Subspace clustering without knowing the number of clusters: A parameter free approach

Vishnu Menon, Gokularam M, Sheetal Kalyani
Department of Electrical Engineering, Indian Institute of Technology Madras
Chennai, India - 600036
Email: ee16s301@ee.iitm.ac.in, ee17d400@smail.iitm.ac.in, skalyani@ee.iitm.ac.in

Abstract—Subspace clustering, the task of clustering high dimensional data when the data points come from a union of subspaces is one of the fundamental tasks in unsupervised machine learning. Most of the existing algorithms for this task involves supplying prior information in form of a parameter, like the number of clusters, to the algorithm. In this work, a parameter free method for subspace clustering is proposed, where the data points are clustered on the basis of the difference in statistical distribution of the angles made by the data points within a subspace and those by points belonging to different subspaces. Given an initial coarse clustering, the proposed algorithm merges the clusters until a true clustering is obtained. This, unlike many existing methods, does not involve the use of an unknown parameter or tuning for one through cross validation. Also, a parameter free method for producing a coarse initial clustering is discussed, which makes the whole process of subspace clustering parameter free. The comparison of algorithm performance with the existing state of the art in synthetic and real data sets, shows the significance of the proposed method.

I. INTRODUCTION

Data Clustering is the problem of categorizing entities in the given dataset into groups called clusters, so that the entities in the same cluster are more similar than those from different clusters. [1] provides comprehensive study of clustering algorithms. Very often, the dataset comprises of points from a Euclidean space and clustering problem reduces to finding the groups which are hidden among those vectors. In most techniques, distance measures are used as a similarity metric for clustering [1]. However, the conventional distance measures become unreliable in high dimensions. In a high dimensional space, the data points are sparsely located. It is shown in [2] that the distance between any two high dimensional points becomes equal as the dimension $n \rightarrow \infty$. Thus, most clustering algorithms which perform reasonably well in lower dimensions, fail in high dimensions. Over the years, several algorithms were developed for clustering data of large dimensions [3].

In many practical scenarios, the high dimensional data points are not uniformly distributed throughout the space, but lie approximately in low dimensional structure [4]. For example, the images of a face under different lighting conditions approximately lie in 9-dimensional subspace, even though they have a very large number of pixels [5]. Principal component analysis (PCA) [6] is a popular technique to retrieve a low dimensional linear subspace in which the high dimensional data points are concentrated. However, when there are multiple categories in the dataset, it is not appropriate to assume that the points lie in a single low dimensional subspace. For instance, if we have images of several faces under varying illumination conditions, then data will be lying in a union of multiple 9-dimensional subspaces. Subspace Clustering addresses this problem by grouping data points such that each group shall contain points from a single subspace of lower dimension [7].

Subspace clustering is used extensively for image representation and compression [8] and computer vision problems like motion segmentation [9], face clustering [10], image segmentation [11] and video segmentation [12]. It also finds applications in other fields including hybrid system identification [13], gene expression analysis [14], metabolic screening of newborns [15], recommendation systems [16] and web text mining [17]. Subspace clustering algorithms can be classified into four main types [7]: (i) algebraic, (ii) iterative, (iii) statistical, (iv) spectral clustering-based.

Spectral clustering-based techniques have gathered a lot of attention in recent years. These methods take two stage approach: First they find ‘affinity matrix’ and then performing spectral clustering [18] on it. Each entry in affinity matrix (sometimes referred as graph) denotes similarity between the corresponding pair of points. The difference between different spectral clustering-based techniques is how the affinity matrix is obtained. In recent years, affinity matrix is obtained using ‘self representation’ of each data point with respect to all the other data points. If the data points $m_i$ are arranged as columns of the matrix $M$, then the self representation is given by $M = MZ$ such that $Z_{ii} = 0$. After obtaining such $Z$, $abs(Z) + abs(Z^T)$ is used as the affinity matrix. Several techniques has been developed based on this idea. Sparse self representation enforces the columns of $Z$ to be sparse, $\ell_1$-minimization (as in Sparse Subspace Clustering (SSC) [19]) or Orthogonal Matching Pursuit (as in SSC(OMP) [20]) can be used to obtain such sparse representation. Thresholding-based Subspace Clustering (TSC) [21] uses least-squares representations in terms of nearest neighbours. Few other techniques utilize low rank self representation like Low-Rank Recovery (LRR) [22] and Low-Rank Subspace Clustering (LRSC) [23].

A most recent work [24] uses block diagonal self-representation for subspace clustering which performs better than several existing approaches. Another work, [25] provides a new approach called Innovation Pursuit which is an iterative method but can be integrated with spectral clustering to provide a new class of spectral clustering-based techniques. Cur-
rently, neural network-based clustering approaches are gaining popularity. Especially, autoencoder architecture is used to obtain sparse and low rank representation for subspace clustering. These techniques can recover non-linear low dimensional structures underlying the data.

Recently, the distributions of angles between data points has been used in to develop a parameter-free technique for outlier detection in high dimensions. So far, no technique has utilized the distributional property of angles for subspace clustering. Motivated by this, we develop a subspace clustering algorithm based on the statistical distribution of angles between high dimensional data points.

A. Motivation

Many clustering algorithms require the user to supply the number of clusters to be formed beforehand. In many situations, fixing the number of clusters apriori is not a good choice, especially when the knowledge about the dataset is limited. For example, in gene expression datasets, the number of clusters to be prefixed is not so clear. There are some clustering algorithms which require one or several parameters, if not the number of clusters. For example, DP-means determines the number of clusters based on a parameter \(\lambda\). When an algorithm requires one or more free parameters, the user has to set them using either cross-validation or prior knowledge about the dataset. However, parameter tuning is a difficult task and any incorrect tuning of parameters would result in huge performance degradation.

There are several techniques in literature to determine number of clusters for conventional distance-based clustering. There are also methods which tune for unknown parameters in the model - for instance, \(\lambda\)-means clustering tunes for the parameter \(\lambda\) in DP-means algorithm. If we use tuning for determining the parameters, the method will become dependent on the dataset and hence can give varying results on different realizations of the same data, in addition to the added computational overhead of tuning.

Most of the subspace clustering algorithms require the number of subspaces and/or other parameters. Especially, self-representation based techniques require regularization parameters to be set along with the number the subspaces. Even neural network-based clustering methods require setting of several hyper-parameters. As suggested in, we can estimate number of subspaces by eigengap heuristic and use it for the algorithm. However, the algorithms require additional parameters which has to be tuned.

Recently, parameter-free approaches have been developed in the areas of high dimensional outlier detection, sparse signal recovery, robust regression and these were shown to have results comparable with those which use the explicit knowledge about the parameters. Hence, we look for a parameter-free method for subspace clustering.

B. Contributions

Given the high-dimensional data points coming from the union of several low-dimensional subspaces, we propose an algorithm to achieve the true clustering without the knowledge of true number of clusters, \(L\). This essentially consists of two steps. First we get an initial clustering with a large number of clusters such that each cluster only contains points from one subspace. In the second stage, the clusters are merged to arrive at an optimal clustering. Given an initial clustering, we propose a method based on the statistical distance between angles of the data points to find the optimal clustering without having to prefix the number of clusters. We also suggest a parameter free approach for initial clustering.

The performance of the proposed algorithm is compared with state-of-the-art subspace clustering algorithms: SSC(OMP) and TSC and another recent algorithm: BDR-Z. We compare the algorithms in terms of Clustering Error (CE) and Normalized Mutual Information (NMI) on synthetic as well as real datasets. It is observed that the proposed algorithm performs on par with other methods even without the need for number of clusters or any other parameters.

C. Organization of the paper

The rest of the paper is organized as follows. In Section II we set up the problem and provide definitions and notations used in the paper. The proposed parameter-free algorithm for subspace clustering is introduced in Section III. In Section IV we provide the analysis of our algorithm under certain assumptions on the data model. Section V provides numerical results on synthetic and real datasets and compares the performance of our algorithm with other existing algorithms.

II. PROBLEM AND ESSENTIAL DEFINITIONS

The problem that we are addressing in this work is to find the optimal clustering of a dataset comprising of high dimensional points coming from a union of subspaces. Suppose we have \(N\) data points \(m_{i} \in \mathbb{R}^{n}, \forall i \in \{1, 2, \ldots, N\}\) and each \(m_{i} \in \mathcal{U}_{1} \cup \mathcal{U}_{2} \cup \ldots \cup \mathcal{U}_{L}\), where \(\mathcal{U}_{k}\)’s, \(k \in \{1, 2, \ldots, L\}\), \(L \ll N\) are subspaces in \(\mathbb{R}^{n}\) with dimensions \(r_{k}\)’s respectively. We assume that there are \(N_{k}\) points from subspace \(\mathcal{U}_{k}\) such that \(N = \sum_{k=1}^{L} N_{k}\).

