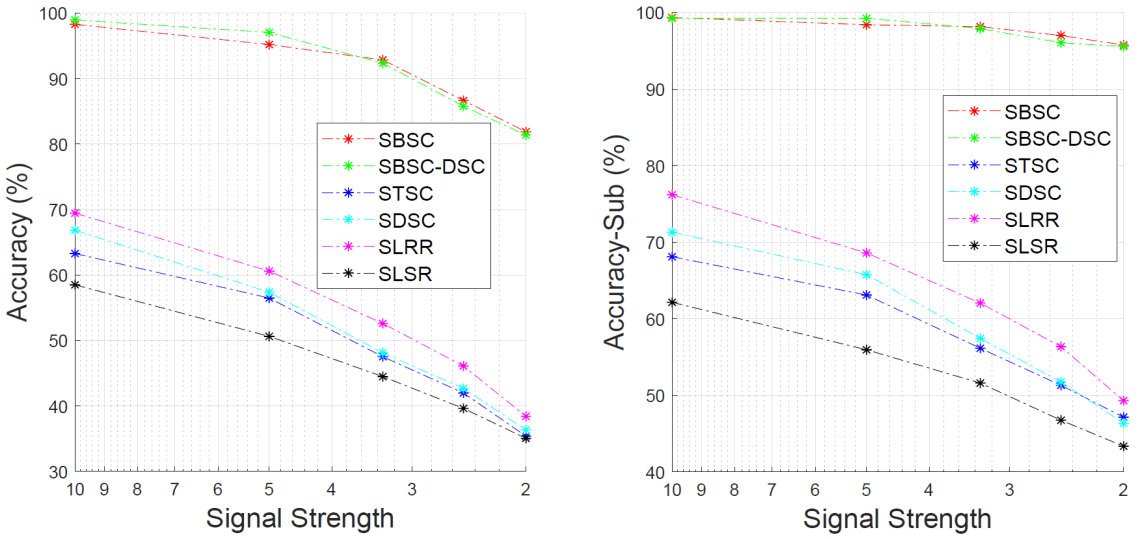
In this section we evaluate tolerance to noise and scalability on synthetic data generated using the model specified in Section 3.1.

4.1.1 Tolerance to Noise

In this section, we test the tolerance to noise of the various algorithms. From (1) we know the un-normalized signal part $\mathbf{U}_k\mathbf{a}_i^{(k)}$ has unit ℓ_2 norm, and the expected squared norm of the noise is $\sigma^2\mathbb{E}[\|\mathbf{e}_i^{(k)}\|_2^2] = \sigma^2 Dd/(d-2)$. Therefore throughout this paper we define $(d-2)/(\sigma^2 Dd)$ as the signal strength (signal to noise ratio). The noise captured the amount of variation of points in $\mathbb{R}^D$.

We change the signal strength from 10 to 2. For each value of signal strength, we simulate 10 datasets with $K = 20$ subspaces, where each subspace contains $N_i = 10000$ data points. For all the sampling based algorithms we fixed the sampling size as $n = 200$.

The results are presented in Figure 1. Accuracy and accuracy-sub are plotted on the left and right hand side panels respectively. The small discrepancy between two sides shows both sampling based algorithms can deliver consistent results between in-sample clustering and out-of-sample classification. At the same time, the SBSC based algorithms constantly deliver much higher accuracy-sub than the other sampling based algorithms, this means for the synthetic datasets, borrowing information from the whole dataset significantly enhanced the clustering results for subset.



(a) (b)

Figure 1: **Tolerance to Noise:** Plot of accuracy (left panel) and accuracy on subsets (right panel) for algorithms applied to synthetic datasets of Section 4.1.1. The x-axis is the signal strength and the y-axis is the accuracy averaged over 10 different datasets. SBSC performs very well.

4.1.2 Scalability

In this section, we test the scalability of SBSC. Specifically, we randomly generate $K = 20$ subspaces in an ambient space with dimension $D = 30$, each of the subspaces has dimension $d = 5$. We increase N_k from 100 to 51200, so the corresponding N increases from 2000 to 1024000. The sampling size n is $\lfloor 2K \log(N) \rfloor$.

The result is presented in Figure 2. On the right hand side y-axis, we show the average accuracy, which is around 95% across all experiments, against the number of data points N, this could justify our choice of n. On the left hand side y-axis, we show the scale plot between running time and N, the linear pattern here agrees with our complexity analysis. As we increase the number of data points N, the accuracy on the whole dataset slightly gets higher, this implies our algorithm is particularly useful for large datasets.

4.2 Results on Benchmark Datasets

In this section, we test SBSC on three benchmark datasets. These datasets were selected to have small, medium and large data size respectively. As expected, the advantage of SBSC over other state-of-the-art algorithms changes from marginal to significant as the size of the dataset increases.

4.2.1 The Extended Yale B dataset

The Extended Yale B dataset (YaleB) contains $N = 2432$ face images of $K = 38$ individuals. Each image is a front view photo of the corresponding individual with different illumination

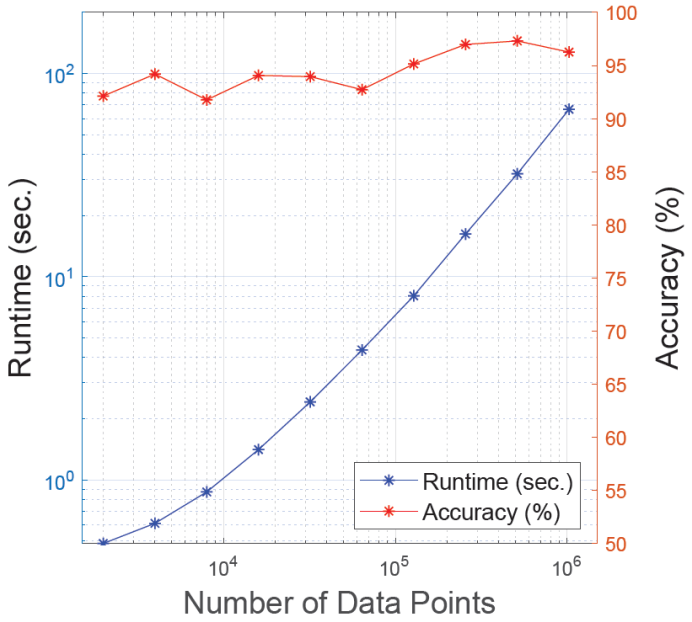


Figure 2: **Scalability:** Plot of running time (blue curve) and accuracy (red curve), averaged over 10 independent datasets, versus the number of data points for SBSC applied to the data of Section 4.1.2. Shows that the algorithm scales well.

condition. To speed up the running time, a dimension reduction step is taken to pre-process the dataset (Rahmani and Atia, 2017), hence in our experiment $D = 500$.

From Table 1, we see that the DSC algorithm delivered the highest accuracy and NMI. As expected, for this small dataset sampling based algorithms did not perform as well. The reason is, there is not enough observations to form sufficiently large homogeneous sub-clusters. This issue is worse for datasets with large number of clusters.

4.2.2 The Zipcode dataset

The Zipcode dataset (LeCun et al., 1990) is a medium-size dataset with $N = 9298$ data points and $D = 256$, each point represents an image of handwritten digit, hence $K = 10$.

From Table 2 we see, that SBSC delivers the best results in all metrics except running time. However the differences in running time are marginal for sampling based algorithms. The accuracy-sub of SBSC is again better than that of traditional sampling based algorithms (see SBSC versus STSC, and SBSC-DSC versus SDSC).

4.2.3 The MNIST dataset

The MNIST dataset (MNIST) contains $N = 70000$ data points, each point represents an image of handwritten digit. The original data was transferred into $\mathbb{R}^{500}$ by convolutional neural network and PCA (You et al., 2016c). Again $K = 10$.

From Table 3 we see, that SBSC with bagging described in Appendix A.2 dominates in nearly every aspect. The large data size of MNIST makes the sampling based algorithms

Method	Accuracy (%)	Accuracy-Sub (%)	NMI (%)	Runtime (sec.)
SBSC	26.53 (1.8)	31.56 (1.94)	41.67 (1.44)	27
SBSC-DSC	60.46 (1.62)	**62.76 (1.88)**	70.15 (0.97)	35
STSC	17.45 (1.27)	21.72 (1.7)	29.17 (1.28)	**0.8**
SDSC	52.18 (1.96)	60.78 (2.04)	54.61 (1.85)	3
SLRR	18.35 (0.64)	28.6 (1.64)	26.33 (0.57)	7
SLSR	26.48 (1.98)	37.78 (2.33)	35.21 (2.22)	1.5
TSC	26.19	NA	39.31	1
DSC	**91.69**	NA	**93.43**	45
SSC	52.96	NA	60.15	169
SSC-OMP	73.88	NA	80.1	1.51
SSC-ENSC	60.81	NA	69.4	3

Table 1: **Performance on Extended Yale B:** The results of sampling based algorithms are averaged over 10 independent runs and the corresponding standard deviations are presented in parentheses. The remaining algorithms do not use random subsampling and were run only once. The metric reported as "NA" is not defined for these algorithms. The best result of each performance metric is in **bold**. DSC delivers the highest accuracy and NMI.

run much faster than traditional methods. Due to their slow speed we did not use bagging on non-sampling algorithms.

5. Conclusion and Future Research

While the idea of subsampling was discussed before (Peng et al., 2015), the main contribution of this paper is finding neighborhood points in the whole dataset and using cluster-wise distance to cluster points in the subset. This results in a higher clustering accuracy.

In calculating cluster-wise distances and classifying out-of-sample points, ridge regression seems to be the most direct method. However, the algorithm itself is highly flexible. Users are encouraged to try different distance functions, classification methods, and metrics in finding neighboring points.

We propose the following directions for future research:

1. In this paper we select the subsamples that are used for initial clustering uniformly at random. It would be interesting to investigate if selecting these points using a more deterministic method such as Coresets (Agarwal et al., 2005) or a quasi-random method such as a Langevin based method (Roberts et al., 1996) could improve the performance of the algorithm.

Method	Accuracy (%)	Accuracy-Sub (%)	NMI (%)	Runtime (sec.)
SBSC	**69.4 (5.17)**	**72.04 (5.37)**	70.3 (1.75)	10
SBSC-DSC	60.84 (2.87)	64.92 (3.37)	62.92 (0.65)	71
STSC	55.28 (4.25)	60.86 (3.8)	53.1 (2.49)	**2**
SDSC	45.62 (6.43)	51.16 (7.31)	45.99 (3.88)	3
SLRR	63.21 (3.96)	65.16 (4.03)	66.09 (1.39)	10
SLSR	58.66 (0.99)	59.85 (0.98)	62.54 (1.38)	4
TSC	65.73	NA	**78.97**	115
DSC	60.92	NA	68.43	800
SSC	48.16	NA	52.37	2165
SSC-OMP	23.58	NA	25.51	3
SSC-ENSC	44.65	NA	50.08	36

Table 2: **Performance on Zipcode:** See the caption of Table 1 for description. We see SBSC delivers the highest accuracy and accuracy-sub, while TSC delivers the highest NMI.

2. While the correct neighborhood property developed in Theorem 2 assures sub-clusters from the same subspaces are close to each other. It would be interesting to further explore, from the theoretical perspective, the impacts of this property on clustering accuracy. For example, it would be interesting to explore the relationship between the identifiability discussed in Aragam et al. (2020) and the correct neighborhood property in this paper.

3. It would be interesting to extend our algorithm to non-linear manifold clustering problems. For example, one could project the dataset into a RKHS and apply our algorithm on the projected features.

Appendix A. Practical Recommendations for Parameter Setting

In Algorithm 1 (SBSC), we assume the number of clusters is known. Several methods have been developed for the estimation of the number of clusters from data (Ng et al., 2002). Intuitively, n should be large enough to represent the structure of the whole dataset while still being relatively small to reduce the computational complexity. In our numerical experiments, we choose n to be linear in $K \log N$.

Ideally, each sub-cluster $\mathbf{Y}_{\mathcal{C}_i}$ should well represent the subspace it belongs to, i.e., contain at least one basis of that subspace. Therefore we want $d_{\max}$ to be larger than $\max_{k=1,\dots,K} d_k$ which is unknown. For this reason we set $d_{\max}$ to grow linearly with D. Similarly the residual minimization parameter m should also be linear in D.

Method	Accuracy (%)	Accuracy-Sub (%)	NMI (%)	Runtime (sec.)
SBSC(1)	95.74 (0.28)	**96.44 (1.14)**	89.9 (0.47)	38
SBSC(6)	**97.15 (0.16)**	95.25 (1.78)	**92.6 (0.3)**	246
STSC(1)	30.2 (2.13)	67.8 (3.95)	11.52 (2.12)	**28**
STSC(6)	40.12 (2.84)	65.23 (2.2)	22.53 (2.36)	172
SLRR(1)	79.5 (1.19)	79.46 (1.3)	79.9 (1.52)	59
SLRR(6)	81 (0.67)	79.6 (0.44)	83.75 (0.74)	378
SLSR(1)	75.06 (6.11)	74.62 (5.99)	76.21 (3.63)	54
SLSR(6)	79.64 (0.85)	76.43 (1.85)	81.24 (0.95)	326
TSC	84.63	NA	87.47	1184
SSC (DC1)*	96.55	NA	NA	5254
SSC (DC2)*	96.1	NA	NA	4390
SSC (DC5)*	94.9	NA	NA	1596
SSC-OMP	81.51	NA	84.45	232
SSC-ENSC	93.79	NA	88.8	500

Table 3: **Performance on MNIST:** See description in Table 1. Additionally, the results of methods with star marks are copied from the original paper that did not report NMI. The number in the parenthesis next to the algorithm name is the number of bags. We see SBSC dominates other algorithms in nearly every aspect.

A.1 Threshold Selection

The spectral clustering algorithm can deliver exact clustering result (Von Luxburg, 2007) if the graph induced by the affinity matrix ($\mathbf{D}+\mathbf{D}^T$) has no false connections; and has exactly K connected components. For a large threshold parameter $t_{\max}$ on the affinity matrix more entries in $\mathbf{D}$ will be kept and our algorithm is more likely to have false connections, while small $t_{\max}$ eliminates false connections but might incur non-connectivity.

Let us consider a heuristic situation: the subset we sampled contains exactly the same points (hence $\frac{n}{K}$ points) for each cluster. Then if we choose the threshold index $t_{\max}$ to be $\frac{n}{2K}$, the induced graph from our affinity matrix will have no false connection (given that points from same subspace have bigger similarities between each other) and the clusters themselves will be connected, therefore the spectral clustering algorithm will deliver the exact clustering result (Luxburg et al., 2005).

In reality clusters do not usually have same points in $\hat{\mathbf{Y}}$, hence we choose $t_{\max}$ to start from a relatively large number $\frac{n}{0.5K}$ and gradually increase it. Based on different threshold values, we can generate different label vectors on the subset $\hat{\mathbf{Y}}$, intuitively label vectors

that can deliver highly accurate results should be similar to each other or stable. Based on this intuition, we developed a simple adaptive algorithm for finding an "optimal" affinity threshold $t_{\max}$; see supplementary code for details. Based on our observation, choosing t_{max} adaptively works well with datasets where each cluster has large amount of points.

A.2 Combining Runs of the Algorithm

Thanks to the speed of our algorithm, we can conduct several independent runs for one experiment (for sampling based algorithms, the results between independent runs might be different) with acceptable running time. In order to make full use of such advantage, we designed an algorithm to combine the results via bagging from several runs of SBSC (Breiman, 1996). Unlike the classification problem, we need to conduct label switching, see Algorithm 2 for details on how label switching is addressed. Please note that bagging can be used for any clustering algorithms. In Section 4 and Appendix E we report the results, both with and without bagging, for all sampling based algorithms.

input : The label vectors $\{\mathbf{l}_j\}_{j=1}^b \in \mathbb{R}^N$ from b independent runs. The number of clusters K. Note that each entry of $\mathbf{l}_j$ is a positive integer from 1 to K.
output: The final label vector $\mathbf{l}_0 \in \mathbb{R}^N$.
for $m = 1$ **to** b **do**
  Write $\mathcal{M}_j = \{r : \mathbf{l}_m(r) = j\}$, $j = 1, ..., K$.
  for $i = 1$ **to** b ***and*** $i \neq m$ **do**
    1. Write $\mathcal{I}_q = \{r : \mathbf{l}_i(r) = q\}$, $q = 1, ..., K$. Let $\mathbf{S} \in \mathbb{R}^{K \times K}$ be a score matrix where

$$[\mathbf{S}]_{jq} = \frac{\mathbf{card}(\mathcal{M}_j \cap \mathcal{I}_q)}{\min(\mathbf{card}(\mathcal{M}_j),\ \mathbf{card}(\mathcal{I}_q))},\ 1 \leq j, q \leq K.$$

    2. Switch the labels in $\mathbf{l}_i$ based on score matrix $\mathbf{S}$:
    for $k = 1$ **to** K **do**
      Let $q = \arg\max_j [\mathbf{S}]_{jk}$. For $\forall r \in \mathcal{I}_q$ set $\mathbf{l}_i(r) := k$.
    end
  end
end
for $n = 1$ **to** N **do**
  Set $\mathbf{l}_0(n) := \mathbf{mode}(\{\mathbf{l}_j(n)\}_{j=1}^b)$.
end

Algorithm 2: Bagging of Clustering Labels

The Step 2 of Algorithm 2 has a subtle issue: there might exists an integer q such that $q = \arg\max_j [\mathbf{S}]_{jk} = \arg\max_j [\mathbf{S}]_{jr}$. Please see our supplementary codes on how to tackle this problem.

Appendix B. Clustering Accuracy and Normalized Mutual Information

The clustering accuracy measures the percentage of correctly labeled data points (You et al., 2016c). It is calculated by

$$accr = \max_{\pi} \frac{100}{N} \sum_{i,j} l^{est}_{\pi(i)j} l^{true}_{ij}, \ 1 \le i \le K, \ 1 \le j \le N.$$

Here π is the permutation of K labels. The estimated label indicator $l^{est}_{\pi(i)j}$ equals to 1 if and only if we assign label $\pi(i)$ to the j-th point, and 0 otherwise. The ground-truth label indicator l^{true}_{ij} equals to 1 if and only if the j-th point has label i, and 0 otherwise.

The normalized mutual information (Strehl and Ghosh, 2002) is calculated by

$$\mathrm{NMI}(\mathbf{l}^{est}, \mathbf{l}^{true}) = \frac{I(\mathbf{l}^{est}, \mathbf{l}^{true})}{\sqrt{H(\mathbf{l}^{est}) H(\mathbf{l}^{true})}}.$$

Here $\mathbf{l}^{est}$ and $\mathbf{l}^{true}$ are estimated/ground-truth label vectors, respectively. We use $I(\mathbf{l}^{est}, \mathbf{l}^{true})$ to denote the mutual information between $\mathbf{l}^{est}$ and $\mathbf{l}^{true}$, and $H(\mathbf{l}^{est})$ to denote the entropy of $\mathbf{l}^{est}$. Similarly for $H(\mathbf{l}^{true})$.

Appendix C. Proofs of Main Theorems

In this section, we will prove the theorems from Section 3. The following Lemmas are used to prove Theorem 1.

Lemma 1 *Let* $\mathbf{b}$ *be a vector sampled uniformly from* $\mathbb{S}^{d-1}$*, and* λ_k $(k = 1, .., d)$ *be constants such that* $1 \ge \lambda_1 \ge \lambda_2 \ge ... \ge \lambda_d \ge 0$*. For constant* $g_1 \in (\lambda_d, \lambda_1)$*, we write* $r_i = (g_1^2 - \lambda_i^2)_+$ *and* $s_i = (g_1^2 - \lambda_i^2)_-$*. Assuming that* $\sum_{i=1}^d r_i > \sum_{i=1}^d s_i$*, then*

$$\mathbb{P}\left[\sum_{i=1}^{d} (\lambda_i b_i)^2 < g_1^2\right] \ge 1 - 2e^{-\epsilon^2},$$

where

$$\epsilon = \frac{\sum_{i=1}^d (r_i - s_i)}{\left(\sqrt{\sum_{i=1}^d r_i^2} + \sqrt{\sum_{i=1}^d s_i^2}\right) + \sqrt{\left(\sqrt{\sum_{i=1}^d r_i^2} + \sqrt{\sum_{i=1}^d s_i^2}\right)^2 + 2s_1 \sum_{i=1}^d (r_i - s_i)}}.$$

Proof We write $b_i = \frac{z_i}{\sqrt{\sum_{j=1}^d z_j^2}}$, where $\{z_i\}_{i=1}^d$ are i.i.d. $\mathcal{N}(0,1)$ random variables. The goal is to bound

$$\mathbb{P}\left[\sum_{i=1}^{d} \left(g_1^2 - \lambda_i^2\right)_- \cdot z_i^2 \ge \sum_{i=1}^{d} \left(g_1^2 - \lambda_i^2\right)_+ \cdot z_i^2\right] = \mathbb{P}\left[\sum_{i=1}^{d} s_i \cdot z_i^2 \ge \sum_{i=1}^{d} r_i \cdot z_i^2\right].$$

Note that $g_1 \in (\lambda_d, \lambda_1)$, hence both $\sum_{i=1}^d r_i$ and $\sum_{i=1}^d s_i$ are strictly positive.

Now we write $X = \sum_{i=1}^{d} s_i \cdot z_i^2$ and $Y = \sum_{i=1}^{d} r_i \cdot z_i^2$. Applying Lemma 1 in Laurent and Massart (2000) we have for positive constants ϵ_a and ϵ_b the following inequalities are true

$$\mathbb{P}\left[X \geq \sum_{i=1}^{d} s_i + 2\sqrt{\sum_{i=1}^{d} s_i^2 \epsilon_a + 2s_1 \epsilon_a^2}\right] \leq e^{-\epsilon_a^2}, \quad \mathbb{P}\left[Y \leq \sum_{i=1}^{d} r_i - 2\sqrt{\sum_{i=1}^{d} r_i^2 \epsilon_b}\right] \leq e^{-\epsilon_b^2}.$$

We set $\epsilon_a = \epsilon_b$ and

$$\sum_{i=1}^{d} s_i + 2\sqrt{\sum_{i=1}^{d} s_i^2 \epsilon_a + 2s_1 \epsilon_a^2} = \sum_{i=1}^{d} r_i - 2\sqrt{\sum_{i=1}^{d} r_i^2 \epsilon_b}.$$

Solving the above quadratic equation we have

$$\epsilon_a = \epsilon_b = \frac{\sum_{i=1}^{d} (r_i - s_i)}{\left(\sqrt{\sum_{i=1}^{d} r_i^2} + \sqrt{\sum_{i=1}^{d} s_i^2}\right) + \sqrt{\left(\sqrt{\sum_{i=1}^{d} r_i^2} + \sqrt{\sum_{i=1}^{d} s_i^2}\right)^2 + 2s_1 \sum_{i=1}^{d} (r_i - s_i)}}.$$

Consequently

$$\begin{aligned}\mathbb{P}[X \geq Y] &\leq \mathbb{P}\left[X \geq \sum_{i=1}^{d} s_i + 2\sqrt{\sum_{i=1}^{d} s_i^2 \epsilon_a + 2s_1 \epsilon_a^2}\right] + \mathbb{P}\left[Y \leq \sum_{i=1}^{d} r_i - 2\sqrt{\sum_{i=1}^{d} r_i^2 \epsilon_b}\right] \\ &\leq e^{-\epsilon_a^2} + e^{-\epsilon_b^2}.\end{aligned}$$

Substituting ϵ_a and ϵ_b into the inequality above yields the result. ■

The following bound on F-distributed random variables follows from Lemma 1.

Corollary 1 *Let $X \sim F(m, n)$, and $m, n \geq 2$. Then for constant $q > 1$, we have*

$$\mathbb{P}[X \geq q] \leq 2e^{-\epsilon^2},$$

where $\epsilon = \frac{1}{2}\left[-\left(\sqrt{m} + \frac{qm}{\sqrt{n}}\right) + \sqrt{\left(\sqrt{m} + \frac{qm}{\sqrt{n}}\right)^2 + 2m(q-1)}\right]$.

Proof We write $b_i = \frac{z_i}{\sum_{i=1}^{m+n} z_i^2}$, and $X = \frac{(\sum_{i=1}^{m} z_i^2)/m}{(\sum_{i=m+1}^{m+n} z_i^2)/n}$, where $\{z_i\}_{i=1}^{m+n}$ are i.i.d. $\mathcal{N}(0,1)$ random variables. Then we have

$$\mathbb{P}[X \geq q] = \mathbb{P}\left[\sum_{i=1}^{m} \frac{1}{mq} \cdot z_i^2 \geq \sum_{i=m+1}^{m+n} \frac{1}{n} \cdot z_i^2\right].$$

The corollary follows by selecting $\lambda_i = \sqrt{\frac{1}{2} + \frac{1}{mq}}$ for $i = 1, ..., m$, $\lambda_i = \sqrt{\frac{1}{2} - \frac{1}{n}}$ for $i = m+1, ..., m+n$, and $g_1 = \sqrt{\frac{1}{2}}$ in Lemma 1. ■

Lemma 2 states a bound on the order statistics of Beta distributed random variables.