**Definition 1.** A clustering of the dataset with \(K \geq 1\) clusters is defined as \(C_{K} = \{I_{1}, I_{2}, \ldots, I_{K}\}\), where \(I_{j}\)’s are mutually disjoint index sets such that \(\forall j = 1, 2, \ldots, K, I_{j} \subseteq \{1, 2, \ldots, N\}\) and \(I_{j} \neq \emptyset\) with \(\bigcup_{j=1}^{K} I_{j} = \{1, 2, \ldots, N\}\). We will call \(I_{j}\)’s as constituent clusters. If \(i \in I_{j}\), we say that \(j\) is the cluster label of the point \(m_{i}\).

**Definition 2.** The true clustering of the dataset is defined as the clustering \(C^{*}_{L} = \{I_{1}, I_{2}, \ldots, I_{L}\}\), where \(\forall j = 1, 2, \ldots, L, I_{j} = \{i \mid m_{i} \in \mathcal{U}_{k} \text{ for some } k \in \{1, 2, \ldots, L\}\}\) and \(|I_{j}| = N_{k}\), where \(|\cdot|\) denotes the cardinality of the set. i.e., each constituent cluster contains indices of all the points from a subspace and only the points from that subspace.

Here, we will be working with angles subtended by high dimensional data points and their distributions. We will be using the normalized data points \(x_{i} = \frac{m_{i}}{\|m_{i}\|_{2}}\), where \(\|\cdot\|_{2}\) denotes the \(\ell_{2}\) norm. These points \(x_{i} \in \mathbb{R}^{n}, \forall i \in \{1, 2, \ldots, N\}\) will
lie in the high dimensional hypersphere $\mathbb{S}^{n-1}$. Let $\theta_{ij}$ denote the principal angle between two data points $\mathbf{m}_i$ and $\mathbf{m}_j$, i.e.,

$$\theta_{ij} = \cos^{-1}(\mathbf{x}_i^T \mathbf{x}_j) \quad \text{where } \theta_{ij} \in [0, \pi]$$  

(1)

Also, let $S^{(j)}$, $j = 1, 2, \ldots, |S|$ denote the $j^{th}$ element in a set $S$. The following are few essential definitions.

**Definition 3.** The within cluster angle set for constituent cluster $I_k$ is defined as

$$W_k = \{ \theta_{ij} \mid i, j \in I_k, \; i < j \}$$

(2)

If $|I_k| = t_k$, then $|W_k| = \binom{t_k}{2}$ is the number of unique angles in the set.

**Definition 4** (Within cluster estimates). *Given a within cluster angle set $W_k$ for $I_k$, suppose $W_k^t \subseteq W_k$ with $|W_k^t| = t$, then the estimates corresponding to this subset is given as*

$$\mu_{wkt} = \frac{1}{t} \sum_j W_k^{t,(j)}$$

$$\sigma_{wkt}^2 = \frac{1}{t-1} \sum_j \left(W_k^{t,(j)} - \mu_{wkt}\right)^2$$

(3)

Here, $\mu_{wkt}$ and $\sigma_{wkt}^2$ are respectively sample mean and sample variance of elements of the set $W_k^t$ and $W_k^{t,(j)}$ is the $j^{th}$ element of the set $W_k^t$.

**Definition 5.** The between cluster angle set for two constituent clusters $I_k$ and $I_l$ is defined as

$$B_{kl} = \{ \theta_{ij} \mid i \in I_k, \; j \in I_l \}$$

(4)

Clearly, $|B_{kl}| = t_k t_l$, the number of possible cross angles.

**Definition 6** (Between cluster estimates). *Given a between cluster angle set $B_{kl}$ for $I_k$ and $I_l$, suppose $B_{kl}^t \subseteq B_{kl}$ with $|B_{kl}^t| = t$, then the estimates corresponding to this subset is given as*

$$\mu_{bkt} = \frac{1}{t} \sum_j B_{kl}^{t,(j)}$$

$$\sigma_{bkt}^2 = \frac{1}{t-1} \sum_j \left(B_{kl}^{t,(j)} - \mu_{bkt}\right)^2$$

(5)

Here, $\mu_{bkt}$ and $\sigma_{bkt}^2$ are respectively sample mean and sample variance of elements of the set $B_{kl}^t$ and $B_{kl}^{t,(j)}$ is the $j^{th}$ element of the set $B_{kl}^t$.

**Definition 7.** Given two constituent clusters $I_k$ and $I_l$, the distance $d_{kl}$ between them is defined as the Bhattacharyya distance $D_B$ between the distribution of angles in $W_k$ and $B_{kl}$, i.e.,

$$d_{kl} = D_B(\theta_{W_k}, \theta_{B_{kl}}),$$

where $\theta_{W_k}$ is the statistical distribution of angles in $W_k$ and $\theta_{B_{kl}}$ is the statistical distribution of angles in $B_{kl}$ and $D_B(\cdot, \cdot)$ is the Bhattacharyya distance between them.

We will also define a few measures for clusterings.

**Definition 8.** Score of a constituent cluster $I_j$ in a clustering $C_K$ is given by

$$\eta_j = \min_{l=1,2,\ldots,K, \; l \neq j} d_{jl}$$

(6)

Also, we define the partner of a cluster as the one which produces its score. i.e., if $J^* = \arg\min_{l=1,2,\ldots,K} d_{jl}$, then $I_{J^*}$ is the partner of $I_j$.

**Definition 9.** Score of a clustering $C_K$ is given by

$$\gamma_K = \min_{i=1,2,\ldots,K} \eta_i$$

(7)

Let $i^* = \arg\min_{i=1,2,\ldots,K} \eta_i$ and also let $I_{i^*}$ be the partner of $I_{i^*}$.

Then, we call $(I_{i^*}, I_{J^*})$ as a mergeable pair of the clustering $C_K$.

**Other Notations:** $\mathbb{E}[Y]$ denotes the expectation, $\text{var}(Y)$ the variance of the random variable $Y$. $N(\mu, \sigma^2)$ denote the normal distribution with mean $\mu$ and variance $\sigma^2$. Let $F_N(\cdot)$ denote the standard normal cdf, $F_N(y) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{y} e^{-x^2/2} dx$.

Let $\Gamma(\cdot)$ denote the gamma function, $\chi_{k}^2$ denote the chi-squared distribution with $k$ degrees of freedom, $\beta(a, b)$ denote standard beta distribution with parameters $a$ and $b$ and $\beta(c, d)[a, b]$ denote non-standard beta distribution with parameters $a$ and $b$ having support $[c, d]$. $\beta'(a, b)$ denote beta prime distribution with parameters $a$ and $b$. Also, w.p. indicates with probability’. $\lfloor x \rfloor$ is the floor of $x$, $\lceil x \rceil$ is the ceiling of $x$ and round$(x)$ is the closest integer to $x$. $O(\cdot)$ denotes the Big O notation for complexity and abs$(x)$ denotes the absolute value of $x$.

## III. Algorithm

In this section, we will explain the proposed algorithm for clustering. First, we will assume that some process gives us an initial coarse clustering $C_P$ and develop an algorithm for merging. Then, we will discuss a method which will select the appropriate clustering from the set of outputs after the merging process such that the final clustering estimate is close to the true clustering. We will also discuss possible initial clustering methods.

### A. Algorithm for merging

Suppose we have a coarse clustering $C_P$, with $P \gg L$. The proposed algorithm runs through from $P$ to 2, by merging clusters and then selects the appropriate clustering through methods described later. First, we will see the merging process starting with $K = P$.

1. Let the current clustering be $C_K$ with constituent clusters $I_1, I_2, \ldots, I_K$. For each $I_j$, calculate the distances between constituent clusters as per Definition 2 and then for each $I_j$, find its score $\eta_j$ using (6) and also get their respective partners.
2. Calculate the score of the clustering $\gamma_K$ as in (7) and find the mergeable pair in the current clustering $C_K$. Let the mergeable pair be $(I_{k1}, I_{k2})$.
3. Merge clusters $I_{k1}$ and $I_{k2}$ and form the new clustering $C_{K-1}$ with $K - 1$ constituent clusters.
4. Repeat steps 1 - 3 until $K = 2$.

In this merging, we form a series of clusterings $C_P, C_{P-1}, \ldots, C_2$ where each subsequent clustering is formed by merging the mergeable pair in previous clustering.
or in other words we simply combine the closest clusters in terms of the distance between distributions of the within cluster angles and between cluster angles.