Lemma 2 *Suppose T_l satisfy Assumption A1. For any $k = 1, ..., K$, let $\{B_{(i)}\}_{i=1}^{N_k-1}$ be the order statistics from a sample of $(N_k - 1)$ i.i.d $\beta(\frac{1}{2}, \frac{d-1}{2})$ random variables, then*

$$\mathbb{P}\left[B_{(N_k-1)} \geq \frac{1}{2}\right] \leq 2(N_k - 1)e^{-{\epsilon_2}^2},$$

and

$$\mathbb{P}\left[B_{(N_k-d_{\max})} \leq T_l^2\right] \leq \frac{(N_k - d_{\max})}{d_{\max}\left(N_k + 1\right)\left(N_k^{\rho} - 1\right)^2}.$$

Here ϵ_2 is defined in (11).

Proof Let $B \sim \beta(\frac{1}{2}, \frac{1}{d-1})$. Then we can write $B = \frac{z_1^2}{\sum_{i=1}^{d} z_i^2}$, where $\{z_i\}_{i=1}^{d}$ are i.i.d. $\mathcal{N}(0, 1)$ random variables. Select $\lambda_1 = 1$, $\lambda_i = 0$ $(i = 2, .., d)$ and and $g_1 = \frac{1}{\sqrt{2}}$ in Lemma 1. Note the following fact

$$\epsilon_2 = \frac{\sqrt{d-1} - 1}{2 + \frac{1}{\sqrt{d-1}+1}} \leq \frac{d-2}{\left(\sqrt{d-1} + 1\right) + \sqrt{\left(\sqrt{d-1} + 1\right) + 2(d-2)}}.$$

From Lemma 1 we have $\mathbb{P}\left[B \geq \frac{1}{2}\right] \leq 2e^{-{\epsilon_2}^2}$. Therefore by union bound inequality we have

$$\mathbb{P}\left[B_{(N_k-1)} \geq \frac{1}{2}\right] \leq 2(N_k - 1)e^{-{\epsilon_2}^2}.$$

This proves the first part of Lemma 2.

Next we prove the second part of Lemma 2. Let $U_{(i)} = F_{(\frac{1}{2}, \frac{d-1}{2})}(B_{(i)})$, here $F_{(\frac{1}{2}, \frac{d-1}{2})}$ is the CDF of the Beta distribution $\beta(\frac{1}{2}, \frac{d-1}{2})$. Note that $\{U_{(i)}\}_{i=1}^{N_k-1}$ are the order statistics of the uniform distribution.

From Assumption *A*1 we know $F_{(\frac{1}{2}, \frac{d-1}{2})}(T_l^2) \leq 1 - \frac{d_{\max}}{N_k^{1-\rho}}$ and hence

$$\mathbb{P}\left[B_{(N_k-d_{\max})} \leq T_l^2\right] \leq \mathbb{P}\left[U_{(N_k-d_{\max})} \leq 1 - \frac{d_{\max}}{N_k^{1-\rho}}\right]. \tag{12}$$

By Chebyshev's inequality and basic properties of the uniform order statistics we have

$$\mathbb{P}\left[U_{(N_k-d_{\max})} \leq 1 - \frac{d_{\max}}{N_k^{1-\rho}}\right] \leq \frac{Var\left[U_{(N_k-d_{\max})}\right]}{\left(\frac{d_{\max}}{N_k} - \frac{d_{\max}}{N_k^{1-\rho}}\right)^2} = \frac{(N_k - d_{\max})}{d_{\max}(N_k + 1)(N_k^{\rho} - 1)^2}. \tag{13}$$

Combine (12) and (13) we know

$$\mathbb{P}\left[B_{(N_k-d_{\max})} \leq T_l^2\right] \leq \frac{(N_k - d_{\max})}{d_{\max}(N_k + 1)(N_k^{\rho} - 1)^2}.$$

This completes the proof. ■

Lemma 3 to Lemma 5 are used to prove Theorem 2.

Lemma 3 *Let* $\mathbf{v}$ *be a random vector that uniformly distributed on* $\mathbb{S}^{d-1}$*. Then we can decompose* $\mathbf{v}$ *into* $\mathbf{v} = \left[\sqrt{g}s, \sqrt{1-g}\mathbf{u}\right]$*, where* $g \sim \beta(\frac{1}{2}, \frac{d-1}{2})$*,* $\mathbf{u} \sim U(\mathbb{S}^{d-2})$ *and* $\mathbb{P}[s=1] = \mathbb{P}[s=-1] = 0.5$ *are three independent random variables.*

Proof It is straightforward to see $\langle \mathbf{v}, \mathbf{v} \rangle = [v_1^2, ..., v_d^2]$ follows the Dirichlet distribution with parameters $\boldsymbol{\alpha} = (\frac{1}{2}, ..., \frac{1}{2}) \in \mathbb{R}^d$. We can decompose $\langle \mathbf{v}, \mathbf{v} \rangle$ into the following concatenation of two random components

$$\left[v_1^2, ..., v_d^2\right] = \left[v_1^2, (1-v_1^2)\frac{\langle \mathbf{v}_{-1}, \mathbf{v}_{-1} \rangle}{1-v_1^2}\right].$$

Since Dirichlet distribution is completely neutral (Lin, 2016), we know that v_1^2 is independent of $\frac{\langle \mathbf{v}_{-1}, \mathbf{v}_{-1} \rangle}{1-v_1^2}$, where $v_1^2 \sim \beta(\frac{1}{2}, \frac{d-1}{2})$ and $\frac{\langle \mathbf{v}_{-1}, \mathbf{v}_{-1} \rangle}{1-v_1^2} \sim Dir(\boldsymbol{\alpha}_{-1})$. From symmetry, we can set $\sqrt{g}s := v_1$ and $\mathbf{u} := \frac{\mathbf{v}_{-1}}{\sqrt{1-v_1^2}}$, where the distributions of g, $\mathbf{u}$ and s are specified in the statement of Lemma 3. This completes the proof.

■

Let $\{\mathbf{a}_i\}_{i=1}^{N_k-1}$ be $(N_k - 1)$ vectors that are uniformly sampled from $\mathbb{S}^{d-1}$. From Lemma 3, we know that for any $i = 1, ..., N_k - 1$, the value of a_{i1} is independent of $\frac{[a_{i2}, .., a_{id}]}{\sqrt{1-a_{i1}^2}}$. The following corollary is then a direct result from this fact.

Corollary 2 *Let* $\{\mathbf{a}_{(i)}\}_{i=1}^{N_k-1}$ *be a permutation of* $\{\mathbf{a}_i\}_{i=1}^{N_k-1}$ *sorted in ascending order of the absolute value of the first coordinate. Then we can write*

$$\mathbf{a}_{(i)} = \left[a_{(i)1}, \sqrt{1-a_{(i)1}^2}\mathbf{b}_{N_k-i}\right],$$

where $\{\mathbf{b}_i\}_{i=1}^{N_k-1}$ *are i.i.d. uniform samples on* $\mathbb{S}^{d-2}$*.*

Lemma 4 *(Lerman et al., 2012, Lemma B.3) Let* $\{\mathbf{b}_i\}_{i=1}^{d_{max}}$ *be i.i.d. uniform samples from* $\mathbb{S}^{d-2}$*,* $d \geq 3$*. Then for any* $t \geq 0$

$$\inf_{\|\mathbf{u}\|_2=1} \sum_{i=1}^{d_{max}} |\langle \mathbf{u}, \mathbf{b}_i \rangle| \geq \sqrt{\frac{2}{\pi}}\frac{d_{max}}{\sqrt{d-1}} - 2\sqrt{d_{max}} - t\sqrt{\frac{d_{max}}{d-2}},$$

with probability at least $1 - e^{-t^2/2}$*.*

Corollary 3 *Use the same definition of* $\{\mathbf{b}_i\}_{i=1}^{d_{max}}$ *from Lemma 4. Then for any* $t \geq 0$*:*

$$\sup_{\|\mathbf{u}\|_2=1} \sum_{i=1}^{d_{max}} \langle \mathbf{u}, \mathbf{b}_i \rangle \leq 2\sqrt{d_{max}} + t\sqrt{\frac{d_{max}}{d-2}},$$

with probability at least $1 - e^{-t^2/2}$*.*

Proof Note that $\mathbb{E}[\langle \mathbf{u}, \mathbf{b} \rangle] = 0$ for any $\mathbf{b} \sim U(\mathbb{S}^{d-2})$ and $\mathbf{u} \in \mathbb{R}^{d-1}$. Therefore by Lemma 6.3 in Ledoux and Talagrand (2013) we have:

$$\mathbb{E}\left[\sup_{\|\mathbf{u}\|_2=1} \sum_{i=1}^{d_{max}} \langle \mathbf{u}, \mathbf{b}_i \rangle\right] \leq 2 \sup_{\|\mathbf{u}\|_2=1} \left[\mathbb{E}\left\|\sum_{i=1}^{d_{max}} \epsilon_i \mathbf{b}_i\right\|^2\right] = 2\sqrt{d_{max}}.$$

Here $\{\epsilon_i\}_{i=1}^{d_{max}}$ are i.i.d. Rademacher random variables. The lemma is proved by following similar steps after equation (B.11) in Lerman et al. (2012). ■

Lemma 5 *Suppose Assumption A3. Write* $\mathbf{a}_0 = [1, 0, ..., 0] \in \mathbb{R}^d$, *and use the definitions for* $\{\mathbf{a}_i\}_{i=1}^{N_k-1}$ *and* $\{\mathbf{a}_{(i)}\}_{i=1}^{N_k-1}$ *from Corollary 2. Let* $\mathbf{B} \in \mathbb{R}^{d\times(d_{max}+1)}$ *be a matrix where its first column is* $\mathbf{a}_0$ *and its i-th column* ($2 \le i \le d_{max}+1$) *is* $\mathbf{a}_{(N_k-i+1)}$. *Let the largest d singular values of* $\mathbf{B}$ *be* $s_1 \ge s_2 \ge \cdots \ge s_d$. *Then we have*

$$\mathbb{P}\left[s_d^2 \ge q_0\right] \ge 1 - \frac{2}{N^{t^2/2}} - \frac{(N_k - d_{\max})}{d_{\max}(N_k+1)(N_k^\rho - 1)^2} - 2(N_k-1)e^{-{\epsilon_2}^2},$$

where ϵ_2 *is defined in* (11).

Proof From Corollary 2, we know $\mathbf{B}$ can be re-written as

$$\mathbf{B} = \begin{pmatrix} 1, & a_{(N_k-1)1}, & ... & a_{(N_k-d_{max})1} \\ \mathbf{0}, & \sqrt{1-a_{(N_k-1)1}^2}\mathbf{b}_1, & ... & \sqrt{1-a_{(N_k-d_{max})1}^2}\mathbf{b}_{d_{max}} \end{pmatrix},$$

where $\{\mathbf{b}_i\}_{i=1}^{d_{max}}$ are i.i.d. uniform samples from $\mathbb{S}^{d-2}$.