The intuition behind this merging process is as follows. We hypothesize that the angles between points from the same subspace come from a statistical distribution and the angles between points from different subspaces follow a different distribution. A theoretical treatment of this hypothesis and the motivation behind it is provided in Sections IV-A and IV-B. This hypothesis implies that, when there are constituent clusters with points from the same subspace, the mergeable pair will contain at the bigger picture of a clustering, whenever a clustering has a low value close to 0. This hypothesis implies that, when there are constituent clusters with points from the same subspace to those in another, the partner of the constituent cluster will either of the distributions of angles in the constituent cluster and between the constituent clusters would be close to 0, provided we have enough angle samples in each set. To clarify, suppose we look at the clustering \( \hat{C}_P \). Take the constituent cluster \( I_1 \) and suppose that \( I_2 \) contains only points from the same subspace, then the angles in \( W_1 \) and \( B_{1k} \) come from the same distribution. This means that we have a low value close to 0 for \( d_{1k} \), which is the measure of divergence between the distributions of angles in \( W_1 \) and \( B_{1k} \). On the other hand, if say \( I_2 \) is a constituent cluster with points from a different subspace to those in \( I_1 \), then the angles in \( W_1 \) and \( B_{ij} \) come from different distributions and hence the distance \( d_{ij} \) will be high and bounded away from 0. We calculate \( \eta_i \) as the minimum distance made by the constituent cluster \( I_1 \). When the clustering contains at least another constituent cluster with only points from the same subspace as \( I_1 \), then the partner of that constituent cluster will be one of those clusters with only points from the same subspace, i.e., suppose for the above case, let \( I_k \) and \( I_m \) have only points from the same subspace as \( I_1 \), then the partner of \( I_1 \) will either of \( I_k \) and \( I_m \). Hence, when we look at the bigger picture of a clustering, whenever a clustering contains at least a pair of constituent clusters having only points from the same subspace, the mergeable pair will contain only points from the same subspace and the clustering score, \( \gamma_K \) will be very close to 0. In each step, we merge the most closest clusters in terms of \( d_{ij} \), which ensures that points from the same subspace get clustered together as we keep merging. The merging process is summarized in Algorithm 1.

**Algorithm 1** Procedure for Merging

**Input:** Initial clustering \( C_P \), normalized data matrix \( X \)

**Initial calculation:** Calculate \( \theta_{ij} \) \( \forall i,j = 1,2,\ldots,N \) as in \( (1) \)

**Initialization:** \( K = P \)

1: Calculate \( \gamma_K \) for the current clustering \( C_K \) as in \( (7) \)
2: Merge the mergeable pair in \( C_K \) and form \( C_{K-1} \)
3: Repeat steps 1 - 2 until \( K = 2 \)

**Output:** Clusterings \( C_P, C_{P-1} \ldots C_2 \)

**B. Selecting optimal clustering**

If we start with a clustering \( C_P \) such that each constituent cluster in \( C_P \) only contains indices of points from one subspace, then through the merging process described in the previous subsection, at some point we will arrive at a true clustering \( C_L^* \). At this point, no two constituent clusters have points from the same subspace, which means that no cluster pair can form a distance \( d_{ij} \) that is close to 0. Hence, the cluster score \( \gamma_L \) will take a jump from a value close to zero to a higher magnitude. This is what we try to exploit in our algorithm to find \( \hat{L} \), where we have a true clustering.

Also note that, here we compute the statistical distances using the angle samples we have in the within cluster and between cluster sets, i.e., for constituent clusters \( I_i \) and \( I_j \), \( \theta_{W_i} \) and \( \theta_{B_{ij}} \) are estimated distributions and the \( d_{ij} \) is an empirical Bhattacharyya distance. Suppose we have \( t \) angle samples, we can state the following on the behaviour of \( \gamma_K \):

a) Suppose for a clustering \( C_K \), \( I_i \) and \( I_j \) contains only points from the same subspace \( U_a \) such that the angles between the points in \( U_a \) are distributed with a distribution \( p_{U_a} \), then as the number of angle samples \( t \rightarrow \infty \), \( \theta_{W_i} \rightarrow p_{U_a} \) and \( \theta_{B_{ij}} \rightarrow p_{U_a} \Rightarrow d_{ij} \rightarrow 0 \). In other words, \( d_{ij} \) will be very close to 0 if one has large number of angle samples for estimating the distribution.

b) Hence in \( C_K \), for \( I_i \) and \( I_j \) as described in a), then \( \eta_i \rightarrow 0 \) given large number of angle samples. This is because of the definition of \( \eta_i \) which is the minimum distance that a constituent cluster makes.

c) From the above, for a clustering \( C_K \), if there exists at least one such \( I_i, I_j \) pair as a), then \( \gamma_K = \min_{t=1,2,\ldots,K} \eta_t \rightarrow 0 \) at large enough number of angle samples.

d) Suppose that such a pair as a) does not exist in \( C_K \), i.e., for any \( i \) and \( j \), \( I_i \) and \( I_j \) contain points from different subspaces, say \( U_a \) and \( U_b \) respectively. Then, as \( t \rightarrow \infty \), \( \theta_{W_i} \rightarrow p_{U_a} \) and \( \theta_{B_{ij}} \rightarrow q \), where \( q \) is the distribution of angles between points from different subspaces. Then, as \( t \) increases, \( d_{ij} \rightarrow D_B(p_{U_a}, q) \) and since these are different distributions, \( d_{ij} \) is bounded away from 0 \( \forall i,j \), which in turn leads to \( \gamma_K \) being bounded away from 0.

e) Also note that, whenever \( I_i \) and \( I_j \) contains a mixture of points from different subspaces, we cannot make an assertion on the nature of distributions in each set and hence the distance metric \( d_{ij} \) becomes unpredictable.

Hence while merging, suppose we arrive at a true clustering at \( K = L \), then there exists no more mergeable pair which have points from the same subspace. So, we can state the following assuming that \( t \rightarrow \infty \):

- For \( K > L \), \( \gamma_K \rightarrow 0 \).
- For \( K = L \), \( \gamma_K \) is bounded away from zero.
- For \( K < L \), \( \gamma_K \) behaviour is unknown.

The behavior of \( \gamma_K \) as we do merging in random subspace model based dataset is illustrated in Fig. 3. Now, we will describe two methods that can identify \( \hat{L} \) from the calculated \( \gamma_K \)’s.

**Algorithm 2** Using difference of \( \gamma_K \)’s

**Input:** \( C_P, C_{P-1} \ldots C_2 \) and associated \( \gamma_K \)’s

1: Find \( D_K = \gamma_K - \gamma_{K+1} \)
2: \( \hat{L} = \arg\max_{K=2,3,\ldots,P-1} D_K \)
3: \( \hat{C}_L = C_{\hat{L}} \)

**Output:** Estimated optimal clustering \( \hat{C}_L \).
The first method based on looking at the differences between successive $\gamma_K$ values is given in Algorithm 2. We know from the above discussion that $\gamma_{L+1} \to 0$ but $\gamma_L$ is bounded away from 0. Hence at $K = \hat{L}$, $D_L = \gamma_L - \gamma_{L+1}$ achieves a high value. When $K > \hat{L}$, both $\gamma_K$ and $\gamma_{K+1} \to 0$, which means that the difference $D_K \to 0$. Hence, we choose that instance, where the scores make the biggest jump, as our $\hat{L}$, as seen in step 2 of the algorithm. This method works well in case of large number of samples and especially when the number of samples used for estimation $t \to \infty$. But note that even in this case, the behaviour of $\gamma_K$ when $K < \hat{L}$ is not well defined. This is because in this case, more merging would have resulted in mergeable pairs having points coming from different subspaces and the behaviour of the statistical distance between these mergeable pairs becomes unpredictable. Although we expect these distances to be away from 0, there is a risk that we might pick a clustering with more merging than required as our estimate. Hence, we propose another scheme which also minimizes the number of clusters. Here we will use a method based on closeness of points, while keeping a minimum of 3 points per cluster, which is the only assumption made. The following steps describe the algorithm for initial clustering.

1. For each data point, find the two most closest points in terms of the acute angles between them, i.e. $\cos^{-1}(\text{abs}(x_i^T x_j))$. Let us call them allies of a point.
2. Starting from any point chosen randomly, form clusters with the point and its two allies. Run through the points forming such clusters, avoiding repetition of points in clusters. Here, a new cluster is formed only when a point and both its allies are not already allocated to a cluster. Hence, in this run, a lot of points go unallocated.
3. For all the points left out, add them to the cluster of its closest ally, if it has a cluster allocated or else add them to the cluster of its second ally. After this run, all points are added to some cluster. This forms our initial clustering, with at least 3 points in each constituent cluster.

IV. THEORETICAL ANALYSIS

To provide theoretical results, we require a model for data points from different clusters. In this section, we will describe the assumptions that we make on the data model and the motivation behind this assumption. Then, under this model, we analyse the algorithm theoretically and derive the threshold $\zeta_K$.