Given the dimensions of $\mathbf{B}$, we know $s_d = \inf_{\|\mathbf{x}\|_2=1} \left\|\mathbf{B}^T\mathbf{x}\right\|_2$. For convenience, we write

$$\mathbf{x}' = \frac{1}{\sqrt{1-x_1^2}}\left[x_2, ..., x_d\right],$$

where $\|\mathbf{x}'\|_2 = 1$, $c_i = \langle \mathbf{x}', \mathbf{b}_i\rangle$, $a_{(N_k)1} = 1$. Let $\mathcal{E}_1$ be the event that $\{s_d^2 \ge q_0\}$, and $\mathcal{E}_2$ be the event that $\{a_{(N_k-i)1}^2 \in \left[T_l^2, \frac{1}{2}\right], \forall i = 1, .., d_{max}\}$. From Lemma 2 we know

$$\mathbb{P}\left[\mathcal{E}_2\right] \ge 1 - \frac{(N_k - d_{\max})}{d_{\max}(N_k+1)(N_k^\rho - 1)^2} - 2(N_k-1)e^{-{\epsilon_2}^2}. \tag{14}$$

Conditioning on $\mathcal{E}_2$, we have the following relations

$$\begin{aligned}
\left\|\mathbf{B}^T\mathbf{x}\right\|_2^2 =& \left\| \begin{pmatrix} 1, & a_{(N_k-1)1}, & ... & a_{(N_k-d_{max})1} \\ \mathbf{0}, & \sqrt{1-a_{(N_k-1)1}^2}\mathbf{b}_1, & ... & \sqrt{1-a_{(N_k-d_{max})1}^2}\mathbf{b}_{d_{max}} \end{pmatrix}^T \mathbf{x} \right\|_2^2 \\
=& \sum_{i=0}^{d_{max}} \left(a_{(N_k-i)1}x_1 + \sqrt{\left(1-a_{(N-i)1}^2\right)\left(1-x_1^2\right)}c_i \right)^2 \\
=& \sum_{i=0}^{d_{max}} a_{(N-i)1}^2 x_1^2 + 2\sum_{i=0}^{d_{max}} \sqrt{a_{(N-i)1}^2(1-a_{(N_k-i)1}^2)(1-x_1^2)}c_i x_1 \\
&+ \sum_{i=1}^{d_{max}} (1-a_{(N_k-i)1}^2)(1-x_1^2)c_i^2 \\
\ge& T_l^2 d_{max}\cdot x_1^2 - \sqrt{(1-x_1^2)x_1^2} \sup_{\|u\|_2=1} \sum_{i=1}^{d_{max}} \langle \mathbf{u}, \mathbf{b}_i\rangle
\end{aligned}$$

$$+ \frac{1-x_1^2}{2} \inf_{\|u\|_2=1} \sum_{i=1}^{d_{max}} \langle \mathbf{u}, \mathbf{b}_i \rangle^2. \tag{15}$$

From Lemma 4 and Corollary 3 and conditional on $\mathcal{E}_2$, we have the following inequality

$$(15) \geq (1-x_1^2)C_2 - \sqrt{(1-x_1^2)x_1^2}C_1 + T_l^2 d_{max} \cdot x_1^2, \tag{16}$$

with probability at least $1 - \frac{2}{N^{t^2/2}}$. Since $1 - x_1^2 \leq 1$, a lower bound of the RHS of (16) is

$$\left(T_l^2 d_{max} - C_2\right) x_1^2 - C_1 x_1 + C_2 \geq \frac{\left(T_l^2 d_{max} - C_2\right) C_2 - \frac{C_1^2}{4}}{T_l^2 d_{max}} \geq q_0,$$

where the q_0 comes from Assumption A3. Finally, note the following fact

$$\begin{aligned} \mathbb{P}\left[\mathcal{E}_1\right] &\geq \mathbb{P}\left[\mathcal{E}_1|\mathcal{E}_2\right] + \mathbb{P}\left[\mathcal{E}_2\right] - 1 \\ &= 1 - \frac{2}{N^{t^2/2}} - \frac{(N_k - d_{\max})}{d_{\max}(N_k+1)(N_k^\rho - 1)^2} - 2(N_k - 1)e^{-{\epsilon_2}^2}. \end{aligned} \tag{17}$$

This completes the proof.

■

Proof of Theorem 1 Let the event $\mathcal{E}_{1i} = \{\mathbf{Y}_{\mathcal{C}_i}$ only contains points in same subspace$\}$, then $\mathcal{E}_1 = \cap_{i=1}^n \mathcal{E}_{1i}$ is the event that Algorithm 1 has sub-cluster preserving property. Let the event $\mathcal{E}_2 = \{\sigma \left\| \mathbf{e}_i^{(k)} \right\|_2 < g_2, \forall i, k\}$, where g_2 is from Assumption $A2$. Our goal is to find a lower bound on $\mathbb{P}\left[\mathcal{E}_1\right]$.

Note the following fact

$$\mathbb{P}\left[\mathcal{E}_1\right] \geq 1 - \sum_{i=1}^n \mathbb{P}\left[\mathcal{E}_{1i}^{\complement} | \mathcal{E}_2\right] + \mathbb{P}\left[\mathcal{E}_2\right] - 1 = \mathbb{P}\left[\mathcal{E}_2\right] - \sum_{i=1}^n \mathbb{P}\left[\mathcal{E}_{1i}^{\complement} | \mathcal{E}_2\right]. \tag{18}$$

Therefore, it suffices to find a lower bound on $\mathbb{P}\left[\mathcal{E}_2\right] - \sum_{i=1}^n \mathbb{P}\left[\mathcal{E}_{1i}^{\complement} | \mathcal{E}_2\right]$.

We start by finding a preliminary upper bound on $\mathbb{P}\left[\mathcal{E}_{11}^{\complement} | \mathcal{E}_2\right]$. WLOG assume that $\mathbf{y_1}^{(1)}$ is one of the sampled points, and $\mathbf{Y}_{\mathcal{C}_1}$ is the sub-cluster associated with it. Recall that in Step 2 of Algorithm 1, we use $|\langle \mathbf{y}_1^{(1)}, \mathbf{y}_i^{(k)} \rangle|$ to measure the affinity between $\mathbf{y}_1^{(1)}$ and $\mathbf{y}_i^{(k)}$, the nearest $(d_{max}+1)$ points are then used to construct the sub-cluster associated with $\mathbf{y}_1^{(1)}$. Write $\hat{A}^k = \{|\langle \mathbf{y}_1^{(1)}, \mathbf{y}_i^{(k)} \rangle|\}_{i=1}^{N_k}$, for $\mathcal{E}_{11}$ to happen we need the largest $(d_{\max}+1)$ values among $\cup_{k=1}^K \hat{A}^k$ to be from the set $\hat{A}^1$. Mathematically this means

$$\mathcal{E}_{11}^{\complement} = \left\{ \hat{A}^1_{(N_1 - d_{\max})} \leq \max_{k \neq 1} \max_{i=1,..,N_k} \hat{A}_i^k \right\},$$

where $\hat{A}_i^k$ is the i-th element in $\hat{A}^k$ and $\hat{A}_{(i)}^k$ is the i-th smallest element in $\hat{A}^k$.

Recall from (1) that $\mathbf{y_i}^{(k)} = \frac{\mathbf{U_k}\mathbf{a}_i^{(k)}+\sigma\mathbf{e}_i^{(k)}}{\left\|\mathbf{U}_k\mathbf{a}_i^{(k)}+\sigma\mathbf{e}_i^{(k)}\right\|_2}$. The triangle inequality tells us that

$$\left\|\mathbf{U}_k\mathbf{a}_i^{(k)}\right\|_2 - \left\|\sigma\mathbf{e}_i^{(k)}\right\|_2 \leq \left\|\mathbf{U}_k\mathbf{a}_i^{(k)}+\sigma\mathbf{e}_i^{(k)}\right\|_2 \leq \left\|\mathbf{U}_k\mathbf{a}_i^{(k)}\right\|_2 + \left\|\sigma\mathbf{e}_i^{(k)}\right\|_2.$$

Therefore conditional on $\mathcal{E}_2$, we know the normalizing constants $\left\|\mathbf{U}_k\mathbf{a}_i^{(k)}+\sigma\mathbf{e}_i^{(k)}\right\|_2$ are bounded in $[1-g_2, 1+g_2]$. We can write $A_i^k = \left\|\mathbf{y}_1^{(1)}\right\|_2 \cdot \left\|\mathbf{y}_i^{(k)}\right\|_2 \cdot \hat{A}_i^k$. It is fairly straightforward to get the following relation

$$\mathbb{P}\left[A^1_{(N_1-d_{\max})} \leq \frac{1+g_2}{1-g_2}\max_{k\neq 1}\max_{1\leq i\leq N_k} A_i^k \middle| \mathcal{E}_2\right] \geq \mathbb{P}\left[\mathcal{E}_{11}^{\complement}\middle|\mathcal{E}_2\right]. \tag{19}$$

Conditioning on $\mathcal{E}_2$ and write $B_i = \langle \mathbf{a}_1^{(1)}, \mathbf{a}_i^{(1)}\rangle^2$, $i = 2, ..., N_1 - 1$. We have the following inequalities

$$\begin{aligned}
A^1_{(N_1-d_{\max})} =& \left|\sqrt{B_{(N_1-d_{\max})}} + \sigma\langle\mathbf{U}_1\mathbf{a}_1^{(1)}, \mathbf{e}_i^{(1)}\rangle + \sigma\langle\mathbf{U}_1\mathbf{a}_i^{(1)}, \mathbf{e}_1^{(1)}\rangle + \sigma^2\langle\mathbf{e}_1^{(1)}, \mathbf{e}_i^{(1)}\rangle\right| \\
\geq& \sqrt{B_{(N_1-d_{\max})}} - \sigma\left\|\mathbf{e}_i^{(1)}\right\|_2 - \sigma\left\|\mathbf{e}_1^{(1)}\right\|_2 - \sigma^2\left\|\mathbf{e}_1^{(1)}\right\|_2 \max_{i\neq 1}\left\|\mathbf{e}_i^{(1)}\right\|_2 \\
\geq& \sqrt{B_{(N_1-d_{\max})}} - 2g_2 - g_2^2.
\end{aligned}$$

Similarly we have

$$\begin{aligned}
\max_{k\neq 1}\max_{1\leq i\leq N_k} A_i^k =& \max_{k\neq 1}\max_{1\leq i\leq N_k}\left|\langle\mathbf{U}_1\mathbf{a}_1^{(1)}, \mathbf{U}_k\mathbf{a}_i^{(k)}\rangle + \sigma\langle\mathbf{U}_1\mathbf{a}_1^{(1)}, \mathbf{e}_i^{(k)}\rangle + \sigma\langle\mathbf{U}_k\mathbf{a}_i^{(k)}, \mathbf{e}_1^{(1)}\rangle + \sigma^2\langle\mathbf{e}_1^{(1)}, \mathbf{e}_i^{(k)}\rangle\right| \\
\leq& \max_{k\neq 1}\max_{1\leq i\leq N_k}\left|\langle\mathbf{U}_1\mathbf{a}_1^{(1)}, \mathbf{U}_k\mathbf{a}_i^{(k)}\rangle\right| + \sigma\max_{k\neq 1}\max_{1\leq i\leq N_k}\left\|\mathbf{e}_i^{(k)}\right\|_2 \\
&+ \sigma\left\|\mathbf{e}_1^{(1)}\right\|_2 + \sigma^2\left\|\mathbf{e}_1^{(1)}\right\|_2 \max_{k\neq 1}\max_{1\leq i\leq N_k}\left\|\mathbf{e}_i^{(k)}\right\|_2 \\
\leq& \max_{k\neq 1}\max_{1\leq i\leq N_k}\left|\langle\mathbf{U}_1\mathbf{a}_1^{(1)}, \mathbf{U}_k\mathbf{a}_i^{(k)}\rangle\right| + 2g_2 + g_2^2.
\end{aligned}$$

Pick T from Assumption $A2$, then the LHS of (19) has the following upper bound

$$\mathbb{P}\left[T \leq Q|\mathcal{E}_2\right] + \mathbb{P}\left[B_{(N_1-d_{\max})} \leq T^2\middle|\mathcal{E}_2\right], \tag{20}$$

where

$$Q = \left(1+\frac{1+g_2}{1-g_2}\right)\left(2g_2+g_2^2\right) + \frac{1+g_2}{1-g_2}\max_{k\neq 1}\max_{1\leq i\leq N_k}\left|\langle\mathbf{U}_1\mathbf{a}_1^{(1)}, \mathbf{U}_k\mathbf{a}_i^{(k)}\rangle\right|.$$

Now we are going to complete our proof in 3 steps.
Step 1: For the first term in (20) we have

$$\mathbb{P}\left[T \leq Q|\mathcal{E}_2\right] = \mathbb{P}\left[g_1 \leq \max_{k\neq 1}\max_{1\leq i\leq N_k}\left|\langle\mathbf{U}_1\mathbf{a}_1^{(1)}, \mathbf{U}_k\mathbf{a}_i^{(k)}\rangle\right|\right].$$