A. Assumptions used and their motivations

The distribution of angle between two high dimensional points is studied in detail in [30]. When independent points are chosen uniformly at random from $S^{n-1}$, the distribution of the angles between any two of them is approximately Gaussian with mean $\frac{\pi}{2}$ and variance $\frac{1}{n-2}$ [30]. Even if all the points are independent, the angles which involve the same point are only pairwise independent [46]. Let us look at the following model:

Model 1. The subspaces $U_i$’s, $i = 1, 2, \ldots, L$ are chosen uniformly at random from the set of all $r_i$ dimensional subspaces respectively and the normalized points in each subspace are sampled uniformly at random from the intersection of $U_i$ and $S^{n-1}$.

In this model, the following can be said about the distribution of angles $\theta_{ij}$.

Lemma 1 (Lemma 12 from [30]). Let $x_1, x_2, \ldots, x_n \in S^{n-1}$ be random points independently chosen with uniform distribution in $S^{n-1}$, and let $\theta_{ij}$ be defined as in (1), then $\theta_{ij}$’s are identically distributed with the pdf given by:

$$h(\theta) = \frac{1}{\sqrt{\pi}} \frac{\Gamma \left( \frac{3}{2} \right)}{\Gamma \left( \frac{n}{2} \right)} (\sin \theta)^{n-2} \quad \theta \in [0, \pi]$$

(8)

Lemma 2. $\theta_{ij}$’s converges in distribution to $N \left( \frac{\pi}{2}, \frac{1}{n-2} \right)$ as $n \to \infty$ and the rate of convergence is of $O \left( \frac{1}{n} \right)$.

Proof. Please refer to Appendix A.

Remark 1. $h(\theta)$ in (8) can be approximated by the pdf of Gaussian distribution with mean $\frac{\pi}{2}$ and variance $\frac{1}{n-2}$, for higher dimensions specifically for $n \geq 5$ as validated in [30].
We can use results proved in [31] to state the following:

**Lemma 3.** Suppose for Model [1], we have a true clustering \(C^*_L\) we have a true clustering

a) When \(i, j \in I_k\), with \(I_k\) corresponding to the subspace \(U_k\), \(\theta_{ij}\)’s are identically distributed with an expected value of \(\frac{\pi}{2}\) and its pdf is given by 
\[
\frac{1}{\sqrt{\pi} \frac{\Gamma\left(\frac{n}{2}\right)}{\Gamma\left(\frac{n+1}{2}\right)}} (\sin \theta)^{n-2}, \quad \theta \in [0, \pi].
\]

b) When \(i \in I_k, j \in I_l\), then \(\theta_{ij}\)’s are identically distributed with an expected value of \(\frac{\pi}{2}\) and its pdf is given by 
\[
\frac{1}{\sqrt{\pi} \frac{\Gamma\left(\frac{n}{2}\right)}{\Gamma\left(\frac{n+1}{2}\right)}} (\sin \theta)^{n-2}, \quad \theta \in [0, \pi].
\]

**Proof.** Please refer to Appendix A.

Through Lemma 3 and Remark 1, we can see that in Model 1 the angles between points coming from the same subspace and the angles between points of different subspaces both follow Gaussian distribution with different variances. Fig. 1 shows the histogram of within cluster angles and between clusters angles in Model 1. Here, we have considered two 7-dimensional subspaces in \(\mathbb{R}^{100}\). We could see that the Gaussian approximation holds good for both the distributions. The within cluster angles are more spread out (with approximate variance \(1/5\)) than between cluster angles (whose variance is approximately \(1/98\)). The empirical Bhattacharyya distance between distribution of within cluster angles and distribution of between cluster angles (with Gaussian approximation) is 0.4137, which is bounded away from 0.

A similar result can also be derived on the correlations between points in high dimensions and is given in Appendix C. We could look at the distribution of either correlations or angles as a discriminating factor and in this work we have focused on the distribution of angles. Model 1 gives us a framework to work with. However, it is too restrictive.

Hence, we generalize this model: We assume that the angles between points in the same subspace are approximately Gaussian distributed with some mean \(\mu_w\) and variance \(\sigma^2_w\), and those coming from points of different subspaces also approximately Gaussian distributed with some other mean \(\mu_b\) and variance \(\sigma^2_b\), with all of the parameters varying according to the data used and the pair of clusters considered. Essentially, the distributions of angles subtended by points in the same subspace and those from different subspaces have different distributions, both of which can be well approximated by Gaussian distributions. We can see that this assumption holds in many cases. Fig. 2 shows the histogram of within cluster angles and between clusters angles in Gene Expression Cancer RNA-Seq dataset [47] [48], which are approximately Gaussian distributed with different parameters [4]. This is the motivation behind making the following assumption in this work.

**Assumption 1.** For \(x_1, x_2, \ldots, x_{N_a} \in U_a\), the angles between them, \(\theta_{ij}\)’s are identically distributed as \(N(\mu_1, \sigma^2_1)\). For any \(i, j, k, l\) and \(1, \) the angles \(\theta_{ij}, \theta_{lk}\) and \(\theta_{kl}\) are not mutually independent, but are pairwise independent, i.e., \(\theta_{ij}\) and \(\theta_{ik}\) are independent, so are \(\theta_{ij}\) and \(\theta_{kl}\).

Also note that, trivially the angles are mutually independent if they do not have a common point involved. We assume throughout this work that the angles formed by points within a subspace are distributed as above and those between points coming from different subspaces have a different distribution. As seen in the previous section, the proposed algorithm for finding the optimal clustering from a given clustering with large number of constituent clusters is based on distances between distributions of angles within and between constituent clusters. We will model the angles using Assumption 1 and derive the threshold \(\zeta_K\) theoretically in the next subsection.

**B. Theoretical behaviour of scores**

For theoretical analysis, we will work under Assumption 1. Consider a clustering \(C_K\). Let \(I_i\) and \(I_j\) be two constituent

1For this illustration, we have considered the clusters - LUAD and PRAD from the dataset.
clusters in $C_K$. Let $|I_i| = \omega^i$ and $|I_j| = \omega^j$. Here, we will do the following:

i. Characterize $d_{ij}$ mathematically looking at its statistical properties when $I_i$ and $I_j$ contains only points from the same subspace $U_t$.

ii. We will use this to develop the threshold $\zeta_K$ which is a high probability upper bound on $\gamma_K$ whenever there exists some $I_i, I_j \in C_K$ which contains only points from the same subspace.

When we use Assumption 1 $d_{ij}$ between constituent clusters $I_i$ and $I_j$ is the Bhattacharyya distance between two normal distributions. Also, since we use estimates for its calculation, this is an empirical Bhattacharyya distance. The properties of the empirical Bhattacharyya distance between normal distributions is studied in [49]. $d_{ij}$ can be written as:

\[
d_{ij} = \frac{1}{4} \left[ \frac{(\mu_{w,t,i} - \mu_{b,t,j})^2}{\sigma_{w,t,i}^2 + \sigma_{b,t,j}^2} + \log_e \left( \frac{1}{4} \left( \frac{\sigma_{w,t,i}^2 + \sigma_{b,t,j}^2}{\sigma_{w,t,i}^2 + \sigma_{b,t,j}^2} + 1 \right) \right) \right]
\]  

(9)

As stated, we will look the case when $I_i$ and $I_j$ contains points only from the same subspace. Under Assumption 1 we have only assumed pairwise independence of angles and not mutual independence when they involve the same data point. In this section, we will ensure that the estimates are generated by independent angles by designing the subsets $W_{ij}^t$ and $B_{ij}^t$ as such. We want to ensure that the samples used for the estimates $\mu_{w,t,i}$ and $\mu_{b,t,j}$ are independent samples and also ensure that $\mu_{w,t,i}$ and $\mu_{b,t,j}$ are independent with respect to each other. For this, we have to ensure that we only pick at most 2 angles formed by a point in these clusters. Lemma 7 in Appendix B designs such subsets $W_{ij}^t$ and $B_{ij}^t$ with $t_i = \lfloor \frac{t}{M} \rfloor$ and $t_j = \min(\omega^i, \omega^j)$, where the samples in these subsets are independent.

Let $t = \min(t_i, t_j) = \min(\lfloor \frac{t}{M} \rfloor, \min(\omega^i, \omega^j))$. Further in the analysis, we will assume that we use only $t$ samples each from the sets $W_{ij}^t$ and $B_{ij}^t$ for getting our estimates. This helps in simplifying the analysis without compromising on its crux. Let these sets be $W_i^t$ and $B_i^t$ and the corresponding estimates be $\mu_{w,t,i}, \sigma_{w,t,i}^2$ and $\mu_{b,t,j}, \sigma_{b,t,j}^2$. We will first look at the distribution and properties of these estimates.

Lemma 4. When $I_i$ and $I_j$ contains only points from the same subspace $U_t$, we can say the following about the estimates using $W_i^t$ and $B_i^t$ under Assumption 1:

\[
\mu_{w,t,i}, \mu_{b,t,j} \sim N \left( \nu_{a, w}, \frac{\rho_a^2}{t} \right)
\]

(10)

\[
\sigma_{w,t,i}^2, \sigma_{b,t,j}^2 \sim \frac{\rho_a^2}{t-1} \chi^2(t-1)
\]

(11)

Also, the estimates $\mu_{w,t,i}$ and $\sigma_{w,t,i}^2$ are independent and so too are $\mu_{b,t,j}$ and $\sigma_{b,t,j}^2$.

Proof. Since we are working with independent samples as designed in Lemma 7 sampled from a Gaussian distribution with mean $\nu_a$ and variance $\rho_a^2$ as per Assumption 1 this is a straightforward result on sample mean and variance of a Gaussian random sample as in Theorem 5.3.1 in [50].

Through the choosing of the subsets on which the estimates are calculated and lemmas stated above, we have established that the estimates using $W_i^t$ and $B_i^t$ are all independent of each other and has distributions as in (10) and (11). Now, we will look at $d_{ij}$, given in (9). Let us denote $X_{ij} = (\mu_{w,t} - \mu_{b,t})$, $Y_{ij} = \mu_{w,t}^2 + \mu_{b,t}^2$, $Z_{ij} = \sigma_{w,t}^2 / \sigma_{b,t}^2$ and also $U_{ij} = \frac{X_{ij}^2}{Y_{ij}}$ and $V_{ij} = Z_{ij} + \frac{1}{Z_{ij}}$. Then, $d_{ij} = \frac{1}{4} \left[ U_{ij} + \log_e \left( \frac{V_{ij}}{4} + \frac{1}{2} \right) \right]$. We will first look at $U_{ij}$.

Lemma 5. When $I_i$ and $I_j$ contains only points from the same subspace $U_t$ under Assumption 1:

\[
U_{ij} \leq \frac{2}{\sqrt{t-1}} \text{ w.p. } F_{\beta\left(\frac{1}{2}, t-1\right)} \left( \frac{t}{(t-1)^{1/2}} \right),
\]

(12)

where $F_{\beta(a,b)}$ is the cdf of a beta prime distribution with parameters $a$ and $b$.

Proof. Please refer to Appendix B.

Lemma 6. Suppose $c = 4 \left( \frac{\sqrt{c^2 - 4}}{2} \right)$. When $I_i$ and $I_j$ contains only points from the same subspace $U_t$ under Assumption 1:

\[
\log_e \left( \frac{V_{ij}}{4} + \frac{1}{2} \right) \leq \frac{2}{\sqrt{t-1}} \text{ w.p. } F_{\beta\left(\frac{1}{2}, t-1\right)} \left( \frac{1}{c} \right) - F_{\beta\left(\frac{1}{2}, t-1\right)} \left( \frac{e^c - 4}{c} \right)
\]

(13)

Proof. Please refer to Appendix B.

Now we have all the results with which we can prove the main theorem. It is stated as follows.

Theorem 1 (Derivation of $\zeta_K$). Suppose clustering $C_K$ with mergeable pair $(I_1, I_2)$ contains at least one pair of constituent clusters which contain only indices of points from the same subspace. Consider that we have used independent samples for estimation as in Lemma 7 with $t_K = \min(\lfloor \frac{t}{M} \rfloor, \omega^i, \omega^j)$, where $|I_1| = \omega^i$ and $|I_2| = \omega^j$. Then under Assumption 1 $\gamma_K \leq \zeta_K$ w.p $F_{\beta\left(\frac{1}{2}, t_K-1\right)} \left( \frac{t_K}{(t_K-1)^{1/2}} \right) + F_{\beta\left(\frac{1}{2}, t_K-1\right)} \left( \frac{e^{c\sqrt{\frac{t_K}{c^2 - 4}}}}{2} \right) - 1$, where $c = 4 \left( \frac{e\sqrt{t_K - 1}}{1 - \frac{1}{2}} \right)$, when:

\[
\zeta_K = \frac{1}{\sqrt{t_K-1}}
\]

(14)

Proof. Note that $d_{ij} \rightarrow 0$ when $I_i$ and $I_j$ have only points from the same subspace and is bounded away from zero otherwise. When a clustering $C_K$ contains at least one pair of constituent clusters which contain only indices of points from the same subspace, then the mergeable pair will be one of those pairs with points from the same subspace. Hence, $I_1$ and $I_2$ contains only indices of points from the same subspace. We know $\gamma_K = d_{i^*, j^*}$. So,

\[
\gamma_K = \frac{1}{4} \left[ U_{i^*, j^*} + \log_e \left( \frac{V_{i^*, j^*}}{4} + \frac{1}{2} \right) \right]
\]
Further, using Lemma 5: \[ \log_\epsilon \left( \frac{V^{1.5} + \frac{1}{2}}{2} \right) \leq \frac{2}{\sqrt{K-1}} \text{ w.p } \epsilon. \]

Further, using Lemma 5: \[ F_{\beta}(t_{K-1}^2) - F_{\beta}(t_{K-1}^2) \leq \frac{c}{t_{K-1}^2} \]

where \( c = 4 \left( 1 + \sqrt{t_{K-1}^2 - \frac{1}{4}} \right) \).

We know that \( \mathbb{P}(A \cap B) \geq \mathbb{P}(A) + \mathbb{P}(B) - 1 \).

Hence, we can say \( \mathbb{P} \geq F_{\beta}(t_{K-1}^2) - \frac{t_{K-1}}{K-1} \) + \( F_{\beta}(t_{K-1}^2) - \frac{t_{K-1}}{K-1} \) - 1,

\[ \gamma_K \leq \frac{1}{4} \left[ \frac{2}{\sqrt{t_{K-1}^2}} + \frac{2}{\sqrt{t_{K-1}^2}} \right] = \frac{1}{\sqrt{t_{K-1}^2}} \]

We call \( \zeta_K = \frac{1}{\sqrt{K-1}}. \)

Table I gives a numerical perspective of the bound and its probability. As seen, the probability increases with \( t_K \) and \( \zeta_K \) falls to 0.

TABLE I

| \( t_K \) | 10 | 50 | 100 | 150 |
|----------|----|----|-----|-----|
| \( \zeta_K \) | 0.3162 | 0.1414 | 0.1 | 0.0816 |
| \( \mathbb{P}(\gamma_K \leq \zeta_K) \) | 0.95511 | 0.99948 | 0.99997 | 0.99999 |

Remark 2. Suppose we are given an initial clustering \( C_P \) such that each constituent cluster in \( C_P \) contains only indices of points from the same subspace. Then under Assumption 7 an estimate \( \hat{C}_L \), which is arrived at by the merging process in Algorithm 1 and the selection process in Algorithm 3, is a true clustering with a high probability.

The basis for this remark is as follows. We have seen that Algorithm 1 always selects the closest clusters in terms of \( d_{ij} \) to merge and hence the merging process keeps merging constituent clusters with only indices of points from the same subspace until it reaches a true clustering. Through Theorem 1 we have shown that until then \( \gamma_K \leq \zeta_K \) with a high probability. And at this point, since the mergeable pair will contain points from different subspaces, \( \gamma_K \) will be high and we select this crossover instance as the estimated clustering in Algorithm 3. We demonstrate this behaviour of \( \gamma_K \) in Fig. 3 through a simulation. Here, we have considered Model 1 and the data points are drawn from \( L = 6 \) subspaces, each of dimension 7 in \( \mathbb{R}^{100} \). It can be seen that \( \gamma_K \) is close to 0 when \( K > 6 \) and when \( K = 6 \), \( \gamma_K \) is bounded away from 0 and \( \gamma_K > \zeta_K \).

C. A Note on Complexity

The complexity for computing all the \( \binom{N}{2} \) angles between \( n \) dimensional points is \( O(N^2 n) \). For \( P \) constituent clusters in the initial clustering, one has to compute the statistical distances for all possible combinations. When these are computed as described in the previous sections, one could pre compute the sum of angles and the squared sum of angles for all possible combinations, leading to a complexity of \( O(P^2) \). In the merging process, using the pre computed values, one can update the distances using simple arithmetic, leading to a complexity of \( O(P) \). Hence, the overall complexity is \( O(\max(N^2 n, P^2)) \). Runtime comparisons are provided in the next section.