From singular value decomposition we can write

$$\langle \mathbf{U}_1\mathbf{a}_1^{(1)}, \mathbf{U}_k\mathbf{a}_i^{(k)}\rangle = \mathbf{a}_1^{(1)T}\mathbf{W}_{1k}\mathbf{\Lambda}_{1k}\mathbf{V}_{1k}^T\mathbf{a}_i^{(k)} := \mathbf{b}_k^T\mathbf{\Lambda}_{1k}\mathbf{V}_{1k}^T\mathbf{a_i}^{(k)},$$

where both $\{\mathbf{b}_k\}_{k=2}^K$ and $\{\mathbf{V}_{1k}^T\mathbf{a}_i^{(k)}\}_{k=2}^K$ are sampled uniformly from $\mathbb{S}^{d-1}$. Therefore

$$\begin{aligned}
\mathbb{P}\left[g_1 \le \max_{k\neq 1}\max_{1\le i\le N_k}\left|\langle \mathbf{U}_1\mathbf{a}_1^{(1)}, \mathbf{U}_k\mathbf{a}_i^{(k)}\rangle\right|\right] =& \mathbb{P}\left[g_1^2 \le \max_{k\neq 1}\max_{1\le i\le N_k}\left(\mathbf{b}_k^T\mathbf{\Lambda}_{1k}\mathbf{V}_{1k}^T\mathbf{a}_i^{(k)}\right)^2\right] \\
&\le \sum_{k=2}^K \mathbb{P}\left[g_1^2 \le \max_{1\le i\le N_k}\left(\mathbf{b}_k^T\mathbf{\Lambda}_{1k}\mathbf{V}_{1k}^T\mathbf{a}_i^{(k)}\right)^2\right] \qquad (21)\\
&\le \sum_{k=2}^K \mathbb{P}\left[g_1^2 \le \sum_{i=1}^d\left(\lambda_i^{(1k)}b_{ki}\right)^2\right] \qquad (22)\\
&\le \sum_{k=2}^K \mathbb{P}\left[g_1^2 \le \sum_{i=1}^d\left(\lambda_i^{(1)}b_{ki}\right)^2\right], \qquad (23)
\end{aligned}$$

where inequality (21) uses the union bound inequality, (22) comes from Cauchy-Schwarz inequality, and (23) uses Definition 1. Since $\{\mathbf{b}_k\}_{k=2}^K \sim U(\mathbb{S}^{d-1})$, we can write (23) as

$$(K-1)\mathbb{P}\left[g_1^2 \le \sum_{i=1}^d\left(\lambda_i^{(1)}b_i\right)^2\right],$$

where $\mathbf{b}$ is uniformly distributed on $\mathbb{S}^{d-1}$. Now we apply Lemma 1 directly to the quantity above and get $\mathbb{P}\left[g_1^2 \le \sum_{i=1}^d\left(\lambda_i^{(1)}b_i\right)^2\right] \le 2e^{-\epsilon'^2}$ where

$$\begin{aligned}
\epsilon' &= \frac{\sum_{i=1}^d(r_i-s_i)}{(\sqrt{\sum_{i=1}^d r_i^2}+\sqrt{\sum_{i=1}^d s_i^2})+\sqrt{(\sqrt{\sum_{i=1}^d r_i^2}+\sqrt{\sum_{i=1}^d s_i^2})^2+2s_1\sum_{i=1}^d(r_i-s_i)}}\\
&= \frac{-(\sqrt{\sum_{i=1}^d r_i^2}+\sqrt{\sum_{i=1}^d s_i^2})+\sqrt{(\sqrt{\sum_{i=1}^d r_i^2}+\sqrt{\sum_{i=1}^d s_i^2})^2+2s_1\sum_{i=1}^d(r_i-s_i)}}{2s_1}\\
&\ge \frac{-(\sqrt{\sum_{i=1}^d r_i^2}+\sqrt{\sum_{i=1}^d s_i^2})+\sqrt{(\sqrt{\sum_{i=1}^d r_i^2}+\sqrt{\sum_{i=1}^d s_i^2})^2+2\sum_{i=1}^d(r_i-s_i)}}{2} \qquad (24)\\
&= \frac{\sum_{i=1}^d(r_i-s_i)}{(\sqrt{\sum_{i=1}^d r_i^2}+\sqrt{\sum_{i=1}^d s_i^2})+\sqrt{(\sqrt{\sum_{i=1}^d r_i^2}+\sqrt{\sum_{i=1}^d s_i^2})^2+2\sum_{i=1}^d(r_i-s_i)}}\\
&\ge \frac{\sum_{i=1}^d(r_i-s_i)}{2\sqrt{\sum_{i=1}^d r_i^2}+\sqrt{4\sum_{i=1}^d r_i^2+2\sum_{i=1}^d r_i}} \ge \epsilon_1.
\end{aligned}$$

Here ϵ_1 is defined in (8), r_i and s_i are defined in Lemma 1, and (24) comes from the following fact for positive constants a, b and $s \in (0,1)$

$$\frac{-a+\sqrt{a^2+2sb}}{2s} \ge \frac{-a+\sqrt{a^2+2b}}{2}.$$

Therefore we have

$$\mathbb{P}\left[g_1 \leq \max_{k\neq 1}\max_{1\leq i\leq n_k}\left|\langle \mathbf{U}_1\mathbf{a}_1^{(1)}, \mathbf{U}_k\mathbf{a}_i^{(k)}\rangle\right|\right] \leq 2(K-1)e^{-\epsilon_1^2}.$$

Step 2: For the second term of (20), we just need to use Lemma 2. Note that for fixed $\mathbf{a}_1^{(1)}$, one can show that $B_i = \langle \mathbf{a}_1^{(1)}, \mathbf{a}_i^{(1)}\rangle^2$ can be treated as a sample from a Beta distribution with parameters $(\frac{1}{2}, \frac{d-1}{2})$. From Lemma 2 and Assumption $A2$ we have

$$\mathbb{P}\left[B_{(N_1-d_{\max})} \leq T^2 \middle| \mathcal{E}_2\right] \leq \frac{(N_1-d_{\max})}{d_{\max}(N_1+1)(N_1^{\rho}-1)^2}.$$

Combine the results above we know

$$\mathbb{P}\left[\mathcal{E}_{11}^{\complement}\middle|\mathcal{E}_2\right] \leq 2(K-1)e^{-{\epsilon_1}^2} + \frac{(N_1-d_{\max})}{d_{\max}(N_1+1)(N_1^{\rho}-1)^2}. \tag{25}$$

Step 3: Now we are going to find the lower bound on $\mathbb{P}[\mathcal{E}_2]$. Let $\mathbf{e}$ be an independent copy of $\mathbf{e}_1^{(1)}$, note that $\|\mathbf{e}\|_2^2/D \sim F_{D,d}$. From Corollary 1 we have

$$\mathbb{P}\left[g_2 \leq \sigma||\mathbf{e}||_2\right] = \mathbb{P}\left[\frac{g_2^2}{D\sigma^2} \leq \frac{||\mathbf{e}||_2^2}{D}\right] \leq 2e^{-t^2},$$

where t can be calculated from Corollary 1. Using Assumption $A2$ we have

$$t > \frac{D\left(\frac{g_2^2}{D\sigma^2}-1\right)}{2\left(\sqrt{D}+\frac{g_2^2}{\sigma^2\sqrt{d}}+\sqrt{d}\right)} = \frac{\sqrt{d}}{2}\left(1-\frac{1+\frac{d}{D}+\sqrt{\frac{d}{D}}}{1+\frac{d}{D}+\frac{g_2^2}{D\sigma^2}}\right) \geq \left(1+\frac{\eta}{2+\eta}\right)\sqrt{\log N}.$$

Therefore we have $\mathbb{P}\left[g_2 \leq \sigma||\mathbf{e}||_2\right] \leq \frac{2}{N^{\left(1+\frac{\eta}{2+\eta}\right)^2}}$. Now we note that

$$\begin{aligned}\mathbb{P}\left[g_2 > \sigma \max_{k=1,\ldots,K}\max_{1\leq i\leq N_k}\left\|\mathbf{e}_i^{(k)}\right\|_2\right] &= \Pi_{i=1}^N\left(1-\mathbb{P}\left[g_2 \leq \sigma\left\|\mathbf{e}_i^{(k)}\right\|_2\right]\right)\\ &\geq (1-2e^{-t^2})^N \geq 1-\frac{2N}{N^{\left(1+\frac{\eta}{2+\eta}\right)^2}},\end{aligned}$$

where the last inequality comes from the Bernoulli's inequality. Therefore

$$\mathbb{P}\left[\mathcal{E}_2\right] \geq 1-\frac{2N}{N^{\left(1+\frac{\eta}{2+\eta}\right)^2}}. \tag{26}$$

Finally, the above arguments hold for any $\mathbf{y}_i^{(k)}$. Putting (25) and (26) together and applying the union bound inequality yields the result

$$\mathbb{P}[\mathcal{E}_1] \geq 1-\sum_{k=1}^K \frac{n_k(N_k-d_{\max})}{d_{\max}(N_k+1)(N_k^{\rho}-1)^2} - 2(K-1)ne^{-\epsilon_1^2} - \frac{2N}{N^{\left(1+\frac{\eta}{2+\eta}\right)^2}}. \tag{27}$$

■

To prove Theorem 2, we will use the following equation

$$(\mathbf{W}^T\mathbf{W} + \lambda\mathbf{I}_{d_2})^{-1}\mathbf{W}^T = \mathbf{W}^T(\mathbf{W}\mathbf{W}^T + \lambda\mathbf{I}_{d_1})^{-1}, \tag{28}$$

where $\mathbf{W} \in \mathbb{R}^{d_1 \times d_2}$ and λ is a positive constant (Murphy, 2012, Chapter 4). Throughout the proof of Theorem 2, the subscript of identity matrix $\mathbf{I}$ will be omitted as its dimension is clear from the context.

Proof of Theorem 2 Similar to the proof of Theorem 1, let $\mathcal{E}_1$ be the event that correct neighborhood property holds for all $\{\mathbf{Y}_{\mathcal{C}_j}\}_{j=1}^n$, let $\mathcal{E}_2$ be the event $\{\sigma\|\mathbf{e}_i^{(k)}\|_2 < g, \forall i,k\}$ (g is from Assumption $A4$), $\mathcal{E}_3$ is the event that the smallest singular value of $\mathbf{B}\mathbf{B}^T$ is at least q_0, $\forall i = 1, ..., n$, and $\mathcal{E}_4$ is the event that the sub-cluster preserving property is satisfied.

Define $\mathcal{I} = \{(i,j) :$ the i-th and the j-th sampled points belong to different clusters, $1 \le i < j \le n\}$, and $\mathcal{J} = \{(i,j) :$ the i-th and the j-th sampled points belong to the same cluster, $1 \le i < j \le n\}$. Conditional on $\mathcal{E}_4$, we know that $\mathbf{Y}_{\mathcal{C}_i}$ and $\mathbf{Y}_{\mathcal{C}_j}$ belong to different clusters if $(i,j) \in \mathcal{I}$, and belong to the same cluster if $(i,j) \in \mathcal{J}$.

We will show that conditioning on $\mathcal{E}_2$, $\mathcal{E}_3$ and $\mathcal{E}_4$, there exists a constant l such that

$$\mathbb{P}\left[\mathcal{E}_1|\mathcal{E}_2,\mathcal{E}_3,\mathcal{E}_4\right] \ge \mathbb{P}\left[d(\mathbf{Y}_{\mathcal{C}_i},\mathbf{Y}_{\mathcal{C}_j})_{\forall(i,j)\in\mathcal{I}} > l \middle| \mathcal{E}_2,\mathcal{E}_3,\mathcal{E}_4\right] \ge 1 - \sum_{\forall(i,j)\in\mathcal{I}} \mathbb{P}\left[d(\mathbf{Y}_{\mathcal{C}_i},\mathbf{Y}_{\mathcal{C}_j}) \le l \middle| \mathcal{E}_2,\mathcal{E}_3,\mathcal{E}_4\right].$$

Then we obtain an upper bound on $\mathbb{P}[d(\mathbf{Y}_{\mathcal{C}_i},\mathbf{Y}_{\mathcal{C}_k}) \le l \mid \mathcal{E}_2,\mathcal{E}_3,\mathcal{E}_4]$, $\forall(i,k) \in \mathcal{I}$. The theorem will follow by using the union bound inequality.

WLOG assume that $\mathbf{Y}_{\mathcal{C}_1}$ and $\mathbf{Y}_{\mathcal{C}_2}$ belong to $\mathcal{S}_1$, and $\mathbf{Y}_{\mathcal{C}_3}$ belongs to $\mathcal{S}_2$. The distance function $d(\mathbf{Y}_{\mathcal{C}_1},\mathbf{Y}_{\mathcal{C}_2})$ can be explicitly written as

$$\left\|\mathbf{Y}_{\mathcal{C}_1} - \mathbf{Y}_{\mathcal{C}_2}(\mathbf{Y}_{\mathcal{C}_2}^T\mathbf{Y}_{\mathcal{C}_2} + \lambda\mathbf{I})^{-1}\mathbf{Y}_{\mathcal{C}_2}^T\mathbf{Y}_{\mathcal{C}_1}\right\|_F + \left\|\mathbf{Y}_{\mathcal{C}_2} - \mathbf{Y}_{\mathcal{C}_1}(\mathbf{Y}_{\mathcal{C}_1}^T\mathbf{Y}_{\mathcal{C}_1} + \lambda\mathbf{I})^{-1}\mathbf{Y}_{\mathcal{C}_1}^T\mathbf{Y}_{\mathcal{C}_2}\right\|_F. \tag{29}$$

Conditional on $\mathcal{E}_4$, we can write $\mathbf{Y}_{\mathcal{C}_1} = \mathbf{U}_1\hat{\mathbf{B}}_1 + \hat{\mathbf{E}}_1$, where $\|[\mathbf{U}_1\hat{\mathbf{B}}_1]_j + [\hat{\mathbf{E}}_1]_j\|_2 = 1$. Let $\mathbf{B}_1$ and $\mathbf{E}_1$ be the "un-normalized" version of $\hat{\mathbf{B}}_1$ and $\hat{\mathbf{E}}_1$ respectively. Here each column of $\mathbf{B}_1$ is a sample from the uniform distribution on $\mathbb{S}^{d-1}$. We have the following relation

$$[\hat{\mathbf{B}}_1]_j = \frac{[\mathbf{B}_1]_j}{\|[\mathbf{U}_1\mathbf{B}_1]_j + [\mathbf{E}_1]_j\|_2},\ [\hat{\mathbf{E}}_1]_j = \frac{[\mathbf{E}_1]_j}{\|[\mathbf{U}_1\mathbf{B}_1]_j + [\mathbf{E}_1]_j\|_2},\ j = 1, ..., d_{max}+1.$$

Similarly we can write $\mathbf{Y}_{\mathcal{C}_2} = \mathbf{U}_1\hat{\mathbf{B}}_2 + \hat{\mathbf{E}}_2$ and $\mathbf{Y}_{\mathcal{C}_3} = \mathbf{U}_2\hat{\mathbf{B}}_3 + \hat{\mathbf{E}}_3$. Using Equation (28), the first term in (29) can be rewritten as

$$\begin{aligned}
&\left\|\mathbf{Y}_{\mathcal{C}_1} - \mathbf{Y}_{\mathcal{C}_2}(\mathbf{Y}_{\mathcal{C}_2}^T\mathbf{Y}_{\mathcal{C}_2} + \lambda\mathbf{I})^{-1}\mathbf{Y}_{\mathcal{C}_2}^T\mathbf{Y}_{\mathcal{C}_1}\right\|_F \\
=&\left\|\mathbf{Y}_{\mathcal{C}_1} - (\mathbf{Y}_{\mathcal{C}_2}\mathbf{Y}_{\mathcal{C}_2}^T + \lambda\mathbf{I} - \lambda\mathbf{I})(\mathbf{Y}_{\mathcal{C}_2}\mathbf{Y}_{\mathcal{C}_2}^T + \lambda\mathbf{I})^{-1}\mathbf{Y}_{\mathcal{C}_1}\right\|_F \\
=&\lambda\left\|(\mathbf{Y}_{\mathcal{C}_2}\mathbf{Y}_{\mathcal{C}_2}^T + \lambda\mathbf{I})^{-1}\mathbf{Y}_{\mathcal{C}_1}\right\|_F \\
<&\lambda\left\|[(\mathbf{Y}_{\mathcal{C}_2}\mathbf{Y}_{\mathcal{C}_2}^T + \lambda\mathbf{I})^{-1} - (\mathbf{U}_1\hat{\mathbf{B}}_2\hat{\mathbf{B}}_2^T\mathbf{U}_1^T + \lambda\mathbf{I})^{-1}]||_F||\mathbf{Y}_{\mathcal{C}_1}\right\|_F \\
&+\lambda\left\|(\mathbf{U}_1\hat{\mathbf{B}}_2\hat{\mathbf{B}}_2^T\mathbf{U}_1^T + \lambda\mathbf{I})^{-1}\mathbf{Y}_{\mathcal{C}_1}\right\|_F
\end{aligned}$$

$$
\begin{aligned}
<&\lambda \left\|(\mathbf{Y}_{\mathcal{C}_2}\mathbf{Y}_{\mathcal{C}_2}^T + \lambda \mathbf{I})^{-1} - (\mathbf{U}_1\hat{\mathbf{B}}_2\hat{\mathbf{B}}_2^T\mathbf{U}_1^T + \lambda \mathbf{I})^{-1}\right\|_F \sqrt{d_{\max}+1} \\
&+ \lambda \left\|(\mathbf{U}_1\hat{\mathbf{B}}_2\hat{\mathbf{B}}_2^T\mathbf{U}_1^T + \lambda \mathbf{I})^{-1}\mathbf{U}_1\hat{\mathbf{B}}_1\right\|_F \\
&+ \lambda \left\|(\mathbf{U}_1\hat{\mathbf{B}}_2\hat{\mathbf{B}}_2^T\mathbf{U}_1^T + \lambda \mathbf{I})^{-1}\right\|_F \left\|\hat{\mathbf{E}}_1\right\|_F . \qquad (30)
\end{aligned}
$$

Now we are going to complete our proof in 3 steps. Unless specified otherwise, the following Step 1 to Step 3 are derived conditioning on $\mathcal{E}_2$, $\mathcal{E}_3$ and $\mathcal{E}_4$.

Step 1: We can rewrite the first term in (30) as the following term

$$\lambda \left\|(\mathbf{G}_2 + \mathbf{H})^{-1} - \mathbf{H}^{-1}\right\|_F \sqrt{d_{\max}+1},$$

where $\mathbf{H} = \mathbf{U}_1\hat{\mathbf{B}}_2\hat{\mathbf{B}}_2^T\mathbf{U}_1^T + \lambda\mathbf{I}$, and $\mathbf{G}_2 = \mathbf{Y}_{\mathcal{C}_2}\mathbf{Y}_{\mathcal{C}_2}^T - \mathbf{U}_1\hat{\mathbf{B}}_2\hat{\mathbf{B}}_2^T\mathbf{U}_1^T = \hat{\mathbf{E}}_2\hat{\mathbf{B}}_2^T\mathbf{U}_1^T + \mathbf{U}_1\hat{\mathbf{B}}_2\hat{\mathbf{E}}_2^T + \hat{\mathbf{E}}_2\hat{\mathbf{E}}_2^T$. Note that the normalizing constant of each column of $\{\hat{\mathbf{E}}_i\}_{i=1}^n$ are bounded in $[1-g, 1+g]$. We then have the following relations

$$
\begin{aligned}
\|\mathbf{G}_2\|_F &\le \left\|\hat{\mathbf{E}}_2\right\|_F \left\|\hat{\mathbf{B}}_2^T\mathbf{U}_1^T\right\|_F + \left\|\mathbf{U}_1\hat{\mathbf{B}}_2 + \hat{\mathbf{E}}_2\right\|_F \left\|\hat{\mathbf{E}}_2^T\right\|_F \\
&\le \frac{(2g-g^2)(d_{max}+1)}{(1-g)^2}. \qquad (31)
\end{aligned}
$$

The above analysis used triangle inequalities and the bounds of normalizing constants.

Using the fact that

$$\|\mathbf{H}^{-1}\|_F < \sqrt{\frac{d(1+g)^4}{q_0^2} + \frac{D-d}{\lambda^2}}$$

and inequality (31), we have the following inequalities

$$\left\|\mathbf{H}^{-1}\mathbf{G}_2\right\|_F \le \left\|\mathbf{H}^{-1}\right\|_F \|\mathbf{G}_2\|_F = \frac{(2g-g^2)\,(d_{max}+1)}{2(1-g)} \cdot \sqrt{\frac{d(1+g)^4}{q_0^2} + \frac{D-d}{\lambda^2}} =: f(d) < \frac{1}{2}.$$

Therefore $\lim_{m\to\infty}(\mathbf{H}^{-1}\mathbf{G}_2)^m = \mathbf{0}$. From Theorem 4.29 in Schott (2016) we know $(\mathbf{I} + \mathbf{H}^{-1}\mathbf{G}_2)^{-1} = \sum_{j=0}^{\infty}(\mathbf{H}^{-1}\mathbf{G}_2)^j$ and

$$
\begin{aligned}
\left\|(\mathbf{G}_2 + \mathbf{H})^{-1} - \mathbf{H}^{-1}\right\|_F &= \left\|\mathbf{H}^{-1}\mathbf{G}(\mathbf{I} + \mathbf{H}^{-1}\mathbf{G}_2)^{-1}\mathbf{H}^{-1}\right\|_F \\
&\le \left\|\sum_{j=1}^{\infty}(\mathbf{H}^{-1}\mathbf{G}_2)^j\right\|_F \left\|\mathbf{H}^{-1}\right\|_F \\
&< \frac{f(d)}{1-f(d)}\sqrt{\frac{d(1+g)^4}{q_0^2} + \frac{D-d}{\lambda^2}}.
\end{aligned}
$$

We then have for the first term in (30)

$$\lambda \left\|(\mathbf{G}_2 + \mathbf{H})^{-1} - \mathbf{H}^{-1}\right\|_F \sqrt{d_{\max}+1} < \frac{f(d)\sqrt{d_{max}+1}}{1-f(d)} \cdot \sqrt{\frac{d(1+g)^4\lambda^2}{q_0^2} + D - d}.$$

For the second term in (30) we have

$$\lambda \left\| \left(\mathbf{U}_1 \hat{\mathbf{B}}_2 \hat{\mathbf{B}}_2^T \mathbf{U}_1^T + \lambda \mathbf{I} \right)^{-1} \hat{\mathbf{B}}_1 \right\|_F = \left\| \hat{\mathbf{B}}_1 - \hat{\mathbf{B}}_2 \hat{\mathbf{B}}_2^T (\hat{\mathbf{B}}_2 \hat{\mathbf{B}}_2^T + \lambda \mathbf{I})^{-1} \hat{\mathbf{B}}_1 \right\|_F$$
$$\leq \left\| \mathbf{I} - \hat{\mathbf{B}}_2 \hat{\mathbf{B}}_2^T (\hat{\mathbf{B}}_2 \hat{\mathbf{B}}_2^T + \lambda \mathbf{I})^{-1} \right\|_F \left\| \hat{\mathbf{B}}_1 \right\|_F \leq \frac{\lambda (1+g)^2 \sqrt{d(d_{max}+1)}}{q_0 (1-g)}.$$

For the third term in (30) we have

$$\lambda \left\| (\mathbf{U}_1 \hat{\mathbf{B}}_2 \hat{\mathbf{B}}_2^T \mathbf{U}_1^T + \lambda \mathbf{I})^{-1} \right\|_F \left\| \hat{\mathbf{E}}_1 \right\|_F \leq g \sqrt{\frac{d(1+g)^4 \lambda^2}{q_0^2} + D - d}.$$

Hence by our assumption, Equation (30) can be upper bounded by the following term

$$\frac{3\lambda (1+g)^2 \sqrt{d(d_{max}+1)}}{q_0 (1-g)},$$

which is deterministic and does not depend on the choices of $\{\mathbf{B}_i\}_{i=1}^n$ and $\{\mathbf{U}_k\}_{k=1}^K$. The distance function in (29) has two parts which are symmetric, therefore we set

$$l := \frac{6\lambda (1+g)^2 \sqrt{d(d_{max}+1)}}{q_0 (1-g)} > d(\mathbf{Y}_{\mathcal{C}_i}, \mathbf{Y}_{\mathcal{C}_j})_{(i,j) \in \mathcal{J}}.$$

Step 2: Now we consider $\mathbb{P}\left[d\left(\mathbf{Y}_{\mathcal{C}_1}, \mathbf{Y}_{\mathcal{C}_3} \right) \leq l | \mathcal{E}_2, \mathcal{E}_3, \mathcal{E}_4 \right]$. We explicitly write $d(\mathbf{Y}_{\mathcal{C}_1}, \mathbf{Y}_{\mathcal{C}_3})$ as

$$\left\| \mathbf{Y}_{\mathcal{C}_1} - \mathbf{Y}_{\mathcal{C}_3} (\mathbf{Y}_{\mathcal{C}_3}^T \mathbf{Y}_{\mathcal{C}_3} + \lambda \mathbf{I})^{-1} \mathbf{Y}_{\mathcal{C}_3}^T \mathbf{Y}_{\mathcal{C}_1} \right\|_F + \left\| \mathbf{Y}_{\mathcal{C}_3} - \mathbf{Y}_{\mathcal{C}_1} (\mathbf{Y}_{\mathcal{C}_1}^T \mathbf{Y}_{\mathcal{C}_1} + \lambda \mathbf{I})^{-1} \mathbf{Y}_{\mathcal{C}_1}^T \mathbf{Y}_{\mathcal{C}_3} \right\|_F. \quad (32)$$