V. NUMERICAL VALIDATIONS

In this section, we validate the performance of our algorithm numerically through simulations in synthetic as well as real datasets. We compare the performance with state-of-the-art subspace clustering algorithms - SSC(OMP) [20] and TSC [21] and another recent algorithm BDR-Z [24]. All these algorithms require us to provide an estimate of the number of clusters \( L \) apriori, which we denote as \( L_{in} \). However, as mentioned in [21], we can estimate \( L \) by eigengap heuristic and use it as \( L_{in} \) for TSC - we call this technique ‘TSC auto \( L \’). Throughout the experiments, we have used the parameters as follows: \( k = 5 \) for SSC(OMP), \( \lambda = 1 \), \( \gamma = 0.1 \), \( \epsilon = 10^{-3} \) for BDR-Z and \( q = \max \left( 3, \frac{\text{round}(\sqrt{2} L_{in})}{20} \right) \) for TSC. For our algorithm we use the initial clustering described in Section III-C. The best performance in each experiment is given in boldface. All the results are obtained in Matlab R2018a on a system with an Intel(R) Core (TM) i7-6700 CPU at 3.40GHz and 3.41GHz, 16GB RAM.

A. Metrics for comparison

We compare the performance in terms of Clustering Error (CE) and Normalized Mutual Information (NMI). Clustering error is defined as the fraction of points misclassified by the algorithm. It is computed as follows. Let \( c_i \) and \( \tilde{c}_i \) be the

2Since not much detail has been provided by the authors on how to tune these parameters, we use the values used in their codes for some datasets like MNIST, which is also used in this paper.
1, 2, . . . , N denote respectively the true cluster label of the point \( \mathbf{m}_i \) and the label assigned to it by the algorithm. Then,

\[
CE = \min_\pi \left( 1 - \frac{1}{N} \sum_{i=1}^{N} I(a_i, \pi(c_i)) \right)
\]

where \( I(a, b) = 1 \) if \( a = b \), 0 otherwise. \( \pi(c_i) \) is the one-one reassignment of the label \( \tilde{c}_i, \pi(\tilde{c}_i) \in \{1, 2, . . . , L\} \) where \( \tilde{c}_i \in \{1, 2, . . . , L\} \). We compute Normalized Mutual Information \( \text{NMI} \) as

\[
\text{NMI} = \frac{\mathcal{I}(C_L; \tilde{C}_L)}{0.5 \left( \mathcal{H}(C_L) + \mathcal{H}(\tilde{C}_L) \right)}
\]

where \( \mathcal{H}(\cdot) \) and \( \mathcal{I}(\cdot; \cdot) \) respectively denote the empirically computed entropy of the cluster and mutual information between clusters. CE close to 0 and NMI close to 1 are desirable.

### B. Results on Synthetic Datasets

#### Random Subspace Models

We first illustrate the results if the data points are from Model 1. It is known that for the points \( x_k \)'s to be uniformly distributed in \( U_k \cap \mathbb{S}^{n-1} \), the individual coordinates of \( \mathbf{m}_i \)'s have to be sampled independently from a standard normal distribution \[30\]. We call this dataset as ‘Subspace-Normal’. We also show the results in random subspace model when we sample individual coordinates of \( \mathbf{m}_i \)’s from a standard uniform distribution. We call this one as ‘Subspace-Uniform’. For both these scenarios, we have taken 1000 points in \( \mathbb{R}^{1000} \), vary \( L \) between 2 and 8 with roughly equal sized true clusters for each \( L \). The dimension of each subspace is chosen randomly between 7 and 15. For each \( L \) we performed 50 trials.

The results are summarized in Table III. It was observed that for these models all the existing algorithms resulted in perfect clustering with CE = 0 and NMI = 1 for every case and every trial when \( L_{in} = L \). However, SSC(OMP) and BDR-Z is bad
when we take $L_{\text{est}} \neq L$ as seen in Table II. TSC auto $L$ performs better than these two, but our method is the only one which results in perfect recovery of subspaces.

Table III provides mean, median and standard error of estimated number of clusters by our algorithm and TSC auto $L$. From these statistics, we can conclude that our algorithm determines the number of clusters exactly, while TSC auto $L$ results in very large number of clusters occasionally.

We compare the average run times of the algorithms on Subspace-Uniform dataset in Figure 4. We could see that our algorithm is faster than others in all the cases. TSC auto $L$ is the slowest taking 20.7227 seconds when $L = 7$. Our algorithm is the fastest taking 0.6144 seconds when $L = 7$. SSC(OMP) and BDR-Z respectively took 1.6687 and 7.6993 seconds for the same $L$.

![Fig. 4. Run time comparison of subspace clustering algorithms on subspace-uniform model dataset](image)

**Dirichlet Process Model**: To illustrate the versatility of our algorithm to other data models, we show the results when the data points obtained from Dirichlet Process (DP) [33]. Results are summarized in Table IV. For each of the listed $\rho/\sigma$, we performed 50 trials, generating 1000 points from $\mathbb{R}^{100}$ each time. Here, $\rho$ represents spread of cluster centroids and $\sigma$ represents spread of points around their respective centroid. Larger $\rho/\sigma$ signifies widely separated dense clusters. For this dataset, we also include results from $\lambda$-means algorithm [39], a parameter tuning algorithm developed for DP model data. It tunes for the parameter $\lambda$ in DP-means algorithm [33]. For each $\rho/\sigma$, we tune for $\lambda$ in the first trial and use that value for remaining trials. It is observed that $\lambda$-means performs bad when $\rho/\sigma$ is small. This is because it is a distance-metric based algorithm and at smaller $\rho/\sigma$, clusters are not well separated. It can be seen from Table IV that even if we provide true $L$ for SSC(OMP), BDR-Z and TSC, they don’t perform well. TSC is the best among these three methods achieving mean CE of 12.46% and mean NMI of 0.8172 when $\rho/\sigma = 17$, while for the same $\rho/\sigma$, ours is the best with mean CE = 0.07% and mean NMI = 0.9914. We can observe that our algorithm perform consistently better for all $\rho/\sigma$ with mean CE $\leq 0.08\%$ and mean NMI $\geq 0.9537$.

**C. Results on Real datasets**

We illustrate the performances of our algorithm and other subspace clustering algorithms on some real datasets. Table V provides the details of the datasets we have used. The first two datasets are gene expression datasets. In such applications, number of clusters in the dataset may not be known apriori. First four datasets are smaller and we perform single trial on these datasets. We have provided the results for SSC(OMP), BDR-Z and TSC when we provide true number of clusters and when we provide $L_{\text{est}} = 6$. We have also provided results for TSC when we use estimate $L$ estimated by eigengap heuristic. The results are given in Table VIA.

We could see that our algorithm outperforms SSC(OMP) and BDR-Z when they are provided with wrong number of clusters in all the datasets. The performance of our algorithm is comparable to other algorithms when we provide right number of clusters to them. Though our algorithm is developed for high dimensional data, its performance in wireless indoor

![Table IV](image)

**Table V**

Details of some real datasets

| Dataset                                | $n$ | $N$ | $L$ |
|----------------------------------------|-----|-----|-----|
| Gene Expression Cancer RNA-Seq [47] [48]| 16383 | 801 | 5   |
| Novartis multi-tissue [52]             | 1000 | 103 | 4   |
| Wireless Indoor Localization [53] [58] | 7    | 2000| 4   |
| Phoneme [59]                           | 256  | 4509| 5   |
| MNIST [65]                             | 784  | 70000| 10  |
TABLE VI
PERFORMANCE OF SUBSPACE CLUSTERING ALGORITHMS ON SOME REAL DATASETS

| Algorithm | Gene Expression Cancer RNA-Seq | Novartis multi-tissue | Wireless Indoor Localization | Phoneme | Average across datasets |
|-----------|--------------------------------|------------------------|----------------------------|---------|------------------------|
| CE        | NMI                            | CE                     | NMI                        | CE      | NMI                    | CE      | NMI |
| Ours      | 0.0087                          | 0.9769                 | 0.1845                     | 0.7247  | 0.2790                 | 0.1720  | 0.7510 |
| SSC(OMP)  | 0.0137                          | 0.9594                 | 0.2524                     | 0.6222  | 0.7445                 | 0.0112  | 0.4154 |
| SSC(OMP)  | 0.0762                          | 0.9248                 | 0.2718                     | 0.7917  | 0.7450                 | 0.0158  | 0.4999 |
| BDR-Z     | 0.0025                          | 0.9901                 | 0.1359                     | 0.7147  | 0.1525                 | 0.6697  | 0.4717 |
| BDR-Z     | 0.0587                          | 0.9454                 | 0.3010                     | 0.6540  | 0.3480                 | 0.5461  | 0.5207 |
| TSC       | 0.0012                          | 0.9948                 | 0.3592                     | 0.6637  | 0.0285                 | 0.9082  | 0.3094 |
| TSC       | 0.0612                          | 0.9416                 | 0.1748                     | 0.8013  | 0.0580                 | 0.8780  | 0.3276 |
| TSC            | 0.0612                          | 0.9441                 | 0.7476                     | 0.6067  | 0.9975                 | 0.3085  | 0.3227 |

TABLE VII
ESTIMATED NUMBER OF CLUSTERS ON SOME REAL DATASETS

| Dataset               | Ours | TSC auto L |
|-----------------------|------|------------|
| Gene Expression Cancer RNA-Seq | 6    | 6          |
| Novartis multi-tissue  | 5    | 21         |
| Wireless Indoor Localization | 11   | 1999       |
| Phoneme               | 5    | 3          |

localization dataset illustrates that we can use it for low dimensional datasets also. In the same dataset, performance of SSC(OMP) is very bad even if true L is provided. This is because the chosen value of parameter k for SSC(OMP) is not suitable for this dataset. Our algorithm, being non-parametric is free from such defects. The averaged results across datasets suggests that our algorithm, with mean CE of 16.1% and mean NMI of 0.7687, performs better than SSC(OMP) and BDR-Z and on par with TSC.