Note the following relation

$$\mathbb{P}\left[d(\mathbf{Y}_{\mathcal{C}_1}, \mathbf{Y}_{\mathcal{C}_3}) \leq l | \mathcal{E}_2, \mathcal{E}_3 \right] \leq \mathbb{P}\left[\left\| \mathbf{Y}_{\mathcal{C}_1} - \mathbf{Y}_{\mathcal{C}_3} (\mathbf{Y}_{\mathcal{C}_3}^T \mathbf{Y}_{\mathcal{C}_3} + \lambda \mathbf{I})^{-1} \mathbf{Y}_{\mathcal{C}_3}^T \mathbf{Y}_{\mathcal{C}_1} \right\|_F \leq \frac{l}{2} \middle| \mathcal{E}_2, \mathcal{E}_3, \mathcal{E}_4 \right]$$
$$+ \mathbb{P}\left[\left\| \mathbf{Y}_{\mathcal{C}_3} - \mathbf{Y}_{\mathcal{C}_1} (\mathbf{Y}_{\mathcal{C}_1}^T \mathbf{Y}_{\mathcal{C}_1} + \lambda \mathbf{I})^{-1} \mathbf{Y}_{\mathcal{C}_1}^T \mathbf{Y}_{\mathcal{C}_3} \right\|_F \leq \frac{l}{2} \middle| \mathcal{E}_2, \mathcal{E}_3, \mathcal{E}_4 \right]. \quad (33)$$

To bound the first term in (32), the following facts come from the triangle inequality

$$\begin{aligned}
&\left\| \mathbf{Y}_{\mathcal{C}_1} - \mathbf{Y}_{\mathcal{C}_3} \left(\mathbf{Y}_{\mathcal{C}_3}^T \mathbf{Y}_{\mathcal{C}_3} + \lambda \mathbf{I} \right)^{-1} \mathbf{Y}_{\mathcal{C}_3}^T \mathbf{Y}_{\mathcal{C}_1} \right\|_F \\
=& \lambda \left\| \left(\mathbf{Y}_{\mathcal{C}_3} \mathbf{Y}_{\mathcal{C}_3}^T + \lambda \mathbf{I} \right)^{-1} \mathbf{Y}_{\mathcal{C}_1} \right\|_F \\
>& \lambda \left\| \left(\mathbf{U}_2 \hat{\mathbf{B}}_3 \hat{\mathbf{B}}_3^T \mathbf{U}_2^T + \lambda \mathbf{I} \right)^{-1} \mathbf{U}_1 \hat{\mathbf{B}}_1 \right\|_F - \lambda \left\| \left(\mathbf{U}_2 \hat{\mathbf{B}}_3 \hat{\mathbf{B}}_3^T \mathbf{U}_2^T + \lambda \mathbf{I} \right)^{-1} \right\|_F \left\| \hat{\mathbf{E}}_1 \right\|_F \\
&- \lambda \left\| \left[\left(\mathbf{Y}_{\mathcal{C}_3} \mathbf{Y}_{\mathcal{C}_3}^T + \lambda \mathbf{I} \right)^{-1} - \left(\mathbf{U}_2 \hat{\mathbf{B}}_3 \hat{\mathbf{B}}_3^T \mathbf{U}_2^T + \lambda \mathbf{I} \right)^{-1} \right] \right\|_F \sqrt{d_{\max} + 1}.
\end{aligned}$$

The last two terms of the line above are upper bounded by $\frac{\lambda (1+g)^2 \sqrt{d(d_{max}+1)}}{q_0 (1-g)}$ as before, and the first term can be bounded by the following relations

$$\lambda \left\| \left(\mathbf{U}_2 \hat{\mathbf{B}}_3 \hat{\mathbf{B}}_3^T \mathbf{U}_2^T + \lambda \mathbf{I} \right)^{-1} \mathbf{U}_1 \hat{\mathbf{B}}_1 \right\|_F$$

$$\geq \left\|\mathbf{U}_1\hat{\mathbf{B}}_1 - \mathbf{U}_2\mathbf{U}_2^T\mathbf{U}_1\hat{\mathbf{B}}_1\right\|_F - \lambda\left\|\mathbf{U}_2\left(\hat{\mathbf{B}}_3\hat{\mathbf{B}}_3^T + \lambda\mathbf{I}\right)^{-1}\mathbf{U}_2^T\mathbf{U}_1\hat{\mathbf{B}}_1\right\|_F$$
$$> \left\|\mathbf{U}_1\hat{\mathbf{B}}_1 - \mathbf{U}_2\mathbf{U}_2^T\mathbf{U}_1\hat{\mathbf{B}}_1\right\|_F - \frac{\lambda(1+g)^2\sqrt{d(d_{max}+1)}}{q_0(1-g)}, \tag{34}$$

where inequality (34) comes from the following relations

$$\lambda\left\|\mathbf{U}_2\left(\hat{\mathbf{B}}_3\hat{\mathbf{B}}_3^T + \lambda\mathbf{I}\right)^{-1}\mathbf{U}_2^T\mathbf{U}_1\hat{\mathbf{B}}_1\right\|_F \leq \lambda\left\|\left(\hat{\mathbf{B}}_3\hat{\mathbf{B}}_3^T + \lambda\mathbf{I}\right)^{-1}\right\|_F\left\|\mathbf{U}_1\hat{\mathbf{B}}_1\right\|_F$$
$$\leq \lambda\frac{\sqrt{d}(1+g)^2}{q_0}\frac{\sqrt{d_{max}+1}}{1-g} = \frac{\lambda(1+g)^2\sqrt{d(d_{max}+1)}}{q_0(1-g)}.$$

For the first term in (34) we have

$$\left\|\mathbf{U}_1\hat{\mathbf{B}}_1 - \mathbf{U}_2\mathbf{U}_2^T\mathbf{U}_1\hat{\mathbf{B}}_1\right\|_F = \sqrt{Tr\left[\hat{\mathbf{B}}_1^T\hat{\mathbf{B}}_1 - \hat{\mathbf{B}}_1^T\mathbf{U}_1^T\mathbf{U}_2\mathbf{U}_2^T\mathbf{U}_1\hat{\mathbf{B}}_1\right]}$$
$$= \left\|\sqrt{\mathbf{I} - \mathbf{\Lambda}_{12}^2}\tilde{\mathbf{B}}_1\mathbf{W}\right\|_F \geq \frac{\left\|\sqrt{\mathbf{I} - \mathbf{\Lambda}_{12}^2}\tilde{\mathbf{B}}_1\right\|_F}{1+g},$$

where $\mathbf{W}$ is the diagonal matrix such that $W_{jj} = \frac{1}{\left\|[\hat{\mathbf{B}}_1]_j\right\|_2}$ ($[\hat{\mathbf{B}}_1]_j$ is the j-th column of $\hat{\mathbf{B}}_1$, $j = 1, .., d_{max}+1$), $\tilde{\mathbf{B}}_1 = \mathbf{V}\mathbf{B}_1$ is a orthogonal transformation of $\mathbf{B}_1$ (here $\mathbf{V}$ is the right orthogonal matrix in the svd of $\mathbf{U}_2^T\mathbf{U}_1$), and $\mathbf{\Lambda}_{12}$ is the diagonal matrix such that $[\mathbf{\Lambda}_{12}]_{ii} = \lambda_i^{(12)}$, $i = 1, ..., d$. Therefore, eventually the first term at the RHS of (33) can be upper bounded by

$$\mathbb{P}\left[\left\|\sqrt{\mathbf{I} - \mathbf{\Lambda}_{12}^2}\tilde{\mathbf{B}}_1\right\|_F \leq \frac{6\lambda(1+g)^2\sqrt{d(d_{max}+1)}}{q_0(1-g)}\right].$$

Using Assumption A4, Lemma 1 and arguments similar to the proof of Theorem 1, we know the quantity above is upper bounded by

$$\mathbb{P}\left[\left\|\sqrt{\mathbf{I} - \mathbf{\Lambda}_{12}^2}\mathbf{v}\right\|_F \leq \sqrt{1 - T_l^2}\right] \leq 2e^{-{\epsilon_1}^2},$$

where ϵ_1 is defined in (8) with g_1 replaced by T_l, and $\mathbf{v}$ is the first column of $\tilde{\mathbf{B}}_1$. Using analogous manipulations we obtain similar results for the second term in (33). Therefore $\mathbb{P}\left[d(\mathbf{Y}_{\mathcal{C}_1}, \mathbf{Y}_{\mathcal{C}_3}) \leq l | \mathcal{E}_2, \mathcal{E}_3, \mathcal{E}_4\right] \leq 4e^{-{\epsilon_1}^2}$.

Step 3: Now we are going to lower bound $\mathbb{P}\left[\mathcal{E}_2, \mathcal{E}_3, \mathcal{E}_4\right]$ from the fact $\mathbb{P}\left[\mathcal{E}_2, \mathcal{E}_3, \mathcal{E}_4\right] \geq 1 - \mathbb{P}[\mathcal{E}_2^{\complement}] - \mathbb{P}[\mathcal{E}_3^{\complement}] - \mathbb{P}[\mathcal{E}_4^{\complement}]$.

Just as in the proof of Theorem 1 we have $\mathbb{P}\left[\sigma\mathbf{e}_i^{(k)} \geq g\right] \leq 2e^{-t^2}$, where

$$t = \frac{D\left(\frac{g^2}{D\sigma^2} - 1\right)}{2\left(\sqrt{D} + \frac{Dg^2}{\sigma^2\sqrt{d}} + \sqrt{d}\right)}.$$

From Assumption $A4$ we know $2e^{-t^2} \leq \frac{2}{N^{\left(1+\frac{\eta}{2+\eta}\right)^2}}$. Using union bound inequality we have

$$\mathbb{P}\left[\mathcal{E}_2^{\complement}\right] \leq \frac{2N}{N^{\left(1+\frac{\eta}{2+\eta}\right)^2}}. \tag{35}$$

From Lemma 5 we have

$$\mathbb{P}\left[\mathcal{E}_3^{\complement}\right] \leq \frac{2n}{N^{t^2/2}} + \sum_{k=1}^{K} n_k \left(\frac{N_k - d_{max}}{d_{max}(N_k+1)(N_k^{\rho}-1)^2} + 2(N_k-1)e^{-\epsilon_2^2}\right). \tag{36}$$

From our assumption we have

$$\mathbb{P}\left[\mathcal{E}_4^{\complement}\right] \leq p_s. \tag{37}$$

Combing (35), (36) and (37) we know

$$\begin{aligned}
\mathbb{P}[\mathcal{E}_1] \geq & 1 - \sum_{\forall (i,j)\in\mathcal{I}} \mathbb{P}\left[d(\mathbf{Y}_{\mathcal{C}_i}, \mathbf{Y}_{\mathcal{C}_j}) \leq l \middle| \mathcal{E}_2, \mathcal{E}_3, \mathcal{E}_4\right] - \mathbb{P}\left[\mathcal{E}_2^{\complement}\right] - \mathbb{P}\left[\mathcal{E}_3^{\complement}\right] - \mathbb{P}\left[\mathcal{E}_4^{\complement}\right] \\
\geq & 1 - 4n(n-1)e^{-\epsilon_1^2} - \frac{2n}{N^{t^2/2}} - \sum_{k=1}^{K} n_k \left(\frac{N_k - d_{max}}{d_{max}(N_k+1)(N_k^{\rho}-1)^2} + 2(N_k-1)e^{-\epsilon_2^2}\right) \\
& - \frac{2N}{N^{\left(1+\frac{\eta}{2+\eta}\right)^2}} - p_s.
\end{aligned}$$

This completes the proof. ■

Appendix D. Residual Minimization by Ridge Regression

In this section we provide the algorithm for classifying the out-of-sample points.

input : $\mathbf{Y}$ to be classified, $\mathbf{R}$ and $\boldsymbol{\ell}$ are the training data and labels, m and λ are the residual minimization and regularization parameters
output: The label vector $\boldsymbol{\ell}$ of all points in $\mathbf{Y}$
1. Generate subsets of training data
for $k = 1$ **to** K **do**
  Uniformly sample m points from the k-th cluster in the training set $\mathbf{R}$, denote this sampled set as $\mathbf{R}_k$;
end
2. Compute the projection matrix for each cluster
for $k = 1$ **to** K **do**
  $\mathbf{P}_k := \mathbf{R}_k(\mathbf{R}_k^T\mathbf{R}_k + \lambda\mathbf{I})^{-1}\mathbf{R}_k^T$
end
3. Compute residuals for points in Y, here N is the number of points in $\mathbf{Y}$
for $i = 1$ **to** N **do**
  for $k = 1$ **to** K **do**
    $\mathbf{r}_i(k) := (\mathbf{I} - \mathbf{P}_k)\mathbf{y}_i$;
  end
end
4. Assign labels through minimum residual
for $i = 1$ **to** N **do**
  $\ell_i = \arg\min_k \mathbf{r}_i(k)$;
end

Algorithm 3: Residual Minimization by Ridge Regression (RMRR) algorithm.

Appendix E. Additional Numerical Results

In this section, we present additional numerical results. Results for some algorithms are omitted for certain datasets due to the limitations on computational resources. Specifically, the additional results are presented in Table 4, Table 5 and Table 6.

Appendix F. Additional Technical Discussion

F.1 The ϵ_1 in Theorem 1

In this section, we will show that under mild conditions, ϵ_1 in (8) grows at least linear in $\sqrt{d}$. For ease of notation, we write $r_i = \left(g_1^2 - \lambda_i^{(1)2}\right)_+$ and $s_i = \left(g_1^2 - \lambda_i^{(1)2}\right)_-$, $i = 1, .., d$. WLOG assume ϵ_1 is evaluated at $k = 1$.

Method	Accuracy (%)	Accuracy-Sub (%)	NMI (%)	Runtime (sec.)
SBSC-SSC	20.99 (1.02)	22.5 (1.29)	34.24 (1.15)	56
SSSC	49.33 (2.51)	56.54 (1,77)	52.82 (2.21)	22
LRR	55.63	NA	64.02	29
LSR	54.11	NA	65.12	8

Table 4: **Additional Results on Extended Yale B**

Method	Accuracy (%)	Accuracy-Sub (%)	NMI (%)	Runtime (sec.)
SBSC(6)	75.16 (3.28)	72.62 (1.52)	78.79 (1.67)	63
SBSC-DSC(6)	64.25 (1.25)	65.34 (1.86)	72.34 (0.76)	388
SBSC-SSC(1)	55.24 (1.34)	61.44 (2.36)	45.18 (1.42)	117
SBSC-SSC(6)	71.07 (0.94)	68.65 (1.21)	67.39 (1.19)	703
STSC(6)	57.76 (1.15)	60.1 (1.8)	60.4 (1.59)	13
SDSC(6)	52 (2.63)	51.59 (1.28)	63.28 (1.26)	51
SSSC(1)	41.52 (5.92)	44.86 (7.06)	38.22 (3.7)	25
SSSC(6)	44.43 (4)	44.06 (2.53)	42.61 (2.23)	150
SLRR(6)	63.7 (3.74)	63.85 (1.74)	69.25 (1.86)	46
SLSR(6)	60.71 (1.04)	59.43 (0.8)	66.39 (1.08)	26
LRR	53.25	NA	53.53	401
LSR	58.91	NA	61.56	192

Table 5: **Additional Results on Zipcode**

Method	Accuracy (%)	Accuracy-Sub (%)	NMI (%)	Runtime (sec.)
SBSC-SSC	84.95 (4.51)	86.48 (4.2)	73.71 (2.06)	834
SSSC(1)	33.26 (2.15)	77.22 (3.9)	13.59 (1.41)	43
SSSC(6)	48.49 (2.75)	79.06 (1.63)	30.41 (2.04)	259

Table 6: **Additional Results on MNIST**

Main result: If there exist constants $c_1 \in (0, g_1]$, $c_2 \in (0,1)$ and $c_3 > 0$ such that $\sum_{i=1}^d r_i \geq c_1 d$, $\frac{\sum_{i=1}^d s_i}{\sum_{i=1}^d r_i} \leq c_2$ and $c_3 d > g_1$, then we have

$$\epsilon_1 \geq \frac{(1-c_2)\sqrt{d}}{2\sqrt{\frac{(c_1+c_3)g_1}{c_1^2}} + \sqrt{\frac{4(c_1+c_3)g_1}{c_1^2} + \frac{2}{c_1}}}.$$

Proof Note that

$$\epsilon_1 = \frac{1 - \frac{\sum_{i=1}^d s_i}{\sum_{i=1}^d r_i}}{2\sqrt{\frac{\sum_{i=1}^d r_i^2}{(\sum_{i=1}^d r_i)^2}} + \sqrt{\frac{4\sum_{i=1}^d r_i^2}{(\sum_{i=1}^d r_i)^2} + \frac{2}{\sum_{i=1}^d r_i}}}. \tag{38}$$

Define $f : \mathcal{V} \to R$, where $f(\mathbf{v}) = \frac{\sum_{i=1}^d v_i^2}{(\sum_{i=1}^d v_i)^2}$, and $\mathcal{V} = \{\mathbf{v} \in [0, g_1]^d : \sum_{i=1}^d v_i = \sum_{i=1}^d r_i\}$. Consider the following $\mathbf{r}^* \in \mathcal{V}$

$$r_i^* = \begin{cases} g_1, & \text{if } i \leq \lfloor \frac{\sum_{i=1}^d r_i}{g_1} \rfloor, \\ \sum_{i=1}^d r_i - \lfloor \frac{\sum_{i=1}^d r_i}{g_1} \rfloor \cdot g_1, & i = \lfloor \frac{\sum_{i=1}^d r_i}{g_1} \rfloor + 1, \\ 0, & \text{otherwise.} \end{cases}$$

We will prove by contradiction that any maximizer of $f(\cdot)$ is a permutation of $\mathbf{r}^*$.

In fact, assume $\mathbf{r}' \in \mathcal{V}$ also maximizes $f(\cdot)$ but is not a permutation of $\mathbf{r}^*$. Assume there are m terms in $\{r_i'\}_{i=1}^d$ that are equal to g_1. Let $r_1' \leq r_2'$ be the two smallest positive terms of $\{r_i'\}_{i=1}^d$. It is straightforward to see $r_2' < g_1$. Consequently, we can find a constant $\delta > 0$ such that $r_1' - \delta, r_2' + \delta \in (0, g_1)$. Note $\mathbf{r}'' = [r_1' - \delta, r_2' + \delta, r_3', ..., r_d'] \in \mathcal{V}$, but $f(\mathbf{r}'') > f(\mathbf{r}')$, which is a contradiction.

Note that $\mathbf{r} \in \mathcal{V}$, we plug $\mathbf{r}^*$ into $f(\cdot)$ and get

$$f(\mathbf{r}) = \frac{\sum_{i=1}^d r_i^2}{(\sum_{i=1}^d r_i)^2} \leq \frac{(\frac{\sum_{i=1}^d r_i}{g_1} + 1)g_1^2}{(\sum_{i=1}^d r_i)^2} \leq \frac{(\frac{c_1 d}{g_1} + 1)g_1^2}{(c_1 d)^2} \leq \frac{(c_1 + c_3)g_1}{c_1^2 d}.$$

Finally, from the inequality above and (38) we have

$$\epsilon_1 \geq \frac{1-c_2}{2\sqrt{\frac{(c_1+c_3)g_1}{c_1^2 d}} + \sqrt{\frac{4(c_1+c_3)g_1}{c_1^2 d} + \frac{2}{c_1 d}}} = \frac{(1-c_2)\sqrt{d}}{2\sqrt{\frac{(c_1+c_3)g_1}{c_1^2}} + \sqrt{\frac{4(c_1+c_3)g_1}{c_1^2} + \frac{2}{c_1}}}.$$

■

References

Pankaj K Agarwal, Sariel Har-Peled, Kasturi R Varadarajan, et al. Geometric approximation via coresets. *Combinatorial and computational geometry*, 52:1–30, 2005.

Bryon Aragam, Chen Dan, Eric P Xing, Pradeep Ravikumar, et al. Identifiability of nonparametric mixture models and bayes optimal clustering. *Annals of Statistics*, 48(4): 2277–2302, 2020.

Mikhail Belkin and Partha Niyogi. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural computation*, 15(6):1373–1396, 2003.

Leo Breiman. Bagging predictors. *Machine learning*, 24(2):123–140, 1996.

Raymond B Cattell. *Factor Analysis: An Introduction and Manual for the Psychologist and Social Scientist*. Harper, 1952.

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

Ehsan Elhamifar and René Vidal. Sparse subspace clustering. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2790–2797. IEEE, 2009.

Ehsan Elhamifar and Rene Vidal. Sparse subspace clustering: Algorithm, theory, and applications. *IEEE transactions on pattern analysis and machine intelligence*, 35(11): 2765–2781, 2013.

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

Wei Hong, John Wright, Kun Huang, and Yi Ma. Multiscale hybrid linear models for lossy image representation. *IEEE Transactions on Image Processing*, 15(12):3655–3671, 2006.

Harold Hotelling. Analysis of a complex of statistical variables into principal components. *Journal of educational psychology*, 24(6):417, 1933.

GJO Jameson. Inequalities for gamma function ratios. *The American Mathematical Monthly*, 120(10):936–940, 2013.

Beatrice Laurent and Pascal Massart. Adaptive estimation of a quadratic functional by model selection. *Annals of Statistics*, 28(5):1302–1338, 2000.

Yann LeCun, Ofer Matan, Bernhard Boser, John S Denker, Don Henderson, Richard E Howard, Wayne Hubbard, LD Jacket, and Henry S Baird. Handwritten zip code recognition with multilayer networks. In *[1990] Proceedings. 10th International Conference on Pattern Recognition*, volume 2, pages 35–40. IEEE, 1990.

Michel Ledoux and Michel Talagrand. *Probability in Banach Spaces: isoperimetry and processes*. Springer Science & Business Media, 2013.

Gilad Lerman, Michael McCoy, Joel A Tropp, and Teng Zhang. Robust computation of linear models, or how to find a needle in a haystack. Technical report, California Inst of Tech Pasadena Dept of Computing and Mathematical Sciences, 2012.

Jiayu Lin. *On the Dirichlet Distribution*. PhD thesis, 2016.

Guangcan Liu, Zhouchen Lin, and Yong Yu. Robust subspace segmentation by low-rank representation. In *Proceedings of the 27th international conference on machine learning (ICML-10)*, pages 663–670, 2010.

Risheng Liu, Ruru Hao, and Zhixun Su. Mixture of manifolds clustering via low rank embedding. *Journal of Information and Computational Science*, 8:725–737, 2011.

Can-Yi Lu, Hai Min, Zhong-Qiu Zhao, Lin Zhu, De-Shuang Huang, and Shuicheng Yan. Robust and efficient subspace segmentation via least squares regression. In *European conference on computer vision*, pages 347–360. Springer, 2012.

Ulrike V Luxburg, Olivier Bousquet, and Mikhail Belkin. Limits of spectral clustering. In *Advances in neural information processing systems*, pages 857–864, 2005.

Kevin P Murphy. *Machine Learning: A Probabilistic Perspective*. MIT press, 2012.

Andrew Y Ng, Michael I Jordan, and Yair Weiss. On spectral clustering: Analysis and an algorithm. In *Advances in neural information processing systems*, pages 849–856, 2002.

Dohyung Park, Constantine Caramanis, and Sujay Sanghavi. Greedy subspace clustering. In *Advances in Neural Information Processing Systems*, pages 2753–2761, 2014.

Xi Peng, Lei Zhang, and Zhang Yi. Scalable sparse subspace clustering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 430–437. IEEE, 2013.

Xi Peng, Huajin Tang, Lei Zhang, Zhang Yi, and Shijie Xiao. A unified framework for representation-based subspace clustering of out-of-sample and large-scale data. *IEEE transactions on neural networks and learning systems*, 27(12):2499–2512, 2015.

Mostafa Rahmani and George K Atia. Subspace clustering via optimal direction search. *IEEE Signal Processing Letters*, 24(12):1793–1797, 2017.

Gareth O Roberts, Richard L Tweedie, et al. Exponential convergence of langevin distributions and their discrete approximations. *Bernoulli*, 2(4):341–363, 1996.

Bernhard Schölkopf, Alexander Smola, and Klaus-Robert Müller. Nonlinear component analysis as a kernel eigenvalue problem. *Neural computation*, 10(5):1299–1319, 1998.

James R Schott. *Matrix Analysis for Statistics*. John Wiley & Sons, 2016.

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

Alexander Strehl and Joydeep Ghosh. Cluster ensembles—a knowledge reuse framework for combining multiple partitions. *Journal of Machine Learning Research*, 3(Dec):583–617, 2002.

Brian St Thomas, Lizhen Lin, Lek-Heng Lim, and Sayan Mukherjee. Learning subspaces of different dimension. *ArXiv preprint arXiv:1404.6841*, 2014.

Rene Vidal, Yi Ma, and Shankar Sastry. Generalized principal component analysis (gpca). *IEEE transactions on pattern analysis and machine intelligence*, 27(12):1945–1959, 2005.

Ulrike Von Luxburg. A tutorial on spectral clustering. *Statistics and computing*, 17(4): 395–416, 2007.

Chong You, Claire Donnat, Daniel P Robinson, and René Vidal. A divide-and-conquer framework for large-scale subspace clustering. In *Signals, Systems and Computers, 2016 50th Asilomar Conference on*, pages 1014–1018. IEEE, 2016a.

Chong You, Chun-Guang Li, Daniel P Robinson, and René Vidal. Oracle based active set algorithm for scalable elastic net subspace clustering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3928–3937. IEEE, 2016b.

Chong You, Daniel Robinson, and René Vidal. Scalable sparse subspace clustering by orthogonal matching pursuit. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3918–3927. IEEE, 2016c.

Pan Zhou, Yunqing Hou, and Jiashi Feng. Deep adversarial subspace clustering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1596–1604, 2018.