Table VII shows the estimated number of clusters by our algorithm and TSC auto L. The results reconfirms our previous analysis that TSC auto L occasionally selects very large number of clusters. For instance, consider the wireless indoor localization dataset. TSC auto L provides 1999 clusters i.e., it considered each point (except a pair) as a cluster. This results in high CE for the same. Here, our algorithm provides 11 clusters and CE is pretty low suggesting that the excess 4 clusters are smaller in size. Also, in Phoneme dataset, our method has recovered the exact number of clusters and it predicted one additional cluster in Gene Expression Cancer RNA-Seq and Novartis multi-tissue datasets. Hence, our algorithm is better at finding number of clusters.

**MNIST dataset:** Since MNIST dataset is large we perform several trials. We vary number of clusters, L from 2 to 10. For each L, we perform 50 trials. In each trial, for each L, we randomly select L out of 10 available clusters with each cluster having 1000 randomly sampled points. For this dataset, we have provided our algorithm with 25 equally sized error-free initial clusters for each trial. The mean CE and NMI are provided in Table VIII. The clustering error of our algorithm, provided with good initial clustering, is on par with the methods SSC(OMP), BDR-Z and TSC which are provided with true number of clusters. It is also observed that our algorithm outperforms all these in terms of mean NMI.

VI. Conclusion

In this paper, we have proposed a parameter free algorithm for subspace clustering, which uses the characteristics of distribution of angles subtended by the points. The algorithm, which works without apriori parameter knowledge, starts with a coarse clustering and merges the clusters iteratively until the clustering score crosses a threshold. We have theoretically analysed the algorithm and derived the threshold under an assumption on the data model and also proposed a parameter free initial clustering method. The performance of the proposed algorithm has been studied numerically in synthetic as well as real datasets. It has been observed that the proposed method performs on par with other existing methods which use true parameter knowledge, in terms of clustering error and estimated number of clusters and outperforms them in many cases, especially when the true parameters are unknown.

APPENDIX A
PROOFS OF RESULTS IN SECTION IV-A

**Proof of Lemma 2** Let $t = \sqrt{n - 2} \left(\theta_{ij} - \frac{\pi}{2}\right)$. Using (8), the log density of t can be obtained as

$$
\log g(t) = \log k_n + (n - 2) \log \cos \left(\frac{t}{\sqrt{n - 2}}\right),
$$

$$
t \in \left[-\sqrt{n - 2} \frac{\pi}{2}, \sqrt{n - 2} \frac{\pi}{2}\right]
$$
where $k_n$ is the normalization term depending on $n$ alone. Using Taylor expansion about $t=0$,
\[
\log q(t) = \log k_n + (n-2) \left[ -\frac{t^2}{2(n-2)} - \frac{t^4}{12(n-2)^2} - \cdots \right] = \log k_n - \frac{t^2}{2} + O\left(\frac{1}{n}\right)
\]
\[
\Rightarrow q(t) \propto e^{-\frac{t^2}{2}} \text{ at the rate of } O\left(\frac{1}{n}\right)
\]
Thus, $t \xrightarrow{D} N(0, 1)$ and hence $\theta_{ij} \xrightarrow{D} N\left(\frac{\pi}{2}, \frac{1}{n-2}\right)$ at the rate of $O\left(\frac{1}{n}\right)$.

**Proof of Lemma 8** Part a) is straight from Lemma 2 in [31]. Note that the angle between two points from different subspaces in Model 1 is statistically same as that between two points chosen uniformly at random from $S^{n-1}$ as in Lemma 9 in [31] and part b) follows Lemma 1 in [31].

**APPENDIX B**

**PROOF OF RESULTS IN SECTION IV-B**

The following lemma is used to design subsets of independent samples for calculation of estimates.

**Lemma 7.** Let the constituent clusters be $I_i = \{i_1, i_2, \ldots, i_{\omega_i}\}$ and $I_j = \{j_1, j_2, \ldots, j_{\omega_j}\}$ and let them contain only indices of points from the same subspace $U_\omega$. Then, $W_i = \{\theta_{ip,i_q} | p, q = 1, 2, \ldots, \omega, p < q\}$ and $B_{ij} = \{\theta_{ip,j_q} | p = 1, 2, \ldots, \omega^i, q = 1, 2, \ldots, \omega^j\}$. Define $W_i^{t} = \{\theta_{ip,j_k} | k = 1, 2, \ldots, \lfloor \frac{\omega^i}{2} \rfloor\}$. Then, $|W_i^{t}| = t_i = \lfloor \frac{\omega^i}{2} \rfloor$. Let $\omega = \min(\omega^i, \omega^j)$. Define $B_{ij}^{t_j} = \{\theta_{ip,j_k} | p = 1, 2, \ldots, \omega\}$. Then, $|B_{ij}^{t_j}| = t_j = \omega$. The following holds on the estimates under Assumption 1

a) $\mu_{w_i,t_i}$ and $\mu_{b_{ij},t_j}$ are calculated using independent angle samples.

b) $\mu_{w_i,t_i}$ and $\mu_{b_{ij},t_j}$ are independent.

c) The corresponding variance estimates $\sigma_{w_i,t_i}^2$ and $\sigma_{b_{ij},t_j}^2$ are also independent.

**Proof.** As seen in the design of the set $W_i^{t_i}$, only one angle is chosen per data point in the set and hence the estimates using this set uses independent samples under Assumption 1. Also, the between angle set $B_{ij}^{t_j}$ contains only one angle per data point which are independent. This proves part a). When we see both the sets together, they contain at most 2 angles formed by a point and under Assumption 1 the angles are pairwise independent if it involves the same point. Hence, $W_i^{t_i} \cup B_{ij}^{t_j}$ contains only independent samples. Hence, the estimates which use disjoint samples from $W_i^{t_i} \cup B_{ij}^{t_j}$ are independent. This proves part b) and c).

**Proof of Lemma 8** Under Assumption 1 we know the distributions of estimates from Lemma 4 $\mu_{w_i,t} - \mu_{b_{ij},t}$ is the difference of independent Gaussian random variables with same mean.

\[
\Rightarrow X_{ij} \sim N\left(0, \frac{2\rho_0^2}{t}\right) \quad \text{and} \quad \frac{t}{2\rho_0^2} X_{ij}^2 \sim \chi_1^2
\]

The last step is just using the definition of a chi-squared random variable. Looking at $Y_{ij}$, we use the result that if $A_1 \sim \chi_{a_1}^2$, $A_2 \sim \chi_{a_2}^2$, then $A_1 + A_2 \sim \chi_{a_1+a_2}^2$. Using this we know, $\sigma_{w_i,t}^2 + \sigma_{b_{ij},t}^2 \sim \frac{\rho_0^2}{t} X_{2}\chi_{2(t-1)}^2$.

\[
\Rightarrow \frac{t-1}{\rho_0^2} X_{ij}^2 \sim \chi_{2(t-1)}^2
\]

Now, $U_{ij} = \frac{X_{ij}^2}{Y_{ij}} = \frac{2(t-1)}{t} A_1 A_2$, where $A_1 \sim \chi_{2}^2$, and $A_2 \sim \chi_{2(t-1)}^2$. Using the result that the ratio of two independent chi-squared random variables follows a beta prime distribution, $\chi_{2}^2 \sim \beta'(\frac{1}{2}, t-1)$. Hence,

\[
\frac{t}{2(t-1)} U_{ij} \sim \beta'\left(\frac{1}{2}, t-1\right)
\]

(15)

So, w.p $F_{\beta'\left(\frac{1}{2}, t-1\right)}\left(\frac{t}{(t-1)^2}\right) \Rightarrow U_{ij} \leq \frac{2}{\sqrt{t-1}}$

**Proof of Lemma 8** Note that $Z_{ij} = \frac{\sigma_{w_i,t}^2}{\sigma_{b_{ij},t}^2} = \frac{A_1}{A_2}$, where $A_1 \sim \chi_{t-1}^2$ and $A_2 \sim \chi_{t-1}^2$. Using the result that the ratio
of two independent chi-squared random variables follows a beta prime distribution \[ Z_{ij} \sim \beta' \left( \frac{\nu - 1}{2}, \frac{\nu - 1}{2} \right) \]. Suppose \( \epsilon = \frac{2}{\sqrt{n}} \). We are looking at the condition,

\[
\log \left( \frac{V_{ij}}{4} + \frac{1}{2} \right) \leq \epsilon \quad \Rightarrow \quad V_{ij} \leq 4 \left( e^\epsilon - \frac{1}{2} \right)
\]

\[
\Rightarrow V_{ij} \leq c \quad \text{where} \quad c = 4 \left( e^\epsilon - \frac{1}{2} \right)
\]

Hence, we are looking at the following probability:

\[
\mathbb{P}(V_{ij} \leq c) = \mathbb{P}(Z_{ij} + \frac{1}{2} \leq c) = \mathbb{P}(Z_{ij}^2 - cZ_{ij} + 1 \leq 0)
\]

Looking at the following equation:

\[
Z_{ij}^2 - cZ_{ij} + 1 = 0 \quad \Rightarrow \quad Z_{ij} = \frac{c \pm \sqrt{c^2 - 4}}{2}
\]

Note that \( \epsilon \geq 0 \quad \Rightarrow \quad e^\epsilon \geq 1 \)

\[
\Rightarrow \quad 16 \left( e^\epsilon - \frac{1}{2} \right)^2 \geq 4 \quad \Rightarrow \quad e^2 \geq 4
\]

Hence, solution always exists. Now, we need to look at the regions where the above condition holds to evaluate the probability. Take a point \( z_0 = \frac{c}{4} \), which lie between the solutions, i.e., \( \frac{c - \sqrt{c^2 - 4}}{2} \leq z_0 \leq \frac{c + \sqrt{c^2 - 4}}{2} \). Evaluating the expression at \( z_0 \):

\[
z^2_0 - cZ_{ij} + 1 = 1 - \frac{c^2}{4}
\]

Looking at \( 1 - \frac{c^2}{4} \). Suppose \( 1 - \frac{c^2}{4} > 0 \), then:

\[
e^2 < 4 \quad \Rightarrow \quad 16 \left( e^\epsilon - \frac{1}{2} \right)^2 < 4 \quad \Rightarrow \quad 0 < e^\epsilon < 1
\]

This cannot hold and we have arrived at a contradiction. Hence, we can say \( 1 - \frac{c^2}{4} \leq 0 \). This means that at a point between the solutions, the expression is always negative. So, the required probability is:

\[
\mathbb{P}(Z_{ij}^2 - cZ_{ij} + 1 \leq 0) = \mathbb{P} \left( \frac{c - \sqrt{c^2 - 4}}{2} \leq Z_{ij} \leq \frac{c + \sqrt{c^2 - 4}}{2} \right)
\]

Using the cdf of \( Z_{ij} \), we get the probability given in (13). \( \square \)

APPENDIX C

DISTRIBUTION OF CORRELATIONS IN MODEL [1]

In this paper, we have used distribution of angles for discriminating clusters. Here, we establish that we can also use the correlations between points instead of angles. The following lemmas provide the distributions of correlation between points under Model [1]

Lemma 8. Let \( x_1, x_2, \ldots \in S^{n-1} \) are random points independently chosen with uniform distribution in \( S^{n-1} \), then the following can be said about the correlations \( \rho_{ij} = x_i^T x_j \):

a) \( \rho_{ij} \)'s are identically distributed as \( \beta_{[-1,1]} \left( \frac{n-1}{2}, \frac{n-1}{2} \right) \)

b) \( \rho_{ij} \)'s converge in distribution to \( \mathcal{N} \left( 0, \frac{1}{n} \right) \) as \( n \to \infty \) and the rate of convergence is of \( O \left( \frac{1}{n} \right) \).

Proof. Let \( w = \cos^2 \left( \frac{\theta_{ij}}{2} \right) \). Using this transformation and equation (3), we obtain \( w \sim \beta \left( \frac{n-1}{2}, \frac{n-1}{2} \right) \). Hence, \( \rho_{ij} = \cos \theta_{ij} = 2w - 1 \) is distributed as \( \beta_{[-1,1]} \left( \frac{n-1}{2}, \frac{n-1}{2} \right) \). This proves part a). Let \( \tau = \sqrt{n} \rho_{ij} \). Using the result from part a), the log density of \( \tau \) is obtained as

\[
\log s(\tau) = \log c_n + \frac{n - 3}{2} \log \left( 1 - \frac{\tau^2}{n} \right), \tau \in [-\sqrt{n}, \sqrt{n}]
\]

where \( c_n \) is the normalization term depending on \( n \) alone. Using Taylor expansion about \( \tau = 0 \),

\[
\log s(\tau) = \log c_n + \frac{n - 3}{2} \left[ -\frac{\tau^2}{n} - \frac{\tau^4}{2n^2} - \cdots \right]
\]

\[
= \log c_n - \frac{\tau^2}{n} + O \left( \frac{1}{n} \right)
\]

\[
\Rightarrow s(\tau) \propto e^{-\frac{\tau^2}{n}} \text{ at the rate of } O \left( \frac{1}{n} \right)
\]

Thus, \( \tau \overset{D}{\to} \mathcal{N}(0, 1) \) and hence \( \rho_{ij} \overset{D}{\to} \mathcal{N} \left( 0, \frac{1}{n} \right) \) at the rate of \( O \left( \frac{1}{n} \right) \). This proves part b). \( \square \)

Lemma 9. Suppose for Model [7] we have a true clustering \( C_\ast \)

a) When \( i, j \in I_k \), with \( I_k \) corresponding to the subspace \( U_k \), the correlations \( \rho_{ij} \)'s are identically distributed as \( \beta_{[-1,1]} \left( \frac{n-1}{2}, \frac{n-1}{2} \right) \)

b) When \( i \in I_k, j \in I_l \), then \( \rho_{ij} \)'s are identically distributed as \( \beta_{[-1,1]} \left( \frac{n-1}{2}, \frac{n-1}{2} \right) \).

Proof. Using part a) in Lemma 8 and arguments similar to those in the proof of Lemma 3, we shall obtain these results. \( \square \)

From lemmas 8 and 9 it can be seen that the correlations between points from the same subspace and the correlations between points of different subspaces are approximately Gaussian distributed with same mean 0 but with different variances \( \frac{1}{r_k} \) and \( \frac{1}{n} \).

REFERENCES

[1] R. Xu and D. C. Wunsch II, “Survey of clustering algorithms,” IEEE Trans. on Neural Networks, vol. 16, no. 3, pp. 645–678, 2005.

[2] K. Beyer, J. Goldstein, R. Ramakrishnan, and U. Shaft, “When is “nearest neighbor” meaningful? in Int. conf. on database theory, Springer, 1999, pp. 217–235.

[3] L. Parsons, E. Haque, and H. Liu, “Subspace clustering for high dimensional data: a review,” Acm Sigkdd Explorations Newsletter, vol. 6, no. 1, pp. 90–105, 2004.

[4] V. S. Cherkassky and F. Muller, Learning from Data: Concepts, Theory, and Methods, 1st ed. New York, NY, USA: John Wiley & Sons, Inc., 1998.

[5] R. Basri and D. W. Jacobs, “Lambertian reflectance and linear subspaces,” IEEE Trans. on Pattern Analysis & Machine Intelligence, vol. 25, no. 2, pp. 218–233, 2003.

[6] I. Jolliffe, Principal component analysis (Springer Series in Statistics). Berlin, Germany: Springer, 2002.

[7] R. Vidal, “Subspace clustering,” IEEE Signal Processing Magazine, vol. 28, no. 2, pp. 52–68, 2011.

[8] W. Hong, J. Wright, K. Huang, and Y. Ma, “Multiscale hybrid linear models for lossy image representation,” IEEE Trans. on Image Processing, vol. 15, no. 12, pp. 3655–3671, 2006.

[9] R. Vidal, R. Tron, and R. Hartley, “Multiframe motion segmentation with missing data using powerfactorization and gpca,” Int. J. of Computer Vision, vol. 79, no. 1, pp. 85–105, 2008.
